Elastic Representations via Hyperbolic Geometry

Arjun Ramesh Kaushik
University at Buffalo, SUNY

Rudrasis Chakraborty
Lawrence Livermore National Lab

Nalini Ratha
University at Buffalo, SUNY

Venu Govindaraju
University at Buffalo, SUNY

Abstract

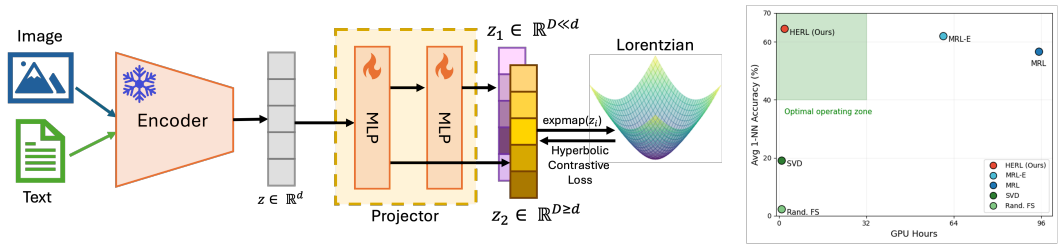
Learned representations (or embeddings) are the foundation of modern machine learning systems, yet they are typically trained as fixed-size embeddings, without accounting for varying downstream resource or task constraints. Matryoshka Representation Learning (MRL) addresses this limitation by learning nested representations, but requires costly end-to-end retraining and does not support post hoc expansion, while Contrastive Sparse Representations (CSR) reduce training overhead via lightweight adaptors but operate at a 4x larger dimensionality. Additionally, all prior works restrict their framework towards compression only. In this work, we introduce **H**yperbolic **E**lastic **R**epresentation **L**earning (HERL), a framework for learning dimension-adaptive embeddings (compression and expansion) through a lightweight adaptor without retraining the base encoder. Our key idea is to exploit the exponential capacity of hyperbolic geometry by learning a non-linear mapping from the encoder space to hyperbolic space. We train a simple MLP to downsample the representations before projecting them onto a Lorentzian and learn adaptive representations that respect the Lorentzian geometry. Our approach is model-agnostic and effective in both supervised and unsupervised settings. Through extensive experiments across image and text modalities, we show that HERL preserves downstream performance across a wide range of embedding sizes, exhibits smooth interpolation between dimensions, and achieves these benefits at a fraction of the computational cost. Code: `https://github.com/arjunrkaushik/herl`.

1 Introduction

Machine learning systems encode data into high-dimensional representations that are often assumed to lie on lower-dimensional manifolds. These learned embeddings serve as the foundation for a wide range of downstream tasks. Despite rapid progress in representation learning, most encoders produce fixed-dimensional embeddings, which limits their adaptability in real-world deployments where compute, memory, and latency constraints vary across settings. Matryoshka Representation Learning (MRL) [26] addresses this limitation by learning nested embeddings that can be truncated at inference time while preserving task performance. However, MRL requires end-to-end retraining of the encoder, which is computationally expensive and often impractical when working with large, pretrained models.

To overcome this, a line of recent work on adaptive representation learning - such as Matryoshka-Adaptor [53], SMEC [56], and Contrastive Sparse Representations (CSR) [52] - proposes lightweight adaptors on top of frozen encoders. While these approaches eliminate the need for retraining, they introduce new limitations: SMEC supports only a single reduced dimensionality, CSR activates a subset of dimensions within an overparameterized embedding, and Matryoshka-Adaptor exhibits performance degradation under aggressive compression. In this work, we revisit adaptive representation learning under frozen encoders and formalize two key desiderata: (i) flexible, continuously

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



(a) Hyperbolic Elastic Representation Learning (HERL)

(b) Comparison of training times

Figure 1: **Overview.** (a) We propose Hyperbolic Elastic Representation Learning (HERL), a post-hoc adaptor (without encoder retraining) that utilizes the inherent exponential volume capacity and faithful structural preservation properties of hyperbolic geometry. (b) HERL enables elastic representations at a fraction of the training cost. Approximately two hours to train representations of size 8 to 8192, on a single H100 GPU, for the ImageNet-1k dataset. On the other hand, Matryoshka Representation Learning (MRL) needs retraining of the encoder taking about sixty hours on four H100 GPUs.

adjustable embedding dimensionality where the active dimension matches the representation size, and (ii) robustness to aggressive compression with minimal loss in downstream performance.

The central design problem for an elastic, post-hoc adaptor is that it must trade dimensionality for fidelity along a continuum: a single encoder output $z \in \mathbb{R}^D$ must be projected to a representation $\phi_d(z)$ for any target budget $d \leq D$, and the resulting d-dimensional embedding must remain useful downstream. Euclidean geometry is a poor host for this trade-off, because the volume of a ball $B_r \subset \mathbb{R}^d$ grows only polynomially as $\Theta(r^d)$, and embedding a balanced binary tree of depth ℓ with bounded distortion already requires dimension $\Omega(\ell/\log \ell)$ [33, 44]. As the budget d shrinks, a Euclidean adaptor must therefore *discard* structure rather than *compress* it. Hyperbolic geometry inverts this scaling. In the d-dimensional hyperbolic space $\mathbb{H}^d$ of constant negative curvature c, the volume of a geodesic ball is $\mathrm{Vol}(B_r^{\mathbb{H}^d}) \propto e^{(d-1)\sqrt{c}r}$ as $r \to \infty$ [43], so capacity grows *exponentially* in radius [39, 44]. In [45], the authors show that *any* finite tree embeds into $\mathbb{H}^2$ with distortion $1+\epsilon$ for any $\epsilon > 0$, using only $O(\log 1/\epsilon)$ bits of precision that no Euclidean analogue can match at any dimension. For an adaptor whose budget shrinks toward $d = 2$ or 3, this exponential headroom means that aggressive compression pushes fine-grained relations outward in radius rather than erasing them.

This capacity advantage is paired with a faithfulness advantage on the structural relations (such as hierarchies, partial orders, graph distances) that frozen Euclidean encoders typically encode only implicitly. Bourgain's theorem shows that embedding an arbitrary n-point metric into Euclidean space incurs $O(\log n)$ distortion, with even complete binary trees requiring $\Omega(\sqrt{\log n})$ distortion [4, 33]. In contrast, the same tree metrics admit near-isometric embeddings in hyperbolic space with distortion $1+\epsilon$ [45, 44]. This gap is most relevant for representation learning, where learned structures often resemble hierarchies such as concept trees, taxonomies, and dependency graphs, precisely the regime where Euclidean embeddings are least efficient. Importantly, this is not merely theoretical: pretrained encoders across modalities exhibit low Gromov δ-hyperbolicity [24], vision-language models naturally organize representations in hyperbolic-like hierarchies [12], and hyperbolic graph networks consistently outperform Euclidean ones on hierarchical benchmarks [9, 34, 1]. Thus, suggesting that the hierarchical structural prior is already baked in the frozen encoder's latent space and the adaptor's role is to *expose* it rather than manufacture it. The hyperbolic embedding additionally endows coordinates with semantic roles [39, 48, 40, 8]. Thus, a post-hoc adaptor that respects this structure can learn nested and elastic embeddings, with minimal degradation in performance at low dimensions.

In summary, we propose a novel technique, **H**yperbolic **E**lastic **R**epresentation **L**earning (HERL), that exploits the exponential volume capacity and structural prior of hyperbolic geometry to learn elastic and nested embeddings without retraining the encoder. HERL satisfies both desiredata, and enables efficient and scalable deployment of pretrained models across diverse resource-constrained environments. Computationally, HERL is approximately 30x cheaper than MRL Fig. 1(b).

2 Background

Riemannian manifolds. A d-dimensional Riemannian manifold $(\mathcal{M}, g)$ equips each tangent space $T_x\mathcal{M}$ with an inner product $\langle \cdot, \cdot \rangle_x$ that varies smoothly with x. The metric induces the geodesic distance $d_{\mathcal{M}}(x, y)$ as the length of the shortest curve between x and y, together with the exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ and its inverse $\log_x : \mathcal{M} \to T_x\mathcal{M}$, which generalise straight-line displacement to curved spaces. We refer the reader to [31] for a comprehensive treatment.

Hyperbolic space. *Hyperbolic space* $\mathbb{H}^d$ is the unique simply-connected complete Riemannian manifold of constant negative sectional curvature [6]. Its defining property is that the volume of a geodesic ball grows *exponentially* with radius rather than polynomially as in $\mathbb{R}^d$ (Euclidean), thereby giving it sufficient room to host tree-structured data with arbitrarily low metric distortion: any finite tree embeds isometrically (up to a $1+\varepsilon$ factor) into $\mathbb{H}^2$ [45]. This property has motivated a substantial line of work using hyperbolic embeddings to capture latent hierarchies in language, graphs, and vision [39, 40, 14, 9, 24]. We exploit this property at the level of *representation compression*: compressing along a hierarchy is far cheaper geometrically when distances scale exponentially, since coarse and fine semantic levels are naturally separated by radius rather than competing for orthogonal directions.

The Lorentz model. Among the equivalent isometric models of $\mathbb{H}^d$ (Poincaré ball, Klein, half-space, Lorentz), we work in the Lorentz (hyperboloid) model, which is numerically more stable for gradient-based optimisation than the Poincaré ball [40, 28]. Let $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denote the Minkowski bilinear form on $\mathbb{R}^{d+1}$,

$$\langle x, y \rangle_{\mathcal{L}} \;=\; -x_0 y_0 + \sum_{i=1}^{d} x_i y_i \tag{1}$$

The Lorentz model with curvature $-c$ ($c > 0$) is the upper sheet of the two-sheeted hyperboloid $\mathbb{L}_c^d \;=\; \big\{\, x \in \mathbb{R}^{d+1} : \langle x, x \rangle_{\mathcal{L}} = -1/c,\; x_0 > 0 \,\big\}$, endowed with the metric obtained by restricting $g^{\mathcal{L}} = \mathrm{diag}(-1, 1, \ldots, 1)$ to the tangent space $T_x\mathbb{L}_c^d = \{v : \langle x, v \rangle_{\mathcal{L}} = 0\}$ [40, 28]. The geodesic distance and exponential map admit the closed forms

$$d_{\mathbb{L}}(x, y) \;=\; \frac{1}{\sqrt{c}} \operatorname{arcosh}\big(-c\, \langle x, y \rangle_{\mathcal{L}}\big) \tag{2}$$

$$\exp_x^c(v) \;=\; \cosh\big(\sqrt{c}\, \|v\|_{\mathcal{L}}\big) x \;+\; \sinh\big(\sqrt{c}\, \|v\|_{\mathcal{L}}\big)\, \frac{v}{\sqrt{c}\, \|v\|_{\mathcal{L}}} \tag{3}$$

where $\|v\|_{\mathcal{L}} = \sqrt{\langle v, v \rangle_{\mathcal{L}}}$ for any spacelike $v \in T_x\mathbb{L}_c^d$. Unlike the Poincaré ball [14], the Lorentz model has no conformal factor that diverges at the boundary, eliminating the dominant source of numerical instability in hyperbolic deep learning.

Adaptive Representation Learning. With increasing computational constraints, recent research has pivoted to adaptive representations that can cater to diverse tasks. Early efforts explored context-based architectural adaptations [25], dynamic widths and depths in BERT [20], and random layer dropping during training to improve pruning robustness [13]. More recently, Matryoshka Representation Learning (MRL) [26] introduced a novel technique for creating flexible, nested substructures within embeddings, enabling fine-grained control over the trade-off between latency and accuracy. This concept has since been extended to various modalities and applications, including large language models [41, 54], diffusion models [17], and multimodal models [5, 21]. SMEC [56] and Matryoshka-Adaptor [53] represent a new trend in Adaptive Representation Learning by introducing a post-hoc adaptor. However, both methods are limited by their generalizability or robustness under aggressive compression. Contrastive Sparse Representations (CSR) [52] is the first work that offers a post-hoc sparse coding alternative but produces fixed-size representations that are four times bigger than the encoder embedding, limiting its applicability to encoder-dependent downstream tasks. *Towards this end, our method trains a lightweight adaptor on any frozen encoder that explicitly controls the geometric resolution as a function of target dimensionality while remaining adaptable to any dimension.*

3 Methodology

We consider a practical post-hoc setting where the encoder $E : \mathcal{X} \to \mathbb{R}^D$ is frozen. Given an input $x \in \mathcal{X}$ with embedding $z = E(x)$, our goal is to learn a lightweight adaptor $A : \mathbb{R}^D \to \mathbb{R}^{D'}$, where $D' \in \mathbb{N}$, such that the transformed representation $\tilde{z} = A(z)$ yields nested, truncatable embeddings, where any prefix $\tilde{z}_{1:d}$. To address this challenging problem, we propose a **H**yperbolic **E**lastic **R**epresentation **L**earning (HERL), which consists of two components: (i) a learnable projector that adapts frozen Euclidean embeddings into a task- and geometry-compatible latent space, and (ii) a hyperbolic optimization module that aligns representations with the Lorentz manifold using contrastive supervision. Together, these components enable the construction of nested embeddings that are both dimensionally truncatable and geometrically coherent, without retraining the base encoder.

(i) Projector (P): Since pretrained encoders are generally optimized in Euclidean space, their representations do not naturally align with the geometric properties of hyperbolic manifolds. Therefore, we introduce a learnable projector that serves as a bridge between Euclidean and hyperbolic spaces, enabling a smooth and structured transformation. Concretely, this projector learns to reparameterize the encoder output $z \in \mathbb{R}^D$ into a form that is compatible with hyperbolic geometry before applying manifold operations.

Beyond geometric alignment, this transformation also addresses a fundamental limitation of high-dimensional embeddings: they often contain redundant or noisy components that are not essential for downstream tasks, especially under strict dimensionality or resource constraints. The projector thus plays a dual role: (i) it filters and compresses the Euclidean representation by selectively preserving task-relevant information, and (ii) it reshapes the representation to better exploit the inductive biases of hyperbolic space, such as its ability to model hierarchical and graph-structured relationships. By jointly performing denoising and geometric adaptation, the projector ensures that the subsequent hyperbolic transformation operates on a compact, informative representation. We realize this projector module P via an MLP as shown in Fig. 1(a).

The final layer of P enables elasticity, and is parameterized as $\mathbb{R}^{D_{\max} \times D}$. For any $d \leq D_{\max}$, we define: $\bar{z}_{1:d} = (\bar{z}_1, \ldots, \bar{z}_d)$, where $\bar{z} = P(E(x)) \cdot W^T$ which corresponds to applying the submatrix $W_{1:d,:}$. This induces a *prefix-consistent* structure, where lower-dimensional embeddings reuse parameters from higher-dimensional ones. Combined with hyperbolic geometry, this encourages early coordinates to capture global structure (low-radius information) and later coordinates to encode fine-grained details (high-radius refinements), aligning with the exponential volume property.

The key clarification is that our adaptor (Projector and W) does not compress in Euclidean space and then separately map into hyperbolic space as two independent stages. The entire adaptor, is trained end-to-end under a loss computed in the Lorentz model, so every gradient it receives flows back through the exponential map and the Lorentzian contrastive objective. The MLP is therefore never optimized for Euclidean reconstruction - it learns to produce tangent-space coordinates whose image under expmap best separates the data on the manifold, meaning compression and geometry are co-designed rather than sequential. This matters because our claim is not that any computation in Euclidean coordinates destroys structure, but the narrower point that an adaptor trained under a Euclidean objective must discard structure as d shrinks, whereas training the same low-dimensional bottleneck against a Lorentzian objective gives the network a target geometry with exponential headroom. Note that the exponential map at the origin is a smooth, invertible reparameterization and thus introduces no information loss.

(ii) Optimization using Hyperbolic Geometry: We optimize the adaptor (Projector and W) in the Lorentz model by first mapping the Euclidean representation $\bar{z}_{1:d} \in \mathbb{R}^{d \leq D_{\max}}$ onto the hyperboloid manifold using the exponential map defined at the origin (Eq. 3): $u = \text{expmap}(\bar{z}_{1:d}) \in \mathbb{H}^{d \leq D_{\max}}$ where $\mathbb{H}^{D_{\max}} \subset \mathbb{R}^{D_{\max}+1}$. Subsequently, all pairwise similarity computations are performed in hyperbolic space using the Lorentzian metric, ensuring geometric consistency with the manifold structure. Note that the hyperbolic projection is only performed during training. During inference, $\bar{z}_{1:d}$ is directly used for downstream tasks requiring a d-dimensional representation.

In the absence of labels, we construct supervision signals from the frozen encoder space. Specifically, for each sample x_i, we utilize its encoder embedding $z_i = E(x_i)$ and define a neighborhood using cosine similarity (Note the change in notation, i corresponds to sample index and not dimensionality).

Subsequetly, we select the top-k nearest neighbors as positives: $\mathcal{NN}_k(i) = \mathrm{TopK}_{j \neq i}\big(\mathrm{sim}(z_i, z_j)\big)$. This choice is motivated by the fact that high-dimensional representations from pretrained encoders are approximately concentrated on a hyperspherical manifold, making cosine similarity a more reliable measure of semantic proximity than Euclidean distance.

We define a contrastive objective by treating samples in $\mathcal{NN}_k(i)$ as positives and all other samples in the batch as negatives. For an anchor i and a sampled positive $j \in \mathcal{NN}_k(i)$, the unsupervised InfoNCE-style loss [55] is given as follows, where τ is a temperature parameter:

$$\mathcal{L}_{\text{unsup}} = -\log \frac{\exp\left(-d_{\mathbb{H}}(u_i, u_j)/\tau\right)}{\sum_{k \neq i} \exp\left(-d_{\mathbb{H}}(u_i, u_k)/\tau\right)},$$

This formulation leverages the exponential volume growth of hyperbolic space, which naturally supports large separation between dissimilar samples while preserving tight clustering of neighbors, making it particularly suitable for modeling structured semantic relationships. Further, the random sampling over training iterations allows the model to learn faithful local structures.

When class labels y_i are available, supervision is incorporated through a class-conditioned metric structure in hyperbolic space. Specifically, for an anchor sample x_i, we define its positive set as all samples sharing the same semantic label, $\mathcal{P}(i) = \{j \neq i \mid y_j = y_i\}$, and implicitly treat all remaining samples in the batch as negatives, i.e., $\mathcal{N}(i) = \{k \neq i \mid y_k \neq y_i\}$. This induces a class-aware partitioning of the representation space that aligns with semantic structure. The supervised loss for each anchor i is then formulated as a normalized log-softmax over the full batch:

$$\mathcal{L}_{\text{sup}}^{(i)} = -\frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp\left(-d_{\mathbb{H}}(u_i, u_j)/\tau\right)}{\sum_{k \neq i} \exp\left(-d_{\mathbb{H}}(u_i, u_k)/\tau\right)}.$$

Once can interpret this as maximizing the likelihood of correctly identifying class-consistent samples under a hyperbolic Gibbs distribution induced by the Lorentz geometry. The normalization over all non-anchor samples ensures that class separation is globally enforced across the batch, while the averaging over positives prevents collapse in cases of class imbalance or multi-instance classes.

Naively using this opitmization objective across dimensions can cause gradient conflict, leading to a drop in performance across both smaller and larger dimensions. Inspired by SMEC [56], we adopt a sequential training strategy that first optimizes the highest-dimensional representation and progressively refines lower-dimensional embeddings.

4 Projector Complexity vs. Embedding Dimension

We observe empirically that when the target dimension is small ($d \lesssim 32$–64), a lightweight (single-layer) projector P leads to a noticeable degradation in performance, whereas a deeper multi-layer projector recovers strong results. Interestingly, beyond this regime, the trend reverses: increasing projector complexity yields diminishing or even negative returns. We provide a principled explanation for this phenomenon through three interacting geometric effects, with detailed proofs and estimators deferred to Appendix A.

Setup. Let $E : \mathcal{X} \to \mathbb{R}^D$ denote a frozen encoder, and $W \in \mathbb{R}^{D_{\max} \times D}$ the elastic weight matrix. For an input $x \in \mathcal{X}$, let $z = E(x)$ and define the projected representation $\bar{z}_{1:d} = P(z)\, W_{1:d,:}^{\top}$. Let $\mathcal{S} \subseteq \mathbb{R}^D$ denote the *semantic subspace* induced by E, i.e., the minimal subspace that preserves class-conditional structure, with $s = \dim(\mathcal{S})$. We identify three failure modes that emerge when d is small:

(i) Rank bottleneck. The submatrix $W_{1:d,:}$ has rank at most d, which limits the effective degrees of freedom in the projected space. When $\mathcal{S}$ is not aligned with the coordinate axes and $d < s$, no linear transformation can fully reorient the representation to align with the dominant directions of W (see Proposition A.1, A.1).

Category	Model	Dim	Text Classification Top-1 Acc (%)↑			Text Clustering Top-1 Acc (%)↑			Text Retrieval NDCG@10 (%)↑		
			MTOPIntent	Banking77	TweetSentiment	BiorxivP2P	BiorxivS2S	TwentyNews	FiQA2018	NFCorpus	SciFACT
Full Rep	NV-Embed-V2	4096	93.58	92.20	79.73	53.61	49.60	64.82	62.65	43.97	77.93
MRL	Stella-1.5B-v5	256	90.45	86.14	76.75	50.81	46.42	60.07	55.59	36.97	**77.48**
	Jina-v3	256	78.81	84.08	73.81	38.14	34.39	51.96	55.73	36.63	66.63
	Nomic-Embed-V1.5	256	72.47	83.69	59.20	38.19	31.83	48.56	35.00	32.54	68.24
	Gecko-Embed-004(Google)	256	77.82	86.01	72.97	36.28	33.09	50.60	55.54	37.81	70.86
	Text-Embed-3-L (OpenAI)	256	70.45	83.19	58.98	35.43	33.86	54.24	50.33	37.94	73.10
	Arctic-Embed-L-V2	256	67.69	80.99	59.06	34.25	34.07	30.06	44.69	35.02	69.51
	M2V-Base-Glove	256	59.26	72.39	50.02	32.26	22.34	25.38	11.82	23.15	50.66
	Jina-V3	64	68.12	67.98	71.18	36.89	33.57	50.22	44.18	33.66	68.84
	Nomic-Embed-V1.5	64	62.77	80.63	55.23	34.81	44.61	48.06	10.22	18.96	36.55
	Potion-Base-2M	64	42.50	65.17	52.52	25.78	14.94	27.07	32.08	30.72	64.28
Ours	Unsup. HERL	32	94.39	92.88	79.51	48.18	48.47	62.85	50.70	45.67	55.70
(w/ NV-Embed-V2)	Unsup. HERL	8192	93.72	93.21	79.34	78.43	75.22	64.70	**61.11**	**55.23**	71.62
	Sup. HERL	32	**94.53**	**94.86**	**79.88**	63.28	75.82	80.14	-	-	-
	Sup. HERL	8192	93.49	93.15	79.60	**80.12**	**76.29**	**82.77**	-	-	-

Table 1: **Results on textual representations.** We compare the downstream utility of representations across three tasks: classification, clustering and retrieval. HERL (Ours) consistently outperforms existing MRL-based representations. Importantly, in some cases, HERL enables a 32 dimensional representation to surpasses full representation (4096 dimension) performance, achieving 128x efficiency in terms of dimension size.

(ii) Volume collapse. The representational capacity of hyperbolic space grows exponentially with dimension:

$$N(R, \varepsilon, d) \propto \left(\frac{\sinh(\sqrt{c}R)}{\sinh(\sqrt{c}\varepsilon)} \right)^{d-1}$$

Below a critical volumetric threshold d_{vol}, the space cannot separate more than K semantic clusters with margin ε in $\mathbb{H}^d_c$, causing the geometry to effectively collapse toward Euclidean behavior (Proposition A.3, A.2).

(iii) Norm collapse. Due to concentration effects in moderate dimensions, $\mathbb{E}[\|\bar{z}\|_2] \approx \sigma\sqrt{d}$. When $d < d_{\text{norm}}$, embeddings concentrate near the Lorentz origin, where hyperbolic distances degenerate:

$$d_{\mathbb{L}}(u_i, u_j) \approx \|\bar{z}_i - \bar{z}_j\|_2$$

In this regime, the hyperbolic inductive bias is effectively lost (Proposition A.6, A.3).

Main result. We define the *critical dimension*: $d^* = \max\{s,\, d_{\text{vol}},\, d_{\text{norm}}\}$.

Theorem 4.1 (Projector complexity threshold; proof in Appendix A.10)**.**

(i) ***Low-dimensional regime*** *($d < d^*$)**:** At least one of the failure modes above is active. In this regime, a linear projector is insufficient to recover faithful hyperbolic structure, while a sufficiently expressive multi-layer projector (depth ≥ 2, width $\geq D$) can jointly mitigate rank misalignment, volume collapse, and norm degeneration.*

(ii) ***High-dimensional regime*** *($d \geq d^*$)**:** All three failure modes are inactive. Additional projector capacity primarily increases optimization variance without improving expressivity. Consequently, simpler projectors are preferred due to reduced overfitting and improved stability.*

Practical estimation. For an ImageNet-scale encoder ($D = 768$, $K = 1000$, $c = 1$), Appendix A.11 provides closed-form estimates of s, d_{vol}, and d_{norm}, yielding $d^* \approx 32$–64, which closely matches our empirical observations (Section 5) and aligns with trends reported in [26].

5 Results and Analysis

We evaluate HERL across two modalities - image and text. For image representations, we evaluate the downstream capability of the representations by measuring Top-1 accuracy via linear probing and 1-NN retrieval accuracy on the ImageNet-1k dataset, following MRL [26]. On the textual representations, we evaluate across three tasks: classification, clustering and retrieval on the MTEB dataset [38]. All implementation and specific dataset details can be found in the appendix. All baseline comparisons directly follow CSR's [52] protocol. For fair comparisons, we do not include CSR [52], whose ambient dimension is consistently 4x the full representation, in our baselines.

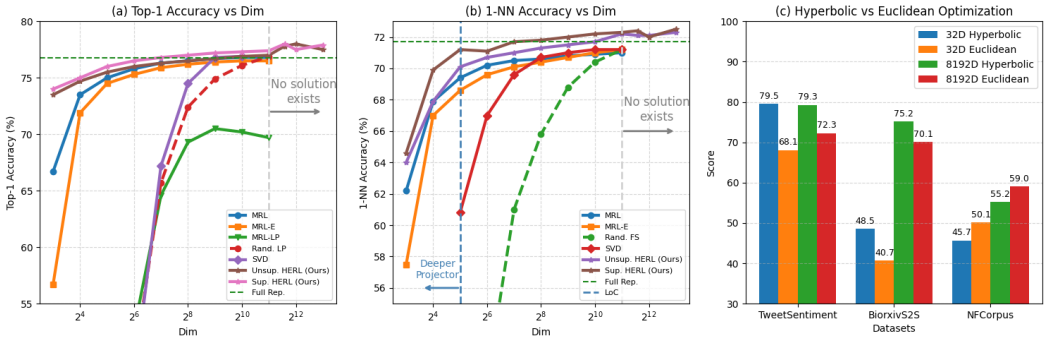


Figure 2: **Performance of HERL on ImageNet-1k and comparison with Euclidean geometry. (a) Top-1 classification accuracy** and **(b) 1-NN retrieval accuracy** over embedding dimensions, comparing our supervised and unsupervised HERL adaptors against baselines. The dashed green line ("Full Rep.") denotes the frozen encoder's native-dimension performance. Both HERL variants show improved performance over baselines. The "No solution exists" annotation marks the dimensionality support region that does not exist, whereas HERL extends naturally into this regime. The "Deeper Projector / LoC (Line of Control)" annotation in (b) indicates the dimension below which the adaptor requires a deeper projection head. **(c) Hyperbolic vs. Euclidean optimization of the adaptor at 32D and 8192D** on three tasks: classification (TweetSentiment), clustering (BiorxivS2S) and retrieval (NFCorpus). Hyperbolic geometry dominates classification and clustering tasks, but struggles with retrieval. Solely due to the fact that the evaluation pipeline normalizes embeddings onto a hypersphere, killing the primal component of distinguishability in hyperbolic geometry - the radius.

5.1 Image Representations

Baselines. We compare our method against several representative approaches: (1) **MRL / MRL-E** [26], which use a ResNet-50 (RN50) backbone where either all nested representations get an independent classifier head (MRL) or a single classifier head is shared amongst nested representations (MRL-E); (2) **SVD**, where we perform a low-rank approximation of the 1000-class classification layer of RN50 with rank 1000; (3) **Rand-LP**, which trains a linear classifier on randomly selected feature subsets; and (4) **Rand-FS**, which randomly subsamples RN50 features for 1-NN classification.

Evaluation metrics. Following CSR [52] and MRL [26], we adopt 1-NN accuracy, implemented using FAISS [23] with exact L2 search. This approach provides an efficient and cost-effective way to evaluate the utility of learned representations for downstream tasks, as 1-NN accuracy requires no additional training. In detail, we use the training set with 1.3M samples as the database and the validation set with 50K samples as the query set. We also report linear probing and few-shot results using Top-1 accuracy.

Analysis. Fig. 2(a) and Fig. 2(b) compare HERL against existing baselines across different embedding dimensions for both retrieval Top-1 classification accuracy and 1-NN retrieval accuracy. Traditional projection approaches such as MRL-LP, random linear projections (Rand. LP) exhibit severe degradation under aggressive compression, particularly below 128 dimensions, indicating their inability to preserve semantic structure in compact spaces. While MRL and MRL-E remain competitive at moderate dimensions, they are limited by expensive encoder retraining. In contrast, HERL maintains strong and stable performance across the entire dimensional range, with both unsupervised and supervised variants consistently outperforming competing approaches under extreme compression. Notably, HERL also allows user to go beyond full representation sizes while improving performance. Thus, indicating that our adaptor is learning the latent space rather than merely concatenating zeros. These findings suggest that hyperbolic geometry is an elastic alternative to conventional Euclidean projection methods.

5.2 Text Representations

Baselines. We evaluate several models that provide MRL-style embeddings on the MTEB benchmark [38], including Stella-1.5B-v5 [57], Jina-V3 [47], Nomic-Embed-V1.5 [41], GeckoText-Embedding-004 [30], OpenAI-Text-Embedding-3-L[42], Arctic-Embed-L-V2.0 [54], and Potion-Base-2M [49].

Evaluation Metrics. We adopt the standard evaluation metrics of the MTEB benchmark [38]. Classification and clustering performance are measured by Top-1 accuracy, while retrieval performance is measured by nDCG@10 [22], the normalised discounted cumulative gain over the top-10 ranked list, which is the standard ranking metric in information retrieval and recommendation.

Analysis. Table 1 shows that HERL achieves strong performance across text classification and clustering tasks while operating at significantly lower dimensions than prior MRL-based approaches. In the highly compressed 32D setting, unsupervised HERL already surpasses / matches the full 4096D NV-Embed-V2 representation on MTOPIntent (94.39 vs. 93.58) and TweetSentiment (79.51 vs. 79.73), while supervised HERL further improves performance to 94.53 and 79.88 respectively. More importantly, HERL demonstrates exceptional performance on clustering benchmarks, where the supervised 8192D variant achieves state-of-the-art accuracy on BiorxivP2P (80.12), BiorxivS2S (76.29), and TwentyNews (82.77), substantially outperforming both the full representation and all MRL baselines. These results indicate that hyperbolic optimization preserves semantic hierarchy and global structure more effectively than Euclidean methods, particularly under extreme compression. Interestingly, even the compact 32D HERL embeddings remain highly competitive with their 8192D counterparts, suggesting that the hyperbolic adaptor efficiently allocates representational capacity in low-dimensional spaces. HERL exhibits a modest reduction in performance on retrieval tasks (FIQA2018 and SciFACT). We argue that this does not reflect weaker hyperbolic representations, but rather an incompatibility between hyperbolic geometry and conventional retrieval evaluation pipelines. Retrieval protocols projects embeddings onto the unit hypersphere before similarity computation, effectively discarding radial information, which is the principal carrier of semantic separation in hyperbolic space. Thereby, introducing an evaluation bias. This is further elucidated in Sec. 6.

6 Ablation studies

Why the Lorentz Model? The Poincaré ball, Klein, half-space, and Lorentz models of $\mathcal{H}^d$ are mutually isometric: any geometric quantity computed in one can in principle be computed in another. Numerically and computationally, however, the choice matters substantially for an elastic adaptor that must operate across a wide range of dimensions and embedding magnitudes. We claim that the Lorentz model is the right choice for our setting, for two concrete reasons.

(i) Numerical stability away from the boundary: Points deep in the hierarchy lie near the boundary of the Poincaré ball, where $\|x\|_2 \to 1^-$. The hyperbolic distance involves the factor $1 - \|x\|^2$ in the denominator, which suffers catastrophic cancellation in float32 once $\|x\| > 1 - 10^{-7}$. This is precisely the regime our adaptor must operate in when d is small and the exponential capacity forces points outward. The Lorentz model parameterises the same point as $\boldsymbol{x} = (x_0, \boldsymbol{x}_s) \in \mathbb{R}^{1,d}$ with the constraint $-x_0^2 + \|\boldsymbol{x}_s{}^2\| = -1$ and $x_0 > 0$, and the distance reduces to an inner product (Eq. 2), admitting a stable evaluation up to the limits of arcosh itself. Empirically, [40] demonstrates that Lorentz-model optimization is much more stable than Poincaré training, and [37] formalize this.

(ii) Compatibility with frozen Euclidean encoders: The frozen encoder produces $z \in \mathbb{R}^D$ in standard Euclidean coordinates. A Lorentz-model adaptor lifts z to $\mathbb{H}^d$ via $\boldsymbol{x} = \exp_{\boldsymbol{o}}(0, \pi(z))$ where $\boldsymbol{o} = (1, \mathbf{0})$ is the apex of the hyperboloid and $\pi : \mathbb{R}^D \to \mathbb{R}^d$ is a learned linear projection; the lift is exact in closed form because the tangent space at $\boldsymbol{o}$ is canonically identified with $\mathbb{R}^d$. The Poincaré lift requires an additional $\tanh$ rescaling that flattens gradients near the boundary and introduces a second source of numerical error. Empirically [14] and theoretically [46], the Lorentz lift produces better-conditioned Jacobians. Our experiments substantiate these insights (Fig. 3(a)).

Euclidean vs. Hyperbolic We perform an ablation (Fig. 2(c)) comparing hyperbolic and Euclidean optimization of the adaptor at both low-dimensional (32D) and full-dimensional (8192D) settings across three representative tasks: classification (TweetSentiment), clustering (BiorxivS2S), and retrieval (NFCorpus). Hyperbolic optimization consistently outperforms Euclidean optimization on classification and clustering tasks, achieving particularly strong gains under extreme compression at 32D (e.g., 79.5 vs. 68.1 on TweetSentiment and 48.5 vs. 40.7 on BiorxivS2S), demonstrating the ability of hyperbolic geometry to preserve semantic and hierarchical structure in compact embedding spaces. Notably, these improvements persist even at 8192D, suggesting that the benefits arise not merely from compression efficiency but from the inductive bias of hyperbolic manifolds themselves, which allocate representational capacity more effectively through exponential volume growth. In contrast, Euclidean optimization slightly outperforms hyperbolic optimization on retrieval

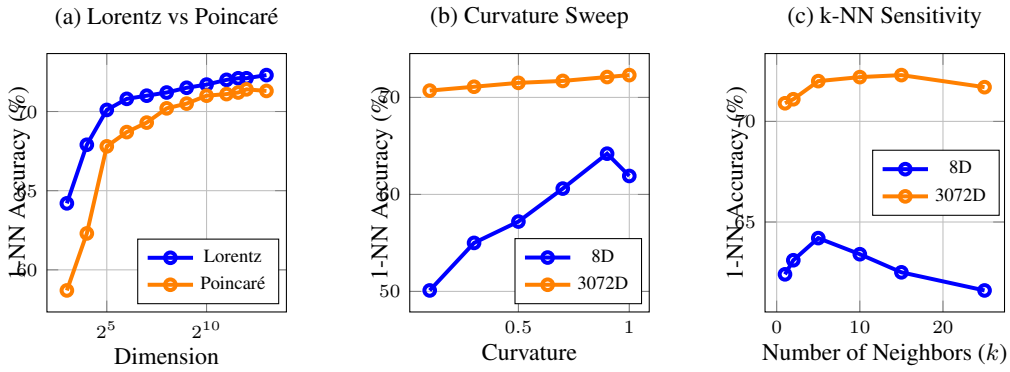


Figure 3: **Ablation studies across geometry, curvature, and neighborhood size.** (a) Lorentz embeddings outperform Poincaré across dimensions, especially in low-dimensions. (b) Increasing curvature significantly improves compact representations (8D), with limited impact in high dimensions (3072D). (c) Low-dimensional embeddings are highly sensitive to neighborhood size k that is used in the contrastive loss. All experiments have been performed on ImageNet-1k.

(NFCorpus) at both dimensions. We attribute this behavior *not to* inferior hyperbolic representations, but to a mismatch between hyperbolic geometry and standard retrieval evaluation protocols, which normalize embeddings onto the unit hypersphere prior to similarity computation. Such normalization removes radial information, the primary carrier of semantic distinguishability in hyperbolic space, thereby collapsing an important degree of freedom that encodes hierarchy, specificity, and semantic separation. Consequently, existing retrieval pipelines may inherently favor Euclidean embeddings, underestimating the representational quality of hyperbolic embeddings in retrieval settings. That said, retrieval performance in hyperbolic space can be brought closer to the Euclidean by careful tuning of curvature, and temperature in the contrastive loss. The Euclidean experiments have been fine-tuned to extract best performance under the same architectural constraints.

Effect of curvature We ablate the effect of curvature on our unsupervised model in Fig. 3(b) by sweeping values in the range $(0, 1]$. Empirically, higher curvature significantly improves performance in low-dimensional representations, while yielding negligible gains at larger dimensions. This behavior aligns with the exponential capacity of hyperbolic space, which is particularly beneficial in compact embeddings, as well as its ability to capture hierarchical structure. Performance degrades at high curvature ($c = 1$) due to a trade-off between hierarchical capacity and local fidelity. While increasing curvature enhances the ability to separate hierarchical structure via exponential volume growth, it also amplifies distances near the boundary, leading to instability and over-separation of nearby points. This degrades local neighborhood preservation, which is critical for nearest-neighbor evaluation. Additionally, extreme curvature introduces optimization challenges and mismatches the partially non-hierarchical structure of real data, resulting in a performance drop.

Number of Nearest Neighbors Fig. 3(c) varies k, the number of nearest neighbours from which positives are sampled for the contrastive loss, and exhibits a bias-variance trade-off in the resulting training signal. At k=1 each anchor is paired with its nearest point, so the supervisory signal is dominated by whichever neighbour happens to be closest in the frozen encoder's geometry. This magnifies the effect of noise and local distortions, producing high-variance gradients and unstable accuracy. Moderate values $k \in [5, 10]$ aggregate the contrastive signal over several consistent neighbours, dampening the influence of a noisy positive while still concentrating the loss on local structure, and consequently yields peak accuracy. As k grows, the sampled neighbourhood begins to seep across class boundaries. The contrastive objective is then implicitly pulled toward cross-class pairs, biasing the learned geometry toward smoother, less discriminative embeddings, explaining the small accuracy drop at the right of the curve. The magnitude of this effect exacerbates at d=8, due to the comparatively limited capacity and accuracy depends strongly on k. Higher dimension representations have more room, thereby allowing more robustness across a wide range of k.

Retrieval under the native hyperbolic metric. Fig. 4(a) revisits the retrieval gap observed in Fig. 2(c). Standard retrieval protocols rank documents by cosine similarity, which projects every embedding onto the unit hypersphere and discards its norm. For HERL, the norm of determines

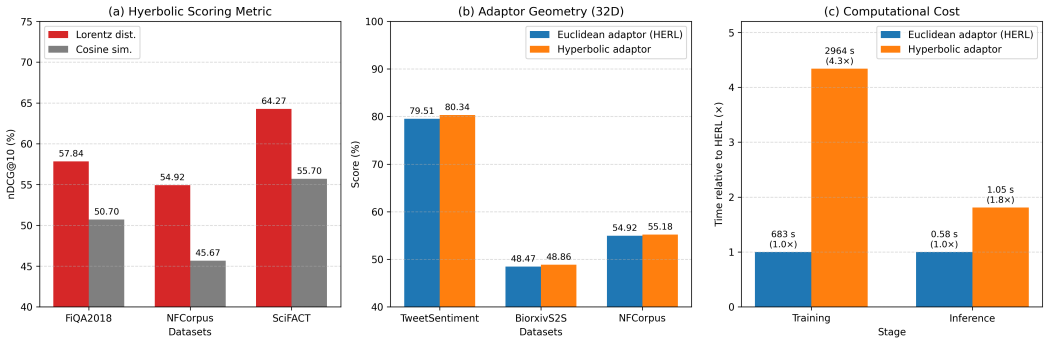


Figure 4: **More ablation studies. (a) Retrieval with Lorentz distance:** Cosine scoring projects embeddings onto the unit hypersphere and discards the hyperbolic radius. Evaluation with the native hyperbolic (Lorentz) distance improves results without retraining. **(b) Fully hyperbolic adaptor:** As expected a fully hyperbolic adaptor architecture outperforms a Euclidean adaptor, both trained with a hyperbolic objective. However, the negligible gains come at a high computational cost. **(c) Computational cost:** The hyperbolic adaptor is $4.3\times$ slower to train and $1.8\times$ slower at inference.

the hyperbolic radius, which is where the adaptor places coarse versus fine-grained information. Therefore, we re-evaluate the 32-D unsupervised HERL embeddings, ranking documents by Lorentz distance $d_{\mathcal{L}}$ (Eq. 2) instead of cosine similarity. No retraining is involved, only the scoring function changes. Scoring with $d_{\mathcal{L}}$ improves HERL across all three retrieval datasets. Thus, the earlier deficit came mainly from the evaluation discarding radial information, not from weaker representations.

Euclidean versus fully hyperbolic adaptor. HERL works with a Euclidean adaptor and uses hyperbolic geometry only in the training loss. A natural alternative is to make the adaptor itself hyperbolic, trained with the same objective. Fig. 4(b,c) compares the two on 32-D embeddings, using Lorentz distance for retrieval in both cases. The fully hyperbolic adaptor gives at most marginal gains at the cost of $4.3\times$ longer training (2964 s vs. 683 s) and $1.8\times$ slower inference (1.05 s vs. 0.58 s per batch of 256, because every layer requires manifold operations. We conclude that the benefit of hyperbolic geometry comes mainly from the training objective, not from the adaptor's layers.

7 Conclusion

In this work, we introduced **H**yperbolic **E**lastic **R**epresentation **L**earning (HERL), a lightweight and model-agnostic framework for learning dimension-adaptive embeddings without retraining the encoder. Unlike prior approaches that are limited to compression or require expensive end-to-end optimization, HERL supports both compression and expansion through a simple post hoc adaptor trained in an unsupervised manner. By leveraging the exponential representational capacity of hyperbolic geometry and learning adaptive embeddings on the Lorentz manifold, HERL preserves downstream performance across a wide spectrum of embedding dimensions while maintaining smooth interpolation between representations. Most importantly, HERL saves massive hours of encoder retraining. Specifically, one can train a representations of 8 to 8192, in approximately two hours on a single GPU, whereas MRL would take about 60 hours on 4 GPUs. These results highlight the potential of hyperbolic geometry as an effective foundation for scalable and flexible representation learning under varying resource constraints. Extending HERL to additional modalities and more challenging downstream tasks, particularly generative modeling, is an exciting direction for future.

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] Balazevic, I., Allen, C., Hospedales, T.: Multi-relational Poincaré graph embeddings. In: Advances in Neural Information Processing Systems. pp. 4463–4473 (2019)

[2] Boteva, V., Gholipour, D., Sokolov, A., Riezler, S.: A full-text learning to rank dataset for medical information retrieval. In: European Conference on Information Retrieval (2016)

[3] Bottou, L.: Stochastic gradient descent tricks. In: Neural Networks: Tricks of the Trade. pp. 421–436. Springer (2012)

[4] Bourgain, J.: On lipschitz embedding of finite metric spaces in hilbert space. Israel Journal of Mathematics **52**(1–2), 46–52 (1985)

[5] Cai, M., Yang, J., Gao, J., Lee, Y.J.: Matryoshka multimodal models. Proceedings of the International Conference on Learning Representation (2025)

[6] Cannon, J.W., et al.: Hyperbolic geometry. Cambridge University Press (1997)

[7] nigo Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., Vulić, I.: Efficient intent detection with dual sentence encoders. In: Workshop on Natural Language Processing for Conversational AI (2020)

[8] Chami, I., Gu, A., Nguyen, D.P., Re, C.: Horopca: Hyperbolic dimensionality reduction via horospherical projections. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 1419–1429 (18–24 Jul 2021), https://proceedings.mlr.press/v139/chami21a.html

[9] Chami, I., Ying, Z., Ré, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 4869–4880 (2019)

[10] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989)

[11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database . In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 248–255. IEEE Computer Society, Los Alamitos, CA, USA (2009), https://doi.ieeecomputersociety.org/10.1109/CVPR.2009.5206848

[12] Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, R.: Hyperbolic image-text representations. In: International Conference on Machine Learning (2023)

[13] Fan, A., Grave, E., Joulin, A.: Reducing transformer depth on demand with structured dropout (2019), https://arxiv.org/abs/1909.11556

[14] Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic neural networks. In: Advances in neural information processing systems. pp. 5345–5355 (2018)

[15] Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, 3rd edn. (1996)

[16] Gu, A., Sala, F., Gunel, B., Ré, C.: Learning mixed-curvature representations in products of model spaces. In: International Conference on Learning Representations (2019)

[17] Gu, J., Zhai, S., Zhang, Y., Susskind, J., Jaitly, N.: Matryoshka diffusion models (2024), https://arxiv.org/abs/2310.15111

[18] Hamenstädt, U.: Entropy-rigidity of locally symmetric spaces of negative curvature. Annals of Mathematics **131**, 35–51 (1990)

[19] Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Networks **4**(2), 251–257 (1991)

[20] Hou, L., Huang, Z., Shang, L., Jiang, X., Chen, X., Liu, Q.: Dynabert: Dynamic bert with adaptive width and depth (2020), https://arxiv.org/abs/2004.04037

[21] Hu, W., Dou, Z.Y., Li, L.H., Kamath, A., Peng, N., Chang, K.W.: Matryoshka query transformer for large vision-language models (2024), https://arxiv.org/abs/2405.19315

[22] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002)

[23] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus (2017), https://arxiv.org/abs/1702.08734

[24] Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., Lempitsky, V.: Hyperbolic image embeddings (2020), https://arxiv.org/abs/1904.02239

[25] Kim, G., Cho, K.: Length-adaptive transformer: Train once with length drop, use anytime with search (2021), https://arxiv.org/abs/2010.07003

[26] Kusupati, A., et al.: Matryoshka representation learning. In: Advances in Neural Information Processing Systems (2022)

[27] Lang, K.: NewsWeeder: Learning to filter netnews. Tech. rep., International Conference on Machine Learning (1995)

[28] Law, M., Liao, R., Snell, J., Zemel, R.: Lorentzian distance learning for hyperbolic representations. In: International Conference on Machine Learning (2019)

[29] Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: Nv-embed: Improved techniques for training llms as generalist embedding models (2025), https://arxiv.org/abs/2405.17428

[30] Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W., Luan, Y., Duddu, S.M.K., Abrego, G.H., Shi, W., Gupta, N., Kusupati, A., Jain, P., Jonnalagadda, S.R., Chang, M.W., Naim, I.: Gecko: Versatile text embeddings distilled from large language models (2024), https://arxiv.org/abs/2403.20327

[31] Lee, J.M.: Introduction to Riemannian Manifolds. Springer, 2nd edn. (2018)

[32] Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., Mehdad, Y.: MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In: Conference of the European Chapter of the Association for Computational Linguistics (2021)

[33] Linial, N., London, E., Rabinovich, Y.: The geometry of graphs and some of its algorithmic applications. Combinatorica **15**(2), 215–245 (1995)

[34] Liu, Q., Nickel, M., Kiela, D.: Hyperbolic graph neural networks. In: Advances in Neural Information Processing Systems. pp. 8230–8241 (2019)

[35] Maggie, Culliton, P., Chen, W.: Tweet Sentiment Extraction. Kaggle competition (2020), https://kaggle.com/competitions/tweet-sentiment-extraction

[36] Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., Balahur, A.: WWW'18 open challenge: Financial opinion mining and question answering. In: Companion Proceedings of the Web Conference (2018)

[37] Mishne, G., Wan, Z., Wang, Y., Yang, S.: The numerical stability of hyperbolic representation learning (2024), https://arxiv.org/abs/2211.00181

[38] Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive text embedding benchmark. In: Conference of the European Chapter of the Association for Computational Linguistics (2023)

[39] Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations (2017), https://arxiv.org/abs/1705.08039

[40] Nickel, M., Kiela, D.: Learning continuous hierarchies in the lorentz model of hyperbolic geometry (2018), https://arxiv.org/abs/1806.03417

[41] Nussbaum, Z., Morris, J.X., Duderstadt, B., Mulyar, A.: Nomic embed: Training a reproducible long context text embedder (2025), https://arxiv.org/abs/2402.01613

[42] OpenAI: New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/ (2024), released January 25, 2024

[43] Ratcliffe, J.G.: Foundations of Hyperbolic Manifolds. Springer, 2nd edn. (2006)

[44] Sala, F., De Sa, C., Gu, A., Ré, C.: Representation tradeoffs for hyperbolic embeddings. In: International Conference on Machine Learning (2018)

[45] Sarkar, R.: Low distortion delaunay embedding of trees in hyperbolic plane. In: Proceedings of the 19th International Conference on Graph Drawing. p. 355–366. GD'11, Springer-Verlag, Berlin, Heidelberg (2011), https://doi-org.gate.lib.buffalo.edu/10.1007/978-3-642-25878-7_34

[46] Shimizu, R., Mukuta, Y., Harada, T.: Hyperbolic neural networks++. In: International Conference on Learning Representations (2021)

[47] Sturua, S., Mohr, I., Kalim Akram, M., Günther, M., Wang, B., Krimmel, M., Wang, F., Mastrapas, G., Koukounas, A., Wang, N., Xiao, H.: Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V. p. 123–129. Springer-Verlag, Berlin, Heidelberg (2025), https://doi-org.gate.lib.buffalo.edu/10.1007/978-3-031-88720-8_21

[48] Tifrea, A., Bécigneul, G., Ganea, O.E.: Poincaré GloVe: Hyperbolic word embeddings. In: International Conference on Learning Representations (2019)

[49] Tulkens, S., van Dongen, T.: Model2vec: Fast state-of-the-art static embeddings (2024), https://github.com/MinishLab/model2vec

[50] Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press (2018)

[51] Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. In: Conference on Empirical Methods in Natural Language Processing (2020)

[52] Wen, T., Wang, Y., Zeng, Z., Peng, Z., Su, Y., Liu, X., Chen, B., Liu, H., Jegelka, S., You, C.: Beyond matryoshka: Revisiting sparse coding for adaptive representation (2025), https://arxiv.org/abs/2503.01776

[53] Yoon, J., Sinha, R., Arik, S.O., Pfister, T.: Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions (2024), https://arxiv.org/abs/2407.20243

[54] Yu, P., Merrick, L., Nuti, G., Campos, D.: Arctic-embed 2.0: Multilingual retrieval without compromise (2024), https://arxiv.org/abs/2412.04506

[55] Yue, Y., Lin, F., Yamada, K.D., Zhang, Z.: Hyperbolic contrastive learning (2023), https://arxiv.org/abs/2302.01409

[56] Zhang, B., Chen, L., Liu, T., Zheng, B.: Smec: Rethinking matryoshka representation learning for retrieval embedding compression (2025), https://arxiv.org/abs/2510.12474

[57] Zhang, D., Li, J., Zeng, Z., Wang, F.: Jasper and stella: distillation of sota embedding models (2025), https://arxiv.org/abs/2412.19048

A Full Analysis: Projector Complexity vs. Embedding Dimension

This section provides the complete propositions, definitions, corollaries, theorem, and proof sketch supporting Sec. 4 of the main paper. Throughout, $E : \mathcal{X} \to \mathbb{R}^D$ denotes the frozen encoder, P is the projector MLP, $W \in \mathbb{R}^{D_{\max} \times D}$ the elastic weight matrix, $z = E(x)$, and $\bar{z}_{1:d} = \mathrm{P}(z)\, W_{1:d}^\top$ the d-dimensional embedding.

A.1 Rank Bottleneck and Semantic Subspace Misalignment

Proposition A.1 (Rank constraint on linear projections)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^D$ be the* semantic subspace *of E, i.e. the minimal subspace such that the restriction $E|_{\mathcal{S}}$ preserves class-conditional structure, and let $s = \dim(\mathcal{S})$. If $d < s$, then no linear map $W_{1:d} \in \mathbb{R}^{d \times D}$ can recover $\mathcal{S}$ without information loss. Moreover, rank sub-multiplicativity [15] gives*

$$\mathrm{rank}(W_{1:d} \circ L) \;\le\; \min(d, \mathrm{rank}(L)) \;\le\; d$$

for any linear $L \in \mathbb{R}^{D \times D}$, so composing with a linear P does not increase the effective rank. When $\mathcal{S}$ is not axis-aligned, the optimal rank-d approximation is the projection onto the leading d eigenvectors of the data covariance restricted to $\mathcal{S}$ [15].

Corollary A.2 (Necessity of a nonlinear projector at low d)**.** *When $d < s$ and $\mathcal{S}$ is not axis-aligned, a linear P cannot align $W_{1:d}$ with the semantically relevant directions. A nonlinear P of depth ≥ 2 and hidden width $\ge D$ can approximate the optimal rotation onto $\mathcal{S}$ to arbitrary precision on compact sets by the universal approximation theorem [10, 19], reducing the residual problem for W to a near-axis-aligned projection. At $d \ge s$, the rank budget exceeds $\dim(\mathcal{S})$ and a linear P suffices; over-parameterizing P then only introduces optimization noise.*

The practical implication is consistent with the theory of Matryoshka Representation Learning [26]: coarse prefixes must encode maximum semantic content, placing the heaviest geometric burden on $W_{1:d}$ precisely when d is small.

A.2 Hyperbolic Volume Collapse at Low Dimension

The key motivation for working in $\mathbb{L}_c^d$ is the exponential volume growth of geodesic balls in $\mathbb{H}^d$,

$$\mathrm{Vol}_d(r) \;\propto\; \sinh^{d-1}(\sqrt{c}\, r) \;\approx\; \tfrac{1}{2^{d-1}}\, e^{(d-1)\sqrt{c}\, r}, \quad r \to \infty \tag{4}$$

compared with $\mathrm{Vol}_d^{\mathbb{R}}(r) \propto r^d$ in Euclidean space [6, 43]. When d is small, the factor $(d-1)$ in the exponent substantially diminishes this separation advantage.

Proposition A.3 (Effective packing capacity in $\mathbb{H}_c^d$)**.** *Fix curvature $c > 0$, ball radius $R > 0$, and separation margin $\varepsilon > 0$. The maximum number of ε-separated points packable in a hyperbolic ball of radius R satisfies*

$$N(R, \varepsilon, d) \;\le\; \left(\frac{\sinh(\sqrt{c}\, R)}{\sinh(\sqrt{c}\, \varepsilon)} \right)^{d-1} \tag{5}$$

following from a standard covering argument using (4) *[18]. The ratio $N(R, \varepsilon, d)/N(R, \varepsilon, d')$ grows exponentially in $d - d'$ for fixed R, ε, confirming that packing capacity is highly sensitive to d at small values.*

Definition A.4 (Volumetric dimension threshold)**.** *Given K classes, the* volumetric threshold d_{vol} *is the minimum dimension required for K classes to admit an ε-separated packing inside a hyperbolic ball of radius R, i.e. the smallest integer d satisfying $N(R, \varepsilon, d) \ge K$. Applying* (5) *and solving yields*

$$d_{\mathrm{vol}} \;=\; 1 \;+\; \left\lceil \frac{\log K}{\log \sinh(\sqrt{c}\, R) - \log \sinh(\sqrt{c}\, \varepsilon)} \right\rceil \tag{6}$$

Corollary A.5 (Diminishing geometric margin at low d)**.** *For $d < d_{\mathrm{vol}}$, $\mathbb{H}_c^d$ cannot geometrically accommodate K ε-separated class clusters regardless of P. The loss landscape becomes geometrically closer to Euclidean, and the exponential inter-class separation that motivates hyperbolic embeddings is substantially weakened. A complex projector compensates by engineering explicit class separation* before *the exponential map, so that the limited hyperbolic volume is used efficiently.*

This is consistent with [44], who demonstrate that hyperbolic embedding fidelity degrades rapidly below $d \approx 10$–50 for tree-structured data, and with [16], who observe that effective curvature must be increased to compensate for small d.

Failure mode	Active when	Root cause	Fix (low d)
Rank bottleneck	$d < s$	Linear P cannot rotate $\mathcal{S}$ to axis-aligned	Deep nonlinear P
Volume collapse	$d < d_{\text{vol}}$	K classes not ε-separable in $\mathbb{H}^d_c$	Class-aware P
Norm collapse	$d < d_{\text{norm}}$	$\|\bar{z}\|_2 < r^*$; geometry degenerates to Euclidean	Scale-amplifying P

Table 2: Summary of the three failure modes, their dependence on target dimension d, and the corresponding projector remedy.

A.3 Norm Collapse and the Flat Neighbourhood of the Origin

The Lorentz exponential map at the origin $o = (1/\sqrt{c},\, \mathbf{0}) \in \mathbb{L}^d_c$ is

$$\exp^c_o(v) \;=\; \cosh\!\big(\sqrt{c}\,\|v\|_{\mathcal{L}}\big)\, o \;+\; \sinh\!\big(\sqrt{c}\,\|v\|_{\mathcal{L}}\big)\, \frac{v}{\sqrt{c}\,\|v\|_{\mathcal{L}}} \tag{7}$$

where $\|v\|_{\mathcal{L}} = \sqrt{\langle v, v\rangle_{\mathcal{L}}}$ controls the radial position on $\mathbb{L}^d_c$.

Proposition A.6 (Norm concentration at low dimension)**.** *Let $\bar{z} \in \mathbb{R}^d$ have i.i.d. coordinates with zero mean and variance σ^2. By concentration of measure [50], $\big|\, \|\bar{z}\|_2 - \sigma\sqrt{d}\,\big| \;\le\; O(\sigma)$ with high probability, so $\mathbb{E}[\|\bar{z}\|_{\mathcal{L}}] \approx \sigma\sqrt{d}$. For $d \ll D$, this quantity is small, placing $\exp^c_o(\bar{z})$ near o where the sectional curvature experienced by the embedding is negligible [28].*

Definition A.7 (Norm dimension threshold)**.** *Let $\delta > 0$ control the minimum acceptable non-Euclidean distortion. Define the* critical radius

$$r^* \;=\; \frac{1}{\sqrt{c}}\,\operatorname{arccosh}(1+\delta) \tag{8}$$

the boundary of the flat neighbourhood of o. The norm threshold *is*

$$d_{\text{norm}} \;=\; \left\lceil \frac{(r^*)^2}{\sigma^2} \right\rceil \;=\; \left\lceil \frac{\operatorname{arccosh}^2(1+\delta)}{c\,\sigma^2} \right\rceil \tag{9}$$

For $d < d_{\text{norm}}$, we have $\mathbb{E}[\|\bar{z}\|_2] < r^*$, so embeddings are expected to lie inside the flat neighbourhood.

Remark A.8 (Degeneration to Euclidean metric in the flat neighbourhood)**.** *For $\|\bar{z}\|_{\mathcal{L}} < r^*$, the Taylor expansions $\cosh(r) \approx 1 + r^2/2$ and $\sinh(r) \approx r$ give $\exp^c_o(\bar{z}) \approx o + \bar{z}$, so the Lorentz geodesic distance reduces to $d_{\mathbb{L}}(u_i, u_j) \approx \|\bar{z}_i - \bar{z}_j\|_2$. The contrastive loss therefore carries no hyperbolic inductive bias in this regime, negating the motivation for the Lorentz model.*

Corollary A.9 (Projector must amplify norms at low d)**.** *For $d < d_{\text{norm}}$, a complex P must explicitly scale embeddings so that $\|P(z)\|_{\mathcal{L}} \ge r^*$. A linear layer will not learn this non-trivial amplification reliably, because the contrastive loss gradient does not distinguish Euclidean-regime collapse from correct hyperbolic placement when embeddings are near o. At $d \ge d_{\text{norm}}$, the norm $\sigma\sqrt{d} \ge r^*$ intrinsically, so a simple projector suffices. A practical diagnostic is to monitor u_0, if $u_0 \approx 1/\sqrt{c}$ at low d, norm collapse is confirmed.*

$$u_0 \;=\; \cosh\!\big(\sqrt{c}\,\|\bar{z}\|_{\mathcal{L}}\big) \;\gg\; \frac{1}{\sqrt{c}}; \tag{10}$$

A.4 Formal Threshold Theorem and Proof

Theorem A.10 (Projector complexity threshold)**.** *Let $s = \dim(\mathcal{S})$ with $\mathcal{S}$ not axis-aligned, and let d_{vol}, d_{norm} be as in Definitions A.4–A.7. Define the* critical dimension

$$d^* \;=\; \max\!\big(s,\; d_{\text{vol}},\; d_{\text{norm}}\big) \tag{11}$$

(i) ***Necessity of a complex projector below*** *d^*. For any $d < d^*$, a linear P is insufficient: at least one of the following active failure modes prevents faithful hyperbolic embedding regardless of W:*

- Rank bottleneck: $\mathrm{rank}(W_{1:d} \circ L) \leq d < s$ *for any linear* L *(Corollary A.2);*
- Volume collapse: $N(R, \varepsilon, d) < K$ *(Corollary A.5);*
- Norm collapse: $\mathbb{E}[\|\bar{z}\|_2] < r^*$*, so* $d_{\mathbb{L}} \approx \|\cdot\|_2$ *(Corollary A.9).*

A projector of depth ≥ 2 *and hidden width* $\geq D$ *resolves all three by the universal approximation theorem [10, 19].*

(ii) ***Sufficiency of a simple projector above*** d^***.** *For* $d \geq d^*$*, all failure conditions are absent. The SGD bias-variance decomposition [3] gives* $\mathrm{MSE}(\hat{g}) = \mathrm{bias}^2 + \sigma^2_{n_P}/N$ *where* $\mathrm{bias} = 0$ *and* $\sigma^2_{n_P}$ *grows monotonically with projector parameter count* n_P*; minimising* n_P *therefore minimises total error.*

Proof. (i) Necessity. The three failure modes are independent; $d < d^*$ activates at least one by definition of d^* as their minimum. For the rank bottleneck: if $d < s$, rank sub-multiplicativity [15] gives $\mathrm{rank}(W_{1:d} \circ L) \leq d < s$ for any linear L, so a linear P cannot align $W_{1:d}$ with $\mathcal{S}$. For volume collapse: if $d < d_{\mathrm{vol}}$, Definition A.4 yields $N(R, \varepsilon, d) < K$ directly from (5), so K-class separation does not exist in $\mathbb{H}^d_c$ regardless of P. For norm collapse: if $d < d_{\mathrm{norm}}$, Definition A.7 gives $\mathbb{E}[\|\bar{z}\|_2] < r^*$, and Remark A.8 shows the Lorentz metric degenerates to Euclidean. A depth ≥ 2, width $\geq D$ projector resolves each failure mode by the universal approximation theorem [19].

(ii) Sufficiency. When $d \geq d^*$, all three failure conditions are absent by definition. The bias is therefore zero and the $\mathcal{O}(n_P/N)$ variance term from [3] grows with projector capacity, making a simple P strictly preferable. □

Remark A.11 (Practical estimation of d^*)**.** *Given a pretrained encoder* E*,* s *can be estimated as the numerical rank of the empirical covariance* $\hat{\Sigma} = \frac{1}{N}\sum_i z_i z_i^\top$ *at a relative tolerance of* 10^{-2} *[15].* d_{vol} *is computable from* K*,* R *(set to the* 95*th percentile of* $\|u_i\|$ *in a pilot run), and* ε *(set to the minimum intra-class geodesic distance in a pilot run).* d_{norm} *requires only* c*,* σ^2 *(estimated from pilot embeddings), and tolerance* δ *(we suggest* $\delta = 0.1$*). For ImageNet-scale encoders with* $D = 768$*,* $K = 1000$*, and* $c = 1$*, these formulas yield* $d^* \approx 32$–64*, consistent with the experimental observations in Sec. 5 of the main paper.*

Dim	Full Rep.	Rand. FS	SVD	MRL	MRL-E	Unsup. HERL	Sup. HERL	CSR
8	-	2.36	19.14	62.19	57.45	63.90	**64.60**	67.78
16	-	12.06	46.02	67.91	67.05	67.89	**69.90**	69.17
32	-	32.91	60.78	69.46	68.60	70.20	**71.23**	70.15
64	-	49.91	67.04	70.17	69.61	70.71	**71.14**	70.94
128	-	60.91	69.63	70.52	70.12	71.00	**71.72**	70.99
256	-	65.75	70.67	70.62	70.36	71.20	**71.89**	71.31
512	-	68.77	71.06	70.82	70.74	71.52	**72.05**	71.29
1024	-	70.41	71.22	70.89	71.07	71.70	**72.20**	71.30
2048	71.19	71.19	71.21	70.97	71.21	72.00	**72.29**	71.20
3072	-	-	-	-	-	72.11	**72.49**	72.04
4096	-	-	-	-	-	**72.19**	72.09	71.87
8192	-	-	-	-	-	72.34	**72.56**	72.11

Table 3: **1-NN accuracy.** Comparison of different methods on ImageNet-1k for 1-NN retrieval accuracy. Note that, CSR [52] uses an ambient dimension of 8192 (4x), activating only certain dimensions. Hence, has been omitted from comparison. The highest score has been **bolded**.

B Experiments on image representations

B.1 Datasets

We utilize ImageNet-1K [11], a large-scale visual dataset widely used for developing and benchmarking computer vision models. It comprises 1,000 object categories, each containing a diverse collection of images. In total, the dataset includes 1,281,167 training images, 50,000 validation images, and 100,000 test images.

B.2 Implementation details

We follow MRL [26], and utilize ResNet50 weights to extract encoder representations for our HERL model, adopting the same preprocessing procedures as well. The projector module is realized through a small MLP consisting of four layers ($2048 \to 1024 \to 512 \to 256$). The output of the projector is used to learn elastic representations. All experiments have been performed on a single H100 GPU. Note that, we also clip the norm of the representations to 5.0 before Lorentzian projection to avoid numerical instability errors. For larger dimensions, we perform graident accumulation to allow operating at the batch size. All hyperbolic operations have been performed using the geoopt library. See Table 4 for more specific details.

B.3 Results

Exact numerical values for Top-1 classification accuracy and 1-NN retrieval are shown in Table 5 and Table 3.

C Experiments on text representations

C.1 Datasets

We evaluate on a broad slice of the Massive Text Embedding Benchmark [38], covering three task families that stress different aspects of an elastic embedding: classification, clustering, and retrieval. All datasets are accessed through the MTEB release; for brevity we cite their original sources below and refer the reader to [38] for the exact splits and evaluation protocols.

Classification. We report accuracy on three datasets spanning intent detection, sentiment, and topic prediction in both English and multilingual settings. MTOPINTENT [32] is a multilingual task-oriented parsing benchmark with $\sim$100,000 utterances over six languages and eleven domains. BANKING77 [7] contains 13,083 customer-service queries labelled with 77 fine-grained banking intents. TWEETSENTIMENTEXTRACTION [35] provides ternary sentiment labels (positive / neutral / negative) on tweets, released originally as a Kaggle competition.

Clustering. We further evaluate on three clustering tasks at varying granularity and document length. BIORXIVCLUSTERINGS2S, and BIORXIVCLUSTERINGP2P [38] cluster papers by category using either titles alone (S2S) or title-plus-abstract (P2P), exposing the adaptor to short and long inputs respectively. TWENTYNEWSGROUPSCLUSTERING [27] clusters 20-class newsgroup titles across ten splits of 1,000 to 10,000 documents.

Retrieval. We measure nDCG@10 on three retrieval tasks covering domain QA, fact-checking, argumentation, and paraphrase identification. FIQA-2018 [36] is an aspect-based sentiment and opinion QA benchmark in the financial domain. NFCORPUS [2] contains 3,244 natural-language

Backbone	Input dim	LR	Epochs	Batch size	Optimizer	Curvature
ResNet50	2048	1e-3	10	2048	RiemannianAdam	0.9

Table 4: Implementation details for experiments on image representations with ImageNet-1k

Dim	Full Rep.	Rand. LP	SVD	MRL	MRL-E	Unsup. HERL	Sup. HERL	CSR
8	-	4.56	2.34	66.63	56.66	73.52	**74.16**	73.62
16	-	11.29	7.17	73.53	71.94	74.75	**75.04**	74.75
32	-	27.21	20.46	75.03	74.48	75.48	**76.02**	75.44
64	-	49.47	48.10	75.82	75.35	76.01	**76.53**	75.88
128	-	65.70	67.24	76.30	75.80	76.29	**76.88**	76.24
256	-	72.43	74.59	76.47	76.22	76.55	**77.01**	76.25
512	-	74.94	76.78	76.65	76.36	76.78	**77.18**	76.34
1024	-	76.10	76.87	76.76	76.48	76.99	**77.31**	76.54
2048	76.87	76.87	-	76.80	76.51	77.03	**77.45**	76.52

Table 5: **Top-1 accuracy.** Comparison of different methods on ImageNet-1k using a linear probe for classification. Note that, CSR [52] uses an embedding dimension of 8192 (4x) and activate only certain dimensions. Hence, has been omitted from comparison. The highest score has been **bolded**.

medical queries with 169,756 relevance judgments over 9,964 documents. SCIFACT [51] pairs 1,400 expert-written scientific claims with evidence-containing abstracts annotated for veracity and rationales.

C.2 Implementation Details

We follow CSR [52] and utilize NV-Embed-V2 [29] to extract encoder representations for our HERL model, adopting the same preprocessing procedures as well. All experiments have been performed on a single H100 GPU. Note that, we also clip the norm of the representations to 5.0 before Lorentzian projection to avoid numerical instability errors. For larger dimensions, we perform graident accumulation to allow operating at the batch size. All hyperbolic operations have been performed using the geoopt library. See Table 6 for more specific details. For classification tasks, the projector module is an MLP made of four layers ($4096 \to 2048 \to 1024 \to 512$). Clustering tasks use a smaller network ($4096 \to 2048 \to 1024$). On the other hand, retrieval tasks utilize the same architecture as classification tasks.

C.3 Results

Exact numerical values for all tasks: classification, clustering and retrieval are shown in Table 7.

Dataset	Backbone	Input dim	LR	Epochs	Batch size	Optimizer	Curvature
MTOPIntent Banking77 TweetSentiment	NV-Embed-V2	4096	4e-5	10	128	RiemannianAdam	0.9
BiorxivP2P	NV-Embed-V2	4096	1e-3	10	128	RiemannianAdam	0.9
BiorxivS2S			1e-3				
TwentyNews			1e-3				
FIQA	NV-Embed-V2	4096	1e-3	10	256	RiemannianAdam	1.5
NFCorpus			1e-3				0.8
SciFACT			1e-3				1.0

Table 6: Implementation details for experiments on text representations across nine datasets.

Category	Model	Dim	Text Classification Top-1 Acc (%)↑			Text Clustering Top-1 Acc (%)↑			Text Retrieval NDCG@10 (%)↑		
			MTOPIntent	Banking77	TweetSentiment	BiorxivP2P	BiorxivS2S	TwentyNews	FiQA2018	NFCorpus	SciFACT
Full Rep	NV-Embed-V2	4096	93.58	92.20	79.73	53.61	49.60	64.82	62.65	43.97	77.93
MRL	Stella-1.5B-v5	256	90.45	86.14	76.75	50.81	46.42	60.07	55.59	36.97	**77.48**
	Jina-v3	256	78.81	84.08	73.81	38.14	34.39	51.96	55.73	36.63	66.63
	Nomic-Embed-V1.5	256	72.47	83.69	59.20	38.19	31.83	48.56	35.00	32.54	68.24
	Gecko-Embed-004(Google)	256	77.82	86.01	72.97	36.28	33.09	50.60	55.54	37.81	70.86
	Text-Embed-3-L (OpenAI)	256	70.45	83.19	58.98	35.43	33.86	54.24	50.33	37.94	73.10
	Arctic-Embed-L-V2	256	67.69	80.99	59.06	34.25	34.07	30.06	44.69	35.02	69.51
	M2V-Base-Glove	256	59.26	72.39	50.02	32.26	22.34	25.38	11.82	23.15	50.66
	Jina-V3	64	68.12	67.98	71.18	36.89	33.57	50.22	44.18	33.66	68.84
	Nomic-Embed-V1.5	64	62.77	80.63	55.23	34.81	44.61	48.06	10.22	18.96	36.55
	Potion-Base-2M	64	42.50	65.17	52.52	25.78	14.94	27.07	32.08	30.72	64.28
Ours (w/ NV-Embed-V2)	Unsup. HERL	32	94.39	92.88	79.51	48.18	48.47	62.85	50.70	45.67	55.70
	Unsup. HERL	8192	93.72	93.21	79.34	78.43	75.22	64.70	**61.11**	**55.23**	71.62
	Sup. HERL	32	**94.53**	**94.86**	**79.88**	63.28	75.82	80.14	-	-	-
	Sup. HERL	8192	93.49	93.15	79.60	**80.12**	**76.29**	**82.77**	-	-	-
Sparse	CSR	32	89.86	91.02	78.55	53.49	49.13	63.05	57.54	38.06	71.17

Table 7: **Results on textual representations.** We compare the downstream utility of representations across three tasks: classification, clustering and retrieval. Note that, CSR [52] uses a 4x ambient dimension, activating only certain dimensions. Hence, has been omitted from comparison. The highest score has been **bolded**.

D Effect of Projector Module

We also explored a tied encoder–decoder architecture without the projector module. Empirically, this configuration led to a substantial degradation in performance for both low-dimensional representations and expanded high-dimensional representations beyond the encoder dimensionality, even under extensive fine-tuning. This suggests that a single linear weight matrix lacks the expressive capacity required to effectively map Euclidean representations onto hyperbolic manifolds, thereby limiting the model's ability to fully exploit the geometric structure and representational advantages of hyperbolic space.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: We propose a novel framework, Hyperbolic Elastic Representation Learning (HERL) to learn nested and elastic representations without retraining the encoder, unlike Matryoshka Representation Learning that requires encoder retraining.

   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: HERL requires aggressive finetuning on retrieval tasks due to the metric mismatch. To keep consistent with evaluation metrics that normalize representations (project to hypersphere) before computing 1-NN accuracy, we report scores using the same protocol. However, normalizing hyperbolic optimized representations erase radial information, which is a primal source of separation in hyperbolic geometry. This has been substantiated in the main paper.

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: Theory backed by emperical results has been shown in the main paper and appendix.

   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: Yes, we provide dataset-specific details to allow full reproducibility. Further, code will be released upon acceptance.

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 do not propose or create a new dataset. All datasets are publicly available and code will be released upon acceptance.

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: Yes, dataset-specific hyperparameters have been discussed in appendix.

   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: We have carefully followed the protocols from prior works in reporting all metrics.

   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: All experiments have been completed on a single H100 GPU, this has been mentioned in the appendix as well.

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, we adhere to the NeurIPS Code of Ethics and preserve anonymity throughout the paper.

   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: We have shown throughout the paper that HERL allows efficient and effective representations, saving hours of training time. Thereby, drastically reducing energy requirements.

    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: N/A

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: Yes, we have cited all dataset sources, code and models.

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: We do not release asset along with the paper.

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: No experiments involved crowdsourcing or 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: N/A

    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: [N/A]

    Justification: N/A

    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.