ProCLIP: Product Space Multimodal Contrastive Alignment

Jiakai Chen[1], Hangke Sui[2]

1 Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign

2 Electrical and Computer Engineering, University of Illinois Urbana-Champaign

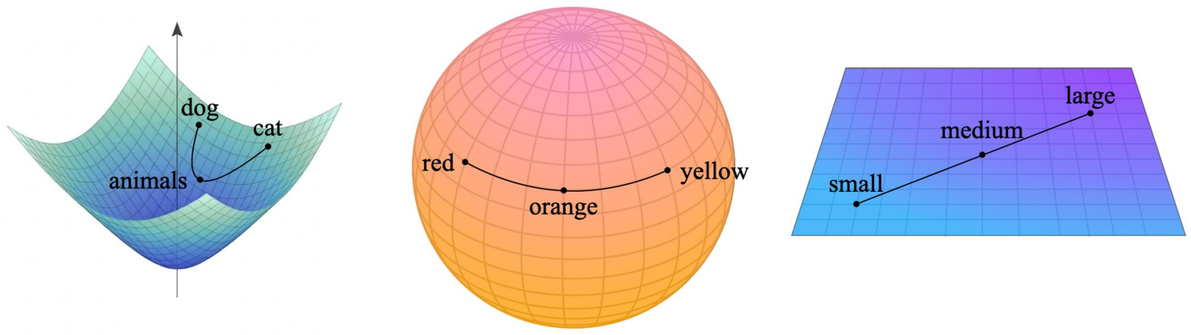
TL;DR We propose ProCLIP, a geometry-aware contrastive framework that embeds multimodal data into a mixed-curvature product space. By combining hyperbolic, spherical, and Euclidean geometries, ProCLIP captures heterogeneous semantic structures and improves cross-modal alignment beyond single-manifold embeddings.

Why Geometry Matters in Multimodal Learning

Contrastive learning (e.g., CLIP) has become the standard paradigm for multimodal representation learning. However, most existing methods embed all modalities into a single latent geometry, typically Euclidean with cosine similarity. This assumption is fundamentally limiting: **Not all semantics live in the same geometry.**

Real-world multimodal semantics exhibit heterogeneous structure:

- Hierarchical relations (e.g., animal → dog)
- Directional similarity (e.g., color hue)
- Continuous variation (e.g., size, intensity)



A single geometry cannot simultaneously capture these diverse structures, leading to representation mismatch and suboptimal alignment. **We need a representation space that can model multiple geometric structures simultaneously.**

=> Can a mixture of geometries provide a better inductive bias?

Core Idea: Product Space Representation

To address this, we represent multimodal embeddings in a product of geometries:

$$z \in \mathcal{Z} := \mathbb{H}^{r_H} \times \mathbb{R}^{r_E} \times \mathbb{S}^{r_S}$$

Each component captures a different semantic structure:

- Hyperbolic space → hierarchy
- Spherical space → directional similarity
- Euclidean space → continuous variation

This decomposes representation into multiple geometry factors, and enables geometry-aware alignment beyond single-manifold representations.

Hyperbolic Component

We follow Lorentz model to represent the hyperbolic space with a hyperboloid.

Minkowski metric:

$$\langle \mathbf{u}, \mathbf{v} \rangle_L = \sum_{k=1}^{r_H} u_k v_k - u_0 v_0.$$

r_H-dimensional hyperbolic space with curvature c:

$$\mathbb{H}_c^{r_H} = \left\{ \mathbf{z} \in \mathbb{R}^{r_H+1} : \langle \mathbf{z}, \mathbf{z} \rangle_L = -\frac{1}{c}, \;\; z_0 > 0 \right\}.$$

Geodesic distance for $\mathbf{z_1}, \mathbf{z_2} \in \mathbb{H}_c^{r_H}$:

$$d_{\mathbb{H}}(\mathbf{z_1}, \mathbf{z_2}) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(-c \langle \mathbf{z_1}, \mathbf{z_2} \rangle_L\right).$$

Spherical Component

r_S-dimensional sphere:

$$\mathbb{S}^{r_S} = \{ \mathbf{s} \in \mathbb{R}^{r_S+1} : \|\mathbf{s}\|_2 = R \}.$$

Geodesic distance for $\mathbf{z_1}, \mathbf{z_2} \in \mathbb{S}_c^{r_S}$:

$$d_{\mathbb{S}}(\mathbf{s_1}, \mathbf{s_2}) = R \arccos\left(\frac{\langle \mathbf{s_1}, \mathbf{s_2} \rangle}{R^2}\right).$$

Euclidean Component

$$d_{\mathbb{E}}(\mathbf{e_1}, \mathbf{e_2}) = \|\mathbf{e_1} - \mathbf{e_2}\|_2, \mathbf{e_1}, \mathbf{e_2} \in \mathbb{R}^{r_E}.$$

Weighted Product Metric

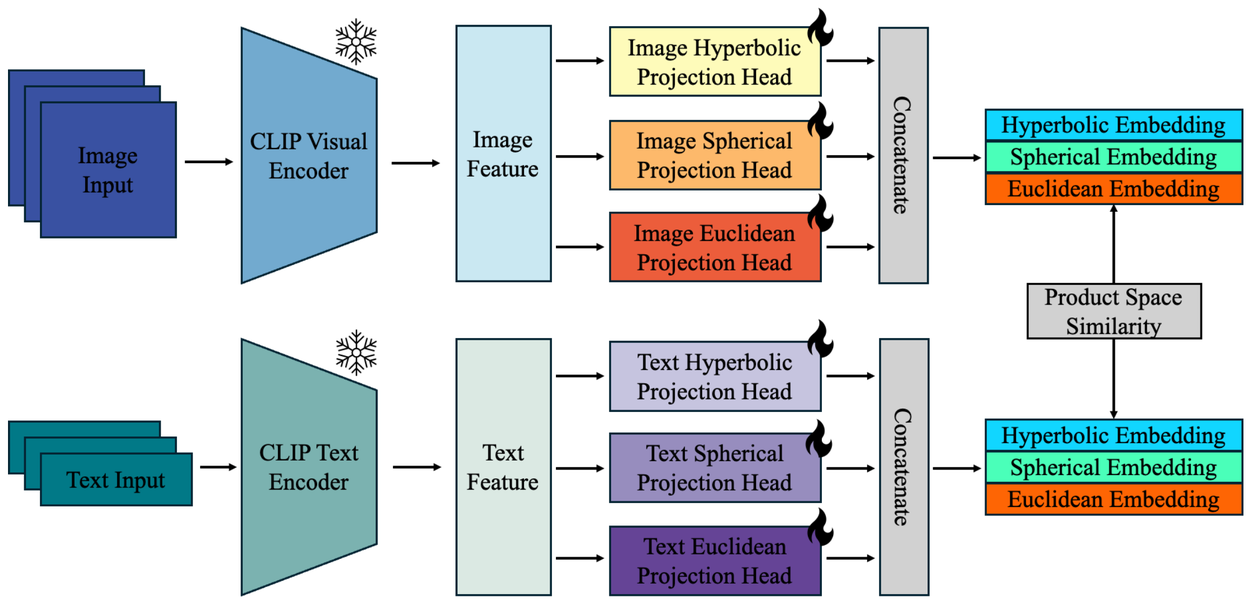
$$d_{\mathcal{Z}}^2(\mathbf{z}_1, \mathbf{z}_2) := \alpha_H \, d_{\mathbb{H}}^2\left(\mathbf{z}_1^{(\mathrm{H})}, \mathbf{z}_2^{(\mathrm{H})}\right) + \alpha_E \left\|\mathbf{z}_1^{(\mathrm{E})} - \mathbf{z}_2^{(\mathrm{E})}\right\|_2^2 + \alpha_S \, d_{\mathbb{S}}^2\left(\mathbf{z}_1^{(\mathrm{S})}, \mathbf{z}_2^{(\mathrm{S})}\right).$$

Contrastive Alignment

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp\left(s(\mathbf{z_i}, \widetilde{\mathbf{z}}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s(\mathbf{z_i}, \widetilde{\mathbf{z}}_j)/\tau\right)} - \sum_{i=1}^{N} \log \frac{\exp\left(s(\mathbf{z_i}, \widetilde{\mathbf{z}}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s(\mathbf{z_j}, \widetilde{\mathbf{z}}_i)/\tau\right)}$$

Experiments and Results

We build on frozen CLIP embeddings and extend them with geometry-specific projection heads. We compute similarity using product-space distance and train with a standard contrastive objective.



- Dataset: Flickr30k, MSCOCO
- Backbone: CLIP ViT-B/32 (frozen)

ProCLIP consistently outperforms single-geometry baselines with the same total dimensions.

Table 1: **Retrieval on Flickr30k and MSCOCO** (Recall@K). We report R@1 and R@10 for image-to-text and text-to-image. Best results within each dataset and dimension block are in **bold**.

Dim.	Model	Flickr30k				MSCOCO			
		Image → Text		Text → Image		Image → Text		Text → Image	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
64×3	ProCLIP	**0.8080**	0.9780	**0.6428**	**0.9380**	**0.6198**	**0.9266**	**0.4703**	**0.8344**
	Spherical	0.7740	**0.9810**	0.6392	0.9354	0.5962	0.9136	0.4448	0.8284
	Hyperbolic	0.7390	0.9720	0.6024	0.9164	0.5952	0.9094	0.4388	0.8064
	Euclidean	0.6890	0.9460	0.5432	0.8914	0.5326	0.8756	0.3886	0.7718
128×3	ProCLIP	**0.8070**	**0.9880**	**0.6634**	**0.9386**	**0.6448**	**0.9282**	**0.4849**	**0.8446**
	Spherical	0.7770	0.9830	0.6480	0.9360	0.5856	0.9152	0.4439	0.8258
	Hyperbolic	0.7860	0.9760	0.6298	0.9270	0.5932	0.9070	0.4387	0.8086
	Euclidean	0.7220	0.9550	0.5950	0.9064	0.5548	0.8818	0.4013	0.7830
256×3	ProCLIP	**0.8320**	**0.9870**	**0.6752**	**0.9496**	**0.6696**	**0.9408**	**0.5072**	**0.8578**
	Spherical	0.7710	0.9820	0.6498	0.9344	0.5970	0.9186	0.4479	0.8296
	Hyperbolic	0.7780	0.9780	0.6400	0.9324	0.5990	0.9096	0.4446	0.7172
	Euclidean	0.7480	0.9600	0.5994	0.9134	0.5650	0.8900	0.4114	0.7905

Takeaway

Modeling representations in a mixed-curvature space provides a more expressive and geometry-aware inductive bias, leading to improved cross-modal alignment over single-manifold embeddings.

Geometry > Dimensionality