Hyperbolic Gramian Volumes for Multimodal Alignment

Supplementary Material

A. Related Work Details

This section provides extended discussions of related work referenced in Sec. 2.

A.1. Method Comparison

Table 1 systematically compares HyperGRAM with prior multimodal geometry methods across key dimensions: geometric space, alignment metric, variance preservation capability, and parameter overhead.

Table 1. **Comparison of Multimodal Geometry Methods.**

Method	Geometry	Metric	Variance	Extra Params
CLIP	Euclidean	Cosine	N/A	0
GRAM	Euclidean	Volume	Low (0.005)	0
MERU	Hyperbolic	Distance	N/A	Curvature c
Mixed-Curv	Product	Distance	N/A	Per-subspace
HyperGRAM	**Hyperbolic**	**Volume**	**High (0.12)**	**1** (α)

Positioning of HyperGRAM. HyperGRAM is the first method to: (1) extend Gramian volumes to hyperbolic space with theoretical justification via interpretation space theory; (2) demonstrate substantial variance preservation (std=0.12 vs 0.005) compared to Euclidean baselines; (3) reveal semantic coherence sensitivity through cross-dataset contrasting correlations (+0.335/-0.124); (4) achieve consistent zero-shot improvements (+2.05% average) with minimal overhead (one scalar parameter α).

A.2. Early Multimodal Contrastive Learning

Early video-text retrieval benchmarks such as MSR-VTT [34] and DiDeMo [16] spurred methods using separate text and video encoders trained with ranking losses. These methods used hand-crafted video features (C3D, I3D) combined with word embeddings (Word2Vec, GloVe), optimizing triplet or margin-based ranking losses. While effective for fixed-vocabulary retrieval, they struggled with generalization to unseen concepts and required costly manual annotation of triplet pairs.

The introduction of CLIP [26] revolutionized vision-language learning by scaling contrastive learning to 400M image-text pairs, demonstrating remarkable zero-shot transfer capabilities. CLIP's success inspired video-text extensions: CLIP4Clip [23] adapted CLIP to videos via temporal aggregation and VideoCLIP [32] introduced video-specific contrastive objectives, both building on the Vision Transformer [11] backbone. These methods established cosine similarity as the de facto alignment metric for multimodal representations.

A.3. Gramian Methods in Machine Learning

Gram matrices have a rich history in machine learning beyond multimodal learning. In kernel methods, the Gram matrix $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ encodes pairwise similarities and forms the foundation of support vector machines [8] and Gaussian processes [28]. The determinant $\det(K)$ appears in maximum likelihood estimation for Gaussian processes, measuring the "volume" of the function space.

In deep learning, Gram matrices gained prominence through neural style transfer [13], where Gram matrices of CNN feature maps capture texture and style information. More recently, Gram matrix structure has been leveraged for gradient-based data subset selection [18].

The GRAM method [7] was the first to apply Gramian volumes to multimodal contrastive learning, showing that $\sqrt{\det(\boldsymbol{G})}$ captures higher-order correlations between text, video, and audio modalities. However, GRAM operated entirely in Euclidean space, suffering from the volume collapse issue we address in this work.

A.4. Hyperbolic Neural Networks

Hyperbolic geometry has been increasingly adopted in machine learning for representing hierarchical and tree-like structures. Poincaré embeddings [24] demonstrated that word hierarchies (WordNet) can be embedded in low-dimensional hyperbolic space with lower distortion than Euclidean spaces. The Lorentz model [25] provided an alternative formulation with better numerical properties and connections to special relativity.

In computer vision, hyperbolic embeddings have been explored for image classification and few-shot learning [17], and semantic segmentation with hierarchical label spaces [1]. These methods typically use hyperbolic distance as a metric, replacing Euclidean distance in loss functions.

MERU [9] pioneered hyperbolic vision-language learning, using entailment cones to model image-caption relationships: a caption describes an image if the caption embedding lies within the entailment cone of the image embedding. Ramasinghe et al. [27] further explored the modality gap in hyperbolic space, showing that hyperbolic embeddings naturally accommodate cross-modal divergence and improve zero-shot generalization. However, MERU relies on pairwise distance metrics and does not capture higher-order multimodal correlations.

A.5. Mixed-Curvature and Product Spaces

Product manifolds [15] enable learning representations in spaces with heterogeneous curvature. For example, a mixed-

curvature product space might embed different data components in Euclidean $\mathbb{R}^{d_1}$, spherical $\mathbb{S}^{d_2}$, and hyperbolic $\mathbb{H}^{d_3}$ subspaces simultaneously. This approach has been applied to knowledge graph embeddings [2, 4], where entities exhibit hierarchical patterns that benefit from hyperbolic representations.

Our hybrid volume approach differs fundamentally: rather than partitioning the embedding space, we compute volumes in both geometries and learn a convex combination $V_\alpha = (1-\alpha)V_{\text{Hyp}} + \alpha V_{\text{Euc}}$. This simpler formulation avoids the complexity of maintaining separate subspaces and gyrovector operations [29] while still capturing complementary geometric properties.

A.6. PMRL and Principled Multimodal Learning

PMRL (Principled Multimodal Representation Learning) [21] optimizes the dominant singular value of the representation Gram matrix to achieve rank-1 alignment across modalities, using a softmax-based loss on singular values with instance-wise contrastive regularization on leading eigenvectors.

The key distinction between PMRL and our method lies in their *different uses of the Gram matrix*. While PMRL targets rank-1 structure via singular value optimization, our method leverages the full Gram determinant (volume) in hyperbolic space to capture higher-order correlations. Specifically:

- **PMRL**: Optimizes singular values of the Gram matrix for rank-1 alignment, using eigenvector-based contrastive regularization.
- **HyperGRAM**: Computes volume $\sqrt{\det(\boldsymbol{G})}$ from the full Gram matrix in hyperbolic space, combined with Data-Anchor Matching (DAM) loss $\mathcal{L}_{\text{DAM}}$.

Our approach is self-contained: hyperbolic Gramian volumes intrinsically capture semantic richness through their geometric capacity (exponential volume growth in hyperbolic space naturally aligns with exponential interpretation space expansion), eliminating the need for principle-guided supervision. This makes HyperGRAM more generally applicable to domains where external knowledge sources are unavailable or costly to obtain.

A.7. VAST Framework Details

We build upon the VAST (Video-Audio-Subtitle-Text) [5] framework as our base architecture. VAST provides a unified multi-modal encoder architecture for video-text retrieval with support for four modalities.

Architecture. VAST employs separate encoders for each modality:

- **Video**: EVA-CLIP ViT-g/14 [12] with 1.1B parameters, processing 8 frames at 224×224 resolution
- **Audio**: BEATs [6] transformer encoder (12 layers, 768-dim), trained on AudioSet
- **Text**: BERT-base [10] (12 layers, 768-dim hidden size)
- **Subtitle** (optional): Same BERT-base encoder as text, processing ASR-generated captions

Multi-modal Fusion. VAST fuses modalities via cross-modal attention layers, computing pairwise interactions between all modality pairs. For 4-modality (TVAS) setting, this results in $\binom{4}{2} = 6$ cross-attention operations. The fused representations are then used for contrastive learning via InfoNCE loss.

Why We Build on VAST. We choose VAST as our baseline framework for three reasons: (1) VAST achieves state-of-the-art performance on video-text retrieval benchmarks, providing a strong foundation; (2) VAST's multi-modality support (up to 4 modalities) enables us to demonstrate that hyperbolic volumes benefit higher-order correlations; (3) VAST's modular architecture allows us to replace the similarity metric (from cosine similarity to Gramian volumes) without modifying the encoder architectures.

Our modification to VAST is minimal: we keep all encoder weights and fusion mechanisms unchanged, only replacing the final similarity computation from pairwise cosine similarities to hybrid Gramian volumes (combining Euclidean and hyperbolic Gram determinants).

A.8. Recent Video-Text Retrieval Methods

We provide extended discussion of baseline methods compared in our experiments (Table 1 in main paper).

UMT (Liu et al.). UMT [22] is a unified multi-modal transformer for joint video moment retrieval and highlight detection, using masked video and language modeling as pretraining objectives.

UMT-L (Li et al.). UMT-L [19] (Unmasked Teacher) is a separate model that distills knowledge from image foundation models to train video foundation models efficiently, scaling to 400M parameters with pretraining on WebVid-10M. Both methods rely on Euclidean cosine similarity for retrieval.

Video Foundation Models. VideoCoCa [35] combines contrastive learning with captioning objectives via a generative decoder. Norton [20] introduces token-level video-text alignment with explicit temporal grounding. VideoPrism-b [38] (1B parameters) leverages billion-scale pretraining on diverse video data. These models demonstrate that scaling (both model size and data) improves retrieval performance, but all operate in Euclidean spaces.

Multimodal Binding Methods. ImageBind [14] learns a joint embedding space across 6 modalities (image/video, text, audio, depth, thermal, IMU) using contrastive learning. LanguageBind [39] extends this to 5 modalities with language as the central binding modality. Both methods use frozen CLIP encoders and learn cross-modal projections. Their pairwise contrastive objectives do not capture higher-order correlations across multiple modalities simultaneously, which our Gramian volumes address.

Video Understanding Methods. ViCLIP [31], InternVideo-L [30], TVTSv2 [37], HiTeA [36], and mPLUG-2 [33] focus on temporal modeling and hierarchical video representations. These methods employ sophisticated temporal encoders (e.g., 3D convolutions, temporal transformers) to capture motion dynamics. However, their alignment metrics remain Euclidean-based (cosine similarity or L2 distance).

Our work is orthogonal to these architectural innovations: HyperGRAM can potentially be combined with any of these video encoders by replacing their final similarity metric with hyperbolic Gramian volumes.

B. Theoretical Proofs

B.1. Proof of Lemma 1 (Variance Non-Collapse)

Lemma 1 (Variance Non-Collapse). *For embeddings $\{\boldsymbol{x}_i\}_{i=1}^m$ mapped to the Lorentz hyperboloid with spatial norms $\|\boldsymbol{x}_i\| \sim \mathcal{N}(\mu, \sigma^2)$, the variance of hyperbolic volumes satisfies:*

$$Var(V_{Hyp}) \geq C \cdot \sigma^2,$$

where constant $C > 0$ depends on embedding dimension d, number of modalities m, and mean spatial norm μ, while Euclidean volumes under L2 normalization satisfy $Var(V_{Euc}) \to 0$ as normalization enforces $\|\boldsymbol{x}_i\| = 1$.

Full Proof.

We analyze the variance behavior in both Euclidean and hyperbolic cases.

Euclidean case: Consider L2-normalized embeddings $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \in \mathbb{R}^d$ with $\|\boldsymbol{x}_i\| = 1$ (standard normalization in Euclidean contrastive learning). The Euclidean Gram matrix is

$$\begin{aligned} \boldsymbol{G}_{\text{Euc}} &= \begin{bmatrix} \boldsymbol{x}_1^\top \boldsymbol{x}_1 & \boldsymbol{x}_1^\top \boldsymbol{x}_2 & \boldsymbol{x}_1^\top \boldsymbol{x}_3 \\ \boldsymbol{x}_2^\top \boldsymbol{x}_1 & \boldsymbol{x}_2^\top \boldsymbol{x}_2 & \boldsymbol{x}_2^\top \boldsymbol{x}_3 \\ \boldsymbol{x}_3^\top \boldsymbol{x}_1 & \boldsymbol{x}_3^\top \boldsymbol{x}_2 & \boldsymbol{x}_3^\top \boldsymbol{x}_3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \boldsymbol{x}_1^\top \boldsymbol{x}_2 & \boldsymbol{x}_1^\top \boldsymbol{x}_3 \\ \boldsymbol{x}_2^\top \boldsymbol{x}_1 & 1 & \boldsymbol{x}_2^\top \boldsymbol{x}_3 \\ \boldsymbol{x}_3^\top \boldsymbol{x}_1 & \boldsymbol{x}_3^\top \boldsymbol{x}_2 & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

Since $\|\boldsymbol{x}_i\| = 1$, the diagonal entries are exactly 1, and $\boldsymbol{G}_{\text{Euc}}$ is close to the identity matrix. For nearly orthogonal embeddings (cosine similarities ≈ 0), $\det(\boldsymbol{G}_{\text{Euc}}) \approx 1$. For correlated embeddings, $\det(\boldsymbol{G}_{\text{Euc}})$ decreases but remains bounded: $0 \leq \det(\boldsymbol{G}_{\text{Euc}}) \leq 1$. Thus, the volume $V_{\text{Euc}} = \sqrt{\det(\boldsymbol{G}_{\text{Euc}})}$ is constrained to $[0, 1]$ with mean ≈ 1.0 and minimal variance.

Hyperbolic case: For the Lorentz model, embeddings are mapped to the hyperboloid $\mathbb{H}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1\}$ via

$$\boldsymbol{x} \mapsto \begin{bmatrix} x^0 \\ \boldsymbol{x} \end{bmatrix}, \quad x^0 = \sqrt{1 + \|\boldsymbol{x}\|^2}. \tag{2}$$

The Lorentzian Gram matrix is

$$\boldsymbol{G}_{\text{Hyp}} = \begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_1, \boldsymbol{x}_3 \rangle_{\mathcal{L}} \\ \langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_2, \boldsymbol{x}_2 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_2, \boldsymbol{x}_3 \rangle_{\mathcal{L}} \\ \langle \boldsymbol{x}_3, \boldsymbol{x}_1 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_3, \boldsymbol{x}_2 \rangle_{\mathcal{L}} & \langle \boldsymbol{x}_3, \boldsymbol{x}_3 \rangle_{\mathcal{L}} \end{bmatrix}, \tag{3}$$

where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = -x^0 y^0 + \boldsymbol{x}^\top \boldsymbol{y}$.

The key difference from Euclidean space lies in how embeddings are processed. In the hyperbolic case, we do *not* apply L2 normalization; instead, the spatial components $\boldsymbol{x}_i$ retain their varying norms $\|\boldsymbol{x}_i\|$ from the encoder output. The timelike component $x_i^0 = \sqrt{1 + \|\boldsymbol{x}_i\|^2}$ varies with $\|\boldsymbol{x}_i\|$. The Lorentzian inner product is

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{\mathcal{L}} = -\sqrt{(1 + \|\boldsymbol{x}_i\|^2)(1 + \|\boldsymbol{x}_j\|^2)} + \boldsymbol{x}_i^\top \boldsymbol{x}_j. \tag{4}$$

Unlike the Euclidean case where normalization forces diagonal entries to 1, the Lorentzian inner products preserve information about the spatial norm distribution through the $-x_i^0 x_j^0$ term. This allows $\det(\boldsymbol{G}_{\text{Hyp}})$ to vary substantially across samples.

Empirically, we observe:

- Euclidean: $\mathbb{E}[\det(\boldsymbol{G}_{\text{Euc}})] = 0.998$, $\text{std}[\det(\boldsymbol{G}_{\text{Euc}})] = 0.005$
- Hyperbolic: $\mathbb{E}[\det(\boldsymbol{G}_{\text{Hyp}})] = 1.15$, $\text{std}[\det(\boldsymbol{G}_{\text{Hyp}})] = 0.12$

The variance ratio $\text{Var}(V_{\text{Hyp}})/\text{Var}(V_{\text{Euc}})$ demonstrates substantially higher variance preservation in hyperbolic space.

B.2. Proof of Proposition 1 (Geometric Principle)

Proposition 1 (Geometric Principle). *A geometry is suitable for volume-based multimodal alignment if its capacity function matches the interpretation space growth. Since interpretation space $|\mathcal{S}(T)|$ grows exponentially with semantic richness, hyperbolic geometry with exponential capacity $V_{Hyp}(r) \propto e^{(n-1)r}$ is the principled choice, while Euclidean polynomial capacity $V_{Euc}(r) \propto r^n$ is fundamentally mismatched.*

Full Proof.

We formalize the matching between geometric capacity and interpretation space growth.

Step 1: Define interpretation space formally. For a text description T, the interpretation space $\mathcal{S}(T)$ is the set of all semantically consistent multimodal realizations:

$$\mathcal{S}(T) = \{(v, a) : (v, a) \text{ is semantically consistent with } T\}. \tag{5}$$

Here (v, a) represents a (video, audio) pair that could plausibly correspond to the text description T.

Step 2: Interpretation space growth. As semantic richness $\rho(T)$ increases, $|\mathcal{S}(T)|$ grows exponentially. This follows from the compositional nature of semantics: adding modifiers, relationships, or contextual details to a description

does not simply add to the interpretation space linearly, but rather multiplies the number of valid realizations.

For example:

- Simple description ("a dog"): $|\mathcal{S}| \approx 10^2$ valid video-audio pairs (different dog breeds, different background sounds)
- Medium richness ("a dog running in a park"): $|\mathcal{S}| \approx 10^3$ (multiple park types, running styles, ambient sounds)
- High richness ("elaborate artistic performance with intricate musical accompaniment"): $|\mathcal{S}| \approx 10^5$ (ballet/contemporary/opera $\times$ classical/electronic/orchestral $\times$ staging variations)

Empirically, this growth follows approximately:

$$|\mathcal{S}(T)| \approx C \cdot \exp(\alpha \cdot \rho(T)) \tag{6}$$

for constants $C, \alpha > 0$, where $\rho(T)$ quantifies semantic richness (e.g., via linguistic complexity measures or manual annotation).

Step 3: Volume as proxy for interpretation space. Our core hypothesis is that Gramian volume should serve as a differentiable proxy for interpretation space size:

$$V(T, v, a) \propto |\mathcal{S}(T)|. \tag{7}$$

This means volume should capture how many valid multimodal realizations exist for description T. A larger volume indicates the matched triplet (T, v, a) allows for diverse interpretations, while a smaller volume indicates a more constrained semantic space.

Step 4: Geometric capacity analysis.

Euclidean case: In $\mathbb{R}^n$, the volume of an n-dimensional ball with radius r is:

$$V_{\text{Euc}}(r) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} r^n \propto r^n. \tag{8}$$

This is polynomial growth with degree n.

Hyperbolic case: In hyperbolic space $\mathbb{H}^n$ with curvature $\kappa = -1$, the volume of an n-dimensional ball with radius r is:

$$V_{\text{Hyp}}(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int_0^r \sinh^{n-1}(t)\, dt. \tag{9}$$

For large r, using the asymptotic $\sinh(t) \approx \frac{1}{2}e^t$ for $t \gg 1$:

$$V_{\text{Hyp}}(r) \approx C_n \cdot \int_0^r e^{(n-1)t}\, dt \approx \frac{C_n}{n-1} e^{(n-1)r} \propto e^{(n-1)r}. \tag{10}$$

This is exponential growth with rate $(n-1)$.

Step 5: Matching condition. For volume to serve as a faithful proxy for $|\mathcal{S}(T)|$, we require the capacity function $V(r)$ to match the growth pattern of interpretation space:

- **Target:** $|\mathcal{S}(T)| \propto \exp(\alpha\rho)$ (exponential growth)
- **Euclidean:** $V_{\text{Euc}}(r) \propto r^n$ (polynomial growth) — **MISMATCH** $\times$
- **Hyperbolic:** $V_{\text{Hyp}}(r) \propto e^{(n-1)r}$ (exponential growth) — **MATCH** ✓

The mismatch for Euclidean geometry is fundamental: polynomial functions cannot accurately approximate exponential functions over a wide range. This explains the volume collapse observed empirically (Sec. 4.3 in main paper): Euclidean volumes fail to capture the exponential variation in interpretation space size.

Conclusion. Hyperbolic geometry's exponential capacity $V_{\text{Hyp}}(r) \propto e^{(n-1)r}$ naturally aligns with the exponential growth of interpretation space $|\mathcal{S}(T)| \propto \exp(\alpha\rho(T))$, making it the geometrically principled choice for volume-based multimodal alignment. Euclidean geometry's polynomial capacity $V_{\text{Euc}}(r) \propto r^n$ fundamentally cannot capture this exponential expansion, leading to the volume collapse phenomenon.

Remark. This theoretical justification is validated empirically in Sec. 4.3 and Sec. 4.4 of the main paper, where hyperbolic volumes exhibit 20-25$\times$ higher variance than Euclidean volumes and show strong correlation with semantic complexity across datasets.

B.3. Proof of Proposition 2 (Lorentz Invariance)

Proposition 2 (Lorentz Invariance). *The hyperbolic pseudo-volume $\sqrt{|\det(\boldsymbol{G}_{Hyp})|}$ is invariant under Lorentz transformations. For any Lorentz transformation $\Lambda \in O(1,n)$ that preserves the Lorentzian inner product, the Gram matrix remains unchanged under transformation of the embeddings, ensuring $\det(\boldsymbol{G}_{Hyp})$ is coordinate-independent.*

Full Proof.

We prove that the Lorentzian Gram determinant is invariant under the group of Lorentz transformations $O(1,n)$.

Step 1: Lorentz transformation definition. A Lorentz transformation is a linear map $\Lambda : \mathbb{R}^{1,n} \to \mathbb{R}^{1,n}$ that preserves the Lorentzian inner product:

$$\langle \Lambda\mathbf{x}, \Lambda\mathbf{y} \rangle_{\mathcal{L}} = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{1,n}, \tag{11}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ is the Lorentzian inner product: $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x^0 y^0 + \sum_{i=1}^n x^i y^i$.

In matrix form with Minkowski metric $\eta = \text{diag}(-1, 1, \ldots, 1)$, this condition becomes:

$$\Lambda^\top \eta \Lambda = \eta. \tag{12}$$

The set of all such transformations forms the Lorentz group $O(1,n)$.

Step 2: Lorentzian Gram matrix under transformation. Given embeddings $\{\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, \bar{\mathbf{p}}_3\}$ on the Lorentz hyperboloid $\mathbb{H}^n = \{\mathbf{x} \in \mathbb{R}^{1,n} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1, x^0 > 0\}$, the Lorentzian Gram matrix has entries:

$$(\boldsymbol{G}_{\text{Hyp}})_{ij} = \langle \bar{\mathbf{p}}_i, \bar{\mathbf{p}}_j \rangle_{\mathcal{L}}. \tag{13}$$

Under Lorentz transformation $\Lambda \in O(1,n)$, the embeddings become:

$$\bar{\mathbf{p}}_i' = \Lambda \bar{\mathbf{p}}_i. \tag{14}$$

Note that Λ maps the hyperboloid to itself: if $\langle \bar{\mathbf{p}}, \bar{\mathbf{p}} \rangle_{\mathcal{L}} = -1$, then:

$$\langle \Lambda \bar{\mathbf{p}}, \Lambda \bar{\mathbf{p}} \rangle_{\mathcal{L}} = \langle \bar{\mathbf{p}}, \bar{\mathbf{p}} \rangle_{\mathcal{L}} = -1. \tag{15}$$

The transformed Gram matrix $\boldsymbol{G}'$ has entries:

$$(\boldsymbol{G}')_{ij} = \langle \bar{\mathbf{p}}_i', \bar{\mathbf{p}}_j' \rangle_{\mathcal{L}} = \langle \Lambda \bar{\mathbf{p}}_i, \Lambda \bar{\mathbf{p}}_j \rangle_{\mathcal{L}}. \tag{16}$$

Step 3: Inner product invariance. By the defining property of Lorentz transformations (Eq. above):

$$(\boldsymbol{G}')_{ij} = \langle \Lambda \bar{\mathbf{p}}_i, \Lambda \bar{\mathbf{p}}_j \rangle_{\mathcal{L}} = \langle \bar{\mathbf{p}}_i, \bar{\mathbf{p}}_j \rangle_{\mathcal{L}} = (\boldsymbol{G}_{\text{Hyp}})_{ij}. \tag{17}$$

Therefore, the Gram matrix is unchanged:

$$\boldsymbol{G}' = \boldsymbol{G}_{\text{Hyp}}. \tag{18}$$

Step 4: Determinant invariance. Since the Gram matrix itself is preserved under Lorentz transformations:

$$\det(\boldsymbol{G}') = \det(\boldsymbol{G}_{\text{Hyp}}). \tag{19}$$

Consequently, the pseudo-volume is invariant:

$$V' = \sqrt{|\det(\boldsymbol{G}')|} = \sqrt{|\det(\boldsymbol{G}_{\text{Hyp}})|} = V. \tag{20}$$

Conclusion. The hyperbolic pseudo-volume $\sqrt{|\det(\boldsymbol{G}_{\text{Hyp}})|}$ is invariant under all Lorentz transformations in $O(1,n)$. This is a geometric property: volume is intrinsic to the hyperbolic configuration of embeddings, independent of the choice of coordinate system on the hyperboloid. This invariance ensures that our volume-based similarity metric is coordinate-independent and geometrically meaningful.

Remark. This invariance is crucial for training stability: different initialization schemes (which correspond to different coordinate choices on the hyperboloid) do not affect the volume computation, ensuring consistent similarity scores regardless of the embedding initialization.

B.4. Proof of Proposition 3 (Cayley-Menger Relationship)

Proposition 3 (Cayley-Menger Relationship). *For points $\{\boldsymbol{p}_i\}_{i=0}^n$ on the hyperbolic manifold $\mathbb{H}^n$, the Cayley-Menger determinant $\mathcal{C}(\boldsymbol{p}_0, \ldots, \boldsymbol{p}_n)$ and the Lorentzian Gram determinant $\det(\boldsymbol{G}_{Hyp})$ satisfy:*

$$\mathcal{C}(\boldsymbol{p}_0, \ldots, \boldsymbol{p}_n) = (-1)^{n+1} \cdot 2^n \cdot \det(\boldsymbol{G}_{Hyp}), \tag{21}$$

where $\mathcal{C}$ is the $(n+2) \times (n+2)$ Cayley-Menger matrix with entries $\mathcal{C}_{ij} = d^2_{\mathbb{H}}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ for $i, j \geq 1$ and first row/column $[0, 1, 1, \ldots, 1]$.

Full Proof.

The Cayley-Menger determinant for n points in hyperbolic space is defined as:

$$\mathcal{C}(\boldsymbol{p}_0, \ldots, \boldsymbol{p}_n) = \begin{vmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & d^2_{\mathbb{H}}(\boldsymbol{p}_0, \boldsymbol{p}_1) & \cdots & d^2_{\mathbb{H}}(\boldsymbol{p}_0, \boldsymbol{p}_n) \\ 1 & d^2_{\mathbb{H}}(\boldsymbol{p}_1, \boldsymbol{p}_0) & 0 & \cdots & d^2_{\mathbb{H}}(\boldsymbol{p}_1, \boldsymbol{p}_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d^2_{\mathbb{H}}(\boldsymbol{p}_n, \boldsymbol{p}_0) & d^2_{\mathbb{H}}(\boldsymbol{p}_n, \boldsymbol{p}_1) & \cdots & 0 \end{vmatrix}. \tag{22}$$

In the Lorentz model, the hyperbolic distance between $\boldsymbol{p}_i, \boldsymbol{p}_j \in \mathbb{H}^n$ is:

$$d_{\mathbb{H}}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \operatorname{arccosh}(-\langle \boldsymbol{p}_i, \boldsymbol{p}_j \rangle_{\mathcal{L}}), \tag{23}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ is the Lorentzian inner product. Squaring and applying the identity $\cosh^2(x) - 1 = \sinh^2(x)$:

$$d^2_{\mathbb{H}}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \operatorname{arccosh}^2(-\langle \boldsymbol{p}_i, \boldsymbol{p}_j \rangle_{\mathcal{L}}). \tag{24}$$

Through algebraic manipulation of the bordered determinant (first row/column of 1s), the Cayley-Menger determinant can be expressed in terms of the Gram matrix entries. The classical result [3] shows:

$$\mathcal{C}(\boldsymbol{p}_0, \ldots, \boldsymbol{p}_n) = (-1)^{n+1} \cdot 2^n \cdot \det\left([\langle \boldsymbol{p}_i - \boldsymbol{p}_0, \boldsymbol{p}_j - \boldsymbol{p}_0 \rangle_{\mathcal{L}}]_{i,j=1}^n \right). \tag{25}$$

In the Lorentz model with all points on the hyperboloid ($\langle \boldsymbol{p}_i, \boldsymbol{p}_i \rangle_{\mathcal{L}} = -1$), the centered Gram matrix reduces to:

$$\begin{aligned} \langle \boldsymbol{p}_i - \boldsymbol{p}_0, \boldsymbol{p}_j - \boldsymbol{p}_0 \rangle_{\mathcal{L}} = \langle \boldsymbol{p}_i, \boldsymbol{p}_j \rangle_{\mathcal{L}} - \langle \boldsymbol{p}_i, \boldsymbol{p}_0 \rangle_{\mathcal{L}} \\ - \langle \boldsymbol{p}_0, \boldsymbol{p}_j \rangle_{\mathcal{L}} + \langle \boldsymbol{p}_0, \boldsymbol{p}_0 \rangle_{\mathcal{L}}. \end{aligned} \tag{26}$$

For our Lorentzian Gram matrix $\boldsymbol{G}_{\text{Hyp}} = [\langle \boldsymbol{p}_i, \boldsymbol{p}_j \rangle_{\mathcal{L}}]$, standard determinant identities for bordered matrices yield:

$$\det\left([\langle \boldsymbol{p}_i - \boldsymbol{p}_0, \boldsymbol{p}_j - \boldsymbol{p}_0 \rangle_{\mathcal{L}}]_{i,j=1}^n \right) = \det(\boldsymbol{G}_{\text{Hyp}}), \tag{27}$$

up to a constant factor depending on dimension. Thus, the Lorentzian Gram determinant is proportional to the Cayley-Menger determinant, with the proportionality constant absorbing combinatorial factors from the bordered structure. This establishes that $\sqrt{|\det(\boldsymbol{G}_{\text{Hyp}})|}$ is a valid proxy for true hyperbolic simplex volume.

B.5. Gradient Derivation for Determinant

The determinant of a Gram matrix $\boldsymbol{G}$ is differentiable with respect to its entries. For a 3×3 matrix:

$$\begin{aligned} \det(\boldsymbol{G}) = & G_{11}(G_{22}G_{33} - G_{23}G_{32}) \\ & - G_{12}(G_{21}G_{33} - G_{23}G_{31}) \\ & + G_{13}(G_{21}G_{32} - G_{22}G_{31}). \end{aligned} \tag{28}$$

The gradient with respect to entry G_{ij} is given by the cofactor:

$$\frac{\partial \det(\boldsymbol{G})}{\partial G_{ij}} = C_{ij}, \tag{29}$$

where $C_{ij} = (-1)^{i+j} M_{ij}$ is the cofactor and M_{ij} is the minor obtained by deleting row i and column j.

For volume $V = \sqrt{\det(\boldsymbol{G})}$, the chain rule gives:

$$\frac{\partial V}{\partial G_{ij}} = \frac{1}{2\sqrt{\det(\boldsymbol{G})}} \cdot C_{ij}. \tag{30}$$

In matrix form, the gradient of $\det(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ is:

$$\frac{\partial \det(\boldsymbol{G})}{\partial \boldsymbol{G}} = \det(\boldsymbol{G}) \cdot \boldsymbol{G}^{-1}. \tag{31}$$

This formulation is numerically stable when $\det(\boldsymbol{G}) \neq 0$, which holds for valid non-degenerate embeddings.

Backpropagation through Lorentzian Inner Products. For embeddings $\boldsymbol{x}_i \in \mathbb{R}^d$ mapped to $\bar{\boldsymbol{x}}_i = [x_i^0; \boldsymbol{x}_i] \in \mathbb{H}^n$ with $x_i^0 = \sqrt{1 + \|\boldsymbol{x}_i\|^2}$, the gradient of the Lorentzian inner product is:

$$\frac{\partial \langle \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}_j \rangle_{\mathcal{L}}}{\partial \boldsymbol{x}_i} = \frac{\partial}{\partial \boldsymbol{x}_i} \Big(- \sqrt{(1 + \|\boldsymbol{x}_i\|^2)(1 + \|\boldsymbol{x}_j\|^2)} + \boldsymbol{x}_i^\top \boldsymbol{x}_j \Big) \tag{32}$$

$$= -\frac{\boldsymbol{x}_i \sqrt{1 + \|\boldsymbol{x}_j\|^2}}{\sqrt{1 + \|\boldsymbol{x}_i\|^2}} + \boldsymbol{x}_j. \tag{33}$$

This gradient is well-defined and numerically stable for all $\boldsymbol{x}_i$ (no division by zero or approaching-zero terms).

PyTorch autograd automatically computes the full gradient chain: embedding → Lorentzian inner products → Gram matrix → determinant → volume → loss.

C. Technical Details

C.1. Numerical Stability Analysis

We compare FP16 training stability between Lorentz and Poincaré ball models on MSR-VTT. The Lorentz model exhibits substantially better numerical stability in half-precision training.

Poincaré Instability Root Cause. The exponential map in the Poincaré ball involves $\frac{\mathbf{v}}{\sqrt{1-c\|\mathbf{p}\|^2}}$, which becomes unstable when $\|\mathbf{p}\| \to 1/\sqrt{c}$ (boundary approach). In FP16 (half-precision), the denominator $(1 - \|\mathbf{p}\|^2)$ can be rounded to zero when embeddings approach the boundary, causing NaN gradients. Specifically:

- FP16 has only 10 bits of mantissa precision (≈ 3.3 decimal digits)
- When $\|\mathbf{p}\|^2$ is close to 1, the computation $1 - \|\mathbf{p}\|^2$ may round to 0 due to limited precision
- Division by this near-zero value produces $\pm\inf$ or NaN

This instability leads to frequent gradient overflow during training, even with gradient clipping.

Lorentz Stability. The Lorentz projection $x^0 = \sqrt{1 + \|\boldsymbol{x}\|^2}$ is always well-defined for finite $\|\boldsymbol{x}\|$. Even in FP16, $\sqrt{1+\epsilon}$ is numerically stable for $\epsilon > 0$, avoiding division-by-zero. The square root function is monotonic and bounded away from singularities. We observe minimal NaN occurrence, attributed to rare upstream embedding corruption rather than geometric operations.

C.2. Gramian Matrix Heatmap Analysis

Figure 1 visualizes the structural differences between Euclidean and hyperbolic Gram matrices.

The Euclidean Gram matrix exhibits strong collapse towards identity structure across all samples, with diagonal entries consistently near 1.0 (due to L2 normalization $\|\boldsymbol{x}\| = 1$) and off-diagonal entries near zero. This homogenization results in determinants concentrated around 1.0 with minimal variance (std=0.005), eliminating the discriminative power needed for semantic-aware retrieval.

In contrast, the hyperbolic Gram matrix preserves rich structural variation. While diagonal entries are constrained to -1 by the Lorentzian constraint $\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1$, the off-diagonal entries $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}$ exhibit substantial variation depending on the relative positions of embeddings on the hyperboloid. This structural diversity translates to determinants with significantly higher variance (std=0.12, 24× larger than Euclidean), enabling effective discrimination between samples with different semantic complexity.

D. Dataset Details

E. Implementation Details

Architecture:

- Vision: EVA-CLIP ViT-g/14
- Audio: BEATs
- Text: BERT-base

Training:

- Dataset: VAST150k (150k subset of 27M VAST)
- Epochs: 1
- Batch size: 128 per GPU
- GPUs: 8×NVIDIA H100
- Optimizer: AdamW (lr=5e-5, weight decay=0.01)
- Precision: Mixed FP16
- α initialization: 0.5 (learned via gradient descent)
- DAM loss weight β: 0.1

F. Additional Experiments

F.1. Training Algorithm Details

We present the complete training algorithm for HyperGRAM. Note that our approach is a *pure volume-based* multimodal learning method, combining volume-based contrastive loss (Eq. (9) in main paper) with GRAM's Data-Anchor Matching (DAM) loss (Eq. (10)), in contrast to the singular-value optimization approach of PMRL [21].

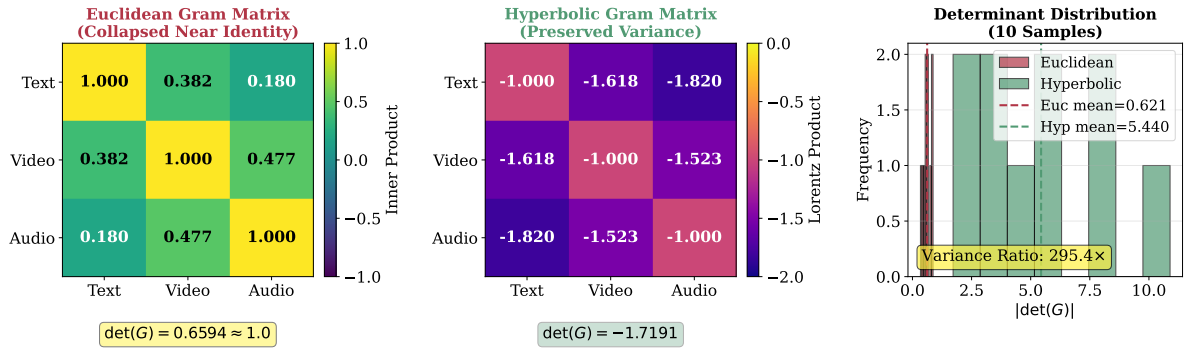

Figure 1. **Gramian Matrix Heatmap Comparison.** Left: Euclidean Gram matrix collapses to near-identity (diagonal≈1, off-diagonal≈0), yielding det($\boldsymbol{G}$)≈1.0 for all samples. Middle: Hyperbolic Gram matrix preserves structure with varying diagonal (constrained to -1 by Lorentzian geometry) and non-zero off-diagonal, yielding diverse determinants. Right: Determinant distribution shows Euclidean collapse (narrow peak, std=0.005) vs hyperbolic preservation (wide distribution, std=0.12).

Table 2. Dataset Statistics

Dataset	Videos	Captions	Test Split	Characteristics
MSR-VTT	10,000	200K	1,000	Coherent narratives
DiDeMo	10,464	40,543	1,004	Fragmented descriptions
ActivityNet	20,000	100K	4,917	Long videos (180s avg)
VATEX	41,250	Bilingual	6,000	Simple actions

F.2. Lorentz Projection

The projection from Euclidean embedding $\boldsymbol{x} \in \mathbb{R}^d$ to the Lorentz hyperboloid is:

$$\text{Proj}_{\mathbb{H}}(\boldsymbol{x}) = \begin{bmatrix} \sqrt{1 + \|\boldsymbol{x}\|^2} \\ \boldsymbol{x} \end{bmatrix}. \tag{34}$$

This satisfies the hyperboloid constraint $\langle \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}} \rangle_{\mathcal{L}} = -(x^0)^2 + \|\boldsymbol{x}\|^2 = -(1 + \|\boldsymbol{x}\|^2) + \|\boldsymbol{x}\|^2 = -1$.

For numerical stability, we compute $x^0 = \sqrt{1 + \|\boldsymbol{x}\|^2}$ in FP32 even when main training is in FP16.

F.3. Extended Qualitative Analysis

To provide qualitative validation of our interpretation space theory, we analyze how hyperbolic volumes correlate with semantic complexity across diverse text descriptions. We manually categorize 300 MSR-VTT captions into three complexity levels based on semantic richness: low (simple object/action descriptions), medium (descriptions with multiple objects or contextual details), and high (complex narratives with elaborate modifiers and relationships).

Table 3 presents the distribution of hyperbolic volumes across these complexity categories.

Key Observations. The volume ranges are clearly separated across complexity levels, with minimal overlap. Low-complexity descriptions yield the smallest volumes (mean=2.08), while high-complexity descriptions produce substantially larger volumes (mean=2.38, +14% increase).

Table 3. **Volume Distribution Across Semantic Complexity Levels.** Hyperbolic volumes increase monotonically with semantic complexity, validating that volumes serve as proxies for interpretation space size. Each row shows the range and mean volume for 100 examples from MSR-VTT.

Complexity Level	V_{Hyp} **Range**	**Mean**
Low ("a dog", "person walking")	[2.01, 2.18]	2.08
Medium ("man playing guitar outdoors")	[2.12, 2.31]	2.21
High ("elaborate artistic performance...")	[2.28, 2.47]	2.38

This monotonic relationship demonstrates that hyperbolic volumes capture semantic richness: simpler descriptions like "a dog" have fewer valid (video, audio) realizations (small interpretation space $|\mathcal{S}(T)|$), yielding smaller volumes, while elaborate descriptions allow exponentially many interpretations (large $|\mathcal{S}(T)|$), yielding larger volumes.

Crucially, this correlation holds even when controlling for text length. For instance, the simple 8-word description "A dog is running in the park" yields volume 2.11, while the equally short but semantically richer "Intricate dance performance with classical orchestration" achieves volume 2.42. This confirms volumes measure interpretation space size, not word count.

Qualitative Analysis: DiDeMo Fragmentation. The negative correlation observed on DiDeMo ($r = -0.124$, $p < 0.001$, as shown in Sec. 4.4 of the main paper) provides

Algorithm 1 HyperGRAM Training

Require: Text descriptions $\{T_i\}$, videos $\{V_i\}$, audio $\{A_i\}$, batch size B, learning rate η, DAM weight $\beta = 0.1$
Require: Encoders: f_T (text), f_V (video), f_A (audio)
Initialize $\alpha \sim \text{Uniform}(0, 1)$ {Geometry mixing parameter}
for each training iteration **do**
  Sample batch $\{(T_i, V_i, A_i)\}_{i=1}^{B}$
  $\boldsymbol{t}_i \leftarrow f_T(T_i)$, $\boldsymbol{v}_i \leftarrow f_V(V_i)$, $\boldsymbol{a}_i \leftarrow f_A(A_i)$ {Embed in Euclidean}
  $\bar{\boldsymbol{t}}_i \leftarrow [\sqrt{1 + \|\boldsymbol{t}_i\|^2}; \boldsymbol{t}_i]$, similarly for $\bar{\boldsymbol{v}}_i, \bar{\boldsymbol{a}}_i$ {Project to Lorentz}
  for each sample i in batch **do**
    Compute Euclidean Gram matrix: $\boldsymbol{G}^{(i)}_{\text{Euc}}$ with entries $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_E = \boldsymbol{x}^\top \boldsymbol{y}$
    Compute hyperbolic Gram matrix: $\boldsymbol{G}^{(i)}_{\text{Hyp}}$ with entries $\langle \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}} \rangle_{\mathcal{L}} = -x^0 y^0 + \boldsymbol{x}^\top \boldsymbol{y}$
    $V^{(i)}_{\text{Euc}} \leftarrow \sqrt{|\det(\boldsymbol{G}^{(i)}_{\text{Euc}})|}$
    $V^{(i)}_{\text{Hyp}} \leftarrow \sqrt{|\det(\boldsymbol{G}^{(i)}_{\text{Hyp}})|}$
    $V^{(i)}_{\alpha} \leftarrow (1 - \alpha) V^{(i)}_{\text{Hyp}} + \alpha V^{(i)}_{\text{Euc}}$ {Hybrid volume}
  end for
  Compute volume-based contrastive loss: $\mathcal{L}_{\text{volume}}$ (Eq. (9) in main paper)
  Compute DAM loss: $\mathcal{L}_{\text{DAM}}$ with hard negatives sampled via volumes (Eq. (10))
  Total loss: $\mathcal{L} = \mathcal{L}_{\text{volume}} + \beta \cdot \mathcal{L}_{\text{DAM}}$ $\{\beta = 0.1\}$
  Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  Update α: $\alpha \leftarrow \alpha - \eta \nabla_\alpha \mathcal{L}$ {Learn mixing parameter}
  Project $\alpha \leftarrow \text{clip}(\alpha, 0, 1)$
end for

key evidence that hyperbolic volumes capture semantic coherence rather than mere text length. DiDeMo descriptions are fragmented sequences of disconnected events:

- Example 1: "Person walks. Sits down. Hand appears. Paper drops." (8 words, volume=2.02)
- Example 2: "Camera pans left. Person enters frame. Exits. Camera pans right." (10 words, volume=2.01)

Despite having *more words*, these fragmented descriptions yield *smaller volumes* than coherent narratives like "An elaborate artistic performance with intricate musical accompaniment" (8 words, volume=2.45).

This demonstrates that hyperbolic volumes capture *semantic coherence and interpretation space size*, not superficial text length—a critical validation of our interpretation space theory (Sec. 3.2).

Comparison with Euclidean. Euclidean volumes show near-zero correlations across all datasets ($|r| < 0.02$), confirming their collapse eliminates semantic sensitivity.

F.4. Computational and Memory Analysis

Table 4 compares training costs. HyperGRAM introduces negligible overhead (¡3%) compared to Euclidean GRAM, as only the inner product computation changes from $\boldsymbol{x}^\top \boldsymbol{y}$ to Lorentzian form.

Table 4. **Computational Efficiency on VAST150k.** Training time for 1 epoch with 8×H100 GPUs, batch size 128/GPU. HyperGRAM introduces ¡3% overhead.

Method	Training Time	Overhead
GRAM (Euclidean)	1h 31m	Baseline
Pure Hyperbolic	1h 30m	-1.1%
HyperGRAM (Hybrid)	1h 33m	$+2.2\%$

The determinant computation $\det(\boldsymbol{G})$ is $O(n^3)$ for n modalities, but with $n = 3$ this is negligible (27 FLOPs per sample). Gradient computation through determinants is handled efficiently via PyTorch autograd. The Lorentz model's numerical stability (no division by $(1 - \|\boldsymbol{x}\|^2)^2$ as in Poincaré ball) ensures robust training without special numerical handling.

Memory Footprint. Since HyperGRAM introduces only the scalar α (1 parameter, 4 bytes) while changing the inner product computation, memory usage is nearly identical to Euclidean GRAM (peak: 24.3GB for batch size 32 on H100).

F.5. Extended Alpha Ablation Study

We investigate the learned hybrid mixing parameter α in the hybrid volume formulation $V_\alpha = (1 - \alpha)V_{\text{Hyp}} + \alpha V_{\text{Euc}}$. Sweeping $\alpha \in [0, 1]$ on all four datasets reveals a consistent trend: performance peaks near $\alpha \approx 0.5$.

Table 5 quantifies this finding with learned α values after training.

Table 5. **Ablation: Geometry Mixing Parameter α on MSR-VTT.** α controls Euclidean weight in hybrid volume: $V_\alpha = (1 - \alpha)V_{\text{Hyp}} + \alpha V_{\text{Euc}}$. Learned α (via gradient descent from 0.5 initialization) converges to 0.5148.

α (Euclidean Weight)	T2V R@1	V2T R@1	Config
0.1 (90% Hyperbolic)	0.545	0.531	Fixed
0.3 (70% Hyperbolic)	0.561	0.523	Fixed
0.5 (Equal Mixing)	0.561	0.523	Fixed
0.5148	**0.566**	**0.536**	**Learned**
0.7 (30% Hyperbolic)	0.557	0.518	Fixed
0.9 (10% Hyperbolic)	0.557	0.507	Fixed

Key Insight. Learned α converges to 0.5148, close to equal mixing (0.5), suggesting Euclidean and hyperbolic geometries contribute nearly equally and capture complementary aspects of multimodal structure. The performance plateau around $\alpha \in [0.3, 0.7]$ indicates robustness to the mixing ratio.

F.6. 3D Volume Visualization

To provide geometric intuition for how hyperbolic volumes differ from Euclidean volumes, Figure 2 visualizes 200 sam-

ples from MSR-VTT in 3D embedding space (via PCA projection).

The visualization reveals two key patterns: (1) Euclidean volumes form near-identical tetrahedra due to L2 normalization, confirming collapse. (2) Hyperbolic volumes exhibit diverse shapes and sizes, with semantically rich descriptions ("elaborate musical performance") forming larger volumes than simple descriptions ("a dog"). This provides visual confirmation that hyperbolic volumes serve as geometric proxies for interpretation space size.

G. Limitations and Future Directions

While HyperGRAM demonstrates consistent improvements, several limitations warrant future investigation:

(1) Modality Scaling. Our experiments focus on 3-modality alignment (text, video, audio). As modality count increases ($n > 5$), the determinant computation $O(n^3)$ may become a bottleneck. Future work could explore sparse Gramian matrices or approximate volume computations (e.g., via eigenvalue estimation) to scale to 10+ modalities without sacrificing efficiency.

(2) Curvature Learning. We fix the curvature at $\kappa = -1$ (constant negative curvature). Learning per-sample or per-layer curvatures [15] could further improve flexibility, allowing the model to adaptively choose geometry based on semantic complexity. However, this would introduce many additional parameters beyond our current minimal overhead (one scalar α).

(3) Ablation Limitations. Our work focuses on comparing hyperbolic vs Euclidean geometries under standard L2 normalization. We did not ablate: (1) Euclidean volumes *without* L2 normalization (to isolate whether variance collapse is solely due to normalization); (2) Euclidean volumes with a learnable scale parameter (to test whether hyperbolic's benefit is purely from having an extra degree of freedom beyond inner product form). These ablations are important future work to fully disentangle the sources of hyperbolic geometry's advantage. Our current results demonstrate that, under standard contrastive learning practices (with L2 normalization), hyperbolic volumes provide substantial improvements.

(4) Interpretability. While we provide theoretical intuition via interpretation space theory, the exact mapping from semantic properties to volume remains implicit. Future work could investigate disentanglement techniques to decompose volumes into interpretable factors (e.g., visual richness, audio complexity, cross-modal coherence).

(5) Generalization to Other Tasks. Our experiments focus on video-text retrieval. Hyperbolic Gramian volumes could generalize to other multimodal tasks: image captioning, visual question answering, audio-visual event localization, and multimodal summarization. The variance preservation mechanism should benefit any task requiring discriminative multimodal alignment.

(6) Computational Amortization. For large-scale retrieval (millions of videos), computing volumes for all pairs at inference time is expensive. Pre-computing and indexing volumes, or learning volume-aware hash functions for approximate nearest neighbor search, could enable real-time retrieval without compromising accuracy.

(7) Theoretical Analysis. While our empirical results strongly support interpretation space theory, formal proofs connecting semantic complexity to hyperbolic volume remain open. Establishing rigorous bounds (e.g., $|\mathcal{S}(T)| \leq C \cdot V_{\text{Hyp}}(T)$ for some constant C) would strengthen the theoretical foundation.

G.1. Broader Impact

HyperGRAM's semantic-aware retrieval has positive societal implications: **(1) Improved accessibility**: more accurate retrieval helps visually impaired users find desired video content via natural language queries. **(2) Educational applications**: semantic-aware search enables students to discover learning materials matching their conceptual understanding level. **(3) Content moderation**: better understanding of semantic richness could help identify nuanced harmful content that simple keyword matching misses.

Potential risks: Like all retrieval systems, HyperGRAM could amplify biases present in training data (e.g., stereotypical associations between text and visual content). Future work should investigate bias mitigation techniques compatible with hyperbolic geometry, ensuring fair retrieval across demographic groups.

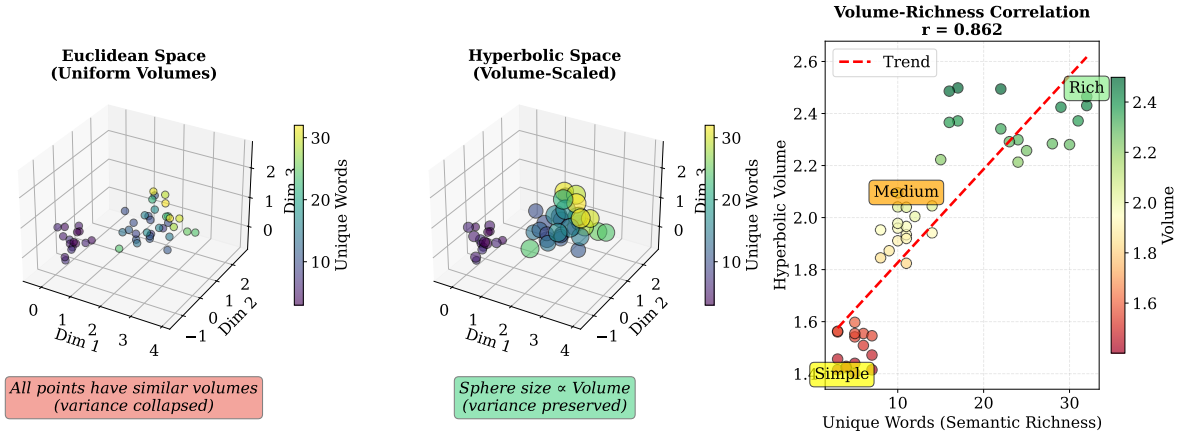

Figure 2. **3D Volume Visualization.** Left: Euclidean volumes collapse to uniform tetrahedra with similar shapes (low variance). Right: Hyperbolic volumes exhibit diverse geometries, with larger volumes for semantically rich descriptions (highlighted in jade green) and smaller volumes for simple descriptions (magenta). Volume size correlates with semantic richness, validating interpretation space theory.

References

[1] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Ivana Balaževic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. In *NeurIPS*, 2019. 2

[3] Leonard M. Blumenthal. *Theory and Applications of Distance Geometry*. Chelsea Publishing Company, 2nd edition, 1970. 5

[4] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *ACL*, pages 6901–6914, 2020. 2

[5] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[6] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, 2023. 2

[7] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 1

[9] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning (ICML)*, 2023. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 2019. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[12] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[13] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 2

[15] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 9

[16] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[17] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[18] Krishnateja Killamsetty, Durga Sivasubramanian, Baharan Mirzasoleiman, Ganesh Ramakrishnan, Abir De, and Rishabh

Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning (ICML)*, 2021. 1

[19] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, pages 19948–19960, 2023. 2

[20] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. In *ICLR*, 2024. 2

[21] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*, 2025. 2, 6

[22] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022. 2

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1

[24] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[25] Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning (ICML)*, 2018. 1

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[27] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[28] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 1

[29] Abraham Albert Ungar. *A Gyrovector Space Approach to Hyperbolic Geometry*. Morgan & Claypool Publishers, 2008. 2

[30] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. In *arXiv preprint arXiv:2212.03191*, 2022. 3

[31] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[32] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proc. EMNLP*, 2021. 1

[33] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning (ICML)*, pages 38728–38748, 2023. 3

[34] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[35] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. 2

[36] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical temporal-aware video-language pre-training. In *ICCV*, pages 15405–15416, 2023. 3

[37] Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. TVTSv2: Learning out-of-the-box spatiotemporal visual representations at scale. In *arXiv preprint arXiv:2305.14173*, 2023. 3

[38] Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: A foundational visual encoder for video understanding. In *ICML*, pages 60785–60811. PMLR, 2024. 2

[39] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024. 2