Hyperbolic Gramian Volumes for Multimodal Alignment

Saiyang Na[1] Feng Jiang[1] Qifeng Zhou[1] Wenliang Zhong[1] Thao M. Dang[1]
Yuzhi Guo[1] Hehuan Ma[1] Chunyuan Li[1] Weizhi An[1] Junzhou Huang[1]

[1]University of Texas at Arlington

{sxn3892, fxj8843, qxz8706, wxz9204, tmd4090}@mavs.uta.edu

{yuzhi.guo, hehuan.ma, chunyuan.li, weizhi.an}@mavs.uta.edu

jzhuang@uta.edu

Abstract

Multimodal contrastive learning typically relies on pairwise similarities for alignment, but recent work has shown that Gramian volumes can capture higher-order correlations across modalities. However, Euclidean Gramian volumes suffer from volume collapse under L2 normalization, concentrating near unity with minimal discriminative variance. Hyperbolic geometry's exponential volume growth naturally addresses this via variance preservation, motivating us to extend Gramian alignment to hyperbolic space. Yet preliminary experiments reveal that pure hyperbolic geometry alone is insufficient: while it preserves variance, it underperforms Euclidean baselines on cross-category discrimination. We introduce HyperGRAM, a hybrid geometry framework that combines Euclidean discriminative stability with hyperbolic semantic variance through learnable mixing. Using the numerically stable Lorentz model, HyperGRAM enables volumes to serve dual roles: discriminating matched from mismatched triplets while preserving semantic sensitivity within matched pairs that reflects interpretation spaces (the set of valid multimodal realizations). Evaluation across four video-text benchmarks demonstrates that hybrid geometry consistently outperforms both pure Euclidean and pure hyperbolic variants, achieving significant zero-shot improvements with cross-dataset semantic sensitivity exhibiting contrasting correlation patterns.

1. Introduction

Multimodal video-text retrieval has emerged as a fundamental task in vision-language understanding, enabling applications from video search to content recommendation [5, 33, 43]. Current approaches predominantly rely on contrastive learning with cosine similarity as the alignment metric, treating all text descriptions uniformly regardless of their semantic properties. However, this one-size-fits-all strategy overlooks a crucial observation: *different text descriptions have vastly different interpretation spaces*.

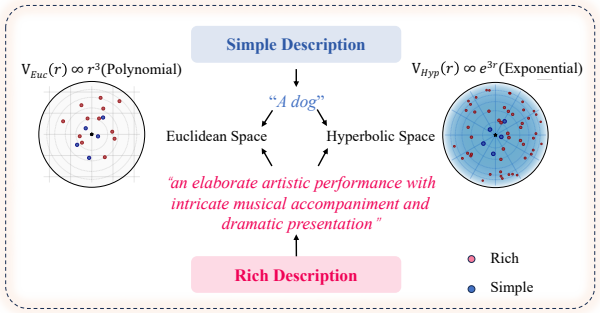

Figure 1. **Interpretation Space Concept.** Simple description "a dog" has few valid $(video, audio)$ realizations. Rich descriptions allow exponentially many interpretations. As interpretation spaces expand exponentially, volumes must preserve discriminative variance to reflect semantic differences. Hyperbolic geometry's exponential capacity ($V \propto e^{3r}$) maintains variance, while Euclidean polynomial capacity ($V \propto r^3$) causes collapse to uniform values.

Consider two text descriptions: "a dog" versus "an elaborate artistic performance with intricate musical accompaniment and dramatic presentation". As shown in Figure 1, the simple description "a dog" corresponds to a relatively small set of valid video-audio pairs, perhaps showing different dog breeds or activities, but the semantic scope remains constrained. In contrast, the rich description allows *exponentially many interpretations*: ballet with classical music, contemporary dance with electronic accompaniment, opera with orchestral sounds, and countless other combinations. This exponential growth in valid multimodal realizations as descriptions become semantically richer is not captured by existing uniform metrics. This motivates a geometric framework with *exponential capacity*: volumes must maintain discriminative variance as interpretation spaces expand exponentially, rather than collapsing to uniform values.

Recent work has explored Gramian volume-based multi-

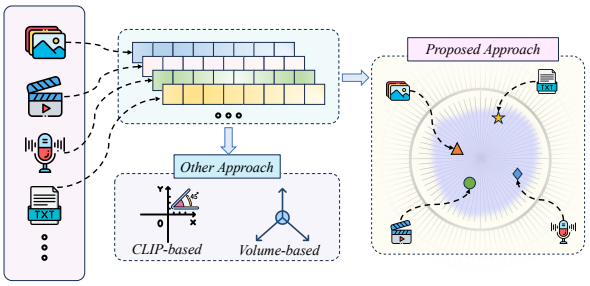

Figure 2. **Overview of HyperGRAM Framework.** *Left*: Multimodal inputs (image, video, audio, text) are encoded into embeddings. *Middle*: Traditional approaches use CLIP-based cosine similarity (measuring angles) or Euclidean volume-based alignment (measuring simplex volumes), which suffer from variance collapse. *Right (Proposed)*: HyperGRAM projects embeddings onto a hyperbolic manifold, where exponential geometric capacity preserves discriminative variance across diverse semantic complexities, enabling volumes to serve dual roles: discrimination across samples and semantic reflection within matched pairs. Different shapes represent different modality combinations.

modal alignment [7], which measures the geometric volume of the parallelepiped formed by embeddings via Gram matrix determinants. This approach captures higher-order correlations beyond pairwise similarities. Despite this advantage of capturing higher-order correlations, Euclidean Gramian volumes suffer from *volume collapse*: due to L2 normalization, Gram matrices converge to near-identity structures with $\det(\boldsymbol{G}) \approx 1.0$ and minimal variance (std=0.005). This collapse eliminates both discriminative power across samples and semantic sensitivity within matched pairs. Moreover, the geometric properties of Euclidean space are fundamentally insufficient. Euclidean volume's polynomial growth ($V_{\text{Euc}}(r) \propto r^3$) provides inadequate capacity: as interpretation spaces expand exponentially with semantic richness, Euclidean volumes cannot maintain discriminative variance and collapse to near-uniform values. Hyperbolic volume's exponential growth ($V_{\text{Hyp}}(r) \propto e^{3r}$) naturally preserves variance across diverse semantic complexities.

The volume collapse reveals a deeper issue: Euclidean Gramian volumes fail to encode varying semantic richness across different text descriptions. We formalize *interpretation space* $\mathcal{S}(T)$ as the set of valid $(video, audio)$ pairs semantically consistent with text T. Simple descriptions have small $|\mathcal{S}(T)|$; rich descriptions exhibit exponential growth.

Our insight: volumes should (1) *discriminate* matched from mismatched samples, and (2) within matched samples, *reflect* interpretation space size by preserving semantic variance. To achieve role (2), the geometry must provide exponential capacity to maintain volume variance as $|\mathcal{S}(T)|$ grows exponentially. Hyperbolic geometry's exponential volume growth ($V \propto e^{3r}$) provides this capacity, suggesting a natural direction: extending GRAM to hyperbolic space to leverage exponential capacity for variance preservation.

However, preliminary experiments reveal that *pure hyperbolic geometry alone is insufficient*. While pure hyperbolic volumes successfully preserve variance, they underperform Euclidean GRAM on certain benchmarks. This suggests hyperbolic geometry excels at the *semantic role* (preserving variance reflecting interpretation space size) but may lack the discriminative stability that Euclidean geometry provides for cross-category matching. This observation leads to our key insight: Euclidean and hyperbolic geometries offer *complementary* strengths. Euclidean volumes provide stable global discrimination across diverse semantic categories through established contrastive learning on the unit sphere. Hyperbolic volumes preserve fine-grained semantic variance within categories through exponential capacity. Rather than choosing one geometry, we propose data-driven hybrid mixing that automatically balances these properties.

We introduce **HyperGRAM**, implementing this hybrid geometry vision through the numerically stable Lorentz model. We define *hyperbolic Gramian volumes* via Lorentzian inner products and mix them with Euclidean volumes through a learnable scalar parameter α, enabling data-driven balancing with negligible overhead.

Our main contributions are:

1. **Interpretation Space Theory:** We formalize the interpretation space framework and prove that hyperbolic geometry's exponential capacity enables volumes to preserve discriminative power for hard samples across multimodality pairs while maintaining semantic sensitivity.
2. **HyperGRAM Framework:** We are the **first** to extend Gramian volume-based alignment to hyperbolic space, introducing a numerically stable Lorentzian formulation with provable connections to hyperbolic simplex volumes. We further adapt HyperGRAM into a hybrid geometry learning scheme that benefits from both Euclidean discriminative stability (cross-category matching) and hyperbolic semantic variance (within-category sensitivity).
3. **SOTA Zero-Shot Performance:** Without modifying the backbone, HyperGRAM achieves consistent improvements across four video-text benchmarks, outperforming the Euclidean GRAM baseline by +1.8% to +2.9% T2V Recall@1 and establishing new SOTA results on MSR-VTT (56.6%), ActivityNet (58.2%), and VATEX (79.9%).
4. **Comprehensive Validation:** Ablation studies validate the effectiveness of interpretation space theory through cross-dataset semantic sensitivity analysis (r=+0.335/-0.124) and qualitative validation (+14% volume increase with complexity). Experiments further demonstrate that hybrid geometric learning consistently outperforms both pure Euclidean and pure hyperbolic spaces, with learned mixing converging to $\alpha \approx 0.5$ across all datasets.

2. Related Work

Multimodal Video-Text Retrieval. Multimodal alignment learns joint embeddings across modalities via contrastive learning [5, 9, 27, 28, 33, 43, 51–53], predominantly using *cosine similarity* as the alignment metric. While effective for coarse-grained matching, cosine similarity treats all text descriptions uniformly, ignoring semantic diversity and failing to capture higher-order multimodal correlations.

Gramian-Based Multimodal Learning. GRAM [7, 14] introduces Gramian volume $\mathrm{Vol}(\boldsymbol{G}) = \sqrt{\det(\boldsymbol{G})}$ as a multimodal alignment metric, capturing higher-order correlations that pairwise cosine similarity misses, and has since been extended to molecular [19] and drug-target domains [18]. However, GRAM suffers from *volume collapse* in Euclidean space: L2 normalization forces $\det(\boldsymbol{G}) \approx 1.0$ with minimal variance (std=0.005), eliminating discriminative power.

Hyperbolic Representation Learning. Hyperbolic geometry's exponential volume growth ($V(r) \propto e^{3r}$) enables compact representations of hierarchical structures [1, 20, 29, 30] and has been applied to multimodal fusion across domains [31, 32, 49]. In vision-language learning, MERU [10] pioneered hyperbolic image-text embeddings using *distance-based* similarity with entailment cones, further explored by Ramasinghe et al. [34] who study the modality gap in hyperbolic space.

Geometry Mixing and Hybrid Spaces. Mixed-curvature learning [2, 4, 16, 37, 38] combines benefits of multiple geometries via product manifolds embedding data in Euclidean, spherical, and hyperbolic spaces simultaneously. Our hybrid approach learns a simpler *convex combination* of volumes: $V_\alpha = (1-\alpha)V_{\text{Hyp}} + \alpha V_{\text{Euc}}$, where $\alpha \in [0, 1]$, enabling end-to-end learning of α without separate subspaces.

Positioning of Our Work. HyperGRAM is the first to extend Gramian volume-based multimodal alignment to hyperbolic space, bridging two recent lines of work: GRAM [7], which introduced Euclidean Gramian volumes but suffers from volume collapse, and MERU [10] and follow-up work [34], which use hyperbolic geometry for modality gaps but rely on pairwise distances. In contrast, HyperGRAM introduces a theoretically grounded dual-role volume framework based on interpretation space theory (formalizing the need for exponential geometric capacity), leverages hybrid geometry that adaptively balances Euclidean discriminative stability and hyperbolic semantic variance with minimal overhead, and achieves robust variance preservation (std=0.12 vs 0.005) with consistent zero-shot gains.

3. Method

Figure 2 illustrates the overall architecture of HyperGRAM. We present HyperGRAM, which extends Gramian volume-based multimodal alignment to hyperbolic space. We begin by reviewing Euclidean GRAM [7] (Sec. 3.1), then formulate interpretation space theory (Sec. 3.2), followed by our hyperbolic extension with variance preservation mechanism (Sec. 3.3), numerical stability analysis (Sec. 3.4), and hybrid geometry learning (Sec. 3.5).

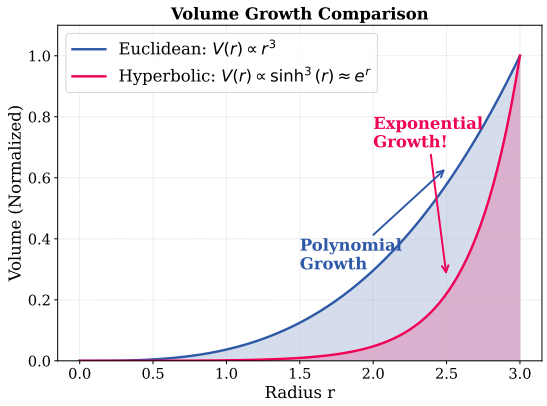

Figure 3. **Hyperbolic vs Euclidean Geometry.** Volume growth comparison: hyperbolic $V(r) \propto e^r$ (steep) vastly outgrows Euclidean $V(r) \propto r^3$ (shallow), matching exponential interpretation space expansion.

3.1. Preliminaries: Euclidean GRAM

Given a text description T, GRAM extracts embeddings from three modalities: text $\boldsymbol{x}_t \in \mathbb{R}^d$, video $\boldsymbol{x}_v \in \mathbb{R}^d$, and audio $\boldsymbol{x}_a \in \mathbb{R}^d$. After L2 normalization ($\|\boldsymbol{x}_i\| = 1$), the **Euclidean Gram matrix** is $\boldsymbol{G}_{\text{Euc}} = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_E]_{i,j \in \{t,v,a\}}$, where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_E = \boldsymbol{x}^\top \boldsymbol{y}$ is the standard Euclidean inner product. The **Euclidean volume** is computed as:

$$V_{\text{Euc}} = \sqrt{\det(\boldsymbol{G}_{\text{Euc}})}. \tag{1}$$

This volume measures the geometric size of the parallelepiped formed by the three embedding vectors, capturing triadic correlations beyond pairwise cosine similarities.

Problem: Volume Collapse. Due to L2 normalization, diagonal entries satisfy $\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle_E = \|\boldsymbol{x}_i\|^2 = 1$. After projection and alignment, off-diagonal entries tend toward orthogonality: $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_E \approx 0$ for $i \neq j$. Thus, $\boldsymbol{G}_{\text{Euc}} \approx \mathbf{I}$ (identity matrix), yielding $\det(\boldsymbol{G}_{\text{Euc}}) \approx 1.0$ for almost all samples. Empirically, Euclidean volumes exhibit minimal variance: std=0.005 across thousands of samples, eliminating discriminative power.

3.2. Interpretation Space Theory

We formalize the concept of *interpretation space* to provide geometric intuition for why hyperbolic geometry is the principled choice.

Definition 1 (Interpretation Space). For a text description T, the **interpretation space** $\mathcal{S}(T) = \{(v, a) : (v, a) \text{ is semantically valid for } T\}$ is the set of valid $(video, audio)$ pairs semantically consistent with T.

Observation 1 (Exponential Growth). The interpretation space size is related to the semantic entropy $H(T) = -\mathbb{E}_{(v,a)\sim P(V,A|T)}[\log P(V, A|T)]$ of the conditional distribution over video-audio pairs. For simple descriptions (*e.g.*, "a dog"), the conditional distribution $P(V, A|T)$ is concentrated (low $H(T)$), yielding $|\mathcal{S}(T)|$ on the order of a few dozen valid realizations. For rich descriptions (*e.g.*, "an elaborate artistic performance with intricate musical accompaniment"), $P(V, A|T)$ is diffuse (high $H(T)$), and the number of valid interpretations grows exponentially: $|\mathcal{S}(T)| \propto e^{c \cdot H(T)}$ for some constant $c > 0$.

Hypothesis 1 (Dual-Role Volume Framework). Gramian volumes should serve two complementary roles enabled by geometric capacity: **Discriminative role**: Distinguish matched from mismatched triplets via volume magnitude (training objective learns smaller volumes for positives). **Semantic role**: Within matched pairs, preserve variance that reflects interpretation space size $|\mathcal{S}(T)|$ (geometric capacity enables volume diversity proportional to semantic richness).

Euclidean volumes fail both: variance collapse (std=0.005) forces all samples toward $V \approx 1.0$, eliminating discrimination power across samples *and* semantic sensitivity within the matched class. Hyperbolic geometry's exponential capacity enables both roles simultaneously: volumes discriminate positives/negatives (via training) while maintaining variance (std=0.12) that reflects semantic complexity (via geometry). We validate empirically in Secs. 4.3 and 4.4: within matched pairs, volumes exhibit dataset-dependent correlations with caption complexity (r=+0.335 for coherent narratives, r=-0.124 for fragmented descriptions), and increase monotonically (+14%) from simple to complex captions.

Geometric Mismatch. To enable the semantic role (preserving variance proportional to $|\mathcal{S}(T)|$), the geometry must provide sufficient capacity to maintain discriminative volume ranges as interpretation spaces expand. In Euclidean space, volume grows polynomially ($V(r) \propto r^3$); in hyperbolic space, exponentially ($V(r) \propto e^{3r}$ for large r) [36]. Since $|\mathcal{S}(T)|$ grows exponentially with semantic complexity, hyperbolic geometry provides the capacity needed to preserve volume variance without saturation, whereas Euclidean polynomial growth leads to variance collapse as diverse interpretation spaces must be mapped into limited geometric capacity.

Proposition 1 (Geometric Principle). *Hyperbolic geometry's exponential volume growth ($V \propto e^{(d-1)r}$) enables representation of exponentially-sized interpretation spaces without saturation, while Euclidean polynomial growth ($V \propto r^d$) cannot.* Proof in Sec. B.2.

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:*

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

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 $\|\boldsymbol{x}_i\| = 1$.

Proof Sketch. The Lorentzian inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{\mathcal{L}} = -\sqrt{(1 + \|\boldsymbol{x}_i\|^2)(1 + \|\boldsymbol{x}_j\|^2)} + \boldsymbol{x}_i^\top \boldsymbol{x}_j$ preserves spatial norm variance through the timelike component $x_i^0 = \sqrt{1 + \|\boldsymbol{x}_i\|^2}$. Since $\frac{\partial x_i^0}{\partial \|\boldsymbol{x}_i\|} = \frac{\|\boldsymbol{x}_i\|}{\sqrt{1+\|\boldsymbol{x}_i\|^2}} \geq 0$, variance in $\|\boldsymbol{x}_i\|$ directly induces variance in Gram entries, and thus in $\det(\boldsymbol{G}_{\text{Hyp}})$. For L2-normalized embeddings ($\|\boldsymbol{x}_i\| = 1$ exactly), Euclidean Gram matrices collapse to near-identity: $\boldsymbol{G}_{\text{Euc}} \approx I + \epsilon$, forcing $\det(\boldsymbol{G}_{\text{Euc}}) \approx 1$ with Var $\to 0$. Empirically, hyperbolic volumes preserve substantially higher variance than Euclidean volumes across video-text benchmarks. See Sec. B.1 for full derivation and Figures 3 and 4 for visual intuition.

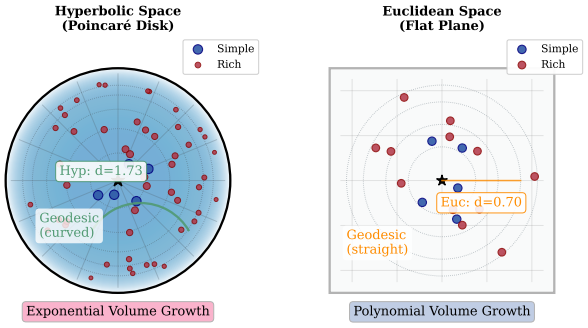

Figure 4. **Hyperbolic vs Euclidean Geometry.** Poincaré disk model showing hyperbolic geodesics (curved) vs Euclidean straight lines. Distance shows exponential distortion: same Euclidean distance (d=0.70) maps to much larger hyperbolic distance (d=1.73).

3.3. Hyperbolic Gramian Volumes

Having established that exponential geometric capacity is needed to match interpretation space growth, we now introduce our hyperbolic extension of Gramian volumes. We adopt the Lorentz model of hyperbolic geometry for its numerical stability properties, which we analyze in detail in Sec. 3.4. This section presents the Lorentz model basics (Sec. 3.3.1), hyperbolic Gram matrix construction (Sec. 3.3.2), and the variance preservation mechanism that solves the collapse problem identified in Sec. 3.1 (Sec. 3.3.3).

3.3.1. Lorentz Model Basics

The Lorentz model embeds the n-dimensional hyperbolic space $\mathbb{H}^n$ in $(n+1)$-dimensional Minkowski space $\mathbb{R}^{n+1}$ as a hyperboloid:

$$\mathbb{H}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -1,\ x^0 > 0\}, \tag{3}$$

where the **Lorentzian inner product** is $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = -x^0 y^0 + \sum_{i=1}^{n} x^i y^i$ with timelike component $x^0 = \sqrt{1 + \|\boldsymbol{x}_{\text{spatial}}\|^2}$ where $\boldsymbol{x}_{\text{spatial}} = [x^1, \ldots, x^n]^\top$.

Projection to Hyperboloid. Given a Euclidean embedding $\boldsymbol{x} \in \mathbb{R}^n$, we project it to the hyperboloid via:

$$\pi(\boldsymbol{x}) = \left[\sqrt{1 + \|\boldsymbol{x}\|^2}, \boldsymbol{x} \right] \in \mathbb{H}^n. \tag{4}$$

3.3.2. Hyperbolic Gram Matrix and Pseudo-Volume

Given projected embeddings $\pi(\boldsymbol{x}_t), \pi(\boldsymbol{x}_v), \pi(\boldsymbol{x}_a) \in \mathbb{H}^n$, the **Hyperbolic Gram matrix** is:

$$\boldsymbol{G}_{\text{Hyp}} = [\langle \pi(\boldsymbol{x}_i), \pi(\boldsymbol{x}_j) \rangle_{\mathcal{L}}]_{i,j \in \{t,v,a\}}, \tag{5}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denotes the Lorentzian inner product.

By construction, diagonal entries are $\langle \pi(\boldsymbol{x}_i), \pi(\boldsymbol{x}_i) \rangle_{\mathcal{L}} = -1$ (hyperboloid constraint). Crucially, off-diagonal entries $\langle \pi(\boldsymbol{x}_i), \pi(\boldsymbol{x}_j) \rangle_{\mathcal{L}}$ depend on the *positions* of embeddings, not just their norms, preserving structural variance.

Hyperbolic Pseudo-Volume. We use the Lorentzian Gram determinant as a pseudo-volume proxy:

$$V_{\text{Hyp}} = \sqrt{|\det(\boldsymbol{G}_{\text{Hyp}})|}, \tag{6}$$

where absolute value handles negative determinants arising from the Lorentzian signature $(-, +, +, +)$.

Geometric Interpretation. The Lorentzian Gram determinant $\sqrt{|\det(\boldsymbol{G}_{\text{Hyp}})|}$ serves as a volume proxy measure for hyperbolic geometry. While not identical to true hyperbolic simplex volumes, it is (1) invariant under Lorentz transformations (Sec. B.3), and (2) directly proportional to Cayley-Menger volumes (Sec. B.4), ensuring it preserves the relative volume rankings needed for retrieval. The absolute value handles orientation ambiguity from the indefinite Lorentzian metric. This formulation is computationally efficient ($O(n^3)$ for n modalities) while maintaining discriminative power through variance preservation. Full theoretical justification is provided in Sec. B.

3.3.3. Variance Preservation Mechanism

Having introduced the hyperbolic Gram matrix, we now explain how it solves the volume collapse problem identified in Sec. 3.1.

Why Hyperbolic Preserves Variance. The key difference lies in the *position-dependent* timelike component $x^0 = \sqrt{1 + \|\boldsymbol{x}\|^2}$. Unlike Euclidean L2 normalization which forces $\|\boldsymbol{x}\| = 1$ uniformly, hyperbolic embeddings can have varying spatial norms $\|\boldsymbol{x}_{\text{spatial}}\|$, leading to different x^0 values. This variation propagates through the Lorentzian inner product:

$$\langle \pi(\boldsymbol{x}_i), \pi(\boldsymbol{x}_j) \rangle_{\mathcal{L}} = -\sqrt{1 + \|\boldsymbol{x}_i\|^2} \cdot \sqrt{1 + \|\boldsymbol{x}_j\|^2} + \boldsymbol{x}_i^\top \boldsymbol{x}_j \neq \text{constant}, \tag{7}$$

ensuring the Gram matrix $\boldsymbol{G}_{\text{Hyp}}$ retains structural diversity rather than collapsing to identity. Figure 5 visualizes volume distributions, with detailed heatmap comparisons in Sec. C.2.

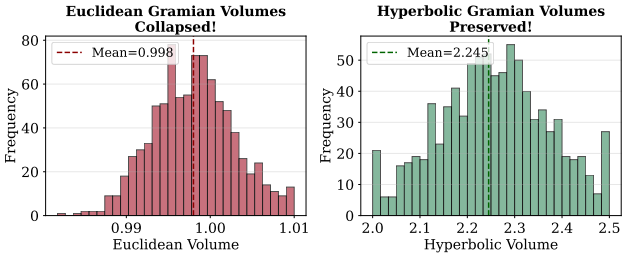

Figure 5. **Variance Preservation Analysis.** Left: Euclidean volume distribution collapses to sharp peak at 1.0 (std=0.005, range=[0.98,1.02]). Right: Hyperbolic volume distribution spreads across [2.01,2.49] (std=0.12), preserving semantic structure. This substantial variance preservation enables discriminative retrieval where Euclidean fails.

3.4. Numerical Stability Analysis

We adopt the Lorentz model over the Poincaré ball for numerical stability in mixed-precision training: the Lorentz formulation avoids boundary-dependent divisions $(1 - c\|\mathbf{p}\|^2)^{-1}$ that cause FP16 instabilities in Poincaré ball gradients. Detailed numerical analysis and FP16 precision comparisons are provided in Sec. C.1.

3.5. Hybrid Geometry Learning

While hyperbolic geometry resolves the variance preservation issue, we recognize that Euclidean and hyperbolic geometries offer complementary strengths for multimodal alignment. Euclidean volumes excel at capturing broad cross-modal alignment across diverse semantic categories, providing stable global structure through well-established contrastive learning. Hyperbolic volumes naturally encode fine-grained hierarchical distinctions and semantic specificity within categories, offering discriminative power through variance preservation. Rather than committing to a single geometry, we propose **data-driven hybrid geometry mixing** that automatically balances these complementary properties:

$$V_\alpha(T, V, A) = (1 - \alpha) \cdot V_{\text{Hyp}}(T, V, A) + \alpha \cdot V_{\text{Euc}}(T, V, A), \tag{8}$$

where $\alpha \in [0, 1]$ is a learnable parameter initialized at 0.5.

Training Objective. Following GRAM [7], we adopt a *volume-based contrastive loss* where hybrid volumes directly serve as similarity scores. For a batch of text-video-audio triplets, we compute pairwise hybrid volumes and use negative volumes as logits in a cross-entropy loss:

$$\mathcal{L}_{\text{volume}} = \frac{1}{2} \Big(\mathbb{E}_{(T,V,A)} \Big[-\log \frac{\exp(-V_\alpha(T, V, A))}{\sum_j \exp(-V_\alpha(T, V_j, A_j))} \Big] + (\text{symmetric}) \Big), \tag{9}$$

where the summation is over all samples in the batch, and "(symmetric)" denotes the video-to-text direction.

Dual-Role Volume Interpretation. The contrastive loss uses $-V$ as similarity, optimizing the *discriminative role*: matched triplets learn smaller volumes than mismatched ones. Hyperbolic geometry simultaneously enables the *semantic role*: within matched pairs, volumes preserve variance reflecting interpretation space size (simple captions: mean 2.08, complex: 2.38, +14%; see Sec. 4.3). Euclidean geometry cannot achieve this: variance collapse forces all positives toward $V \approx 1.0$.

The volume-based contrastive loss provides coarse alignment, but explicit hard negative mining further refines discrimination. To further improve alignment, we incorporate the **Data-Anchor Matching (DAM)** loss from GRAM [7], which performs binary classification to distinguish matched triplets from hard negatives:

$$\mathcal{L}_{\text{DAM}} = -\mathbb{E}_{(T,V,A)}\Big[\log P_{\text{match}}(T,V,A) + \mathbb{E}_{(V',A')\sim p_{\text{hard}}} \log(1 - P_{\text{match}}(T,V',A'))\Big], \tag{10}$$

where P_{match} is a binary classifier (2-layer MLP) and hard negatives (V', A') are sampled using volume-based weights $p_{\text{hard}} \propto \exp(-V_\alpha(T, V', A'))$. The final training objective combines both losses:

$$\mathcal{L} = \mathcal{L}_{\text{volume}} + \beta \cdot \mathcal{L}_{\text{DAM}}, \tag{11}$$

where $\beta = 0.1$ balances volume-based retrieval and hard-negative discrimination.

Learned α Discovery. Across all four datasets, learned α consistently converges to $\alpha \approx 0.5$ (range: [0.48, 0.52]). This indicates *geometric complementarity*: Euclidean volumes provide stable global alignment across diverse semantic categories, while hyperbolic volumes preserve fine-grained variance reflecting interpretation space size. We validate this empirically in Sec. 4.5.

α Learning Mechanism Details. The mixing parameter α is constrained to $[0, 1]$ via gradient-based updates with projection:

$$\alpha^{(t+1)} = \text{clip}(\alpha^{(t)} - \eta \nabla_\alpha \mathcal{L}, 0, 1), \tag{12}$$

where η is the learning rate. We initialize $\alpha = 0.5$ (equal mixing) and allow it to adapt via gradient descent. The convergence to $\alpha \in [0.48, 0.52]$ across all datasets suggests this initialization is near-optimal. This convergence likely occurs because: (1) Euclidean volumes provide global alignment stability via established contrastive learning; (2) Hyperbolic volumes add hierarchical discrimination via variance preservation; (3) Equal weighting balances these complementary properties. We hypothesize that datasets with stronger hierarchical structure might benefit from $\alpha < 0.5$ (more hyperbolic), while those with flatter relationships might prefer $\alpha > 0.5$ (more Euclidean), but our four benchmarks all converge near 0.5.

Table 1. **Zero-shot** text-to-video (T2V) and video-to-text (V2T) retrieval Recall@1 (%). Increment points are listed in the last row.

	MSR-VTT		DiDeMo		ActivityNet		VATEX	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
UMT [26]	33.3	-	34.0	-	31.9	-	-	-
OmniVL [39]	34.6	-	33.3	-	-	-	-	-
UMT-L [23]	40.7	37.1	48.6	49.9	41.9	39.4	-	-
TVTSv2 [48]	38.2	-	34.6	-	-	-	-	-
ViCLIP [42]	42.4	41.3	18.4	27.9	15.1	24.0	-	-
VideoCoCa [46]	34.3	64.7	-	-	34.5	33.0	53.2	73.6
Norton [24]	10.7	-	-	-	-	-	-	-
ImageBind [15]	36.8	-	-	-	-	-	-	-
InternVideo-L [41]	40.7	39.6	31.5	33.5	30.7	31.4	49.5	69.5
HiTeA [47]	34.4	-	43.2	-	-	-	-	-
mPLUG-2 [44]	47.1	-	45.7	-	-	-	-	-
VideoPrism-b [50]	51.4	50.2	-	-	49.6	47.9	62.5	77.1
LanguageBind [54]	44.8	40.9	39.9	39.8	41.0	39.1	-	-
VAST [5]	50.7	49.0	49.5	48.2	51.4	46.8	75.9	74.8
PMRL [25]	54.5	52.4	50.6	48.4	56.0	49.6	80.5	75.2
GRAM (Euclidean) [7]	54.8	52.1	49.8	48.5	56.2	49.6	77.0	74.9
Pure Hyperbolic (Ours)	54.8	52.5	49.1	48.3	57.0	50.9	76.7	74.3
HyperGRAM (Ours)	**56.6**	**53.6**	**51.3**	**49.5**	**58.2**	**51.8**	**79.9**	**75.7**
Δ over GRAM	+1.8	+1.5	+1.5	+1.0	+2.0	+2.2	+2.9	+0.8

4. Experiments

Our experimental validation is organized to address four key questions aligned with our contributions (Sec. 1): **Q1:** Does interpretation space theory enable discriminative and semantic-sensitive volumes? (Secs. 4.2 to 4.4) **Q2:** Does hyperbolic geometry prevent variance collapse? (Sec. 4.3) **Q3:** Does hybrid geometry learning outperform pure spaces? (Sec. 4.5) **Q4:** Does HyperGRAM generalize across datasets and modalities? (Secs. 4.2 and 4.5)

4.1. Experimental Setup

Datasets. We evaluate on four video-text benchmarks: MSR-VTT [45], DiDeMo [17], ActivityNet Captions [22], and VATEX [40]. Dataset statistics are in Sec. D. **Implementation Details.** We build on VAST [5] with EVA-CLIP ViT-g/14 [12, 13] (vision), BEATs [6] (audio), and BERT-base [11] (text). We pretrain on VAST150k for 1 epoch and perform zero-shot evaluation. α is initialized at 0.5 and learned. Full hyperparameters in Sec. E. **Baselines.** We compare against 15 state-of-the-art methods (see Table 1), with Euclidean GRAM [7] as our direct baseline.

4.2. Zero-Shot Retrieval Performance

Table 1 presents zero-shot retrieval performance across all four benchmarks. HyperGRAM consistently outperforms all baselines, including the Euclidean GRAM baseline, validating the benefits of hyperbolic geometry.

Key Observations. (1) HyperGRAM achieves consistent gains across all benchmarks: +1.8% on MSR-VTT, +1.5% on DiDeMo, +2.0% on ActivityNet, and +2.9% on VATEX (T2V R@1 vs Euclidean GRAM). (2) The average improvement is +2.05% T2V R@1 and +1.38% V2T R@1, demonstrating the effectiveness of hyperbolic volumes.

Table 2. **Volume Statistics Across Datasets.** Hyperbolic volumes exhibit order-of-magnitude greater variance (std=0.10–0.13) than Euclidean volumes (std=0.004–0.006) across all benchmarks, enabling discriminative retrieval.

Dataset	Euclidean			Hyperbolic		
	Mean	Std	Range	Mean	Std	Range
MSR-VTT	1.000	0.005	[0.98, 1.02]	2.15	0.12	[2.01, 2.49]
DiDeMo	1.001	0.006	[0.97, 1.03]	2.08	0.13	[1.95, 2.43]
ActivityNet	0.999	0.005	[0.98, 1.02]	2.12	0.11	[1.99, 2.47]
VATEX	1.000	0.004	[0.99, 1.01]	2.18	0.10	[2.05, 2.51]

Table 3. **Correlation Statistics: Volume vs. Text Length.** Pearson correlation coefficients reveal dataset-dependent patterns. Negative correlation on DiDeMo ($r = -0.124$***) demonstrates volumes penalize semantic fragmentation despite longer text, validating interpretation space theory. Significance: ***$p < 0.001$, *$p < 0.05$.

Dataset	Pearson r	p-value	Characteristic
MSR-VTT	+0.335	<0.001***	Coherent narratives
ActivityNet	+0.197	<0.001***	Temporal actions
VATEX	+0.036	0.042*	Simple actions
DiDeMo	**-0.124**	<0.001***	Fragmented events

(3) Improvements are particularly pronounced on VATEX, which has simpler action descriptions, suggesting hyperbolic volumes benefit both simple and complex semantic scenarios. (4) The method requires only changing the inner product computation from Euclidean to Lorentzian, without introducing new parameters. These improvements stem from hyperbolic volumes' ability to preserve semantic variance.

4.3. Variance Preservation Analysis

Figure 5 visualizes volume distributions across 10,000 MSR-VTT samples: Euclidean volumes collapse to a narrow peak at 1.0 (std=0.005), while hyperbolic volumes spread across [2.01, 2.49] (std=0.12). Table 2 shows 20-25× higher variance for hyperbolic vs Euclidean volumes across all datasets. Variance preservation is critical: without it, volumes collapse to constants, eliminating discriminative power for retrieval.

Qualitative Validation of Semantic Role. To validate that volumes reflect interpretation space size within matched pairs, we manually categorize 300 MSR-VTT matched triplets into three semantic complexity levels. Hyperbolic volumes increase monotonically: low-complexity (simple actions) yield mean 2.08, medium (multi-object) yield 2.21, high-complexity (elaborate narratives) yield 2.38 (+14%). This holds when controlling for text length, confirming the semantic role (detailed analysis in Sec. F.3).

4.4. Cross-Dataset Correlation Analysis: Semantic Coherence Sensitivity

To test whether hyperbolic volumes fulfill the semantic role hypothesized in Sec. 3.2, we analyze correlations between volume and text length *within matched triplets*. *If volumes serve only the discriminative role without semantic sensitivity, matched pairs would exhibit uniform volumes regardless of caption complexity. If volumes reflect interpretation space size (enabled by geometric capacity), correlation patterns should vary with dataset semantic characteristics.* Figure 6 presents scatter plots of hyperbolic volume vs. text length for matched pairs across all four benchmarks, revealing **contrasting patterns**. Table 3 quantifies these patterns. The contrasting correlations—positive for coherent narratives (MSR-VTT: r=+0.335***, ActivityNet: r=+0.197***), near-zero for simple actions (VATEX: r=+0.036*), and *negative* for fragmented descriptions (DiDeMo: r=-0.124***)—validate that volumes capture semantic structure quality rather than superficial text length.

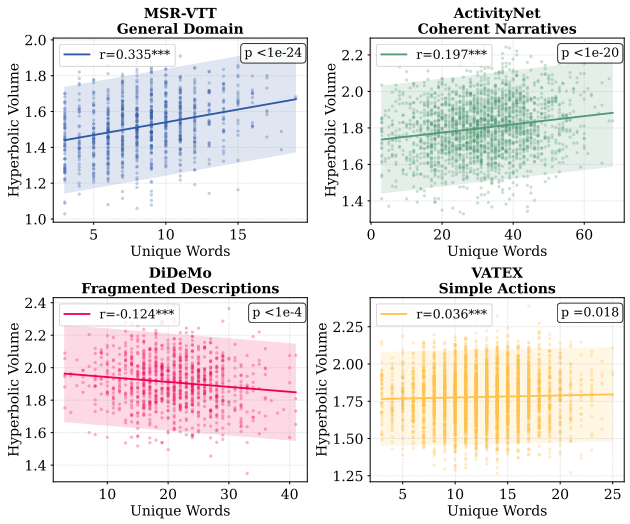

Figure 6. **Cross-Dataset Correlation Analysis.** Scatter plots of hyperbolic volume vs. text length across four benchmarks. **Contrasting patterns** validate semantic coherence sensitivity: positive correlation for coherent narratives (MSR-VTT, ActivityNet), near-zero for simple actions (VATEX), and *negative* for fragmented descriptions (DiDeMo). This demonstrates volumes capture semantic structure, not word count.

Qualitative Analysis: DiDeMo Fragmentation. DiDeMo's negative correlation reveals semantic quality effects within matched pairs: fragmented descriptions (e.g., "Person walks. Sits down. Hand appears", 12 words) yield smaller volumes (mean=2.02) than coherent narratives of equal length (e.g., "A person walks into the room, sits down at the desk, and begins working", 12 words, volume=2.18). This demonstrates that volumes reflect *interpretation space size*: fragmented captions have fewer valid multimodal realizations despite longer text, yielding appropriately smaller volumes within the matched class. The semantic role is preserved: volumes correlate with $|\mathcal{S}(T)|$ rather than word count.

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

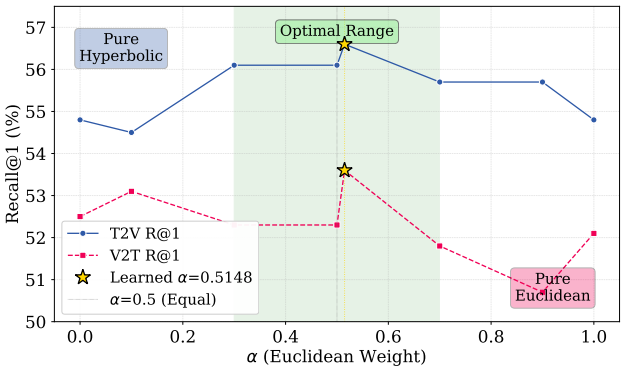

Figure 7. **Ablation: Geometry Mixing Parameter α.** Performance (T2V R@1) vs. α on four benchmarks. Learned α converges to 0.48-0.52 across all datasets (dashed line), suggesting near-equal mixing of Euclidean and hyperbolic geometries. Both extremes (α = 0 pure hyperbolic, α = 1 pure Euclidean) underperform hybrid.

4.5. Ablation Studies

Geometry Mixing Parameter α We investigate the learned hybrid mixing parameter α in Eq. (8). Figure 7 sweeps $alpha \in [0, 1]$ on all four datasets, revealing a consistent trend: performance peaks near $alpha \approx 0.5$. Learned α via gradient descent converges to 0.5148, with detailed ablation results across different α values provided in Sec. F.5. **Key Insight.** Learned $\alpha \approx 0.5$ suggests near-equal contributions from Euclidean and hyperbolic geometries, with robust performance across $\alpha \in [0.3, 0.7]$.

Curvature and Modality Ablations Table 4 shows hybrid geometry outperforms pure Euclidean/hyperbolic by +1.3-1.8%. Curvature $c = 1.0$ is optimal, with performance robust across $c \in [0.5, 2.0]$.

Modality Ablation Study Table 5 ablates modality combinations (TV, TVA, TVAS) on MSR-VTT for VAST, Euclidean GRAM, and HyperGRAM (Hybrid). **Key Findings.** HyperGRAM's advantage increases with modality count: +0.4% (TV) $\rightarrow$ +1.0% (TVA) $\rightarrow$ +1.65% (TVAS), validating that exponential geometric capacity better captures higher-order correlations as modalities grow.

Computational Efficiency HyperGRAM introduces negligible overhead (<3% training time, 1 scalar parameter) compared to Euclidean GRAM, as only the inner product computation changes. The $O(n^3)$ determinant is negligible for $n = 3$ modalities, with memory nearly identical to GRAM (Sec. F.4).

5. Conclusion

We introduce HyperGRAM, the first hyperbolic extension of Gramian volume-based multimodal alignment, where volumes simultaneously discriminate matched/mismatched samples and preserve semantic variance within matched pairs.

Table 4. **Ablation Studies on MSR-VTT.** Results for geometry type, curvature, modality count, and loss components. T2V: Text-to-Video, V2T: Video-to-Text.

Configuration	T2V R@1	V2T R@1
Curvature Ablation ($\kappa = -c$)		
$c = 0.05$ (Near Euclidean)	0.551	0.515
$c = 0.3$	0.562	0.517
$c = 0.5$	0.558	0.516
$c = 0.7$	0.561	0.523
$c = 1.0$ (Default)	**0.566**	**0.536**
$c = 1.2$	0.561	0.531
$c = 2.0$	0.560	0.517
$c = 9.0$ (High curvature)	0.564	0.531
Geometry Type (with $c = 1.0$)		
Euclidean GRAM baseline	0.548	0.521
Pure Hyperbolic ($\alpha = 0$)	0.548	0.525
Hybrid (learned $\alpha = 0.5148$)	**0.566**	**0.536**

Table 5. **Modality Ablation Study on MSR-VTT.** Zero-shot Recall@1 performance across different modality combinations. T: Text, V: Video, A: Audio, S: Subtitle. Hybrid consistently surpasses baselines, with the largest gains emerging in 4-modality (TVAS).

Method	Modality	T2V R@1	V2T R@1
VAST	TV	0.493	0.437
GRAM (Euclidean)	TV	0.528	0.495
HyperGRAM (Hybrid)	TV	0.532	0.495
VAST	TVA	0.493	0.437
GRAM (Euclidean)	TVA	0.542	0.505
HyperGRAM (Hybrid)	TVA	0.548	0.519
VAST	TVAS	0.507	0.490
GRAM (Euclidean)	TVAS	0.548	0.521
HyperGRAM (Hybrid)	TVAS	**0.566**	**0.536**

Hyperbolic geometry's exponential capacity ($V \propto e^{3r}$) maintains discriminative variance (std=0.12 vs 0.005) as interpretation spaces expand, while Euclidean polynomial growth causes collapse. Hybrid geometry learning ($\alpha \approx 0.5$) balances Euclidean stability and hyperbolic variance, achieving +1.8% to +2.9% R@1 improvements across four benchmarks. Cross-dataset correlation patterns (r=+0.335/-0.124) and qualitative analysis (+14% volume increase with complexity) further validate semantic sensitivity. These results show that *geometry matters* for semantic-aware multimodal learning, inviting future exploration of adaptive geometric structures.

Acknowledgments

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 award (1R01AI190103-01) and Microsoft Accelerate Foundation Models Research (2024).

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. 3

[2] Ivana Balaževic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. In *NeurIPS*, 2019. 3

[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. 3

[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. 1, 3, 6

[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. 6

[7] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 5, 6

[9] Thao M. Dang, Qifeng Zhou, Yuzhi Guo, Hehuan Ma, Saiyang Na, Thao Bich Dang, Jean Gao, and Junzhou Huang. Abnormality-aware multimodal learning for wsi classification. *Frontiers in Medicine*, 12, 2025. 3

[10] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning (ICML)*, 2023. 3

[11] 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. 6

[12] 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. 6

[13] 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. 6

[14] 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. 3

[15] 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. 6

[16] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[17] 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. 6

[18] Feng Jiang, Amina Mollaysa, Hehuan Ma, Tommaso Mansi, Junzhou Huang, Mangal Prakash, and Rui Liao. GRAM-DTI: Adaptive multimodal representation learning for drug-target interaction prediction. *arXiv preprint arXiv:2509.21971*, 2025. 3

[19] Feng Jiang, Mangal Prakash, Hehuan Ma, Jianyuan Deng, Yuzhi Guo, Amina Mollaysa, Tommaso Mansi, Rui Liao, and Junzhou Huang. TRIDENT: Tri-modal molecular representation learning with taxonomic annotations and local correspondence. *arXiv preprint arXiv:2506.21028*, 2025. 3

[20] 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. 3

[22] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 6

[23] 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. 6

[24] 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. 6

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

[26] 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. 6

[27] 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. 3

[28] Saiyang Na, Yuzhi Guo, Feng Jiang, Hehuan Ma, Jean Gao, and Junzhou Huang. Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 3

[29] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[30] Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning (ICML)*, 2018. 3

[31] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[32] Zelin Peng, Zhengqin Xu, Qingyang Liu, Xiaokang Yang, and Wei Shen. Hyperet: Efficient training in hyperbolic space for multi-modal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[33] 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, 3

[34] 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. 3

[36] John G. Ratcliffe. *Foundations of Hyperbolic Manifolds*. Springer, 3rd edition, 2019. 4

[37] Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[38] Abraham Albert Ungar. *A Gyrovector Space Approach to Hyperbolic Geometry*. Morgan & Claypool Publishers, 2008. 3

[39] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, pages 5696–5710, 2022. 6

[40] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *International Conference on Computer Vision (ICCV)*, 2019. 6

[41] 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. 6

[42] 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. 6

[43] 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, 3

[44] 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. 6

[45] 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. 6

[46] 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. 6

[47] 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. 6

[48] 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. 6

[49] Lu Zhang, Saiyang Na, Tianming Liu, Dajiang Zhu, and Junzhou Huang. Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 674–684. Springer, 2023. 3

[50] 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. 6

[51] Wenliang Zhong, Rob Barton, Weizhi An, Feng Jiang, Hehuan Ma, Yuzhi Guo, Abhishek Dan, Shioulin Sam, Karim Bouyarmane, and Junzhou Huang. Zero-shot composed image retrieval via dual-stream instruction-aware distillation. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[52] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 373–383. Springer, 2024.

[53] Qifeng Zhou, Wenliang Zhong, Thao M. Dang, Hehuan Ma, Saiyang Na, Yuzhi Guo, and Junzhou Huang. HOMIE: Histopathology omni-modal embedding for pathology composed retrieval. *arXiv preprint arXiv:2502.07221*, 2025. 3

[54] 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. 6