

# Hyperbolic Hierarchical Alignment Reasoning Network for Text-3D Retrieval

Wenrui Li<sup>\*1</sup>, Yidan Lu<sup>\*1</sup>, Yeyu Chai<sup>1</sup>, Rui Zhao<sup>2</sup>, Hengyu Man<sup>1</sup> and Xiaopeng Fan<sup>134†</sup>

<sup>1</sup>Harbin Institute of Technology

<sup>2</sup>Nanyang Technological University

<sup>3</sup>Peng Cheng Laboratory

<sup>4</sup>Harbin Institute of Technology Suzhou Research Institute

liwr618@163.com; 24S103311@stu.hit.edu.cn; 23s136130@stu.hit.edu.cn

zhao.rui@ntu.edu.sg; manhengyu@hotmail.com; fxp@hit.edu.cn

## Abstract

With the daily influx of 3D data on the internet, text-3D retrieval has gained increasing attention. However, current methods face two major challenges: Hierarchy Representation Collapse (HRC) and Redundancy-Induced Saliency Dilution (RISD). HRC compresses abstract-to-specific and whole-to-part hierarchies in Euclidean embeddings, while RISD averages noisy fragments, obscuring critical semantic cues and diminishing the model’s ability to distinguish hard negatives. To address these challenges, we introduce the Hyperbolic Hierarchical Alignment Reasoning Network (H<sup>2</sup>ARN) for text-3D retrieval. H<sup>2</sup>ARN embeds both text and 3D data in a Lorentz-model hyperbolic space, where exponential volume growth inherently preserves hierarchical distances. A hierarchical ordering loss constructs a shrinking entailment cone around each text vector, ensuring that the matched 3D instance falls within the cone, while an instance-level contrastive loss jointly enforces separation from non-matching samples. To tackle RISD, we propose a contribution-aware hyperbolic aggregation module that leverages Lorentzian distance to assess the relevance of each local feature and applies contribution-weighted aggregation guided by hyperbolic geometry, enhancing discriminative regions while suppressing redundancy without additional supervision. We also release the expanded T3DR-HIT v2 benchmark, which contains 8,935 text-to-3D pairs, 2.6 times the original size, covering both fine-grained cultural artefacts and complex indoor scenes.

**Code** — <https://github.com/liwrui/H2ARN>

## Introduction

With the rapid increase in the volume and variety of 3D data available online, text-3D retrieval has attracted significant attention for its broad application potential. Unlike traditional cross-modal retrieval tasks limited to 2D alignment, this area directly maps natural language to rich geometric, topological, and textural information. This capability enables more accurate and actionable content analysis, benefiting applications such as 3D crack recognition (Chen et al. 2022, 2023, 2024b), and multimodal processing (Bai et al. 2024, 2025; Xiao, Li,

<sup>\*</sup>These authors contributed equally.

<sup>†</sup>Corresponding author.

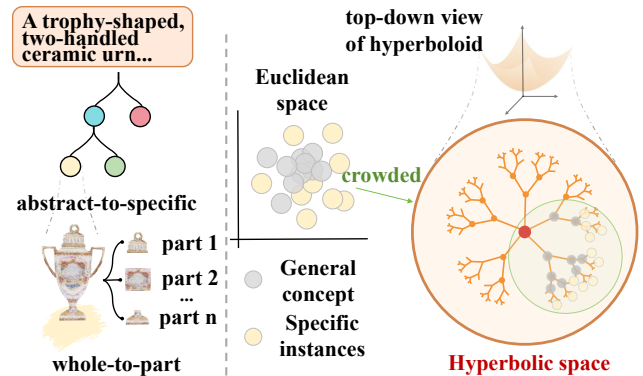


Figure 1: Conceptual illustration of hierarchical data representation. **Left:** The exponentially growing tree structures inherent in both abstract-to-specific semantics and whole-to-part geometry. **Right:** Euclidean space suffers from a "crowding" effect, whereas hyperbolic space preserves the hierarchy.

and Jia 2025; Xiao and Wang 2025; Zhang et al. 2025a,b; Li et al. 2025e,b, 2024, 2023a, 2025c; Bao et al. 2022, 2025).

However, bridging the semantic gap between language and 3D geometry presents substantially greater challenges than traditional cross-modal retrieval tasks (Li et al. 2023b). Both 3D data and natural language exhibit inherent tree-like hierarchies: semantics evolve from abstract concepts to concrete details, while geometry transitions from holistic structures to fine-grained components. This hierarchical structure leads to an exponential growth in the number of nodes with increasing depth. When embedded into Euclidean or conventional Riemannian spaces, which grow at most polynomially with radius (Lee 2018), a "crowding" effect becomes inevitable, as shown in the right panel of Figure 1. Samples that are semantically distinct but structurally similar are compressed into close proximity in high-dimensional embedding spaces. Moreover, real-world 3D data often contain artifacts and texture noise that introduce unavoidable redundancy. Mainstream methods typically use mean pooling to aggregate local fragments into global representations, assuming equal contribution from all parts. As a result, crucial geometric features are often diluted by semantically irrelevant noise. Therefore, the existing text-3D retrieval methods facing two fundamen-

tal challenges: Hierarchy Representation Collapse (HRC) and Redundancy-Induced Saliency Dilution (RISD).

First, HRC disrupts semantic consistency across modalities. Both natural language and 3D geometry inherently follow a tree-structured hierarchy, progressing from abstract to concrete semantics and from global to local geometric structures (Li et al. 2023c). The number of nodes increases exponentially with hierarchical depth. For example, as illustrated in the left panel of Figure 1, the text “A trophy-shaped, two-handled ceramic urn” denotes an abstract concept, while its corresponding 3D point cloud encodes fine-grained details such as the ornate handles and specific surface patterns. These details correspond to more specific instances within the same hierarchy. When this exponentially growing hierarchy is projected into Euclidean or conventional Riemannian spaces (Li et al. 2025a), which expand at most polynomially with radius, intrinsic tree distances become severely distorted. General concepts and specific instances are compressed into crowded regions, resulting in overlap and ambiguity within the embedding space. As a result, the model fails to preserve the intrinsic property that higher-level concepts should encompass more neighbors, whereas lower-level instances should remain distinguishable. This distortion contributes to simultaneous declines in both recall and precision.

Second, RISD amplifies discriminative errors during feature aggregation. Real-world 3D data often contain redundant fragments, such as scanning artifacts and decorative textures, while natural language descriptions may include non-discriminative elements, such as prepositions and function words. When equal-weighted strategies such as mean pooling are employed, the contribution of local fragments to the global representation is uniformly smoothed. Consequently, critical geometric and semantic cues are averaged out, diminishing the embedding’s ability to distinguish hard negative samples.

To address these challenges, we propose the Hyperbolic Hierarchical Alignment Reasoning Network ( $H^2$ ARN), which jointly embeds textual descriptions and 3D point clouds into a Lorentzian hyperbolic space with constant negative curvature. Due to its exponential volume growth with respect to radius (Gromov 1987), this space naturally accommodates tree-structured hierarchies (Figure 1, right panel). The origin represents the most general concepts, and embeddings positioned closer to the origin carry more abstract semantics, which inherently subsume the more specific instances situated farther away. Leveraging this geometric framework, we first impose a cross-modal constraint: text embeddings are required to lie closer to the origin than their corresponding point cloud embeddings, thereby spatially encoding the abstract-to-specific relationship. Next, we introduce a contribution-aware intra-modal aggregation mechanism. Local geometric features and word tokens are treated as leaf nodes and are contextually enriched via a self-attention module. These enriched representations, along with an initial global anchor obtained through mean pooling, are projected into hyperbolic space. The Lorentzian distance is then used to quantify each leaf node’s semantic contribution to the anchor, assigning it an importance weight. This guides a weighted aggregation process that produces a final global representation, seman-

tically cleaner and positioned deeper within the hierarchy. The optimization process incorporates two geometric loss functions. The first is a Lorentzian contrastive loss, which promotes instance-level alignment by pulling matched pairs closer and pushing mismatched pairs farther apart in hyperbolic space. The second is the Hierarchical Ordering Loss, which explicitly encodes the partial ordering of “text entails 3D” through entailment cones. Given a text embedding  $x$  and a point cloud embedding  $y$ , we construct a hyperbolic cone centered at  $x$  with a radially shrinking aperture. If  $y$  lies within the cone, the partial order is satisfied and no penalty is incurred. Otherwise, the loss is proportional to the angular deviation from the cone boundary. This mechanism dynamically adjusts the cone to encompass relevant 3D instances while excluding unrelated ones during training. The main contributions of this paper are summarized as follows:

- We propose the  $H^2$ ARN, which constructs a Lorentzian hyperbolic space with constant negative curvature and introduces a hierarchical ordering loss. By explicitly enforcing partial order constraints via entailment cones in the embedding space, our model effectively mitigates the problem of hierarchical representation collapse.
- We introduce a contribution-aware hyperbolic aggregation mechanism that leverages Lorentzian distance to estimate the semantic contribution of each local fragment. When jointly trained with hierarchical ordering loss, the model improves its ability to distinguish hard negative samples without additional supervision.
- We expand the T3DR-HIT dataset to 2.6 times its original size, increasing the number of text–3D pairs from 3,380 to 8,935. Our model demonstrates superior performance and generalization capabilities on the enlarged dataset. We believe that releasing this expanded dataset will benefit the broader research community.

## Related Work

### Cross-modal Retrieval

Cross-modal retrieval aims to bridge the semantic gap between different data modalities, with feature alignment being the central challenge (Tang et al. 2025; Yang, Li, and Cheng 2025; Yang et al. 2025; Li et al. 2025d). The most prominent subfield is image-text retrieval, where alignment strategies are often categorized into coarse-grained, fine-grained, and hybrid-grained approaches. Coarse-grained methods typically map entire images and sentences to a shared embedding space for holistic comparison (Faghri et al. 2017). Some enhance global representations by modeling intra-modal relationships using graph convolutional networks (Li et al. 2019) or transformers (Messina et al. 2021b; Li and Fan 2022; Li et al. 2023b), while others focus on improving the loss functions with instance-level constraints (Zheng et al. 2020), hierarchical relation modeling (Fu et al. 2023), or adversarial learning (Peng and Qi 2019). Fine-grained methods, in contrast, focus on aligning local features. A seminal work introduced cross-attention to discover latent alignments between visual and textual fragments (Lee et al. 2018). Subsequent approaches refined this core idea by extending alignment to the relation-level (Wei et al. 2020; Messina et al. 2021a), selectively

attending to salient local fragments based on global context (Zhang et al. 2020; Bao et al. 2023; Wang et al. 2025), or introducing innovative techniques such as multi-level contrastive learning (Wu et al. 2019) and probabilistic modeling (Li, Xiong, and Fan 2024). Hybrid-grained methods unify the strengths of both via strategies such as inferring more accurate matching scores through similarity attention filtering (Diao et al. 2021), integrating coarse and fine-grained learning into a unified framework with a consistency-constrained contrastive loss (Liu et al. 2023), or employing intra-modal fusion guidance and inter-modal bidirectional guidance (Chen et al. 2024a).

Inspired by these advances, research has extended to the 3D domain, initially focusing on text-3D shape retrieval. This sub-task focuses on matching textual descriptions with isolated 3D shapes, typically generated models. A pioneering study first achieved end-to-end text-shape retrieval by combining association learning with metric learning (Chen et al. 2018). Subsequent works have largely relied on multi-view renderings, representing 3D shapes as view sequences (Han et al. 2019), learning a trimodal embedding space (Ruan et al. 2024), or using self-and-cross-attention to aggregate multi-view and point cloud features (Lin et al. 2024), often supplemented by hard negative mining strategies (Wu et al. 2024). To move beyond reliance on 2D renderings, some methods have introduced direct matching between 3D shape parts and words (Tang et al. 2023) or proposed unified query transformers for joint understanding (Li et al. 2023c). However, by focusing on isolated objects lacking scene details and often neglecting the underlying data distribution, these methods can be susceptible to learning biases.

Different from previous text-3D benchmarks targeted specific scenarios like synthetic indoor scenes (Yu et al. 2024), a seminal study recently advanced the field by introducing T3DR-HIT, the first large-scale benchmark based on real-world scans of both coarse-grained scenes and fine-grained artifacts (Li et al. 2025a). The study also introduced a Riemannian attention mechanism to enhance retrieval accuracy. Despite these advances, prior methods primarily operate in Euclidean or conventional Riemannian spaces with limited volume growth, making them susceptible to HRC. Concurrently, their common use of equal-weighted aggregation strategies fails to effectively address the RISD problem. Our work directly addresses these gaps by leveraging the exponential capacity of Lorentzian hyperbolic geometry and a contribution-aware aggregation mechanism, thereby enabling more robust and precise text-3D retrieval.

## Method

In this section, we present in detail the modeling architecture and learning objectives of **H<sup>2</sup>ARN**. As illustrated in Figure 2, the H<sup>2</sup>ARN framework is composed of two primary modules: a **Structural Context Encoder** that refines local features in Euclidean space, and a **Hyperbolic Hierarchical Alignment Module** that embeds and aligns them in hyperbolic space. We first introduce the preliminaries of the Lorentz model of hyperbolic geometry. Then, we discuss the architectural components of our model and its optimization strategy.

## Preliminaries of Lorentz Model

Hyperbolic space  $\mathbb{H}^d$  is a Riemannian manifold with constant negative curvature. Its volume grows exponentially with geodesic radius, mirroring the branching of tree-structured data and easing the crowding that arises in Euclidean embeddings. Several coordinate models represent  $\mathbb{H}^d$ , for example the Poincaré ball or the upper half-space, but these embed the manifold in  $\mathbb{R}^d$  at the cost of metric distortion. We adopt the Lorentz model because it provides an isometric embedding in the  $(d + 1)$ -dimensional Minkowski space  $\mathbb{R}^{1,d}$ , preserving distances exactly and enabling stable closed-form geodesic operations. For curvature  $-c < 0$ , the Lorentz model is the future sheet of the two-sheeted hyperboloid in  $\mathbb{R}^{1,d}$ :

$$\mathbb{H}_c^d = \{\mathbf{u} \in \mathbb{R}^{d+1} : \langle \mathbf{u}, \mathbf{u} \rangle_{\mathcal{L}} = -\frac{1}{c}, u_{d+1} > 0\}, \quad (1)$$

where  $\langle \cdot, \cdot \rangle_{\mathcal{L}}$  denotes the Lorentz inner product. For two vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d+1}$  with spatial parts  $\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathbb{R}^d$  and time components  $u_{d+1}, v_{d+1} \in \mathbb{R}$ , the inner product is defined as  $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{L}} = \langle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_E - u_{d+1}v_{d+1}$ , where  $\langle \cdot, \cdot \rangle_E$  is the standard Euclidean inner product. All vectors on this manifold satisfy the constraint  $u_{d+1} = \sqrt{1/c + \|\tilde{\mathbf{u}}\|_E^2}$ . Based on the structure, we use the Lorentzian distance to measure the geometric distance between embedded points. This distance corresponds to the length of the shortest path (geodesic) on the manifold, effectively capturing hierarchical semantic relationships. For any two points  $\mathbf{u}, \mathbf{v} \in \mathbb{H}_c^d$ , their Lorentzian distance is defined as:

$$d_{\mathbb{H}}(\mathbf{u}, \mathbf{v}) = \frac{1}{\sqrt{c}} \operatorname{arccosh}(-c \langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{L}}). \quad (2)$$

To enable feature projection and optimization on the manifold, it is necessary to introduce the tangent space and its mapping to the manifold. The tangent space at a point  $\mathbf{w} \in \mathbb{H}_c^d$  is a  $d$ -dimensional Euclidean space that is orthogonal to  $\mathbf{w}$  under the Lorentz inner product:

$$T_{\mathbf{w}}\mathbb{H}_c^d = \{\mathbf{v} \in \mathbb{R}^{d+1} : \langle \mathbf{v}, \mathbf{w} \rangle_{\mathcal{L}} = 0\}, \quad (3)$$

where  $T_{\mathbf{w}}\mathbb{H}_c^d$  denotes the tangent space at point  $\mathbf{w}$ . The exponential map,  $\exp_{\mathbf{w}}^c : T_{\mathbf{w}}\mathbb{H}_c^d \rightarrow \mathbb{H}_c^d$ , serves as a crucial bridge, projecting a vector from the flat tangent space onto the curved manifold. For a general point  $\mathbf{w}$  and a tangent vector  $\mathbf{v} \in T_{\mathbf{w}}\mathbb{H}_c^d$ , this map is defined as:

$$\exp_{\mathbf{w}}^c(\mathbf{v}) = \cosh(\sqrt{c}\|\mathbf{v}\|_{\mathcal{L}})\mathbf{w} + \frac{\sinh(\sqrt{c}\|\mathbf{v}\|_{\mathcal{L}})}{\sqrt{c}\|\mathbf{v}\|_{\mathcal{L}}}\mathbf{v}, \quad (4)$$

where  $\|\mathbf{v}\|_{\mathcal{L}} = \sqrt{|\langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{L}}|}$  is the Lorentzian norm.

The exponential map serves as the bridge for lifting Euclidean features into the hyperbolic manifold. In practice, we focus on the exponential map centered at the hyperboloid origin  $\mathbf{o} = (0, \dots, 0, 1/\sqrt{c})$ , since any Euclidean feature vector  $\mathbf{v} \in \mathbb{R}^d$  lies in the tangent space  $T_{\mathbf{o}}\mathbb{H}_c^d$ . Its temporal component is zero, ensuring that  $\langle \mathbf{o}, \mathbf{v} \rangle_{\mathcal{L}} = 0$  holds automatically. Substituting  $\mathbf{w} = \mathbf{o}$  into the general Lorentz exponential map yields the embedded point  $\mathbf{u} = \exp_{\mathbf{o}}^c(\mathbf{v})$ , where the spatial component is  $\mathbf{v}$  scaled by a hyperbolic factor and the time coordinate is determined by curvature. This faithfully positions

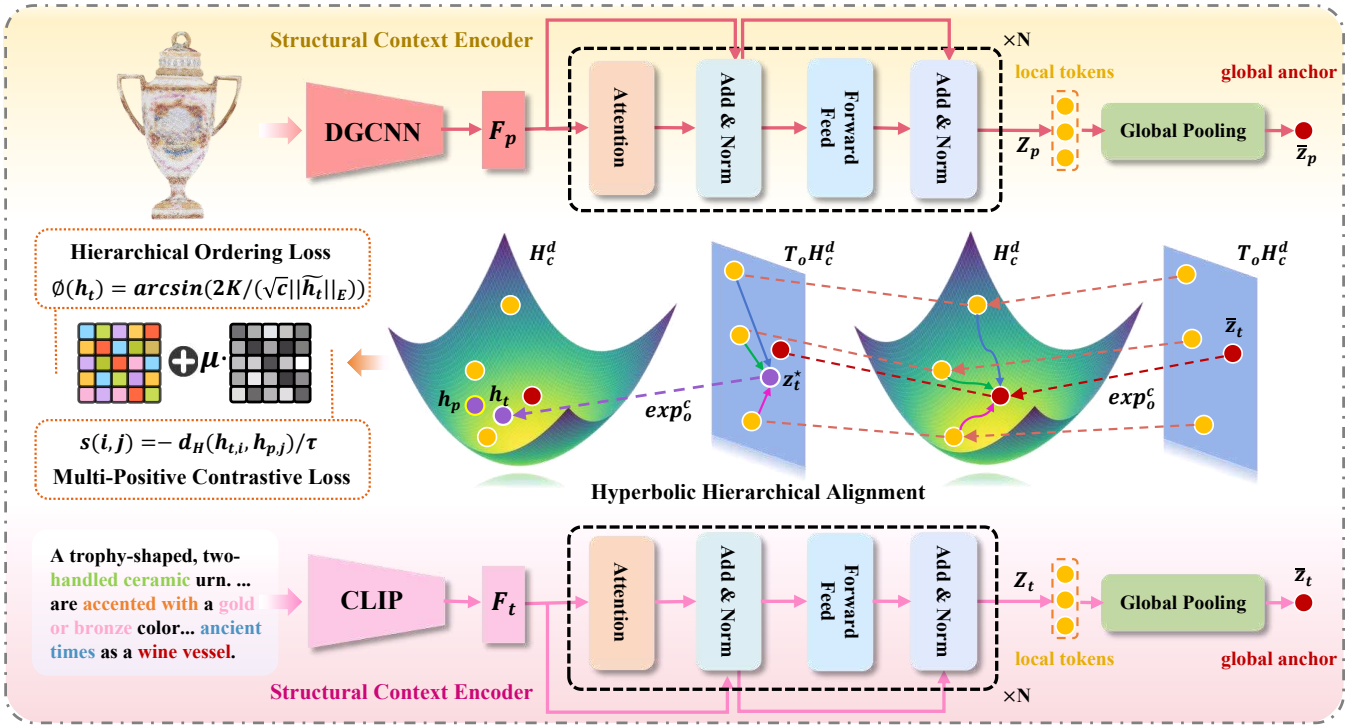


Figure 2: An overview of the  $H^2ARN$  architecture. The Structural Context Encoder first refines local features from each modality in Euclidean space to produce context-aware representations. Subsequently, the Hyperbolic Hierarchical Alignment Module aligns the features in hyperbolic space via a contribution-aware aggregation mechanism and a dual geometric loss, preserving their semantic hierarchy.

the feature on the constant-curvature manifold and facilitates geometry-aware learning, which can be written as:

$$\tilde{\mathbf{u}} = \frac{\sinh(\sqrt{c}\|\mathbf{v}\|_E)}{\sqrt{c}\|\mathbf{v}\|_E} \mathbf{v}, \quad u_{d+1} = \frac{\cosh(\sqrt{c}\|\mathbf{v}\|_E)}{\sqrt{c}}, \quad (5)$$

where the time component  $u_{d+1}$  can be equivalently derived from the spatial component via the hyperboloid constraint.

### Structural Context Encoder

To capture each modality’s intrinsic structure and furnish context-aware representations, we first feed raw inputs to strong modality-specific backbones, yielding local feature sequences  $F_t \in \mathbb{R}^{L_t \times D_t}$  for text and  $F_p \in \mathbb{R}^{L_p \times D_p}$  for point clouds. Text tokens are encoded with CLIP (Radford et al. 2021), whose large-scale vision–language pre-training supplies rich semantic priors for nuanced descriptions. Point clouds are processed by DGCNN (Wang et al. 2019), which dynamically constructs neighbourhood graphs to model point–point relations while integrating colour channels to preserve fine visual detail. Although these encoders excel at local pattern extraction, the resulting sequences lack global awareness of long-range dependencies. We therefore project the features to a shared latent dimension  $d$  and pass them through a stack of Pre-Layer-Norm Transformer blocks. In each block, the sequence is linearly mapped to query, key and value tensors, followed by multi-head scaled dot-product attention that captures cross-token interactions; position-wise

feed-forward networks and residual connections then refine the context. This self-attention pipeline endows both modalities with coherent, context-enhanced embeddings that are ready for subsequent hyperbolic alignment.

### Hyperbolic Hierarchical Alignment Module

This module is the core of our model, specifically designed to address the dual challenges of HRC and RISD. It achieves this through a novel Contribution-Aware Hyperbolic Aggregation mechanism and a Dual Geometric Loss Function that is computed entirely within the hyperbolic geometry.

**Contribution-Aware Hyperbolic Aggregation.** The contribution-aware hyperbolic aggregation mechanism addresses RISD by weighting local tokens based on their geometric relevance to global semantics. Let  $Z_t$  and  $Z_p$  denote the context-enhanced Euclidean token matrices from the previous stage, with each row vector  $z_i$  regarded as a leaf node. We first compute an initial anchor  $\bar{z} = \frac{1}{L} \sum_{i=1}^L z_i$  and map both the anchor and all leaf nodes to  $\mathbb{H}_c^d$  using the exponential map defined in Eq. (5). For each leaf, we calculate its Lorentzian distance to the anchor as defined in Eq. (2). A softmax over the negative distances yields contribution weights  $\omega_i$  that reflect semantic saliency. These weights guide the weighted Euclidean sum  $z^* = \sum_{i=1}^L \omega_i z_i$ . Since  $z^*$  has a smaller norm than any individual  $z_i$ , its hyperbolic image naturally lies closer to the origin, capturing a more abstract and denoised global concept. The resulting

root embeddings,  $\mathbf{h}_t = \exp_{\mathbb{O}}^c(z_t^*)$  and  $\mathbf{h}_p = \exp_{\mathbb{O}}^c(z_p^*)$ , provide semantically purified representations for text and point clouds, effectively countering the dilution of key cues by redundant fragments. To prevent numerical overflow in the exponential map, each local feature matrix  $Z \in \mathbb{R}^{L \times d}$  is scaled by a learnable, modality-specific factor  $\alpha$ , i.e.,  $Z' = \alpha Z$ , prior to aggregation.

**Dual Geometric Loss Function.** Our optimization objective structures the embedding space to simultaneously ensure instance-level discrimination and preserve the inter-modal abstract-to-specific semantic hierarchy. This is achieved through two synergistic components: a multi-positive contrastive loss constructed from negative Lorentzian distance, which aligns instances by pulling positive pairs closer and pushing negative pairs apart, and a hierarchical ordering loss, which explicitly enforces the "text entails 3D" partial order using entailment cones with a radially narrowing scope.

**The Multi-Positive Contrastive Loss.** The multi-positive contrastive loss  $\mathcal{L}_{\text{cont}}$  operates at the instance level. We define the similarity score  $s(i, j)$  between the final global root embeddings of a text instance  $i$  ( $\mathbf{h}_{t,i}$ ) and a 3D instance  $j$  ( $\mathbf{h}_{p,j}$ ) as their negative Lorentzian distance:

$$s(i, j) = -d_{\mathbb{H}}(\mathbf{h}_{t,i}, \mathbf{h}_{p,j}) / \tau, \quad (6)$$

where  $\tau$  is a temperature hyperparameter. Based on this similarity, we construct a symmetric InfoNCE-style loss adapted for scenarios where multiple positive samples may exist for a given query in a batch. The loss from text to point cloud,  $\mathcal{L}_{t \rightarrow p}$ , is formulated as:

$$\mathcal{L}_{t \rightarrow p} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\sum_{j \in \mathcal{P}_i} e^{s(i,j)}}{\sum_{k=1}^B e^{s(i,k)}}, \quad (7)$$

where  $B$  is the batch size and  $\mathcal{P}_i$  is the set of indices of point clouds positive to text  $i$ . To ensure bidirectional alignment, the full contrastive loss  $\mathcal{L}_{\text{cont}}$  is the symmetric average of this and the corresponding point-cloud-to-text loss  $\mathcal{L}_{p \rightarrow t}$ :

$$\mathcal{L}_{\text{cont}} = \frac{1}{2} (\mathcal{L}_{t \rightarrow p} + \mathcal{L}_{p \rightarrow t}). \quad (8)$$

**The Hierarchical Ordering Loss.** The hierarchical ordering loss  $\mathcal{L}_{\text{ord}}$  mitigates HRC by embedding the "text entails 3D" partial order directly in hyperbolic geometry through entailment cones. For every text root embedding  $\mathbf{h}_t \in \mathbb{H}_c^d$  we define a hyperbolic cone whose axis is  $\mathbf{h}_t$  and whose half-aperture  $\phi(\mathbf{h}_t)$  contracts as the vector drifts outward from the origin, thereby capturing the intuition that concepts become more specific and semantically narrower with increasing radius. The half-aperture can be written as:

$$\phi(\mathbf{h}_t) = \arcsin \left( \frac{2K}{\sqrt{c} \|\tilde{\mathbf{h}}_t\|_E} \right), \quad (9)$$

where  $\|\tilde{\mathbf{h}}_t\|_E$  is the Euclidean norm of the spatial component of the text embedding,  $c$  is the curvature, and  $K = 0.1$  caps the maximal cone width for concepts near the origin. Given a paired 3D embedding  $\mathbf{h}_p$ , we compute the exterior angle between the cone's axis  $\mathbf{h}_t$  and  $\mathbf{h}_p$ :

$$\theta(\mathbf{h}_t, \mathbf{h}_p) = \arccos \left( \frac{h_{p,d+1} + c \cdot h_{t,d+1} \langle \mathbf{h}_t, \mathbf{h}_p \rangle_{\mathcal{L}}}{\|\tilde{\mathbf{h}}_t\|_E \sqrt{(c \langle \mathbf{h}_t, \mathbf{h}_p \rangle_{\mathcal{L}})^2 - 1}} \right), \quad (10)$$

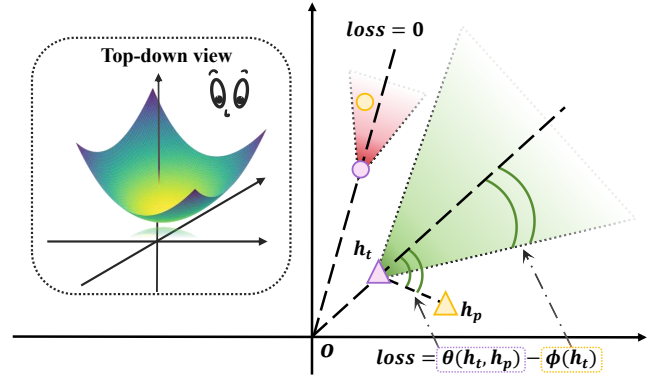


Figure 3: Geometric illustration of the Hierarchical Ordering Loss. The loss enforces the "text entails 3D" partial order by penalizing a 3D embedding  $h_p$  only if it lies outside the entailment cone defined by its corresponding text embedding  $h_t$ . The penalty is proportional to the difference between the exterior angle  $\theta(h_t, h_p)$  and the cone's half-aperture  $\phi(h_t)$ .

where  $h_{.,d+1}$  denotes the time component and  $\langle \cdot, \cdot \rangle_{\mathcal{L}}$  is the Lorentz inner product. The ordering loss penalises only those 3D points that fall outside the cone. As illustrated in Figure 3, if the exterior angle exceeds the aperture, a penalty proportional to their difference is incurred. For a positive pair  $(\mathbf{h}_t, \mathbf{h}_p)$ , the loss is:

$$\mathcal{L}_{\text{ord}} = \max(0, \theta(\mathbf{h}_t, \mathbf{h}_p) - \phi(\mathbf{h}_t)). \quad (11)$$

This nonsymmetric geometric constraint forces the text embedding to occupy a more general, "ancestor" position relative to its specific 3D instance, thereby constructing the desired hierarchy and preventing its collapse. The final training objective is a weighted sum of these two losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cont}} + \lambda \mathcal{L}_{\text{ord}}, \quad (12)$$

where the hyperparameter  $\lambda$  balances discrimination against hierarchical consistency.

## Experiments

### Experimental Settings

**Datasets.** We validate our model using the original T3DR-HIT dataset as well as an expanded version, referred to as T3DR-HIT v2. The original dataset includes 3,380 text-3D pairs, but it presents a notable limitation: fine-grained artifact scenes are described by only a single caption per object, in contrast to indoor scenes, which typically include at least three captions. To correct this imbalance and evaluate scalability, we utilized the LLaVA large language model (llava-v1.6-mistral-7b-hf) to generate three additional, distinct captions for each artifact using a set of diverse prompts. This enhancement results in richer and more comprehensive textual representations. Concurrently, we also expanded the point cloud data by incorporating additional artifacts captured from the Elephant Meta Dataset provided by the Henan Broadcasting and Television Station. The augmented dataset, T3DR-HIT v2, contains a total of 8,935

Datasets	Methods	Backbone		Text $\rightarrow$ PC			PC $\rightarrow$ Text			Rsum	Hyperparameters		
		Text	Point Cloud	R@1	R@5	R@10	R@1	R@5	R@10		Batch Size	Nhead	SA Layers
T3DR-HIT	RMARN <sub>2025</sub>	CLIP	PointNet	13	39	47	-	-	-	99	64	16	6
	RMARN <sub>2025</sub>	CLIP	PointNet	19	50	53	-	-	-	122	32	32	6
	RMARN <sub>2025</sub>	BERT	PointNet++	25	58	62	-	-	-	145	64	32	6
	RMARN <sub>2025</sub>	BERT	PointNet++	31	61	69	-	-	-	161	64	32	8
	<b>H<sup>2</sup>ARN (Ours)</b>	<b>CLIP</b>	<b>DGCNN</b>	<b>32</b>	<b>63</b>	<b>73</b>	-	-	-	<b>168</b>	<b>256</b>	<b>64</b>	<b>6</b>
T3DR-HIT v2	RMARN <sub>2025</sub>	CLIP	PointNet	7.6	25.2	37.7	6.5	20.0	30.3	127.3	64	16	6
	RMARN <sub>2025</sub>	CLIP	DGCNN	13.4	38.3	58.3	18.4	40.9	51.0	220.3	32	32	8
	<b>H<sup>2</sup>ARN (Ours)</b>	<b>CLIP</b>	<b>DGCNN</b>	<b>15.6</b>	<b>43.3</b>	<b>58.6</b>	<b>15.0</b>	<b>37.2</b>	<b>55.2</b>	<b>224.9</b>	<b>256</b>	<b>16</b>	<b>6</b>
	<b>H<sup>2</sup>ARN (Ours)</b>	<b>CLIP</b>	<b>DGCNN</b>	<b>14.7</b>	<b>42.9</b>	<b>59.5</b>	<b>18.5</b>	<b>44.7</b>	<b>54.4</b>	<b>234.7</b>	<b>32</b>	<b>32</b>	<b>8</b>
	<b>H<sup>2</sup>ARN (Ours)</b>	<b>CLIP</b>	<b>DGCNN</b>	<b>16.9</b>	<b>44.4</b>	<b>59.4</b>	<b>16.4</b>	<b>41.0</b>	<b>56.9</b>	<b>235.0</b>	<b>256</b>	<b>32</b>	<b>8</b>
	<b>H<sup>2</sup>ARN (Ours)</b>	<b>CLIP</b>	<b>DGCNN</b>	<b>16.4</b>	<b>44.5</b>	<b>60.6</b>	<b>19.6</b>	<b>42.3</b>	<b>55.1</b>	<b>238.5</b>	<b>256</b>	<b>64</b>	<b>6</b>

Table 1: Performance comparison on the T3DR-HIT and our expanded T3DR-HIT v2 datasets.

text-3D pairs, representing a 2.6-fold increase in size. We partition the dataset into an 80:20 training and testing split and conduct experiments on both versions to thoroughly assess the effectiveness and robustness of our proposed model.

**Evaluation Metrics.** To quantitatively evaluate retrieval performance, we employ two standard metrics widely used in cross-modal retrieval: Recall@K (R@K) and Rsum. R@K is defined as the proportion of queries for which the correct corresponding item is found within the top-K retrieved results. We report R@K for  $K=\{1, 5, 10\}$ , as this reflects performance at different levels of retrieval precision. To provide a single, comprehensive measure of overall performance, we also report Rsum, which is the sum of all R@K values across both retrieval directions (text-to-point cloud and point cloud-to-text). For all metrics, higher values signify better retrieval performance.

**Implementation Details.** We implement our H<sup>2</sup>ARN with the following architectural parameters and training settings. The shared latent dimension  $d$  for all embeddings is set to 512, and the initial feature dimensions for both text ( $D_t$ ) and point clouds ( $D_p$ ) are also 512. The local feature sequence lengths are fixed at  $L_t = 77$  for text and  $L_p = 100$  for point clouds. Key parameters of the hyperbolic space are learnable. To ensure its positivity, the curvature parameter  $c$  is parameterized via its logarithm, i.e., the model learns  $\log(c)$ , and is initialized to  $c = 1.0$ . Similarly, to prevent numerical overflow during the hyperbolic projection, a modality-specific scaling factor  $\alpha$  is applied to the local Euclidean features before aggregation. This factor is also learned via its logarithm and initialized as  $\alpha = 1/\sqrt{d}$ .

We train the model for 100 epochs using a batch size of 256. For optimization, we employ the AdamW optimizer with a learning rate of  $2 \times 10^{-3}$ , and parameters  $\beta_1 = 0.91$ ,  $\beta_2 = 0.9993$ , and  $\epsilon = 10^{-8}$ . A linear learning rate scheduler with a warmup phase over the first 10% of total training steps is used to stabilize training. For the dual geometric loss, the temperature  $\tau$  in the contrastive loss is set to 0.07. The weight  $\lambda$  for the hierarchical ordering Loss is set to 0.2, and the constant  $K$  for the entailment cone is 0.1.

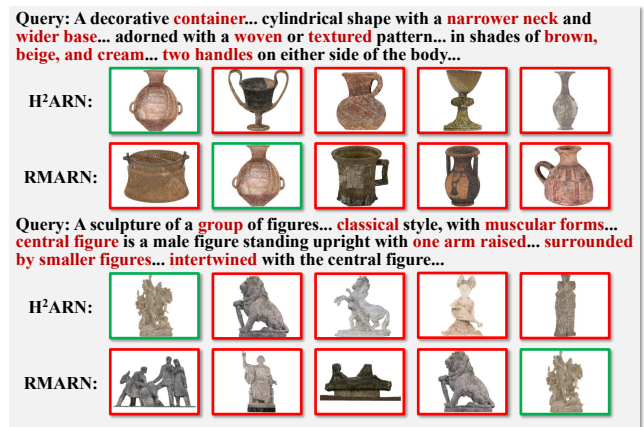


Figure 4: Qualitative comparison of text-to-3D retrieval results on the T3DR-HIT v2 dataset. For each query, the top-5 retrieved point clouds are shown, ranked from left to right by matching score. Green boxes indicate correct matches, while red boxes indicate incorrect ones.

## Performance Comparison

We compare H<sup>2</sup>ARN with RMARN, the method that introduced the T3DR-HIT dataset and remains the only published baseline. On both the original and the expanded T3DR-HIT v2 datasets, H<sup>2</sup>ARN consistently achieves superior quantitative performance, as shown in Table 1. On the original dataset, our model, equipped with CLIP and DGCNN backbones, sets a new state-of-the-art by outperforming RMARN across all R@K metrics for text-to-point cloud retrieval. This performance advantage becomes even more evident on the more challenging T3DR-HIT v2 dataset. When compared under identical backbone configurations, H<sup>2</sup>ARN significantly outperforms RMARN in both retrieval directions, achieving an Rsum of 238.5 compared to RMARN’s 220.3. This marked improvement, independent of feature extractor choice, underscores the effectiveness of our key innovations: the use of hyperbolic geometry to address HRC and the introduc-

Methods	Text $\rightarrow$ PC		PC $\rightarrow$ Text		Rsum
	R@1	R@5	R@1	R@5	
<b>H<sup>2</sup>ARN (Ours)</b>	<b>16.4</b>	<b>44.5</b>	<b>19.6</b>	<b>42.3</b>	<b>238.5</b>
w/o $\mathcal{L}_{ord}$	15.3	40.1	18.4	41.1	229.6
w/o Aggregation	15.2	43.6	16.9	41.2	233.5
w/o both	14.3	41.8	14.5	37.5	222.0

Table 2: Ablation study on the core components of H<sup>2</sup>ARN.

Methods	Text $\rightarrow$ PC		PC $\rightarrow$ Text		Rsum
	R@1	R@5	R@1	R@5	
Eu + MP	10.1	38.8	12.5	32.5	196.3
Eu + CA	12.5	41.0	14.2	36.4	215.1
<b>H<sup>2</sup>ARN (Ours)</b>	<b>16.4</b>	<b>44.5</b>	<b>19.6</b>	<b>42.3</b>	<b>238.5</b>

Table 3: Ablation study on the effect of the hyperbolic space.

tion of contribution-aware aggregation to mitigate RISD. The results confirm our model’s robustness and scalability to diverse scenarios, with qualitative examples in Figure 4 further demonstrating its accurate fine-grained retrieval.

### Ablation Study

In this section, we conduct a series of ablation studies to analyze the impact of our key design choices.

**Effect of Core Components.** We first validate the effectiveness of our two primary contributions: the Hierarchical Ordering Loss ( $\mathcal{L}_{ord}$ ) and the Contribution-Aware Hyperbolic Aggregation mechanism. As shown in Table 2, removing the  $\mathcal{L}_{ord}$  results in a significant performance drop, with the Rsum decreasing from 238.5 to 229.6. This decline highlights the crucial role of  $\mathcal{L}_{ord}$  in explicitly modeling the semantic hierarchy and resolving the HRC problem. The impact of the Contribution-Aware Aggregation is also marked. When we ablate it by using the initial hyperbolic anchor (the mapped mean of leaf nodes) directly as the global representation, the model’s performance degrades, with the Rsum falling to 233.5. This demonstrates that our aggregation mechanism is vital for filtering out redundant information and mitigating the RISD problem. Removing both components leads to a severe performance collapse to an Rsum of 222.0.

**Effect of Hyperbolic Space.** To verify the fundamental importance of hyperbolic geometry, we compare our full model against two Euclidean-based variants, with results in Table 3. Our first baseline, *Eu + MP*, performs mean pooling on context-aware features and uses standard cosine similarity for alignment, yielding the lowest Rsum of 196.3. We then apply our contribution-aware aggregation within the Euclidean framework (*Eu + CA*), where weights are determined by the dot product between leaves and their anchor. This variant significantly improves performance to an Rsum of 215.1, confirming that our aggregation mechanism is effective at mitigating RISD even without hyperbolic geometry. Finally, our full H<sup>2</sup>ARN model, which leverages the hyperbolic embedding space and the hierarchical ordering loss, achieves the

(Nhead, Layers)	Text $\rightarrow$ PC		PC $\rightarrow$ Text		Rsum
	R@1	R@5	R@1	R@5	
(16, 4)	15.4	39.4	15.7	35.6	223.2
(16, 6)	15.6	43.3	15.0	37.2	224.9
(16, 8)	14.5	40.7	14.1	36.6	210.2
(32, 4)	16.7	<b>46.1</b>	14.3	38.6	226.4
(32, 6)	16.5	43.2	14.4	41.7	231.0
(32, 8)	16.9	44.4	16.4	41.0	235.0
(64, 4)	15.2	41.3	<b>20.3</b>	40.5	223.5
(64, 6)	16.4	44.5	19.6	42.3	<b>238.5</b>
(64, 8)	<b>17.0</b>	42.6	18.3	<b>46.4</b>	237.4

Table 4: Ablation study on the number of attention heads and self-attention layers on the T3DR-HIT v2 dataset.

best performance by a large margin. This progression directly validates the superiority of hyperbolic geometry.

**Effect of Self-Attention Layers.** We further analyze the impact of the number of attention heads and layers in the structural context encoder. The results, presented in Table 4, reveal a complex interplay between model width and depth, with no single configuration dominating all individual metrics. The optimal settings for specific R@K metrics are distributed across different architectures. For instance, the best R@5 score for text-to-point cloud retrieval is achieved by a relatively shallow model (32 heads, 4 layers) at 46.1, whereas the point cloud-to-text direction favors a deeper, wider architecture (64 heads, 8 layers) with a score of 46.4. This indicates a clear trade-off between model width and depth depending on the specific evaluation criterion. Despite the varied performance on individual metrics, the configuration with 64 attention heads and 6 layers achieves the highest overall performance, reaching a peak Rsum of 238.5.

## Conclusion

This paper introduces the Hyperbolic Hierarchical Alignment Reasoning Network (H<sup>2</sup>ARN) for text-3D retrieval. By leveraging Lorentzian hyperbolic geometry, H<sup>2</sup>ARN embeds both text and 3D data in a space naturally suited to hierarchical structures, enforcing semantic entailment through a geometric ordering loss. Its contribution-aware aggregation mechanism further enhances discriminative power by emphasizing semantically relevant features. Experimental results on both the original and expanded T3DR-HIT datasets confirm that H<sup>2</sup>ARN significantly outperforms existing methods, demonstrating superior generalization and robustness. These contributions not only advance the state of the art in text-3D retrieval but also lay foundational groundwork for future cross-modal hyperbolic representations.

## Acknowledgments

This work was supported in part by the National Key R&D Program of China (2023YFA1008501) and the National Natural Science Foundation of China (NSFC) under grants 624B2049 and U22B2035.

## References

- Bai, H.; Zhang, J.; Zhao, Z.; Wu, Y.; Deng, L.; Cui, Y.; Feng, T.; and Xu, S. 2025. Task-driven Image Fusion with Learnable Fusion Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7457–7468.
- Bai, H.; Zhao, Z.; Zhang, J.; Wu, Y.; Deng, L.; Cui, Y.; Jiang, B.; and Xu, S. 2024. ReFusion: Learning Image Fusion from Reconstruction with Learnable Loss Via Meta-Learning. *International Journal of Computer Vision*, 1–21.
- Bao, L.; Wei, L.; Zhou, W.; Liu, L.; Xie, L.; Li, H.; and Tian, Q. 2023. Multi-granularity matching transformer for text-based person search. *IEEE Transactions on Multimedia*, 26: 4281–4293.
- Bao, Q.; Liu, F.; Jiao, L.; Liu, Y.; Li, S.; Li, L.; Liu, X.; and Chen, P. 2025. Visual-Language Scene-Relation-Aware Zero-Shot Captioner. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10): 8725–8739.
- Bao, Q.; Liu, F.; Liu, Y.; Jiao, L.; Liu, X.; and Li, L. 2022. Hierarchical Scene Normality-Binding Modeling for Anomaly Detection in Surveillance Videos. In *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, 6103–6112. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.
- Chen, K.; Choy, C. B.; Savva, M.; Chang, A. X.; Funkhouser, T.; and Savarese, S. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, 100–116. Springer.
- Chen, Y.; Huang, J.; Xiong, S.; and Lu, X. 2024a. Integrating multisubspace joint learning with multilevel guidance for cross-modal retrieval of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.
- Chen, Z.; Lai, Z.; Chen, J.; and Li, J. 2024b. Mind Marginal Non-Crack Regions: Clustering-Inspired Representation Learning for Crack Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12698–12708.
- Chen, Z.; Zhang, J.; Lai, Z.; Chen, J.; Liu, Z.; and Li, J. 2022. Geometry-Aware Guided Loss for Deep Crack Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4703–4712.
- Chen, Z.; Zhang, J.; Lai, Z.; Zhu, G.; Liu, Z.; Chen, J.; and Li, J. 2023. The Devil is in the Crack Orientation: A New Perspective for Crack Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6653–6663.
- Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1218–1226.
- Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.
- Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.
- Gromov, M. 1987. Hyperbolic groups. In *Essays in group theory*, 75–263. Springer.
- Han, Z.; Shang, M.; Wang, X.; Liu, Y.-S.; and Zwicker, M. 2019. Y2Seq2Seq: Cross-modal representation learning for 3D shape and text by joint reconstruction and prediction of view and word sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 126–133.
- Lee, J. M. 2018. *Introduction to Riemannian manifolds*, volume 2. Springer.
- Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 201–216.
- Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4654–4662.
- Li, W.; and Fan, X. 2022. Image-text alignment and retrieval using light-weight transformer. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4758–4762. IEEE.
- Li, W.; Han, W.; Chen, Y.; Chai, Y.; Lu, Y.; Wang, X.; and Fan, X. 2025a. Riemann-based multi-scale attention reasoning network for text-3D retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18485–18493.
- Li, W.; Han, W.; Deng, L.-J.; Xiong, R.; and Fan, X. 2025b. Spiking Variational Graph Representation Inference for Video Summarization. *IEEE Transactions on Image Processing*, 34: 5697–5709.
- Li, W.; Ma, Z.; Deng, L.-J.; Fan, X.; and Tian, Y. 2023a. Neuron-Based Spiking Transmission and Reasoning Network for Robust Image-Text Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3516–3528.
- Li, W.; Ma, Z.; Shi, J.; and Fan, X. 2023b. The style transformer with common knowledge optimization for image-text retrieval. *IEEE Signal Processing Letters*, 30: 1197–1201.
- Li, W.; Wang, P.; Wang, X.; Zuo, W.; Fan, X.; and Tian, Y. 2025c. Multi-Timescale Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11): 10772–10786.
- Li, W.; Wang, P.; Xiong, R.; and Fan, X. 2024. Spiking Tucker Fusion Transformer for Audio-Visual Zero-Shot Learning. *IEEE Transactions on Image Processing*, 33: 4840–4852.
- Li, W.; Xiong, R.; and Fan, X. 2024. Multi-layer probabilistic association reasoning network for image-text retrieval. *IEEE transactions on circuits and systems for video technology*, 34(10): 9706–9717.
- Li, W.; Yang, Z.; Han, W.; Man, H.; Wang, X.; and Fan, X. 2025d. Hyperbolic-Constraint Point Cloud Reconstruction from Single RGB-D Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 4959–4967.
- Li, X.; Ding, J.; Chen, Z.; and Elhoseiny, M. 2023c. Uni3dl: Unified model for 3d and language understanding. *arXiv preprint arXiv:2312.03026*.

- Li, Z.; Liao, J.; Tang, C.; Zhang, H.; Li, Y.; Bian, Y.; Sheng, X.; Feng, X.; Li, Y.; Gao, C.; et al. 2025e. USTC-TD: A test dataset and benchmark for image and video coding in 2020s. *IEEE Transactions on Multimedia*.
- Lin, D.; Cheng, Y.; Guo, A.; Mao, S.; and Li, Y. 2024. SCA-PVNet: Self-and-cross attention based aggregation of point cloud and multi-view for 3D object retrieval. *Knowledge-Based Systems*, 296: 111920.
- Liu, C.; Zhang, Y.; Wang, H.; Chen, W.; Wang, F.; Huang, Y.; Shen, Y.-D.; and Wang, L. 2023. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32: 3622–3633.
- Messina, N.; Amato, G.; Esuli, A.; Falchi, F.; Gennaro, C.; and Marchand-Maillet, S. 2021a. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4): 1–23.
- Messina, N.; Falchi, F.; Esuli, A.; and Amato, G. 2021b. Transformer reasoning network for image-text matching and retrieval. In *2020 25th International conference on pattern recognition (ICPR)*, 5222–5229. IEEE.
- Peng, Y.; and Qi, J. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1): 1–24.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Ruan, Y.; Lee, H.-H.; Zhang, Y.; Zhang, K.; and Chang, A. X. 2024. Tricolo: Trimodal contrastive loss for text to shape retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5815–5825.
- Tang, C.; Yang, X.; Wu, B.; Han, Z.; and Chang, Y. 2023. Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6884–6893.
- Tang, H.; Yan, R.; Yin, X.; Zhang, Q.; Zhang, X.; Ma, S.; Gao, W.; and Jia, C. 2025. HGC-Avatar: Hierarchical Gaussian Compression for Streamable Dynamic 3D Avatars. *arXiv preprint arXiv:2510.16463*.
- Wang, D.; Tian, J.; Liang, X.; Tian, Y.; and He, L. 2025. Global-aware Fragment Representation Aggregation Network for image–text retrieval. *Pattern Recognition*, 159: 111085.
- Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.
- Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.
- Wu, H.; Li, R.; Wang, H.; and Xiong, H. 2024. COM3D: Leveraging Cross-View Correspondence and Cross-Modal Mining for 3D Retrieval. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.
- Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W.-Y. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6609–6618.
- Xiao, Z.; Li, Z.; and Jia, W. 2025. Occlusion-Embedded Hybrid Transformer for Light Field Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8700–8708.
- Xiao, Z.; and Wang, X. 2025. Event-based Video Super-Resolution via State Space Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12564–12574.
- Yang, Z.; Li, W.; and Cheng, G. 2025. SHMamba: Structured Hyperbolic State Space Model for Audio-Visual Question Answering. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 3582–3593.
- Yang, Z.; Li, W.; Hou, J.; and Cheng, G. 2025. Multi-modal spiking tensor regression network for audio-visual zero-shot learning. *Neurocomputing*, 629: 129636.
- Yu, F.; Wang, Z.; Li, D.; Zhu, P.; Liang, X.; Wang, X.; and Okumura, M. 2024. Towards cross-modal point cloud retrieval for indoor scenes. In *International Conference on Multimedia Modeling*, 89–102. Springer.
- Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3536–3545.
- Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; and Zhang, L. 2025a. Perceive-IR: Learning to Perceive Degradation Better for All-in-One Image Restoration. *IEEE Transactions on Image Processing*, 1–1.
- Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; and Du, B. 2025b. UniUIR: Considering Underwater Image Restoration as an All-in-One Learner. *IEEE Transactions on Image Processing*, 34: 6963–6977.
- Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23.