This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Curvature-Aware Captioning: Leveraging Geodesic Attention for 3D Scene Understanding

Ziyao He , Yingjie Liu , ZhangYangRui , Mingsong Chen , Xuan Tang , Xian Wei*
East China Normal University
3663 N. Zhongshan Rd., 200062
*xwei@sei.ecnu.edu.cn

Abstract

Accurate 3D scene description is fundamental to robotic navigation and augmented reality, yet current dense captioning methods face significant limitations in processing sparse point cloud data. Existing approaches that apply Euclidean embedding spaces struggle to simultaneously preserve fine-grained local geometric details and model exponentially growing global semantic hierarchies, leading to either inaccurate localization or disjointed, shallow scene descriptions. In this work, we propose a novel **CURVATURE-AWARE CAPTIONING** *framework, integrating novel non-Euclidean geodesic attention mechanisms, to resolve the localization-contextualization conflict. Specifically, self-attention within Oblique space enforces dimensional homogeneity while establishing long-range dependencies. Bidirectional geodesic cross-attention within Lorentz space models hierarchical semantic relationships across scene instances, enabling simultaneous precision in object localization and coherence in scene descriptions. Theoretical analysis confirms that the curvature complementarity between the Oblique manifold and Lorentz hyperboloid resolves the Euclidean-hyperbolic conflict, ensuring feature stability via isotropic optimization while preserving inherent hierarchical relationships. Extensive experiments on ScanRefer and Nr3D benchmarks demonstrate state-of-the-art performance, with significant gains in both localization accuracy and descriptive richness.*

1. Introduction

3D dense captioning, which requires simultaneous object localization and descriptive caption generation within 3D scenes (e.g., point clouds) [13, 29], has seen significant methodological evolution. Early approaches predominantly adopted a "detect-then-describe" pipeline [6, 10, 14, 22, 46], suffering from cumulative errors. End-to-end set prediction frameworks address this by decoupling localization and caption decoding, exemplified by methods like *Vote2Cap-DETR++* [12].

Recent end-to-end transformer models like SpaCap3D [46], UniT3D [15], Vote2Cap-DETR++ [12] and BiCA [25] have achieved state-of-the-art (SOTA) performance in 3D dense captioning by leveraging unified attention mechanisms to integrate multi-modal features. These frameworks dynamically align visual cues with linguistic context through cross-modal attention, effectively resolving spatial-semantic dependencies while maintaining geometric consistency across scales. Notably, the recently proposed BiCA [25] method introduces a *bidirectional attention mechanism* that excels at integrating relationships between instance features and contextual features.

However, while capturing both local object features and global scene context, existing methods still struggle with **geometric conflicts** between fine-grained spatial localization and hierarchical semantic reasoning. Specifically, local object cues (e.g., surface geometry) demand Euclidean-like flat metrics, while global context (e.g., object hierarchies) inherently aligns with hyperbolic curvature for representing exponentially growing semantic distances [5]. Prior works in non-Euclidean deep learning [27] and hierarchical representation [52, 54] suggest that modeling such structural biases requires dedicated geometric spaces:

- **OBLIQUE MANIFOLD** ($\mathcal{O}^{d\times k}$): Ensures stability via $\|\mathbf{W}_{:,i}\|_2 = 1,\ \forall i \in \{1,\dots,k\}$
- **LORENTZ SPACE** ($\mathbb{H}^n_{\mathscr{L}}$): Models hierarchies via $\mathscr{G}_{\mathbb{H}_{\mathscr{L}}}(\mathbf{u},\mathbf{v}) = \operatorname{arcosh}\left(-\langle \mathbf{u},\mathbf{v}\rangle_{\mathscr{L}}\right)$

where column normalization in $\mathcal{O}$ transforms optimization landscapes [1, 40], and $\mathscr{G}_{\mathbb{H}_{\mathscr{L}}}$ denotes the Lorentzian distance metric [20].

Inspired by these geometric advantages and building upon BiCA, we propose **CURVATURE-AWARE CAPTIONING (CAC)**, a novel framework that resolves attentional conflicts through *stage-specific manifold projection*.

The Oblique Manifold projection enables near-linear gradient descent paths, accelerating convergence while pre-

serving geometric integrity, significantly stabilizing bounding box regression and enhancing localization precision. Complementarily, Lorentz Space captures exponentially growing semantic hierarchies through hyperbolic distances, dynamically modeling compositional relationships between objects at varying scales. These geometric transformations collectively resolve spatial-semantic conflicts while maintaining coherence between object localization and descriptive captioning. Our main contributions are summarized as follows:

- **Encoder-Oblique Projection**: In the encoding and object localization stage, we utilize geodesic attention in Oblique manifold, promoting optimization stability.
- **Decoder-Lorentz Contextualization**: During decoding (caption generation), we extend bidirectional contextual attention in Lorentz Space, where hyperbolic geodesics capture hierarchical relations.
- **State-of-the-art performance** on ScanRefer and Nr3D benchmarks, achieving ↑**2.7% CIDEr@0.5** on ScanRefer and ↑**4.6% CIDEr@0.5** on Nr3D the highest reported score to date. Our work pioneers stage-specific manifold learning for 3D vision-language tasks, achieving geometrically consistent scene understanding.

2. Related Work

2.1. 3D Dense Captioning

3D dense captioning, an evolving task demanding precise localization and descriptive generation of objects in complex 3D scenes [16], has progressed through diverse methodological innovations, starting with graph-based techniques that treat detector proposals as nodes in k-nearest neighbor graphs for feature extraction [22, 33]. Subsequent transformer-based methods replaced explicit graph operations with attention mechanisms to better model spatial relationships [6, 23, 46], while recent efforts explore multi-task synergies between dense captioning, 3D visual grounding [8], and 3D question answering [4, 50]. Standard architectures comprise scene encoders employing detectors such as [21, 34, 39] for object-level feature extraction, relational modules modeling inter-object connections through graph-based, transformer-based, or knowledge distillation approaches [11, 12], and feature decoders generating bounding boxes and captions using GRU-based or transformer-based implementations [51].

2.2. Embedding on Oblique Manifold

The Oblique Manifold [40] is an embedded submanifold defined by column-wise unit Euclidean norm constraints. This space finds critical application in independent component analysis [1], where its intrinsic normalization property (analogous to L_2-normalization) facilitates uniform scaling in gradient-based optimization, enhancing convergence efficiency and model accuracy [19]. Geometrically, it transforms elongated contour profiles into near-spherical geometries with isotropic curvature, mitigating directional bias to accelerate convergence while reducing susceptibility to suboptimal solutions. Formal definitions appear in the Appendix.

2.3. Hyperbolic Geometry for Point Clouds

Recent advances leverage hyperbolic embeddings to exploit the hierarchical nature of 3D point clouds, capitalizing on hyperbolic geometry's capacity for efficient representation of high-dimensional data in low-dimensional spaces [7]. Beyond foundational hierarchy modeling with graph-based approaches using InfoNCE [37] and entailment losses [18], domain-specific adaptations have gained prominence: Hyperbolic ProtoNet facilitates few-shot visual classification [24], while HypRe [35] pioneers explicit regularization for part-whole relationships. Extensions enhance 3D processing through attentional mechanisms (PHGT [30]), multispace fusion (HypLiLoc [47]), and Riemannian-geometry modules improving point cloud matching (HECPG [48]). Complementing these, Hypformer [49] demonstrates how Lorentzian temporal dimensions intrinsically induce hierarchical structures to capture inherent relationships within point clouds.

3. Preliminary

This section introduces core mathematical foundations for our architecture: the Oblique Manifold and Lorentz Space. These frameworks overcome Euclidean representation limitations in 3D geometric understanding.

3.1. Oblique Manifold

The Oblique Manifold ($\mathcal{OM}$) constitutes a Riemannian submanifold of Euclidean space defined by column-wise unit Euclidean norm constraints. This geometric structure underpins independent component analysis [1], where intrinsic orthonormality facilitates efficient feature separation through constrained optimization.Formally, $\mathcal{OM}(n, g)$ is defined as:

$$\mathcal{OM}(n, g) = \left\{ \boldsymbol{P} \in \mathbb{R}^{n \times g} : \operatorname{diag}\left(\boldsymbol{P}^T \boldsymbol{P}\right) = \boldsymbol{I}_g \right\}, \quad (1)$$

where $\operatorname{diag}(\cdot)$ denotes the diagonal matrix, n represents the feature dimension of each patch vector.

Projecting point cloud features from Euclidean space to $\mathcal{OM}$ preserves relative geometric relationships [28], satisfying attention mechanisms' fundamental requirement to capture relational dependencies.

$$\boldsymbol{P} := \operatorname{Proj}(\boldsymbol{P}) = \operatorname{Cat}\left(\left\{ \frac{p_i}{\|p_i\|} \right\}_{i=1}^{g} \right), \quad (2)$$

where $\boldsymbol{P} = \{p_1, p_2, \cdots, p_g\}$ is the point cloud patch matrix comprising g patches, each p_i representing a feature vector

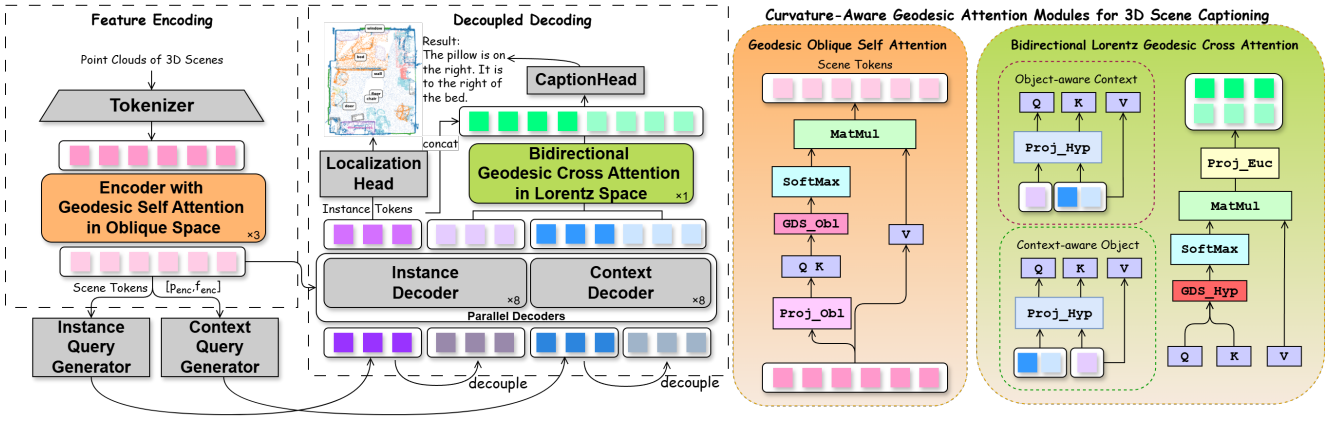


Figure 1. Left: Overview of the proposed Curvature-Aware Captioning (CAC) framework. The processing pipeline is mapped onto non-Euclidean manifolds, with attention computed via geodesic distance metrics. Right: The Geodesic Oblique Self Attention and Bidirectional Lorentz Geodesic Attention.

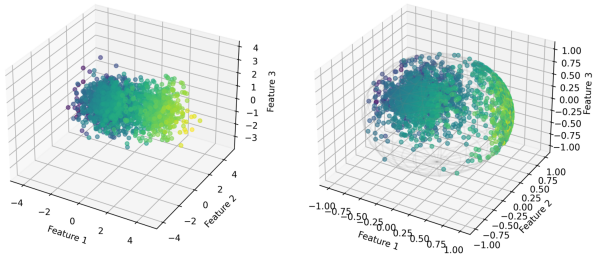


Figure 2. 3D point cloud distribution contrast: Euclidean space (left) versus Oblique manifold projection (right) with column-wise unit Euclidean norm constraints.

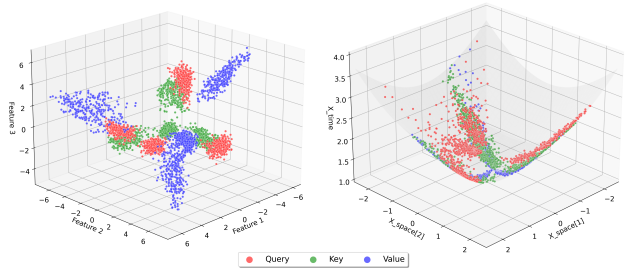


Figure 3. Comparison of 3D point distributions: Euclidean Cartesian space (left) versus Lorentz Space embeddings (right). Hyperbolic projections via exponential mapping reveal hierarchical clustering patterns, with semitransparent boundaries indicating the manifold structure.

of the i-th patch. Here $\mathrm{Cat}(\cdot)$ denotes the concatenation function, $\|\cdot\|$ denotes the squared Frobenius norm in the ambient space, as shown in Figure 2.

After projection, the geodesic distance of the input point pair $\{\boldsymbol{Q}, \boldsymbol{K}\}$ on $\mathcal{OM}$ can be calculated as:

$$\mathrm{dist}(\boldsymbol{Q}, \boldsymbol{K}) = \sqrt{\sum_{i=1}^{n} \arccos^2 \left(\mathrm{diag}\left(\boldsymbol{Q}^T \boldsymbol{K}\right)\right)_i}. \quad (3)$$

3.2. Lorentz Space

Hyperbolic space offers a fundamental non-Euclidean framework for hierarchical representation through constant negative curvature. We employ the Lorentz hyperboloid model [41] to avoid numerical instabilities from exponential volume growth during training, replacing alternatives like Poincaré ball, Klein, and Poincaré half-space models [43] projecting onto space-like hyperplanes.

Hyperbolic space is formalized through the **Lorentz hyperboloid model** defined as:

$$\mathbb{L}^n = \left\{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{L}} = -1/c\right\} (c > 0), \quad (4)$$

where $c > 0$ determines curvature $-c$. Vectors decompose as $\mathbf{x} = [\mathbf{x}_{\text{space}}, \mathbf{x}_{\text{time}}]$ with $\mathbf{x}_{\text{time}} = \sqrt{1/c + \|\mathbf{x}_{\text{space}}\|^2}$. The Lorentzian inner product follows:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = \mathbf{x}_{\text{space}} \cdot \mathbf{y}_{\text{space}} - \mathbf{x}_{\text{time}} \cdot \mathbf{y}_{\text{time}}. \quad (5)$$

To mitigate numerical instability during training [26], we exclusively map through the origin $\mathbf{O} = [\mathbf{0}, 1/\sqrt{c}]$. The exponential map $\exp_{\mathbf{O}} : \mathcal{T}_{\mathbf{O}}\mathbb{L}^n \to \mathbb{L}^n$ projects Euclidean tangent vectors $\mathbf{u} = [\mathbf{u}_{\text{enc}}, 0]$ to the manifold:

$$\mathbf{x}_{\text{space}} = \frac{\sinh(\sqrt{c}\|\mathbf{u}_{\text{enc}}\|)}{\sqrt{c}\|\mathbf{u}_{\text{enc}}\|}\mathbf{u}_{\text{enc}}, \quad (6)$$

where $\mathbf{u}_{\text{enc}} \in \mathbb{R}^n$ represents the Euclidean embedding component of the tangent vector at the origin $\mathbf{O}$. Conversely, the logarithmic map $\log_{\mathbf{O}} : \mathbb{L}^n \to \mathcal{T}_{\mathbf{O}}\mathbb{L}^n$ projects manifold points to Euclidean space:

$$\mathbf{u}_{\text{enc}} = \frac{\operatorname{arcosh}(-\langle \mathbf{x}, \mathbf{O} \rangle_{\mathbb{L}})}{\sqrt{\langle \mathbf{x}, \mathbf{O} \rangle_{\mathbb{L}}^2 - 1}}\mathbf{x}_{\text{space}}, \quad (7)$$

where $\mathbf{u}_{\text{enc}} = \alpha f(T)$ and α (initialized as $1/\sqrt{n}$) is learned in logarithmic space to prevent embedding collapse and numerical overflow [17]. Analogous scaling applies to image embeddings $\mathbf{v}_{\text{enc}} = \alpha_{\text{img}} g(I)$. *Geodesics* in $\mathbb{L}^n$ correspond to hyperplane-origin intersections with distance metric:

$$\mathscr{G}_{\mathbb{H}_{\mathscr{L}}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \cosh^{-1}\left(-c\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}\right). \tag{8}$$

This formulation intrinsically captures hierarchical relationships while maintaining geometric consistency across spatial scales, enabling robust optimization in negatively curved spaces.

4. Methodology

This section introduces **CAC**, a geometrically unified framework addressing spatial and hierarchical representation limitations in 3D dense captioning. We demonstrate curvature complementarity between Oblique and Lorentz spaces resolves Euclidean-hyperbolic conflicts (Appendix derivations). The input 3D point clouds PC $= [p_{\text{in}}; f_{\text{in}}] \in \mathbb{R}^{40,000\times(3+F)}$ comprise $40,000$ points with positional coordinates (xyz) and F-dimensional features.

4.1. Overall Architecture

Current state-of-the-art 3D dense captioning methods inadequately leverage spatial and hierarchical structures, neglecting inherent geometric priors in point clouds. We resolve this limitation through **CAC**, extending Vote2Cap-DETR++ to migrate data into geometric-prior-optimized space, with ablation studies confirming enhanced feature extraction.

Recognizing hyperbolic space's innate hierarchical representation capacity, we introduce hyperbolic multi-head attention that transitions both Object-aware Contexts and Context-aware Objects modules into hyperbolic space, as shown in Figure 1. This enables explicit hierarchical modeling while incorporating geodesic distance mechanisms that amplify inherent geometric priors. Our framework employs geometrically consistent geodesic attention operators, simultaneously leveraging 3D spatial priors and hierarchical structures.

4.2. Encoder on Oblique Manifold

In this work, in order to advance sparse 3D point cloud embedding, we replace the 3DETR encoder's attention module [34] by our *Geodesic Oblique Self Attention* mechanism, as shown in Figure 1. Specifically, following Vote2Cap-DETR++'s preprocess pipeline, PointNet++'s set-abstraction layer [39] tokenizes PC into coordinates $p_{\text{abs}} \in \mathbb{R}^{2,048\times 3}$ and feature representations $f_{\text{abs}} \in \mathbb{R}^{2,048\times 256}$, where $2,048$ denotes the sampled point count and 256 denotes the token embedding dimension. Subsequently processed by the geometrically enhanced 3DETR encoder, this framework outputs encoded scene tokens $p_{\text{enc}} \in \mathbb{R}^{2,048\times 3}$ and $f_{\text{enc}} \in \mathbb{R}^{2,048\times 256}$. These tokens are subsequently processed by the geometrically enhanced 3DETR encoder and downsampled to yield the encoded scene tokens $\mathbf{p}_{\text{enc}} \in \mathbb{R}^{1,024\times 3}$ and $\mathbf{f}_{\text{enc}} \in \mathbb{R}^{1,024\times 256}$.

The *Geodesic Oblique Self Attention* mechanism imposes Oblique Manifold constraints to ensure isotropic attention distribution [31]. The mechanism projects embedded features ($\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{2048\times 256}$) onto the Oblique manifold using equation (2). This projection enables establishing spatial-geometric relationships between queries and keys through pairwise geodesic distances $D \in \mathbb{R}^{2048\times 2048}$ computed via equation (3), which intrinsically respect the manifold curvature. To ensure numerical stability during the computation of these large-scale pairwise distances, we employ a clipping operation with $\epsilon = 10^{-4}$ to constrain the inputs within the valid domain $[-1+\epsilon, 1-\epsilon]$, preventing potential numerical instabilities. Differentiable aggregation modulates softmax-weighted feature fusion using these geodesic similarity scores, boosting object recognition accuracy by capturing intrinsic spatial structures and contextual relationships in 3D point clouds, please refer to Algorithm 1.

Algorithm 1 Geodesic Oblique Self Attention

Require: $X \in \mathbb{R}^{n\times d}$, $Pos \in \mathbb{R}^{n\times 3}$
Ensure: Geodesic attention output $\hat{v} \in \mathbb{R}^{n\times d}$
$q, k \leftarrow \text{Proj}_{\mathcal{OBL}}(\text{Emb}(X, Pos))$ {Via equation(2)}
$v \leftarrow X$
for $i = 1$ to n **do**
  for $j = 1$ to n **do**
    $d_{ij} \leftarrow GDS_{\mathcal{OBL}}(q_i, k_j)$ {Via equation(3)}
  end for
end for
$\hat{v} \leftarrow MatMul\langle \text{softmax}(-D), v\rangle$
return $\hat{v}$ {Geodesic attention output}

4.3. Vote Query

Inspired by Vote2Cap-DETR++,our Vote Query Generator produces queriesthrough a spatial refinement process:

$$[\Delta p_{\text{vote}}; \Delta f_{\text{vote}}] = \text{FFN}(f_{\text{enc}}), \tag{9}$$

$$(p_{vq}, f_{vq}) = \text{SA}(p_{\text{enc}} + \Delta p_{\text{vote}}, f_{\text{enc}} + \Delta f_{\text{vote}}), \tag{10}$$

where $\Delta p_{\text{vote}} \in \mathbb{R}^{1,024\times 3}$ learns center-aligned spatial shifts by a **F**eed **F**orward **N**etwork(FFN), and SA denotes the set-abstraction layer (radius=0.3, 16-point sampling) that aggregates features at refined coordinates, sampling npoint $=$ 256 points.

Crucially extending [12], we employ a decoupled decoding structure that iteratively updates spatial locations across

layers:

$$f^i_{\text{query}} = \text{Layer}_{i-1}\left(f^{i-1}_{\text{query}} + \text{PE}\left(p^i_{vq}\right)\right), \\ p^i_{vq} = p^{i-1}_{vq} + \text{FFN}\left(f^i_{\text{query}}\right), \tag{11}$$

where Layer_{i-1} denotes the i-th decoder layer and $\text{PE}(\cdot)$ represents 3D Fourier positional encoding, mitigating cumulative errors inherent in detect-then-describe pipelines [53, 56].

Both Instance and Context Query Generators employ iterative spatial refinement, with Context processing excluding Δp_{vote} fusion during aggregation.

4.4. Decoder with Bidirectional Lorentz Attention

Our decoder maintains Vote2Cap-DETR++'s decoupled architecture [12], processing both encoder outputs and vote queries from the generator. This dual-decoder framework generates instance and context features, outputting $p_{\text{dec}} \in \mathbb{R}^{npoint\times 2\times 3}$ and $f_{\text{dec}} \in \mathbb{R}^{npoint\times 2\times 256}$ where $npoint = 256$ denotes the number of sampled points during vote query processing and 2 denotes the decoupled outputs. For caption generation, we utilize the first dimension of decoupled instance features $f_{\text{cap}} \in \mathbb{R}^{npoint\times 1\times 256}$ as input to the caption task head to obtain object segmentation information.

As depicted in Figure 1, we enhance cross-attention by constraining Key and Value matrices within hyperbolic geometry [3], replacing dot products with geodesic distances to leverage 3D hierarchical structures. Following feature extraction, hyperbolic bidirectional attention modules implement distinct configurations: Object-aware Context (OAC) employs instance features as **Q** with context features as **K/V**, while Context-aware Object (CAO) utilizes context features as **Q** with instance features as **K/V**, with ablation studies confirming significant performance gains.

For bidirectional attention inputs, we utilize the second dimension of decoupled instance features $f_{\text{ins}} \in \mathbb{R}^{\text{npoint}\times 1\times 256}$ and both dimensions of context features $f_{\text{ctx}} \in \mathbb{R}^{\text{npoint}\times 2\times 256}$. Features undergo Lorentz projection into hyperbolic space (Equation 6, Figure 3), establishing geometric foundations for attention computation. Attention weights derive from pairwise geodesic distances (Equation 8) that intrinsically capture hyperbolic curvature relationships between queries and keys, replacing conventional dot products. To maintain numerical stability during this computation, we implement a clipping operation with $\epsilon = 10^{-15}$ to constrain inputs within the valid domain $[1+\epsilon, \infty]$. This ensures the argument of the $\text{arcosh}(\cdot)$ function remains strictly greater than or equal to unity, thereby preventing numerical instabilities near boundary conditions. These stabilized distances subsequently undergo exponential mapping and temperature-scaled softmax normalization to produce curvature-aware weights. The architecture aggregates contextual information through geodesic-weighted combinations before Euclidean reprojection (Equation 7), completing the attention cycle as detailed in Algorithm 2. Final processing applies mean pooling to Context-aware Object outputs, yielding feature representations f_{ins}, f_{oac}, $f_{\text{cao}} \in \mathbb{R}^{n_{\text{point}}\times 1\times 256}$ that are concatenated as input to the captioning task head.

Algorithm 2 Bidirectional Lorentz Geodesic Attention

Require: $q, k, v \in \mathbb{R}^{n\times b\times d}$
Ensure: Geodesic attention output $\hat{v} \in \mathbb{R}^{n\times b\times d}$
$q_{\mathbb{H}}, k_{\mathbb{H}} \leftarrow \text{Proj}_{\mathcal{HYP}}(q, k)$ {Via equation (6)}
for $i = 1$ to n **do**
  for $j = 1$ to n **do**
    $d_{ij} \leftarrow GDS_{\mathcal{HYP}}(q_{\mathbb{H}i}, k_{\mathbb{H}j})$ {Via equation (8)}
  end for
end for
$A \leftarrow \text{softmax}(\exp(-D/\tau))$ {τ: temperature}
$\hat{v} \leftarrow MatMul\langle A, v\rangle$
return $\hat{v}$

4.5. Training CAC

Our loss builds upon Vote2Cap-DETR++ [12], adopting its *decoupled localization-describing paradigm*. The objective integrates four weighted components: $\mathcal{L}_{\text{vq}}$ supervising point shifts to object centers, $\mathcal{L}_{\text{det}}$ refining proposal assignment via Hungarian algorithm, $\mathcal{L}_{\text{cap}}$ implementing dual training (MLE+SCST), and $\mathcal{L}_{\text{qr}}$ enabling iterative query localization across decoder layers.

Detection Loss

Employing 3DETR's Hungarian algorithm [34] for proposal-ground truth matching, we emphasize 3D GIoU loss to enhance localization:

$$\mathcal{L}_{\text{set}} = \alpha_1\mathcal{L}_{\text{giou}} + \alpha_2\mathcal{L}_{\text{cls}} + \alpha_3\mathcal{L}_{\text{center-reg}} + \alpha_4\mathcal{L}_{\text{size-reg}}, \tag{12}$$

with heuristic weights $\alpha_1 = 10$, $\alpha_2 = 1$, $\alpha_3 = 5$, $\alpha_4 = 1$. This loss aggregates across all $n_{\text{dec-layer}}$ decoder layer.

Caption Loss

Our caption head integrates Maximum Likelihood Estimation (MLE) and *Self-Critical Sequence Training* (SCST) [42]. The MLE objective minimizes:

$$\mathcal{L}_{c_i} = -\sum_{t=1}^{T} \log \hat{P}\left(c_i^{t+1} \mid \mathcal{V}; c_i^{[1:t]}\right), \tag{13}$$

while SCST fine-tuning employs beam search (size k) and greedy decoding:

$$\mathcal{L}_{c_i} = -\sum_{i=1}^{k} \left(R(\hat{c}_i) - R(\hat{g})\right) \cdot \frac{1}{|\hat{c}_i|} \log \hat{P}\left(\hat{c}_i \mid \mathcal{V}\right). \tag{14}$$

Method	$\mathcal{L}_{des}$	IoU = 0.25				IoU = 0.50			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap [14]	MLE	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57
MORE [22]		58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33
SpaCap3d [45]		-	-	-	-	42.76	25.38	22.84	45.66
Contextual [55]		-	-	-	-	42.77	23.60	22.05	45.13
3DJCG [6]		60.86	39.67	27.45	59.02	47.68	31.53	24.28	51.80
3D-VLP* [23]		64.09	39.84	27.65	58.78	50.02	31.87	24.53	51.17
Vote2Cap-DETR [11]		71.45	39.34	28.25	59.33	61.81	34.46	26.22	54.40
Vote2Cap-DETR++ [12]		76.36	41.37	28.70	60.00	67.58	37.05	26.89	55.64
Vote2Cap-DETR++^{R}		74.59	41.35	28.60	59.79	66.06	36.93	26.76	55.39
Ours:									
CAC($\mathcal{O}$)		75.56	40.34	28.55	59.53	68.07	36.53	26.72	55.08
CAC($\mathcal{O}$&$\mathbb{H}$)		**78.67**	**42.11**	**28.79**	**60.22**	**69.92**	**37.67**	**26.89**	**55.62**
χ-Trans2Cap [51]	SCST	58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97
Contextual [55]		-	-	-	-	50.29	25.64	22.57	44.71
Vote2Cap-DETR [11]		84.15	42.51	28.47	59.26	73.77	38.21	26.64	54.71
Vote2Cap-DETR++ [12]		88.28	44.07	28.75	59.89	78.16	39.72	26.94	55.52
Vote2Cap-DETR++^{R}		87.52	44.00	28.71	59.59	77.65	39.59	26.88	55.21
Ours:									
CAC($\mathcal{O}$)		89.51	43.47	28.76	59.51	79.09	38.96	26.85	54.95
CAC($\mathcal{O}$&$\mathbb{H}$)		**90.16**	**44.07**	**28.80**	**60.29**	**80.35**	**39.95**	**26.94**	**55.66**

Table 1. Quantitative comparisons on ScanRefer [8] validation set following Scan2Cap [14]'s protocol. Superscript ^{R} indicates reproduced results. Our framework establishes new state-of-the-art performance across caption supervision paradigms.

Here $R(\cdot)$ represents the CIDEr reward metric [44], with length normalization $|\hat{c}_i|$ ensuring balanced caption weighting, where $\hat{g}$ denotes the greedily decoded baseline caption and $\hat{c}_i$ represents the i-th beam search candidate.

Set-to-Set Training and Integrated Objectives

Our *set-to-set strategy* randomly samples one caption per instance, assigns annotations via Hungarian matching, and averages batch caption losses $\mathcal{L}_{c_i}$ into $\mathcal{L}_{\text{cap}}$. The unified objectives are:

$$\mathcal{L}_{\text{V1}} = \beta_1 \mathcal{L}_{\text{vq}} + \beta_2 \sum_{i=1}^{n_{\text{dec-layer}}} \mathcal{L}_{\text{set}} + \beta_3 \mathcal{L}_{\text{cap}}, \tag{15}$$

$$\mathcal{L}_{\text{V2}} = \mathcal{L}_{\text{V1}} + \beta_4 \sum_{i \in \delta} \mathcal{L}_{\text{qr}}, \tag{16}$$

with $\beta_1 = \beta_4 = 10$, $\beta_2 = 1$, $\beta_3 = 5$ applied to decoder layers δ performing spatial refinement.

5. Experiments

Datasets. This study employs two benchmark datasets for 3D dense captioning: ScanRefer [9] (36,665 descriptions across 7,875 objects in 562 scenes) and Nr3D [2] (32,919 descriptions for 4,664 objects in 511 scenes), both providing human-annotated scene and object descriptions. Training data originates from ScanNet's 1,201 distinct scenes. Evaluation utilizes 9,508 ScanRefer descriptions (2,068 objects in 141 scenes) and 8,584 Nr3D descriptions (1,214 objects in 130 scenes), all drawn from ScanNet's 312-scene validation set.

Evaluation Metrics. To ensure evaluation consistency with prior works [6, 12], we apply Non-Maximum Suppression (NMS) [36] to predictions despite our method's inherent robustness. Post-NMS, we establish ground-truth correspondences by selecting highest-IoU proposals for each annotated instance (b_i, C_i). Performance is evaluated using the **m@kIoU metric** [14]:

$$m@kIoU = \frac{1}{N} \sum_{i=1}^{N} m(\hat{c}_i, C_i) \cdot \mathbb{I}\{IoU(b_i, \hat{b}_i) \geq k\}, \tag{17}$$

combining **localization accuracy** (IoU $\geq$ k) and **caption quality** (metric m). Here N enumerates instances, m denotes caption metrics (CIDEr, METEOR, BLEU-4, ROUGE-L), and $\mathbb{I}\{\cdot\}$ is the indicator function.

Implementation Details. Following [12], our training employs a three-stage pipeline. The initial phase pre-trains the model for 1,080 epochs on ScanNet [16] excluding caption modules, using AdamW optimization [32] with cosine annealing from 5×10^{-4} to 10^{-6} learning rate, 0.1 weight decay, and gradient clipping at batch size 8. Subsequently, 720 epochs of joint training on ScanRefer and Nr3D apply cross-entropy loss with a batch size of 8, with fixed 10^{-6} detector learning rate while decaying the caption head from 10^{-4} to 10^{-6}. The final stage refines the caption head via SCST [42] for 180 epochs at fixed 10^{-6} learning rate and batch size 2, freezing detector parameters. Implemented in PyTorch [38], our framework exhibits distinct performance characteristics across its three training phases. Table 2 summarizes the basic configurations for different training phases. During the pre-training stage, the average iteration

time is 0.878s with a statistical sample size of 16,635 and memory consumption of approximately 18GB, achieving a per-sample computational complexity of 10.61 GFLOPs and a per-iteration complexity of 84.88 GFLOPs. The second phase attains an average iteration time of 1.173s over 5,163 samples, maintaining approximately 18GB memory usage, with a per-sample FLOPs of 11.16 GFLOPs and a per-iteration FLOPs of 89.30 GFLOPs. Finally, the self-critical sequence training (SCST) fine-tuning stage demonstrates an average iteration time of 1.629s across 5,265 samples, with reduced memory footprint of approximately 12GB, and exhibits a per-sample FLOPs of 13.92 GFLOPs and a per-iteration FLOPs of 27.84 GFLOPs, all evaluated on a single RTX4090 GPU.A more detailed comparison will be provided in the Appendix.

Configuration Item	**Pretrain**	**MLE**	**SCST**
Dataset	ScanNet	ScanRefer	ScanRefer
Batch Size	8	8	2
Learning Rate	5e-4	1e-4	1e-6
Max Epochs	1080	720	180
GPU Memory Usage	~18GB	~18GB	~12GB
Captioner	None	Train	Train
Detector	Train	Train	Frozen
Beam Search	Not Used	Used	Used

Table 2. Configuration Comparison of Different Training Stages

5.1. Comparison with Existing Methods

We benchmark CAC($\mathcal{O}$) and CAC($\mathcal{O}\&\mathbb{H}$) against leading methods on ScanRefer [8] and Nr3D [2]. Evaluation applies standard IoU thresholds: 0.25/0.50 for ScanRefer (Table 1) and 0.50 exclusively for Nr3D (Table 3), with C@0.5 as the primary benchmark. Baseline metrics reflect originally reported values, with "-" indicating unreported results.

Our multi-stage attention mechanism outperforms existing methods on both ScanRefer (requiring comprehensive attribute/spatial descriptions) and Nr3D (testing free-form human captions), achieving ↑2.7% CIDEr@0.5 on ScanRefer and ↑4.6% CIDEr@0.5 on Nr3D, the highest reported scores to date. This demonstrates significant improvements in diverse, accurate caption generation (Tables 1, 3) while validating real-world applicability.

5.2. Ablation Study

During our experimental investigation, we implemented BiCA's innovative approach of independently generating contextual and instance queries. However, the reproduced results proved unsatisfactory as detailed in Table 5, with comprehensive reproduction methodology documented in the Appendix. Consequently, we excluded BiCA from comparative analysis in Table 1.

Ablation studies on input **QKV** configurations within our bidirectional decoder framework demonstrate our approach's superiority over BiCA's methodology [25], evidenced by CAC($\mathcal{O}$)^{BiCA} results in Table 5. Comprehensive ablation experiments on alternative input formulations during BiCA reproduction validate our architectural choices, with detailed analysis in the Appendix. Comprehensive seed robustness testing on Nr3D across three random seeds (0, 333, 777) under MLE and SCST optimization confirms consistent performance (Table 4). CAC($\mathcal{O}\&\mathbb{H}$) with SCST shows exceptional stability, with seed 777 achieving $51.18\% \pm 0.09\%$ (C@0.5), $29.58\% \pm 0.05\%$ (B-4@0.5), $26.28\% \pm 0.07\%$ (M@0.5), and $55.92\% \pm 0.06\%$ (R@0.5). The narrow confidence intervals demonstrate minimal variance and high reproducibility.

Method	$\mathcal{L}_{des}$	IoU = 0.50			
		C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap		27.47	17.24	21.80	49.06
SpaCap3d		33.71	19.92	22.61	50.50
D3Net		33.85	20.70	23.13	53.38
Contextual		35.26	20.42	22.77	50.78
3DJCG		38.06	22.82	23.77	52.99
V2C-DETR	MLE	43.84	26.68	25.41	54.43
V2C-DETR++		47.08	27.70	25.44	55.22
V2C-DETR++^{R}		46.10	27.19	25.38	55.22
Ours:					
CAC($\mathcal{O}$)		**50.99**	**28.89**	**26.41**	**56.18**
CAC($\mathcal{O}\&\mathbb{H}$)		49.90	28.70	26.00	55.71
χ-Tran2Cap		33.62	19.29	22.27	50.00
Contextual		37.37	20.96	22.89	51.11
D3Net		38.42	22.22	24.74	54.37
V2C-DETR		45.53	26.88	25.43	54.76
V2C-DETR++	SCST	47.62	28.41	25.63	54.77
V2C-DETR++^{R}		48.13	28.03	25.66	54.74
Ours:					
CAC($\mathcal{O}$)		50.99	29.32	26.13	55.81
CAC($\mathcal{O}\&\mathbb{H}$)		**52.78**	**29.78**	**26.13**	**55.94**

Table 3. Benchmark evaluation on the Nr3D validation set [2] demonstrates our method's superior performance against prior arts under both MLE training and SCST optimization frameworks.

Method	$\mathcal{L}_{des}$	Seed	IoU = 0.50			
			C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
CAC($\mathcal{O}$)	MLE	**0**	50.99	28.89	26.41	56.18
		333	50.81	28.65	26.09	56.00
		777	50.64	28.50	26.29	56.05
	SCST	**0**	50.99	29.32	26.13	55.81
		333	51.46	28.96	25.96	55.60
		777	51.12	28.91	26.01	55.50
CAC($\mathcal{O}\&\mathbb{H}$)	MLE	**0**	49.90	28.70	26.00	55.71
		333	50.07	28.97	26.03	56.16
		777	51.08	29.21	26.22	55.91
	SCST	**0**	52.78	29.78	26.13	55.94
		333	49.34	28.98	26.16	56.12
		777	53.18	29.58	26.28	55.92

Table 4. Based on the Nr3D dataset [2], stability test results across training seeds of 0, 333, and 777 are reported under both the MLE and SCST frameworks.

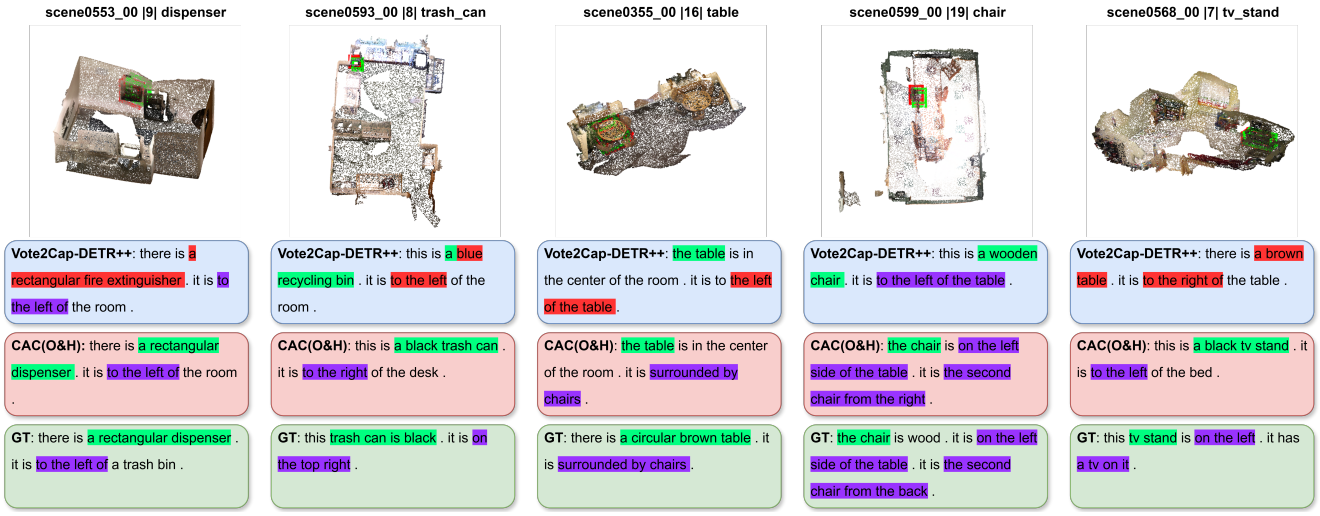


Figure 4. Qualitative results on the ScanRefer benchmark. Green highlights indicate accurate object descriptions, purple highlights denote contextual environmental descriptions, and red highlights identify erroneous descriptions with semantic deviations.

Method	$\mathcal{L}_{des}$	IoU = 0.50			
		C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
BiCA^{R}		65.22	37.59	26.87	55.76
V2C-DETR++^{R}		66.06	36.93	26.76	55.39
CAC$(\mathcal{O})^{BiCA}$	MLE	67.07	36.27	26.49	54.98
CAC$(\mathcal{O})$		68.07	36.53	26.72	55.08
CAC$(\mathcal{O}\&\mathbb{H})$		**69.92**	**37.67**	**26.89**	**55.62**
BiCA^{R}		76.42	38.75	26.70	54.66
V2C-DETR++^{R}		87.52	44.00	28.71	59.59
CAC$(\mathcal{O})^{BiCA}$	SCST	77.86	39.10	27.01	54.83
CAC$(\mathcal{O})$		79.09	38.96	26.85	54.95
CAC$(\mathcal{O}\&\mathbb{H})$		**80.35**	**39.95**	**26.94**	**55.66**

Table 5. Comparative evaluation against BiCA [25] under identical experimental conditions: IoU threshold = 0.50.

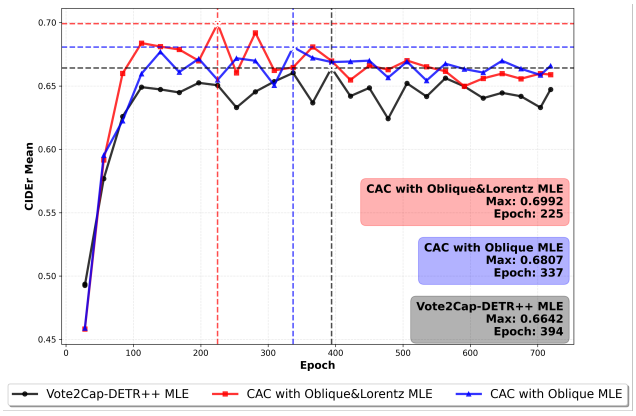


Figure 5. Comparison of CIDEr@0.5 Mean evaluation results of three experiments.

6. Qualitative Analysis

Qualitative comparisons reveal CAC$(\mathcal{O}\&\mathbb{H})$'s significant descriptive improvements over Vote2Cap-DETR++ in accurately capturing object attributes and spatial relationships, as shown in Figure 4. Where Vote2Cap++ misidentifies a black trash can as a "blue recycling bin" and mislabels a dispenser as a "fire extinguisher", CAC correctly identifies these as a "black trash can" and "rectangular dispenser" respectively. More crucially, CAC demonstrates superior spatial awareness, describing furniture arrangements with precise references like "it is the second chair from the right" and contextualizing relationships as "the table is surrounded by chairs", while Vote2Cap++ generates implausible statements such as "the table is to the left of the table" and vague descriptions like "it is to the left of the room". As evidenced in Figure 5, our approach achieves earlier convergence to optimal solutions compared to Vote2Cap++, with dual-space learning in both Oblique and Lorentz Space yielding significantly faster convergence than Oblique-only optimization.

7. Conclusion

We introduce the **CAC** framework, which resolves the localization-contextualization conflict in 3D scene understanding by unifying Oblique-space geodesic attention—enforcing column-wise unit-norm constraints for isotropic stability—and Lorentzian hyperbolic attention ($\kappa_{\mathbb{H}} = -c$) for hierarchical semantics. This geometric synergy prevents dimensional collapse and ensures robust feature learning. State-of-the-art results on ScanRefer and Nr3D benchmarks demonstrate significant gains in localization accuracy and descriptive richness, laying a solid foundation for embodied AI applications.

8. Acknowledgments

This research is supported by the General Program of Shanghai Natural Science Foundation (No.24ZR1419800, No.23ZR1419300), the National Natural Science Foundation of China (No.42130112), the Ministry of Industry and Information Technology of China, Science and Technology Commission of Shanghai Municipality (No.22DZ2229004), and Shanghai Frontiers Science Center of Molecule Intelligent Syntheses.

References

[1] P-A Absil and Kyle A Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In *2006 IEEE international conference on acoustics speech and signal processing proceedings*, pages V–V. IEEE, 2006. 1, 2

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020. 6, 7

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 5

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 2

[5] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 1

[6] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16464–16473, 2022. 1, 2, 6

[7] Edoardo Cetin, Benjamin Chamberlain, Michael Bronstein, and Jonathan J Hunt. Hyperbolic deep reinforcement learning. *arXiv preprint arXiv:2210.01542*, 2022. 2

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2, 6, 7

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. pages 202–221, 2020. 6

[10] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision*, pages 487–505. Springer, 2022. 1

[11] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11124–11133, 2023. 2, 6

[12] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7331–7347, 2024. 1, 2, 4, 5, 6

[13] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18000–18010, 2023. 1

[14] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 1, 6

[15] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18109–18119, 2023. 1

[16] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6

[17] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 4

[18] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pages 1646–1655. PMLR, 2018. 2

[19] Ziyang Guo, Anyou Min, Bing Yang, Junhong Chen, Hong Li, and Junbin Gao. A sparse oblique-manifold nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 2

[20] Jakob Hedicke. Lorentzian distance functions in contact geometry. *Journal of Topology and Analysis*, 16(02):205–225, 2024. 1

[21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 2

[22] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer, 2022. 1, 2, 6

[23] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 2, 6

[24] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020. 2

[25] Minjung Kim, Hyung Suk Lim, Soonyoung Lee, Bumsoo Kim, and Gunhee Kim. Bi-directional contextual attention for 3d dense captioning. In *European Conference on Computer Vision*, pages 385–401. Springer, 2024. 1, 7, 8

[26] Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. Hype: Hyperbolic entailment filtering for underspecified images and texts. *arXiv preprint arXiv:2404.17507*, 2024. 3

[27] Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025. 1

[28] Zhengyu Li, Xuan Tang, Zihao Xu, Xihao Wang, Hui Yu, Mingsong Chen, et al. Geodesic self-attention for 3d point clouds. *Advances in Neural Information Processing Systems*, 35:6190–6203, 2022. 2

[29] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Learning of 3d graph convolution networks for point cloud analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4212–4224, 2021. 1

[30] Lamei Liu and Zhiqiang Liu. Application of hyperbolic space attention mechanisms in 3d point cloud classification. In *2024 6th International Conference on Communications, Information System and Computer Engineering*, pages 658–662. IEEE, 2024. 2

[31] Chuanjiang Long, Zhen Dong, Keqiang Wang, Feng Yu, Chaobin He, and Zhong-Ren Chen. Molecular weight distribution shape approach for simultaneously enhancing the stiffness, ductility and strength of isotropic semicrystalline polymers based on linear unimodal and bimodal polyethylenes. *Polymer*, 275:125936, 2023. 4

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[33] Aihua Mao, Zhi Yang, Wanxin Chen, Ran Yi, and Yong-jin Liu. Complete 3d relationships extraction modality alignment network for 3d dense captioning. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4867–4880, 2023. 2

[34] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2906–2917, 2021. 2, 4, 5

[35] Antonio Montanaro, Diego Valsesia, and Enrico Magli. Rethinking the compositionality of point clouds through regularization in the hyperbolic space. *Advances in Neural Information Processing Systems*, 35:33741–33753, 2022. 2

[36] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition*, pages 850–855. IEEE, 2006. 6

[37] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2

[38] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 4

[40] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8412–8422, 2021. 1, 2

[41] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 3

[42] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 5, 6

[43] Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. 3

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[45] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 6

[46] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 1, 2

[47] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5176–5185, 2023. 2

[48] Yifan Xie, Jihua Zhu, Shiqi Li, Naiwen Hu, and Pengcheng Shi. Hecpg: Hyperbolic embedding and confident patch-guided network for point cloud matching. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 2

[49] Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient transformer fully in hyperbolic space. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3770–3781, 2024. 2

[50] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1772–1786, 2022. 2

[51] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 2, 6

[52] Chengkun Zhang and Junbin Gao. Hype-han: Hyperbolic hierarchical attention network for semantic embedding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3990–3996, 2021. 1

[53] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 5

[54] Wang Zhao, Yan-Pei Cao, Jiale Xu, Yuejiang Dong, and Ying Shan. Assembler: Scalable 3d part assembly via anchor point diffusion. *arXiv preprint arXiv:2506.17074*, 2025. 1

[55] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022. 6

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5