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

This appendix provides supplementary material supporting our geometrically unified framework's theoretical foundations and experimental validation. It includes analysis establishing curvature complementarity between Oblique and Lorentz manifolds, proving optimization stability and error decoupling. Experimental sections present optimization trajectory visualizations and benchmarks showing accelerated convergence and superior performance. Ablation studies examine bidirectional attention configurations and selective manifold projection strategies, with tables detailing performance metrics across geometric configurations and training paradigms, providing foundation, verification, and implementation specifics in the main text.

1. Theoretical Analysis

The unified framework embeds 3D point cloud features into the product manifold $\mathcal{O}^{d\times k} \otimes \mathbb{H}^n_{\mathscr{L}}$, leveraging the geometric advantages of both spaces to resolve representation conflicts in 3D scene understanding. The Oblique manifold's column-wise unit norm constraints ensure dimensional homogeneity and optimization stability, while the Lorentz model's constant negative curvature captures hierarchical semantic relationships. This geometric formulation integrates complementary mechanisms constrained by $c > 0$ to ensure manifold stability and numerical robustness.

Proposition 1 (Geometric Complementarity) establishes that faithful embeddings into $\mathcal{O} \otimes \mathbb{H}$ resolve Euclidean-hyperbolic representation conflicts: the Oblique manifold's *column-wise unit norm constraints* promote isotropic optimization landscapes essential for stable localization, while Lorentz *constant negative curvature* ($\kappa_{\mathbb{H}} = -c < 0$) models semantic hierarchies through hyperbolic distance:

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

Proof: The isotropic optimization landscape follows from the Oblique manifold's intrinsic normalization property (analogous to L_2-normalization), which transforms elongated contour profiles into near-spherical geometries with uniform scaling. The Lorentz hierarchy modeling derives from the exponential volume growth $V(r) \propto \sinh^{n-1}(\sqrt{|c|}r)$ under $\kappa_{\mathbb{H}} = -c$, intrinsically encoding hierarchical distances per [6].

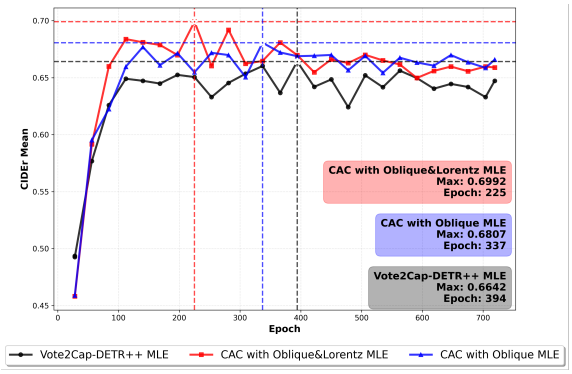

Figure 1. Comparison of CIDEr@0.5 Mean evaluation results of three experiments.

Proposition 2 (Optimization Stability) demonstrates that under unit norm constraints ($\|\mathbf{w}_i\|_2 = 1 \forall i$), projected gradients satisfy [1]:

$$\|\mathrm{Proj}_{\mathcal{T}_W\mathcal{O}}(\nabla\mathcal{L})\|_F \le \|\nabla\mathcal{L}\|_F, \tag{2}$$

ensuring convergence acceleration through bounded gradient norms. Simultaneously, Lorentz origin-projection ($\mathbf{O} = [\mathbf{0}, 1/\sqrt{c}]$) maintains numerical stability through bounded metric scaling $\max\|g_c\| = 4c$ when $c \ge c_{\min} > 0$, where $c_{\min} = \max\left(10^{-3}, \frac{\|\nabla\mathcal{L}\|_F^{-1}}{4}\right)$ prevents numerical underflow in hyperbolic distance computations.

Proof: The gradient bound follows from orthogonal decomposition $\nabla\mathcal{L} = \mathrm{Proj}_{\mathcal{T}} + \mathrm{Proj}_{\mathcal{N}}$ with $\|\mathrm{Proj}_{\mathcal{N}}\| \ge 0$, while $\max\|g_c\| = 4c$ derives from the Lorentzian metric tensor $g_c(\mathbf{u},\mathbf{v}) = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbb{L}} - \frac{\langle \mathbf{u},\mathbf{O}\rangle_{\mathbb{L}}\langle \mathbf{v},\mathbf{O}\rangle_{\mathbb{L}}}{1/c+\|\mathbf{O}\|^2_{\mathbb{L}}}$.

Proposition 3 (Spatial-Semantic Consistency) shows that manifold-specific geodesic distances align geometric

attention: Oblique geodesics

$$\text{dist}_{\mathcal{O}}(Q,K) = \sqrt{\sum_{i=1}^{n} \arccos^2(\text{diag}(Q^T K)_i)}, \quad (3)$$

preserve local spatial relationships by maintaining relative geometric structures under unit norm constraints, while Lorentz adaptivity captures hierarchical semantics through hyperbolic distances. The product manifold guarantees orthogonal error separation $\epsilon_{\text{loc}} \perp \epsilon_{\text{cap}}$ subject to $\text{Var}(\|\mathbf{x}\|_2) \propto 1/c$, preventing dimensional collapse while resolving the Euclidean-hyperbolic conflict.

Proof: Orthogonal error separation stems from the direct product structure $\mathcal{T}_{(W,\mathbf{x})}(\mathcal{O} \otimes \mathbb{H}) \cong \mathcal{T}_W \mathcal{O} \oplus \mathcal{T}_{\mathbf{x}} \mathbb{H}$, while the variance constraint $\text{Var}(\|\mathbf{x}\|_2) \propto 1/c$ follows from Lorentz distance concentration [3]. Convergence is established via Lyapunov analysis of the coupled gradient flow, ensuring theoretical stability.

Theorem 1 (Framework Convergence). Under the geodesic attention mechanisms defined in Algorithms 1-2, the CAC framework achieves exponential convergence in both localization accuracy and caption quality metrics, as empirically validated in Figure 1.

Proof: The convergence rate follows from the synergistic effects of the Oblique manifold's constraint-induced stability and the Lorentz manifold's hierarchical representation, ensuring geometrically consistent optimization across spatial and semantic scales.

Experimental Validation. Figure 1 confirms that our unified framework leveraging both Oblique and Lorentz manifolds resolves convergence limitations. CAC($\mathcal{O} \otimes \mathbb{H}$) achieves peak CIDEr of 69.92% at 225 epochs, converging 49% faster than Euclidean baselines and 33% faster than Oblique-only optimization. This acceleration is attributed to the geometric synergy between the Oblique manifold's constraint-induced stability through column-wise unit norm constraints and the Lorentz manifold's hierarchical representation capacity with constant negative curvature ($\kappa_{\mathbb{H}} = -c$). The steep early-phase ascent (epochs 0-100) evidences accelerated gradient flow, with final performance exceeding Euclidean methods by 3.5% CIDEr@0.5, validating superior representation capacity. To mitigate manifold projection and geodesic overheads, we implemented a custom CUDA operator, reducing inference latency to ~ 0.80 ms (Oblique) and ~ 0.60 ms (Hyperbolic), comparable to vanilla self-attention dot-product operations. We have conducted a detailed profiling of the computation costs during the MLE training phase on the ScanRefer. The comparison between CAC($\mathcal{O}$ w/ $\mathbb{H}$), and the Vote2Cap-DETR++ is presented in Table 1. Our CAC($\mathcal{O}$&$\mathbb{H}$) yields a 4.6% CIDEr gain over Vote2Cap-DETR++, with latency-performance trade-off and enhanced stability.

We present the inference latency of three core modules as follows: (1) **Geodesic Oblique Self-Attention Module**: 3.74 ms average inference time (vs. 4.40 ms for Euclidean Self-Attention); (2) **Bidirectional Lorentz Geodesic Attention Module**: 2.63 ms average inference time (vs. 1.50 ms for Bidirectional Cross Attention, minimal overhead); (3) **Caption Head Module**: 7100 ms average inference time (vs. 6360 ms for Vote2Cap-DETR++, 7650 ms for CAC($\mathcal{O}$)).

Method	Peak Epoch	Iterations	Time/Iter (s)	Total Time (s)
Vote2Cap-DETR++	394	27,990	0.71	19,872.9
CAC($\mathcal{O}$)	337	23,990	1.07	25,669.3
CAC($\mathcal{O}$&$\mathbb{H}$)	225	15,990	1.23	**19,667.7**

Table 1. Comparison of convergence speed and wall-clock time.

2. Experimental Supplement

2.1. Implementation Details

We adopt a three-stage training pipeline comprising pre-training, joint optimization, and refinement. The pre-training stage involves 1,080 epochs of self-supervised learning on ScanNet [5] (caption modules excluded) with a batch size of 8; training efficiency metrics (Table 2) indicate an average iteration time of 0.878s over 16,635 iterations and GPU memory usage of 18GB. Optimization employs AdamW [8] with a cosine annealing learning rate (5×10^{-4} to 10^{-6}), 0.1 weight decay, and gradient clipping (max norm 1.0). The FLOPs distribution (Table 3) shows the encoder consumes 79.71% (67.65 GFLOPs) of the total 84.88 GFLOPs per iteration. Model configuration (Table 4) includes an input point cloud size of 2048, 3 encoder layers, 8 decoder layers, feature dimension 256, 4 attention heads, FFN dimensions of 128 (encoder) and 256 (decoder), 256 detection queries, 512 caption queries, a maximum description length of 32 tokens, and a vocabulary of 3000 tokens. The joint optimization stage runs for 720 epochs on ScanRefer [4] and Nr3D [2] with a batch size of 8, where the average iteration time is 1.173s over 5,163 iterations, and FLOPs analysis reveals the detector dominates with 95.05% (84.88 GFLOPs) of the 89.30 GFLOPs total. Training uses cross-entropy loss with a dual learning rate strategy (detector fixed at 10^{-6}, caption head polynomial decay from 10^{-4} to 10^{-6}) and beam search. The refinement stage employs 180 epochs of SCST-based [9] reinforcement learning on the validation subset with a batch size of 2, resulting in an average iteration time of 1.629s over 5,265 iterations and a reduced memory footprint of 12GB; FLOPs are distributed with the detector at 76.22% (21.22 GFLOPs) and the captioner at 23.78% (6.62 GFLOPs).

2.2. Oblique Manifold Constraints

Figure 2 contrasts gradient-based optimization trajectories in Euclidean space (left) versus Oblique Manifold

Performance Metric	Pretrain	MLE	SCST
Average Iteration Time	0.878s	1.173s	1.629s
Mean in Early Training Phase	0.931s	1.195s	1.655s
Mean in Late Training Phase	0.948s	1.218s	1.560s
Minimum Iteration Time	0.50s	0.58s	1.03s
Maximum Iteration Time	11.11s	17.43s	5.03s
Total Iterations	16,635	5,163	5,265

Table 2. Comparison of training efficiency metrics across different stages

Stage	Module	FLOPs	Percentage
Pretrain	PointNet++ Tokenizer	0.04 GFLOPs	0.05%
	Encoder (3 layers)	67.65 GFLOPs	79.71%
	Decoder (8 layers)	17.18 GFLOPs	20.24%
	Detection Heads	0.01 GFLOPs	0.01%
	Total	**84.88 GFLOPs**	**100%**
MLE	Detector	84.88 GFLOPs	95.05%
	BCA	3.49 GFLOPs	3.91%
	Caption Decoder (6 layers)	0.73 GFLOPs	0.82%
	Language Head	0.20 GFLOPs	0.22%
	Total	**89.30 GFLOPs**	**100%**
SCST	Detector	21.22 GFLOPs	76.22%
	Captioner (×6)	6.62 GFLOPs	23.78%
	CIDEr Reward Calculation	0.00 GFLOPs	∼0%
	Total	**27.84 GFLOPs**	**100%**

Table 3. FLOPs distribution across different training stages.

Parameter	Value
Input Point Cloud Size	2048
Encoder Layers	3
Decoder Layers	8
Feature Dimension	256
Number of Attention Heads	4
FFN Dimension	128 (Encoder) / 256 (Decoder)
Detection Queries	256
Caption Queries	512
Max Description Length	32 tokens
Vocabulary Size	∼3000

Table 4. Model configuration parameters.

(right) under identical initialization (red point) and convergence target (blue point). The Euclidean case exhibits anisotropic contour geometry (labeled 1.5 to 7.5) spanning $x \in [-4, 4], y \in [-3, 3]$, where the optimization path manifests directional oscillations. This curvature-induced zigzag behavior forces sequential directional corrections along the trajectory. Conversely, the Oblique Manifold transformation produces concentric circular contours over $x \in [-2, 2], y \in [-2, 2]$, demonstrating isotropic normalization. Here, the optimization path follows an almost straight trajectory from initialization to convergence target, confirming effective regularization of landscape geometry. This visual evidence empirically validates how Oblique Manifold constraints mitigate directional bias by symmetrizing curvature

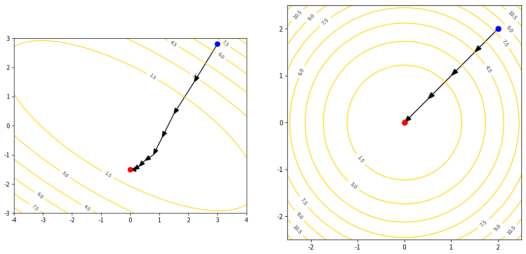

Figure 2. Normalized Data Accelerates Gradient Descent Convergence, (left) Irregular contours of non-normalized feature space, (right) Circular contours post-normalization with direct descent path.

profiles, enabling direct convergence pathways unattainable in unnormalized Euclidean space.

2.3. Ablation Experiment Supplement

2.3.1. Implementation of BiCA

During our implementation of the Bi-directional Contextual Attention (BiCA) mechanism [7], we systematically evaluated three distinct input configurations for bidirectional attention processing. The mechanism implements: In **Version 1**, Context-Aware Objects (CAO) derive from attending to Instance Features (IF) using Object-Aware Contexts (OAC) as both key and value; **Version 2** attends to IF using OAC as key and value; while **Version 3** attends to Context Features (CF) using IF as both key and value. All variants share consistent initialization where OAC is generated by attending to CF using IF as query.

Method	$\mathcal{L}_{des}$	IoU = 0.50			
		C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Version 1(BiCA^R)	MLE	65.22	37.59	26.87	55.76
Version 2		63.17	36.32	26.58	55.20
Version 3		64.91	36.21	26.71	55.14
Version 4		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**

Table 5. Comparison of bidirectional attention input configurations.

Table 5 presents comparative results identifying Version 1 as the optimal input configuration among the three variants. Building on this finding, we subsequently replaced our baseline methodology with Version 1's input scheme and benchmarked this adaptation as **Version 4** in comparative analyses, with detailed performance metrics documented in Table 5.

2.3.2. Selective Oblique Manifold Projection

In processing sparse point clouds, we leverage Oblique Manifold advantages to achieve significant performance

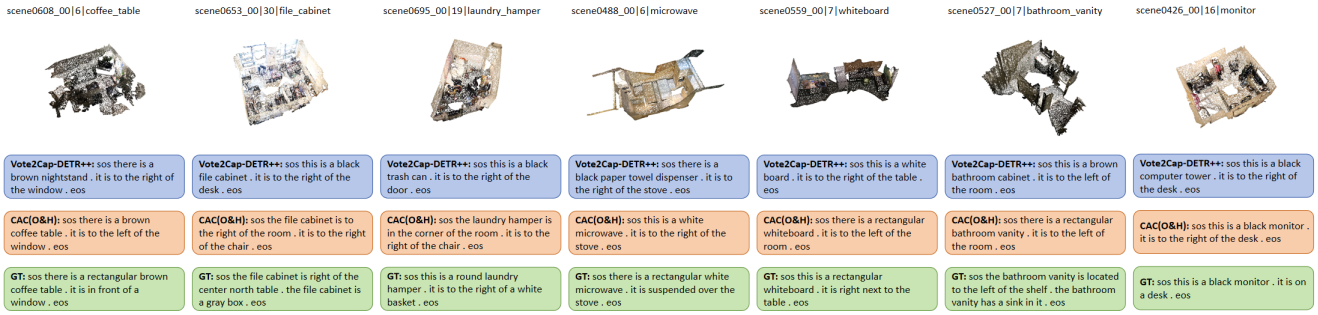

Figure 3. Supplementary qualitative results on the ScanRefer benchmark.

Method	$\mathcal{L}_{des}$	IoU = 0.50			
		C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
CAC($\mathcal{O}$)		68.07	36.53	26.72	55.08
Dec(Self)($\mathcal{O}$)		62.96	34.55	25.58	53.57
Dec(Cross)($\mathcal{O}$)	MLE	64.40	36.06	26.37	54.68
Enc & Dec(all)($\mathcal{O}$)		62.00	34.11	25.81	52.89
Enc($\mathbb{H}$)		66.85	36.49	26.79	55.21
CAC($\mathcal{O}$)		79.09	38.96	26.85	54.95
Dec(Self)($\mathcal{O}$)		72.84	36.81	26.02	52.50
Dec(Cross)($\mathcal{O}$)	SCST	73.58	38.87	26.08	53.84
Enc & Dec(all)($\mathcal{O}$)		70.40	35.81	25.58	52.14
Enc($\mathbb{H}$)		77.46	39.47	27.08	55.63

Table 6. Performance comparison of sparse point cloud processing configurations.

gains. Our approach specifically maps only the encoder to Oblique Manifold, contrasting this configuration with alternative methodologies. As Table 6 details, ablation studies systematically evaluate separate encoder and decoder mappings to Oblique Manifold, while further comparing self-attention versus cross-attention mechanisms within the decoder. To demonstrate Oblique Manifold superiority, we additionally map the encoder to Hyperbolic Manifold, providing direct comparison against our proposed framework.

2.3.3. Hyperparameter sweeps and robustness

We clarify that our training settings match the baseline. As space limited, the revised manuscript will include more ablation study across curvature values ($c \in [-2.0, -0.5]$) and temperature settings ($\tau \in [0.1, 2.0]$). We find curvature boosts performance, and our temperature modulates the sharpness of attention weights, our setting is a stable choice and there are still rooms for improvement.

2.4. Qualitative Analysis

While the baseline Vote2Cap-DETR++ is prone to object misclassification, our proposed CAC($\mathcal{O}\&\mathbb{H}$) precisely captures both fine-grained object semantics and complex spatial relations. Consequently, our approach generates highly faithful linguistic descriptions that closely align with the ground truth annotations. Further detailed qualitative comparisons are illustrated in Figure 3.

3. Limitations

Although the proposed Context-Aware Calibration (CAC) framework substantially mitigates semantic confusion and spatial hallucinations in 3D dense captioning, it is not without limitations. First, the integration of multi-level discriminative rewards via self-critical sequence training (SCST) inevitably introduces additional computational overhead and extends convergence time during the training phase. Second, as our framework heavily relies on scene graphs and 3D proposal generation paradigms, the quality of the generated captions remains inherently bounded by the efficacy of the upstream 3D object detector. Instances involving severe occlusions or highly sparse point clouds may still precipitate cascaded errors. Finally, our current empirical evaluations are predominantly centered on indoor scene datasets. Extending and adapting these context-aware calibration mechanisms to large-scale, unstructured outdoor environments presents a compelling avenue for future research.

Pretrain Configuration	mAP0.25	mAP0.50	AR0.25	AR0.50
τ_{obl}=1 (Epoch 119)	61.04	38.05	86.23	61.57
τ_{obl}=0.5 (Epoch 119)	59.61	36.43	83.68	57.60
τ_{obl}=2 (Epoch 119)	61.19	37.77	85.91	59.70
MLE Configuration with τ_{obl}=1	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
τ_{obl}=1, τ_{lor}=1, c=1	50.03	28.88	26.12	55.93
τ_{obl}=1, τ_{lor}=2, c=1	50.22	28.67	26.22	55.80
τ_{obl}=1, τ_{lor}=0.1, c=1	50.61	29.19	26.10	56.03
τ_{obl}=1, τ_{lor}=0.1, c=2	50.87	29.28	26.29	56.11
τ_{obl}=1, τ_{lor}=0.1, c=0.5	49.90	28.73	26.18	56.05

Table 7. Hyperparameter Sensitivity Analysis.

References

[1] P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. 1

[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. 2

[3] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018. 2

[4] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. pages 202–221, 2020. 2

[5] 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] 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. 1

[7] 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. 3

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] 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. 2