

Article

PCMTRefer: Symmetry-Aware Text-Guided Point Mamba for Referring Segmentation in Indoor 3D Point Clouds

Li Yuan ^{1,2} , Bo Kong ^{1,2} , Chenhao Li ^{1,2}, Anting Guo ^{1,2,*} and Wenjiang Huang ^{3,*} 

¹ School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China; lyuan@ustb.edu.cn (L.Y.); m202320748@xs.ustb.edu.cn (B.K.); m202510550@xs.ustb.edu.cn (C.L.)

² Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China

³ Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

* Correspondence: guoat@ustb.edu.cn (A.G.); huangwj@aircas.ac.cn (W.H.)

Abstract

Indoor 3D referring segmentation aims to identify and segment the target object in a point cloud according to a natural language expression. Although recent advances in multimodal feature fusion have markedly improved this task, existing methods do not jointly model the structural regularities commonly observed in indoor objects, such as approximate symmetry and repetitive local patterns, in conjunction with the directional constraints conveyed by referring expressions. In this work, we formulate symmetry-aware representation as the extraction of direction-consistent structural responses from both forward and backward traversals of the same spatially ordered sequence while preserving direction-sensitive variations introduced by occlusion, point cloud incompleteness, and cluttered scene layouts. Based on this formulation, we propose PCMTRefer, a symmetry-aware text-guided Point Mamba framework for indoor 3D referring segmentation. The input point cloud is first partitioned using an octree and arranged into a spatially coherent sequence via Z-order (Morton) ordering. A bidirectional state-space encoder then aggregates complementary context from both traversal directions, yielding richer representations of regular boundaries, repetitive structures, and approximately bilateral object geometries. In parallel, an asymmetric text-to-point guidance module injects semantic cues—object categories, attributes, and spatial relationships—into point-wise features, while a Background-Relaxation Token offers an auxiliary matching channel for non-target background regions. A Gumbel-Softmax-based semantic primitive learning module further extracts discriminative cues from referring expressions and integrates language semantics with point-level geometric features through a multi-scale decoder. Experimental results on the ScanRefer benchmark show that PCMTRefer achieves an Overall Acc@0.25 of 58.56%, an Overall Acc@0.5 of 54.19%, and an mIoU of 49.97%. Multi-seed validation further confirms the statistical stability of these results, with standard deviations of less than 0.2% across three independent runs.

Keywords: 3D point cloud; referring segmentation; text-guided segmentation; Point Mamba; multimodal fusion; state-space model; symmetry; asymmetry



Academic Editors: Shuang Xu and Zhixun Su

Received: 4 June 2026

Revised: 24 July 2026

Accepted: 31 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Three-dimensional point clouds preserve geometric, depth, and spatial layout information and have become a fundamental representation for indoor scene understanding [1–3]. In 3D referring segmentation, a model must recover the fine-grained boundaries of the target object while identifying the referred object from multiple candidates with similar appearances or semantic categories according to a natural language description [4,5]. Although indoor objects rarely exhibit strict mirror symmetry, they often contain regular contours, repetitive components, and approximately bilateral structures. At the same time, occlusion, partial scanning, and missing points introduce direction-sensitive variations [1]. An effective model must therefore exploit both stable structural regularities and asymmetric variations in complex scenes, while simultaneously incorporating linguistic cues such as object categories, attributes, and spatial relationships.

However, point clouds are inherently unordered, sparse, and non-uniform in density [2,3]. Local features are susceptible to noise and missing observations, whereas descriptions involving relative positions and inter-object relationships demand strong long-range spatial modeling [4,5]. Existing voxel-based methods suffer from empty-cell redundancy and high computational cost [6–9], while point-based approaches struggle to capture long-range structural dependencies. Transformer architectures can model global context, but the quadratic complexity of self-attention becomes a bottleneck [10]. Large background regions and geometrically similar candidates further lead to erroneous cross-modal matching.

We address these challenges with PCMTRefer, an end-to-end framework that couples symmetry-aware spatial modeling with asymmetric text-conditioned guidance. The method first builds a spatially coherent point sequence via octree partitioning and Z-order traversal, then aggregates complementary context through forward and backward state-space scans. For regular boundaries, repetitive local patterns, and approximately bilateral structures, the two traversal directions yield consistent structural responses; occluded, incomplete, and cluttered regions instead preserve direction-sensitive discrepancies. Combined with depthwise separable convolutions, the Point Mamba encoder captures both local geometric patterns and long-range spatial dependencies.

During cross-modal fusion, local point groups are formed by farthest point sampling and k-nearest neighbors, and cross-attention treats point features as queries and language features as keys and values. To suppress spurious alignments between background regions and irrelevant language tokens, a Background-Relaxation Token serves as an independent matching channel for non-target regions. A semantic primitive learning module employs Gumbel-Softmax to extract discriminative cues tied to object categories, attributes, and spatial relationships. These primitives then act as queries in a multi-scale decoder, aligning language semantics with point-level geometric features.

Our main contributions are threefold. First, we propose PCMTRefer, an end-to-end framework that unifies spatial structure modeling and text-conditioned guidance for 3D referring segmentation. Second, we develop a symmetry-aware Point Mamba encoder combining octree partitioning, Z-order traversal, and bidirectional state-space scanning to jointly capture structural consistency and directional variation. Third, we introduce a multimodal fusion mechanism with a Background-Relaxation Token and Gumbel-Softmax-based semantic primitive learning that suppresses background interference and improves discrimination among geometrically similar candidates.

2. Related Work

Visual–language understanding has evolved from simple feature matching to attention-based multimodal representation learning [11–14]. Early approaches encoded visual and textual inputs separately and combined them through feature concatenation, element-

wise multiplication, or similarity scoring. Transformer-based co-attention models and single-stream multimodal encoders later enabled direct interactions between visual regions and language tokens, improving cross-modal reasoning [10–13]. Large-scale contrastive pretraining further strengthened generalization by aligning visual and textual embeddings in a shared semantic space [14]. Most of these methods, however, are trained on 2D image–text pairs and cannot directly model 3D spatial structures. Adapting them to point clouds is nontrivial because depth, occlusion, surface geometry, and object layout are all essential in 3D scene understanding.

Referring image segmentation has laid an important methodological foundation for language-guided segmentation. This task aims to generate a pixel-level mask for the object described by a natural language expression. Existing methods have introduced language-guided attention, cross-modal self-attention, boundary-aware feature fusion, DETR-style query decoding, and prompting strategies based on foundation models [15–19]. These studies suggest that accurate segmentation benefits from early language guidance, repeated multimodal interaction, multi-scale visual features, and query-based mask decoding. Image-based methods are, however, designed for regular 2D grids and cannot be directly transferred to unordered point clouds. Moreover, whereas expressions in 2D scenes often hinge on visual appearance, referring expressions in indoor 3D scenes tend to rely more heavily on spatial relationships, object-support relations, and scene layout.

Point cloud semantic and instance segmentation have likewise seen rapid progress. Approaches have evolved from PointNet-style shared multilayer perceptrons to hierarchical local aggregation, graph neural networks, sparse convolutions, Point Transformers, and query-based mask prediction [2,3,8,9,20–23]. PointNet showed that raw unordered points can be processed directly, while PointNet++ improved local geometric modeling through hierarchical grouping. Point Transformer further strengthened global relational modeling among points but its memory cost can become prohibitive for large-scale point sets. Recent 3D instance segmentation methods employ learnable queries to predict object masks, offering a strong foundation for referring segmentation [18,23]. Still, most segmentation backbones do not incorporate language information during geometric feature encoding, and their queries are typically randomly initialized, category-initialized, or task-specific, rather than being derived directly from referring expressions.

Text-guided 3D understanding was first explored extensively in the context of 3D visual grounding. Given a natural language description, these methods predict the 3D bounding box of the corresponding target object. Public datasets such as ScanRefer and ReferIt3D have advanced 3D language understanding by requiring models to combine object appearance, spatial relationships, and scene context for target localization [4,5]. Bounding boxes, however, capture only coarse object locations and lack the precise point-level geometric information needed for applications such as robotic grasping, interactive scene editing, and fine-grained 3D perception. Research has therefore expanded from 3D visual grounding to 3D referring expression segmentation, where point-level masks provide more precise language-guided 3D understanding. Representative methods include TGNN, InstanceRefer, X-RefSeg3D, 3D-STMN, RefMask3D, SegPoint, and MDIN [24–30]. With the rapid development of large language models and visual foundation models, 3D vision–language tasks have shifted toward more open and fine-grained settings. In 2025–2026, 3D-DRES further extended this direction to Detailed Referring Expression Segmentation, emphasizing fine-grained boundary recovery under complex language descriptions, whereas SEGA3D unified referring segmentation and reasoning segmentation, so that models segment the target object and simultaneously reason about complex scene relationships through language. These developments suggest that 3D language-guided segmentation is moving from conventional cross-modal feature fusion toward more com-

prehensive frameworks that integrate semantic reasoning, open-vocabulary understanding, and fine-grained geometric modeling.

Despite this progress, two key limitations persist. First, most existing methods rely primarily on Transformers or cross-attention for visual–language interaction and incur considerable computational and memory costs when processing large-scale point clouds. Second, existing approaches usually match language tokens directly with point cloud features, with limited attention to noisy semantic alignment caused by the large number of background points in complex indoor scenes. They also lack explicit mechanisms for modeling discriminative linguistic cues. Unlike prior work, PCMTRefer does not merely introduce Point Mamba into 3D referring segmentation. Instead, it introduces a redesigned spatially ordered Point Mamba encoder specifically tailored for this task and further incorporates a Background-Relaxation Token and a Gumbel-Softmax-based semantic primitive learning module. These designs improve the efficiency of long-range geometric modeling while strengthening text–point cross-modal semantic alignment, yielding more accurate point-level referring segmentation.

3. Materials and Methods

3.1. Overall Network Architecture

PCMTRefer adopts an end-to-end architecture that integrates natural language semantics with the geometric information of indoor point clouds, comprising point cloud encoding, text representation, cross-modal feature interaction, and multi-scale mask decoding. Before introducing the individual components, we first formally define the 3D referring segmentation task and the corresponding model inputs and outputs.

Given an indoor point cloud containing N points,

$$\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^N \in \mathbb{R}^{N \times 6}, \quad (1)$$

where

$$\mathbf{p}_i = [x_i, y_i, z_i, r_i, g_i, b_i] \quad (2)$$

denotes the 3D coordinates and color attributes of the i -th point. The corresponding natural language expression is represented as

$$\mathbf{T} = \{t_j\}_{j=1}^L, \quad (3)$$

where L denotes the number of language tokens.

The objective of 3D referring segmentation is to learn a mapping from the point cloud and referring expression to a point-level target mask:

$$f_\theta : (\mathbf{P}, \mathbf{T}) \rightarrow \hat{\mathbf{M}}, \quad (4)$$

where f_θ denotes the proposed network parameterized by θ , and

$$\hat{\mathbf{M}} = [\hat{m}_1, \hat{m}_2, \dots, \hat{m}_N] \in [0, 1]^N \quad (5)$$

represents the predicted point-level mask probabilities. Each element $\hat{m}_i$ indicates the probability that the i -th point belongs to the object referred to by the language expression. During inference, the final binary mask is obtained by applying a threshold δ :

$$M_i = \begin{cases} 1, & \hat{m}_i \geq \delta, \\ 0, & \hat{m}_i < \delta. \end{cases} \quad (6)$$

The corresponding ground-truth mask used during training is denoted as

$$\mathbf{Y} = [y_1, y_2, \dots, y_N] \in \{0, 1\}^N, \quad (7)$$

where $y_i = 1$ indicates that the i -th point belongs to the target object, whereas $y_i = 0$ denotes a background point.

As illustrated in Figure 1, PCMTRefer predicts the target mask from the input pair $(\mathbf{P}, \mathbf{T})$ through four main stages: symmetry-aware point cloud encoding, language encoding and semantic enhancement, asymmetric text-guided multimodal fusion, and multi-scale semantic decoding. The point cloud branch first constructs a spatially ordered sequence and employs forward and backward state-space scans to extract direction-consistent structural information while preserving direction-sensitive variations. The language branch uses RoBERTa to extract contextual token embeddings and a global sentence representation.

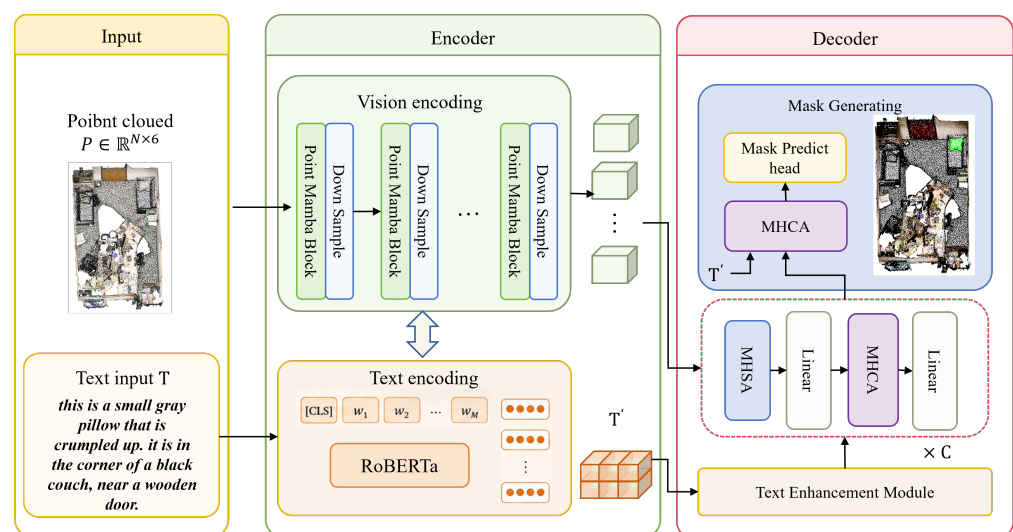


Figure 1. Overall architecture of the proposed symmetry-aware and asymmetric cross-modal PCMTRefer framework.

Across multiple feature scales, point features are conditionally guided by language representations through the asymmetric cross-modal fusion module. The fused vision–language features are then decoded using learned semantic primitives as queries. Finally, a lightweight mask prediction head produces point-wise logits, which are converted into the target segmentation mask.

3.2. Symmetry-Aware Point Mamba Encoder

Figure 2 illustrates the proposed symmetry-aware Point Mamba encoder for 3D referring segmentation. In this work, symmetry awareness does not refer to explicitly estimating an object’s symmetry axis or establishing mirror-point correspondences. Instead, it refers to extracting direction-consistent structural features and direction-sensitive asymmetric variations through a pair of opposite scans over the same spatially ordered reference. The encoder first transforms the unordered point cloud into a spatially coherent one-dimensional sequence via octree partitioning and Z-order sorting. The forward and backward state-space branches then aggregate contextual information along the sequence and its reverse order, respectively. Repetitive components and approximately symmetric structures tend to generate shared responses in both directions, whereas occlusion, missing points, and irregular layouts introduce directional discrepancies. The encoder adopts a hierarchical architecture with progressive downsampling to construct complementary point cloud representations at different spatial scales.

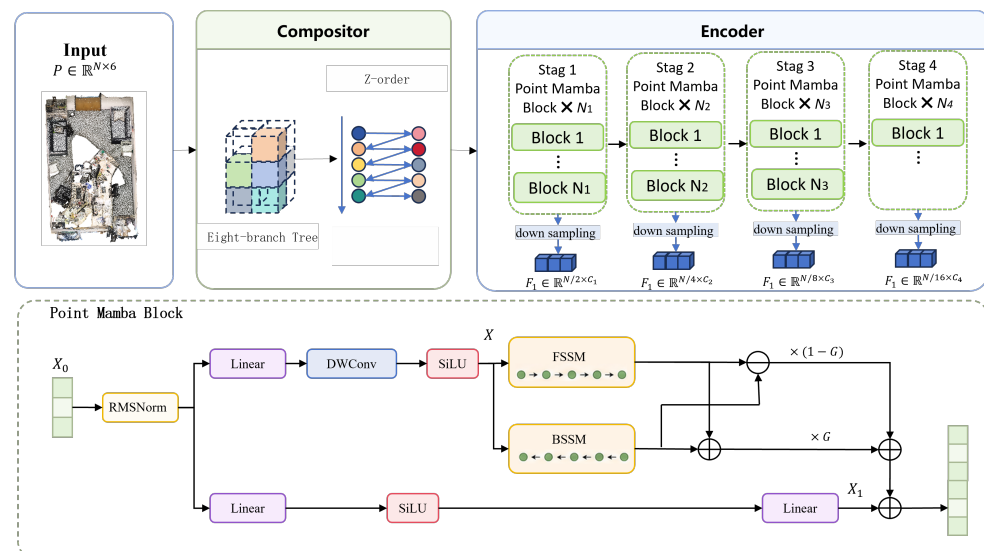


Figure 2. The structure of Symmetry-Aware Point Mamba Encoder.

State-space models require sequential inputs, whereas raw point clouds are inherently unordered sets. Arranging points in a random or arbitrary fixed order may disrupt spatial neighborhood relationships and impair geometric context modeling. We therefore first partition the input scene hierarchically with an octree and then sort the leaf nodes using Z-order (Morton ordering). In this way, the original 3D point cloud is mapped to a one-dimensional sequence with spatial continuity. Compared with conventional coordinate-based sorting strategies, Z-order better preserves local neighborhood relationships, allowing spatially adjacent points to remain relatively close in the resulting sequence [31].

After sorting, the point coordinates and features are projected into a latent feature space through a linear embedding layer. The encoder comprises four hierarchical stages, each containing several symmetry-aware Point Mamba blocks followed by a downsampling layer. The shallow stages capture local surfaces, boundaries, and repetitive neighborhood patterns. The intermediate stages expand the receptive field to model part-level structural correspondences, and the deeper stages aggregate object-level contours and scene-level contextual information. Since structural regularities may occur at the levels of local surfaces, component arrangements, and overall object contours, hierarchical encoding provides a basis for comparing and integrating directional responses across different spatial scales.

Each symmetry-aware Point Mamba block consists of two branches. In the first branch, the input features are normalized and linearly transformed, followed by a depthwise separable convolution and a SiLU activation. The depthwise separable convolution comprises a depthwise convolution and a pointwise convolution [32]. The depthwise convolution performs local convolution independently within each feature channel, and the pointwise convolution uses a 1×1 convolution to exchange information across channels. This branch enhances local geometric modeling at low computational cost.

In the second branch, the normalized input is projected to generate gating features, while the convolution-enhanced features are processed by forward and backward state-space scans. Let the point feature sequence obtained after the depthwise separable convolution be

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}, \quad (8)$$

where $\mathbf{x}_t$ denotes the input feature of the t -th point in the ordered sequence. To reduce the directional bias introduced by a single traversal order, state-space scanning is performed along both the forward and reverse Z-order sequences. The output of the backward branch

is subsequently remapped to the original point indices, thereby establishing point-wise correspondences between the two directional representations.

For the forward scan, the state is sequentially updated for $t = 1, \dots, N$ as follows:

$$\mathbf{h}_t^f = \mathbf{A}\mathbf{h}_{t-1}^f + \mathbf{B}\mathbf{x}_t, \quad (9)$$

$$\mathbf{y}_t^f = \mathbf{C}\mathbf{h}_t^f + \mathbf{D}\mathbf{x}_t, \quad (10)$$

where $\mathbf{h}_t^f$ and $\mathbf{y}_t^f$ denote the forward hidden state and output feature at position t , respectively. Correspondingly, the backward scan is performed for $t = N, \dots, 1$:

$$\mathbf{h}_t^b = \mathbf{A}\mathbf{h}_{t+1}^b + \mathbf{B}\mathbf{x}_t, \quad (11)$$

$$\mathbf{y}_t^b = \mathbf{C}\mathbf{h}_t^b + \mathbf{D}\mathbf{x}_t, \quad (12)$$

where $\mathbf{h}_t^b$ and $\mathbf{y}_t^b$ denote the backward hidden state and output feature aligned with the original point index, respectively. The matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ denote the state-transition and feature-projection parameters. The forward branch aggregates spatial context along the original order, whereas the backward branch captures complementary long-range dependencies from the opposite direction. This bidirectional design alleviates the influence of sequence orientation on point feature representation.

The forward and backward outputs are denoted as

$$\mathbf{Y}^f = [\mathbf{y}_1^f, \mathbf{y}_2^f, \dots, \mathbf{y}_N^f], \quad \mathbf{Y}^b = [\mathbf{y}_1^b, \mathbf{y}_2^b, \dots, \mathbf{y}_N^b]. \quad (13)$$

Based on these representations, the direction-consistent component jointly supported by the two traversal directions is defined as

$$\mathbf{Y}^{\text{sym}} = \mathbf{Y}^f + \mathbf{Y}^b, \quad (14)$$

whereas the response discrepancy between the two directions is defined as

$$\mathbf{Y}^{\text{asym}} = |\mathbf{Y}^f - \mathbf{Y}^b|. \quad (15)$$

The representation $\mathbf{Y}^{\text{sym}}$ remains unchanged when the forward and backward branches are exchanged. It preserves structural information that can be consistently captured from both traversal directions, such as regular boundaries, repetitive local patterns, and continuous surface structures. It thus serves as a direction-consistent representation of point cloud structure. In contrast, $\mathbf{Y}^{\text{asym}}$ measures discrepancies between the directional responses and captures asymmetric variations caused by occlusion, local point loss, scanning noise, and irregular scene layouts.

To adaptively balance the direction-consistent and direction-sensitive components, a gating coefficient is generated from the input features:

$$\mathbf{G} = \sigma(\mathbf{W}_g \mathbf{X} + \mathbf{b}_g), \quad (16)$$

where $\sigma(\cdot)$ denotes the sigmoid function. The final fused representation is computed as

$$\mathbf{Y} = \mathbf{G} \odot \mathbf{Y}^{\text{sym}} + (1 - \mathbf{G}) \odot \mathbf{Y}^{\text{asym}} + \mathbf{X}_1, \quad (17)$$

where $\odot$ denotes element-wise multiplication, $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable gating parameters, and $\mathbf{X}_1$ denotes the residual feature used to preserve the original point representation. The gating mechanism dynamically adjusts the contributions of the direction-consistent and direction-sensitive components according to the structural properties of each point. For

regions with regular contours or repetitive local structures, the model can assign greater importance to direction-consistent information. For regions affected by occlusion, missing observations, or complex neighborhood interference, the model can preserve discriminative directional information.

The symmetry-aware mechanism does not explicitly estimate geometric symmetry axes. Instead, it models stable structural responses and asymmetric variations under opposite traversal directions through paired forward–backward spatial scans and a consistency–discrepancy decomposition. This design enables the encoder to jointly capture fine-grained local geometry and long-range spatial context with linear sequence complexity [31,33].

3.3. Asymmetric Text-Guided Multimodal Feature Fusion

An asymmetric text-guided multimodal fusion module, embedded at multiple feature scales, bridges the modality gap between language and point cloud representations. At each scale, farthest point sampling (FPS) selects representative center points, following the hierarchical point-set abstraction strategy introduced in PointNet++ [3]. We sample 256 center points per scene; different language queries for the same scene share the same sampling result. Let the point coordinates and features at the s -th scale be denoted by

$$\mathbf{P}^{(s)} = \left\{ \mathbf{p}_i^{(s)} \right\}_{i=1}^{N_s}, \quad \mathbf{F}^{(s)} \in \mathbb{R}^{N_s \times d_s}, \quad (18)$$

where N_s and d_s denote the number of points and the feature dimension at the s -th scale, respectively. The representative center points are obtained as

$$\mathbf{C}^{(s)} = \text{FPS}\left(\mathbf{P}^{(s)}, M\right), \quad M = 256. \quad (19)$$

For each center point $\mathbf{c}_i^{(s)}$, its local neighborhood is constructed using k -nearest-neighbor search:

$$\mathcal{N}_i^{(s)} = \text{kNN}\left(\mathbf{c}_i^{(s)}, \mathbf{P}^{(s)}, k\right), \quad k = 32. \quad (20)$$

To eliminate the influence of global coordinate offsets, the coordinates within each local group are normalized relative to the corresponding center point:

$$\tilde{\mathbf{p}}_{ij}^{(s)} = \mathbf{p}_{ij}^{(s)} - \mathbf{c}_i^{(s)}, \quad \mathbf{p}_{ij}^{(s)} \in \mathcal{N}_i^{(s)}. \quad (21)$$

The normalized local coordinates and the point features produced by the preceding encoder stage are then aggregated to obtain grouped point representations. This grouping operation introduces a geometric prior into multimodal interaction, allowing cross-modal attention to be computed over spatially coherent local regions rather than isolated points prone to noise. The grouped point features serve as point-side interaction units, and the language token features provide semantic guidance.

PCMTRefer adopts an asymmetric text-to-point interaction mechanism. Local point cloud features act as queries; language tokens together with the Background-Relaxation Token serve as keys and values. Language information affects point features only through a unidirectional text-to-point pathway—the language representation is never updated by point cloud features. The two modalities thus play distinct roles during feature interaction: the point cloud branch preserves dense geometric evidence (object shape, spatial location, boundaries), and the language branch supplies conditional constraints (object categories, attributes, spatial relationships) that identify regions consistent with the referring expression. The asymmetry in this work therefore concerns cross-modal information flow, feature-update direction, and modality-specific functionality, not geometric structure per se. When multiple candidate objects exhibit similar appearances, approximately symmetric

structures, or belong to the same semantic category, linguistic cues such as color, size, and relative spatial relationships help resolve the ambiguity among geometrically similar candidates, pinpointing the referred target.

The Background-Relaxation Token (BRT) is a key component of the proposed multi-modal fusion module. Indoor 3D scenes usually contain a large number of background regions, such as walls, floors, and unrelated furniture, that are semantically irrelevant to the referring expression. Conventional cross-attention mechanisms nevertheless force all point features to establish correspondences with language tokens, which may introduce noisy semantic alignments. We address this with a learnable Background-Relaxation Token. During cross-modal fusion, the Background Token is concatenated with the language tokens to build an augmented language representation; cross-modal attention is then computed between the augmented language features and local point cloud features. A lightweight classification branch predicts whether each query belongs to foreground or background, routing background queries to the Background Token. When all queries are classified as background, the model outputs an empty mask. Through this mechanism, the Background Token absorbs features irrelevant to the referring expression, reducing noisy semantic alignment and yielding more stable point-level segmentation.

The language-enhanced point features are propagated back to the original point cloud through inverse-distance weighted interpolation, following a strategy similar to feature propagation in hierarchical point-based networks [3]. The propagated features are multiplied element-wise with the original point representations to generate scale-specific vision–language features. By fusing at multiple feature scales, language information guides both coarse-grained structural reasoning and fine-grained geometric localization. This fusion strategy also effectively suppresses background interference in cluttered indoor scenes, particularly when multiple objects belong to the same semantic category.

3.4. Symmetry-Aware Semantic Primitive Text Enhancement

RoBERTa token embeddings provide strong contextual language representations. However, 3D referring segmentation usually depends on a small number of informative words. For example, in the expression “the small round table beside the white chair”, the words small, round, beside, and white convey different types of discriminative information. A single global sentence representation may weaken these fine-grained semantic attributes, while the original token embeddings do not explicitly organize semantic cues for mask decoding.

As illustrated in Figure 3, the module first initializes a set of learnable semantic features $\mathbf{O}$ from a Gaussian distribution. These features are projected into the same dimensional space as the RoBERTa token embeddings and are used as primitive queries. The token embeddings are projected into keys and values, and a similarity matrix is then computed between the primitive queries and the token keys.

To sharpen the discriminative power of the language representations, we introduce a semantic primitive learning mechanism based on Gumbel–Softmax [34]. Learnable primitive queries first interact with the language tokens, after which Gumbel–Softmax performs differentiable approximate discrete selection, encouraging each semantic primitive to concentrate on a handful of discriminative words rather than spreading attention across all tokens.

Unlike conventional Softmax, which distributes continuous attention weights over all tokens, Gumbel–Softmax suppresses semantic interference from irrelevant words and keeps fine-grained semantic information from being diluted across many tokens, while still supporting gradient propagation. After semantic aggregation and linear projection,

different semantic primitives capture complementary information about color, shape, size, spatial position, and relative relationships.

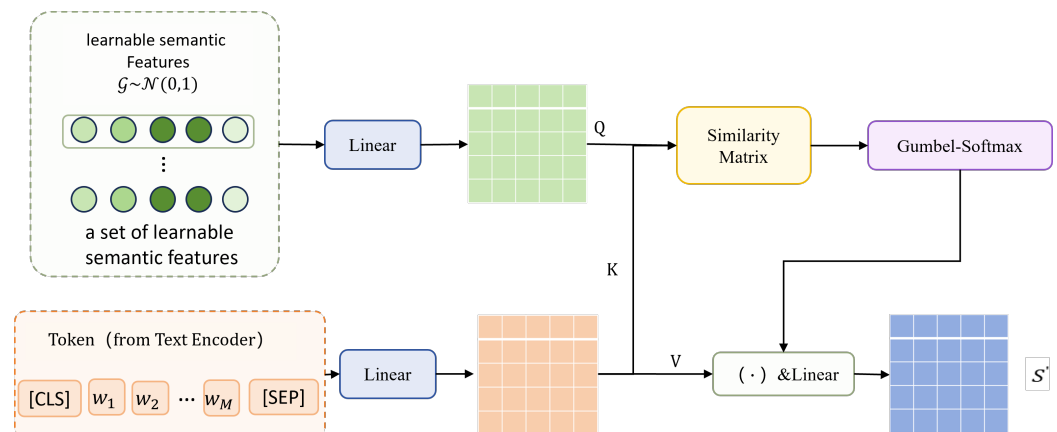


Figure 3. Architecture of the semantic primitive text enhancement module.

The learned primitives then serve as queries in the multi-scale decoder, steering the mapping of language semantics onto point-level geometric features and improving text–point semantic alignment in complex indoor scenes.

The enhanced semantic primitives are finally used as decoder queries. They allow the decoder to attend to point features from different semantic perspectives and combine the resulting representations for final mask prediction.

3.5. Multi-Scale Fusion Decoder and Training Objectives

The decoder maps the enhanced language semantics into the point space through an asymmetric cross-modal interaction process. It takes the semantic primitives as query vectors and the multi-scale vision–language point features as keys and values. Multi-scale cross-attention is sequentially applied to low-, medium-, and high-resolution point features.

Self-attention among the semantic primitives exchanges complementary information across different linguistic attributes. Residual connections and layer normalization stabilize multimodal feature learning. Because each primitive attends to different semantic attributes, the decoder establishes more discriminative correspondences between the textual description and point-level geometric structures.

After several decoding layers, the model obtains visually decoded features aligned with the semantic primitives. In a subsequent aggregation step, the global sentence feature attends to the primitive-aligned visual features as a query. The result is combined with the original features via a residual connection and fed into a lightweight mask prediction head. The prediction head outputs a binary probability for each point, and the final segmentation mask is obtained by thresholding the predicted probabilities.

Viewing the decoder through the lens of joint structural and linguistic modeling, it first receives direction-consistent structural information and asymmetric observation variations from the point cloud branch, then progressively incorporates attribute and directional relationship constraints from the language representation. For multiple candidate objects with similar geometric appearances or approximately symmetric structures, structural information helps preserve complete boundaries and local shapes, whereas language conditions distinguish the referred object from competing candidates, enabling target-specific point-level segmentation.

The training objective consists of a cross-modal semantic alignment loss, a point-level binary segmentation loss, and a Dice loss. The overall loss is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{BCE}} + \lambda_3 \mathcal{L}_{\text{Dice}}, \quad (22)$$

where $\mathcal{L}_{\text{align}}$ denotes the cross-modal contrastive semantic alignment loss, which pulls point cloud query features toward the corresponding semantic tokens in the referring expression—nouns, attribute words, spatial relation words—within a shared semantic space, while pushing them away from irrelevant tokens. $\mathcal{L}_{\text{BCE}}$ supervises point-wise binary mask prediction, and $\mathcal{L}_{\text{Dice}}$ optimizes the overlap between the predicted mask and the ground-truth mask.

All loss weights are held fixed across experiments. This objective jointly considers cross-modal semantic representation learning and point-level segmentation optimization, enabling the model to learn more consistent vision–language representations and produce more accurate target masks.

3.6. Symmetry and Asymmetry Analysis

The proposed framework models symmetry and asymmetry from two complementary perspectives: structural response consistency and cross-modal information flow. From the perspective of geometric encoding, symmetry awareness is defined as the extraction of direction-consistent responses from forward and backward traversals of the same spatially ordered sequence. Octree partitioning and Z-order traversal provide a unified and spatially coherent sequential reference, while bidirectional state-space scanning aggregates contextual information from opposite traversal directions. For objects with regular contours, repetitive local patterns, and approximately bilateral structures, the two traversal directions tend to produce highly consistent structural responses. In contrast, occlusion, point cloud incompleteness, and cluttered scene layouts lead to more pronounced directional discrepancies. Consequently, the bidirectional representations jointly preserve both structural regularities and asymmetric observation characteristics. From the perspective of cross-modal interaction, PCMTRefer adopts an asymmetric text-to-point guidance mechanism, where point cloud features serve as queries and language features together with the Background-Relaxation Token serve as keys and values. The language branch specifies object categories, attributes, and spatial relationships, whereas the point cloud branch provides dense geometric evidence. Through this unified design, PCMTRefer integrates symmetry-aware point cloud representation learning with asymmetric text-guided multimodal interaction within a single coherent framework.

4. Experimental Results

4.1. Datasets

We evaluate PCMTRefer on two public benchmarks, ScanRefer and Multi3DRefer, which represent different 3D referring segmentation settings. Both datasets are built upon the ScanNet v2 indoor 3D scanning dataset [1]. Their official training, validation, and test splits contain 562/141/97 scenes and 707/141/333 scenes, respectively.

ScanRefer provides 51,583 language–object referring samples across 800 indoor scenes [4]. Each expression corresponds to a unique target object and describes its category, appearance, size, color, spatial location, and relationships with surrounding objects. Following the official protocol, the test set is split into Unique (target category appears once) and Multiple (several objects share the same category, demanding finer-grained attribute and spatial-relation reasoning) subsets.

Multi3DRefer extends ScanRefer to more challenging multi-target settings [35], introducing Multi-target, Zero-shot, and Distractor scenarios that stress complex spatial reasoning, multi-object localization, and cross-modal semantic understanding.

Both datasets use the 18 semantic categories defined in ScanNet. Expressions average 8–10 words in ScanRefer and 10–12 words in Multi3DRefer, and cover color, size, shape, spatial relationships, and inter-object relations.

We adopt standard data augmentation during training to improve generalization: random point sampling, coordinate normalization, random rotation, random coordinate perturbation, and color augmentation.

Comparative experiments are conducted on both ScanRefer and Multi3DRefer. Because ablation studies require repeated training at substantial computational cost, all ablations are performed on ScanRefer, which is the more widely adopted benchmark for component analysis.

4.2. Experimental Settings

PCMTRefer is implemented in PyTorch 2.1.0. Experiments run on Ubuntu 22.04 LTS with CUDA 12.1 and cuDNN 8.9, using four NVIDIA GeForce RTX 4090 GPUs (24 GB each).

We train with AdamW for 80 epochs at a batch size of 1. The initial learning rate is 1×10^{-4} , and a MultiStepLR scheduler reduces it by a factor of 10 over the final 40 epochs.

To comprehensively evaluate the statistical stability of the proposed method, all experiments are conducted under three independent random seeds: 42, 123, and 456. For each seed, the random states of Python 3.9, NumPy 1.24.3, and PyTorch 2.0.1 are deterministically fixed. Unless otherwise specified, the reported results represent the mean and standard deviation computed across the three runs. The main dependencies used for data processing, point cloud operations, and language encoding include NumPy 1.24.3, SciPy 1.10.1, Open3D 0.17.0, and Hugging Face Transformers 4.30.2.

4.3. Comparison Experiments

PCMTRefer delivers competitive performance on both ScanRefer and Multi3DRefer, demonstrating its applicability to single-target, multi-target, and complex semantic referring scenarios. The quantitative comparison results on ScanRefer are presented in Table 1, while the comparison results on Multi3DRefer are reported in Tables 2 and 3.

Table 1. Performance comparison on the ScanRefer dataset (%).

Method	U@0.25	U@0.5	M@0.25	M@0.5	O@0.25	O@0.5	mIoU
TGNN [24]	–	–	–	–	37.5	31.4	27.8
TGNN [†] [24]	69.3	57.8	31.2	26.6	38.6	32.7	28.8
InstanceRefer [†] [25]	81.6	72.2	29.4	23.5	40.2	33.5	30.6
X-RefSeg3D [26]	–	–	–	–	40.3	33.8	29.9
3D-STMN [27]	89.3	84.0	46.2	29.2	54.6	39.8	39.5
SegPoint [29]	–	–	–	–	–	–	41.7
RefMask3D [28]	89.6	84.7	48.1	40.8	55.9	49.2	44.9
MDIN [30]	91.0	87.2	50.1	44.9	58.0	53.1	48.3
PCMTRefer (mean ± std)	91.65 ± 0.18	86.36 ± 0.18	51.80 ± 0.22	45.31 ± 0.20	58.53 ± 0.17	54.16 ± 0.18	49.96 ± 0.19

[†] Results reproduced by MDIN [30] under their experimental settings.

On the ScanRefer dataset, PCMTRefer achieves an Overall Acc@0.25 of 58.56% and an Overall Acc@0.5 of 54.19%, outperforming most existing methods. Multi-seed validation across three independent runs (seeds 42, 123, and 456) yields consistent results with standard deviations of 0.17% (O@0.25) and 0.18% (O@0.5), confirming the robustness of these performance gains. On the Multi3DRefer dataset, the model obtains an Overall Acc@0.25 of 68.8%, an Overall Acc@0.5 of 46.4%, and an mIoU of 48.9%. It also performs

well under challenging evaluation protocols, including Multi-target and Single-target with distractors. These results show that the proposed method accurately localizes single targets and also handle complex spatial relationships, multi-object referring expressions, and ambiguity caused by same-category distractors.

Table 2. Performance comparison on Multi3DRefer at Acc@0.25 (%).

Method	Acc@0.25 (%)					Overall
	zt w/ dis	zt w/o dis	st w/ dis	st w/o dis	mt	
ReLA	36.2	72.7	48.3	83.4	73.0	61.8
M3DRef	39.2	81.6	50.8	77.5	66.8	55.7
CLIP3D-STMN	42.6	76.2	49.0	77.8	68.8	60.4
MDIN	47.9	78.8	55.5	84.4	76.3	67.0
PCMTRefer (mean $\pm$ std)	47.00 $\pm$ 0.25	79.27 $\pm$ 0.21	56.67 $\pm$ 0.29	87.07 $\pm$ 0.21	76.97 $\pm$ 0.21	68.77 $\pm$ 0.21

Table 3. Performance comparison on Multi3DRefer at Acc@0.5 and mIoU (%).

Method	Acc@0.5 (%)					Overall	mIoU
	zt w/ dis	zt w/o dis	st w/ dis	st w/o dis	mt		
ReLA	36.2	72.7	20.4	65.5	42.4	37.4	42.8
M3DRef	39.2	81.6	29.4	67.4	41.0	37.5	37.4
CLIP3D-STMN	42.6	76.2	24.6	69.2	43.9	40.9	43.0
MDIN	47.9	78.8	29.5	71.7	46.8	44.7	47.5
PCMTRefer (mean $\pm$ std)	46.57 $\pm$ 0.21	78.90 $\pm$ 0.24	31.10 $\pm$ 0.24	71.90 $\pm$ 0.24	47.57 $\pm$ 0.21	46.40 $\pm$ 0.24	48.90 $\pm$ 0.24

On the Unique-object subset of ScanRefer, PCMTRefer achieves an Acc@0.25 of 91.67% and an Acc@0.5 of 86.39%. Among the compared methods, the Acc@0.25 result is the highest, while the Acc@0.5 result is slightly lower than that of MDIN but still surpasses most representative methods.

The Multiple-object subset of ScanRefer and the Multi-target, Zero-shot, and Distractor settings of Multi3DRefer pose greater challenges. These scenarios usually contain multiple objects from the same category, making it difficult to identify the referred target using geometric appearance alone. The model must also exploit linguistic constraints, including color, spatial position, relative direction, and contextual relationships.

On the Multiple-object subset of ScanRefer, PCMTRefer achieves an Acc@0.25 of 51.80% and an Acc@0.5 of 45.32%, exceeding MDIN by 1.70 and 0.42 percentage points, respectively. On the Multi3DRefer dataset, the model achieves an Acc@0.25 of 77.0% and an Acc@0.5 of 47.6% in the Multi-target setting, with an overall mIoU of 48.9%. These results indicate that, as scene complexity increases, the text-guided cross-modal fusion strategy exploits attribute descriptions and spatial relationships to better distinguish among multiple candidate objects while reducing incorrect matches caused by same-category distractors.

Qualitative results on ScanRefer are shown in Figure 4, and qualitative comparisons on Multi3DRefer appear in Figure 5. The visualizations confirm that PCMTRefer produces more complete, continuous, and accurate target masks across single-target, multi-target, and distractor-containing scenes. In particular, when multiple objects from the same category are present, the model leverages color, position, and spatial-relation descriptions to identify the referred target, whereas the compared methods are more likely to produce missed targets, incomplete boundaries, or erroneous segmentation of distractor objects. These results suggest that the Point Mamba encoder, text-guided multimodal fusion

module, and semantic primitive text enhancement module jointly improve geometric–semantic correspondence and enhance target localization and segmentation quality in complex scenes.

The failure cases on the Multi3DRefer dataset are illustrated in Figure 6. PCMTRefer still exhibits limitations in a small number of highly challenging scenarios. When the referring expression involves zero-shot targets, multiple highly similar objects, or strong semantic ambiguity, the model may still be affected by distractor objects, resulting in localization errors or incorrect segmentation. In particular, when the target category is absent from the training distribution or several candidate objects have highly similar appearances and spatial layouts, the current cross-modal alignment mechanism may be insufficient to fully resolve semantic ambiguity. These observations indicate that the proposed method would benefit from further advances in open-vocabulary 3D understanding, target-existence prediction, and zero-shot generalization.

Taken together, the quantitative comparisons, qualitative visualizations, and failure-case analysis on ScanRefer and Multi3DRefer show that the Point Mamba encoder, text-guided multimodal fusion module, and semantic primitive text enhancement module each contribute distinctively—through geometric representation learning, cross-modal semantic alignment, and fine-grained language modeling, respectively. Multi-seed validation confirms that these gains are statistically stable. Together, these components enable PCMTRefer to achieve robust and competitive performance across different task settings and complex indoor scenes, while also revealing its current limitations in zero-shot semantic understanding and difficult target discrimination.



Figure 4. Qualitative visualization results of PCMTRefer for indoor 3D referring segmentation.

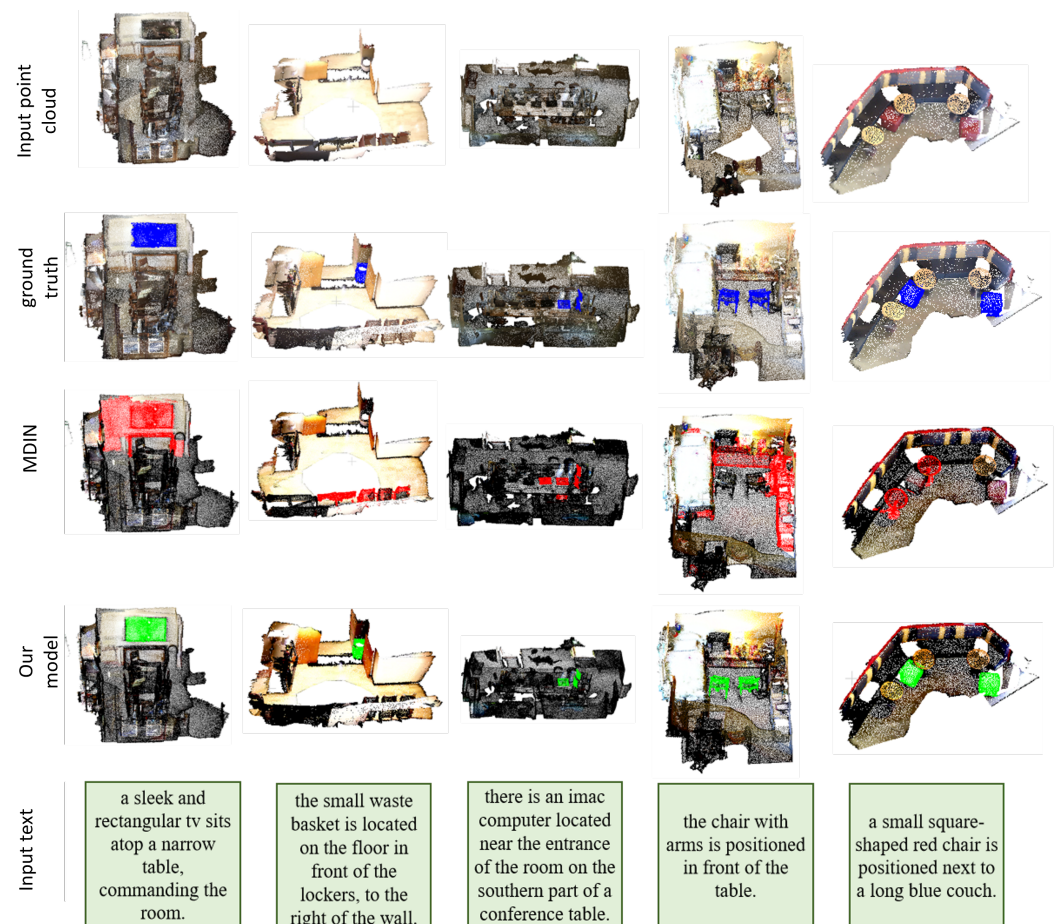


Figure 5. Qualitative visualization results on the Multi3DRefer dataset.

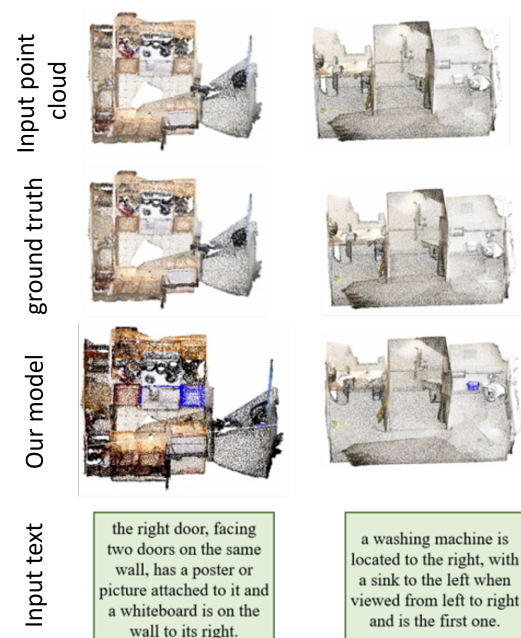


Figure 6. Visualization of typical failure cases on the Multi3DRefer dataset.

4.4. Computational Efficiency Analysis

We assess the computational efficiency of the Point Mamba encoder through a controlled comparison, replacing the Point Mamba module with a Transformer encoder while holding the remaining network architecture, feature dimensions, and training config-

urations constant. Both models are kept at comparable parameter counts to ensure a fair comparison.

Computational cost and inference time are measured over all samples in ScanRefer and Multi3DRefer. Inference time is the per-sample average; peak memory is the maximum GPU memory observed during the full evaluation. Results appear in Table 4.

Table 4. Computational efficiency comparison between the proposed model and the Transformer-based variant.

Method	FLOPs (G)	Params (M)	Inference Time (ms)	Peak Memory (MB)
Transformer	15.66	139.96	415.7	1904
Ours	12.05	140.01	312.8	1904

Table 4 shows that the Point Mamba model reduces FLOPs from 15.66 G to 12.05 G (roughly 23.1%) and cuts average inference time from 415.7 ms to 312.8 ms (roughly 24.8%). Both models share the same peak GPU memory of 1904 MB, so the Point Mamba design does not add memory overhead under identical architecture and feature dimensions.

The two models have nearly identical parameter counts (140.01 M vs. 139.96 M). The efficiency gains therefore do not come from reduced capacity; rather, the Point Mamba encoder achieves lower computational cost and faster inference at a comparable model size.

4.5. Ablation Studies

We evaluate the contributions of key components through ablation experiments on ScanRefer, using Acc@0.5 and mIoU as the primary metrics. Acc@0.5 measures mask overlap with the referred object under a stricter IoU threshold; mIoU captures the average quality of dense point-level segmentation. The ablation results for the point cloud encoder, text enhancement module, and multimodal feature fusion module are reported below.

4.5.1. Ablation Study of the Symmetry-Aware Point Mamba Encoder

Table 5 compares three encoder variants: a 3D U-Net encoder, a Point Mamba block without bidirectional state-space modeling, and the complete symmetry-aware Point Mamba block. All three variants use the same text fusion module, semantic decoder, training strategy, and hyperparameter settings, with only the geometric encoder being replaced.

Table 5. Ablation study of the point cloud encoder on the ScanRefer dataset (%).

Variant	Acc@0.5	mIoU
3D U-Net encoder	49.82	47.54
Remove Bi-SSM	50.27	47.89
Full Point Mamba block	54.19	49.97

The 3D U-Net encoder aggregates spatial neighborhood information through local 3D convolutions [7]. Unlike the Point Mamba encoder, it neither linearizes point-cloud features into a Z-order sequence nor captures long-range structural dependencies via bidirectional scanning. Consequently, replacing Point Mamba with 3D U-Net leaves the model reliant on local convolutional aggregation, limiting its capacity to jointly capture repetitive local patterns, long-range structural correspondences, and directional response variations due to occlusion, missing observations, or irregular scene layouts.

The full model scans the Z-order sequence bidirectionally and jointly represents both the structural information shared across traversal directions and the discrepancies between them. With the Point Mamba encoder, the model achieves Acc@0.5 of 54.19% and mIoU of 49.97%. The gain over the 3D U-Net variant confirms that Point Mamba yields

richer representations of long-range geometric dependencies, approximately symmetric structures, and direction-sensitive variations in complex 3D scenes.

This ablation experiment is designed to evaluate the contribution of each component under a unified setting, not to independently optimize different encoder architectures. Accordingly, for the 3D U-Net variant, we retained the same training strategy, optimizer, learning rate, loss function, decoder architecture, and all other network components as in the complete model, changing only the encoder. This controlled setup ensures a fair comparison and minimizes the influence of unrelated implementation differences.

4.5.2. Ablation Study of the Semantic Primitive Text Enhancement Module

Table 6 evaluates the effects of semantic primitive learning and the Gumbel–Softmax token selection mechanism.

Table 6. Ablation study of the text enhancement module on the ScanRefer dataset (%).

Variant	Acc@0.5	mIoU
Without text enhancement	52.93	48.02
Replace Gumbel–Softmax with Softmax	53.76	49.35
Full text enhancement	54.19	49.97

Removing the text enhancement module forces the model to rely directly on RoBERTa features. With standard Softmax, attention can still diffuse across uninformative words. The complete module, via approximate discrete token selection, steers each semantic primitive toward target attributes and spatial relationships.

The full text enhancement module achieves Acc@0.5 of 54.19% and mIoU of 49.97%, confirming that discriminative semantic extraction strengthens the language branch and improves final segmentation.

4.5.3. Ablation Study of the Text-Guided Multimodal Feature Fusion Module

Table 7 presents the ablation results for the text-guided multimodal feature fusion module. “Without Fusion” indicates that cross-attention is removed and only a linear layer is used for feature alignment, whereas “Without BRT” indicates that the Background-Relaxation Token is removed.

Table 7. Ablation study of the multimodal feature fusion module on the ScanRefer dataset (%).

Method	Acc@0.5	mIoU
Without Fusion	50.73	47.16
Without BRT	53.48	49.12
Ours	54.19	49.97

When cross-attention is replaced with a linear alignment layer, Acc@0.5 decreases from 54.19% to 50.73%, while mIoU decreases from 49.97% to 47.16%. When the Background-Relaxation Token is removed, Acc@0.5 decreases to 53.48% and mIoU decreases to 49.12%.

These results confirm that the cross-attention mechanism effectively injects semantic information and strengthens language guidance over the point cloud. The Background-Relaxation Token provides an additional matching channel for regions unrelated to the referring expression, reducing interference from irrelevant background points. A simple linear fusion layer, by contrast, fails to capture fine-grained geometric–semantic correspondences, weakening multimodal interaction, particularly in complex scenes.

Overall, the ablation results highlight the complementary nature of the proposed components. The Point Mamba encoder strengthens local geometric representation and

long-range contextual modeling via spatially ordered sequences and bidirectional state-space scanning. The text enhancement module sharpens linguistic representations through Gumbel–Softmax-based semantic primitive learning. The text-guided multimodal fusion module enables multi-scale interaction between point cloud geometry and language semantics through local grouping, cross-attention, and the Background-Relaxation Token. Together, these components improve target localization and dense point-level segmentation in complex indoor scenes.

4.6. Multi-Seed Statistical Validation

To assess statistical reliability beyond a single initialization, we train PCMTRefer three times with random seeds 42, 123, and 456. All other settings—optimizer, learning-rate schedule, data augmentation, and network architecture—are held fixed across runs. On ScanRefer (Table 1), the three runs yield means and standard deviations of $58.53\% \pm 0.17\%$ (O@0.25), $54.16\% \pm 0.18\%$ (O@0.5), and $49.96\% \pm 0.19\%$ (mIoU). On Multi3DRefer, the overall metrics at Acc@0.25 and Acc@0.5 show standard deviations of 0.21% and 0.24%, respectively (Tables 2 and 3). The consistently small standard deviations across both datasets show that PCMTRefer converges stably and is not unduly sensitive to random initialization. These results confirm that the performance gains reported above are statistically meaningful, not artifacts of a particular seed.

5. Discussion

Three-dimensional referring segmentation requires the joint modeling of point cloud geometry and linguistic semantics. The experimental results show that spatially ordered point cloud sequence modeling, long-range dependency capture, and text-guided feature fusion jointly contribute to target localization and point-level mask prediction. In particular, when multiple instances from the same category are present in a scene, linguistic information related to position, attributes, and inter-object relationships is critical for distinguishing the referred target.

The Point Mamba encoder transforms unordered point clouds into spatially coherent sequences through Z-order traversal and employs bidirectional state-space modeling to aggregate both local structures and global context [31,33]. The ablation results show that removing Bi-SSM leads to clear performance degradation, suggesting that long-range spatial dependencies aid understanding complex scene relationships. Notably, the proposed method does not explicitly estimate object symmetry axes or impose a symmetry loss. Instead, it relies on local structural regularities and spatial neighborhood continuity rather than strict geometric symmetry.

For language modeling, the semantic primitive module extracts discriminative information related to object categories, attributes, and spatial relationships, and uses the resulting representations as queries to guide point feature decoding. The Background-Relaxation Token provides an additional matching channel for irrelevant regions, reducing unreliable correspondences between background points and language tokens. The corresponding results suggest that fine-grained linguistic information and background suppression contribute positively to point-level mask prediction.

In terms of computational efficiency, the Mamba-based and Transformer-based variants have nearly identical parameter counts and peak memory consumption. However, the Mamba-based model requires fewer FLOPs and achieves a shorter average inference time. This suggests that the efficiency improvement stems from the sequence modeling mechanism of the state-space model rather than from a reduction in model capacity. Furthermore, the multi-seed validation experiments (Section 4.6) demonstrate that the proposed method maintains stable convergence and consistent performance across different random initial-

izations, with standard deviations below 0.25% on all primary metrics. Nevertheless, the current scene scale remains limited, and the scalability of the proposed model to larger point clouds and longer sequences warrants further investigation.

Future work will target zero-shot 3D referring segmentation, target-existence prediction, large-scale 3D vision–language pretraining, and lighter, more efficient Point Mamba architectures. These directions should further improve the robustness, generalization ability, and practical applicability of the model in complex real-world scenes.

6. Conclusions

This paper introduced PCMTRefer, an end-to-end framework for text-guided 3D referring segmentation. The framework employs spatially ordered bidirectional state-space modeling to capture local structural regularities and long-range contextual information from point clouds. It further integrates text-guided multimodal fusion, the Background-Relaxation Token, and semantic primitive learning to facilitate effective interaction between linguistic semantics and point-level geometric features, improving target localization and mask prediction in complex indoor scenes.

Extensive experimental results on the ScanRefer and Multi3DRefer public datasets show that PCMTRefer achieves competitive performance across multiple evaluation metrics. On ScanRefer, the model obtains an Overall Acc@0.25 of 58.56%, an Overall Acc@0.5 of 54.19%, and an mIoU of 49.97%. Multi-seed validation across three independent runs (seeds 42, 123, and 456) yields mean scores of $58.53\% \pm 0.17\%$ (O@0.25), $54.16\% \pm 0.18\%$ (O@0.5), and $49.96\% \pm 0.19\%$ (mIoU), confirming the statistical stability of the reported results. It also demonstrates favorable segmentation performance under the Overall and Multi-target evaluation settings of Multi3DRefer. The ablation results indicate that the Point Mamba encoder, text enhancement module, and text-guided multimodal fusion module all contribute positively to the final performance. The computational efficiency comparison further shows that, with nearly identical parameter counts and peak memory consumption, the Mamba-based variant reduces FLOPs and average inference time by approximately 23.1% and 24.8%, respectively, compared with the Transformer-based variant.

The proposed method still has several limitations. The current evaluation mainly focuses on static indoor scenes, and the diversity of data distributions and scene scales remains limited. In challenging cases involving zero-shot referring expressions, target-absent scenarios, open-vocabulary understanding, or highly similar semantic descriptions, the model may still be affected by semantic ambiguity and distracting objects. Although the multi-seed experiments confirm stable convergence on the current benchmarks, the generalization ability and statistical reliability of the model on larger-scale point clouds, cross-scene distributions, and real-world dynamic environments warrant further investigation.

Future work will target zero-shot 3D referring segmentation, target-existence prediction, large-scale 3D vision–language pretraining, and lighter, more efficient Point Mamba architectures. We also plan to investigate cross-dataset generalization, robustness in complex scenes, and more rigorous statistical evaluation, with the aim of improving the practical applicability of the model in real-world 3D interactive environments.

Author Contributions: Conceptualization, L.Y. and B.K.; methodology, L.Y. and B.K.; software, B.K. and C.L.; validation, C.L. and A.G.; formal analysis, L.Y. and B.K.; investigation, B.K. and C.L.; resources, L.Y. and W.H.; data curation, C.L.; writing—original draft preparation, L.Y. and B.K.; writing—review and editing, L.Y., B.K., C.L. and W.H.; visualization, C.L. and A.G.; supervision, L.Y. and W.H.; project administration, L.Y. and W.H.; funding acquisition, L.Y. and W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the scientific and technological project of the Ministry of Agriculture and Rural Affairs, the Fundamental Research Funds for the Central Universities (FRF-TP-26-016), and the National Natural Science Foundation of China (42201365).

Data Availability Statement: The ScanRefer dataset used in this study is publicly available at <https://github.com/daveredrum/ScanRefer>, accessed on 17 July 2026, and the ScanNet dataset is available at <https://github.com/ScanNet/ScanNet>, accessed on 17 July 2026, both under their respective dataset licenses. The source code and trained models of PCMTRefer are available at <https://github.com/mi0000/PCMTRefer>, accessed on 17 July 2026.

Acknowledgments: The authors acknowledge the high-performance computing platform of the University of Science and Technology Beijing for computational support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5828–5839.
2. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
3. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.
4. Chen, D.Z.; Chang, A.X.; Nießner, M. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 202–221.
5. Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; Guibas, L.J. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 422–440.
6. Maturana, D.; Scherer, S. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 922–928.
7. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Athens, Greece, 17–21 October 2016; pp. 424–432.
8. Graham, B.; Engelcke, M.; van der Maaten, L. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232.
9. Choy, C.; Gwak, J.; Savarese, S. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
11. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
12. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5100–5111.
13. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal Image-Text Representation Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 104–120.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 8748–8763.

15. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H.S. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 18155–18165.
16. Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. CRIS: CLIP-Driven Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 11686–11695.
17. Kim, N.; Kim, D.; Lan, C.; Zeng, W.; Kwak, S. ReSTR: Convolution-Free Referring Image Segmentation Using Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 18145–18154.
18. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
19. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.
20. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* **2019**, *38*, 1–12. [[CrossRef](#)]
21. Thomas, H.; Qi, C.R.; Deschard, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420.
22. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.S.; Koltun, V. Point Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.
23. Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; Leibe, B. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 8216–8223.
24. Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; Liu, T.-L. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual Event, 2–9 February 2021; Volume 35, pp. 1610–1618. [[CrossRef](#)]
25. Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Li, Z.; Cui, S. InstanceRefer: Cooperative Holistic Understanding for Visual Grounding on Point Clouds through Instance Multi-Level Contextual Referring. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1791–1800.
26. Qian, Z.; Ma, Y.; Ji, J.; Sun, X. X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4551–4559.
27. Wu, C.; Ma, Y.; Chen, Q.; Wang, H.; Luo, G.; Ji, J.; Sun, X. 3D-STMN: Dependency-Driven Superpoint-Text Matching Network for End-to-End 3D Referring Expression Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5940–5948.
28. He, S.; Ding, H. RefMask3D: Language-Guided Transformer for 3D Referring Segmentation. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), Melbourne, Australia, 28 October–1 November 2024; pp. 8316–8325.
29. He, S.; Ding, H.; Jiang, X.; Wen, B. SegPoint: Segment Any Point Cloud via Large Language Model. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 349–367.
30. Wu, C.; Liu, Y.; Ji, J.; Ma, Y.; Wang, H.; Luo, G.; Ding, H.; Sun, X.; Ji, R. 3D-GRES: Generalized 3D Referring Expression Segmentation. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), Melbourne, Australia, 28 October–1 November 2024.
31. Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; Wang, H. Point Mamba: A Novel Point Cloud Backbone Based on State Space Model with Octree-Based Ordering Strategy. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
32. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [[CrossRef](#)]
33. Liang, D.; Zhou, X.; Wang, X.; Zhu, X.; Xu, W.; Zou, Z.; Ye, X.; Bai, X. PointMamba: A Simple State Space Model for Point Cloud Analysis. In Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024.

34. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
35. Zhang, Y.; Gong, Z.; Chang, A.X. Multi3DRefer: Grounding Text Description to Multiple 3D Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 15225–15236.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.