MR-COSMO: Visual-Text Memory Recall and Direct CrOSs-MOdal Alignment Method for Query-Driven 3D Segmentation

Chade Li[1,2], **Pengju Zhang**[1*], **Yihong Wu**[1,2*]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.
lichade2021@ia.ac.cn, {pengju.zhang@ia, yhwu@nlpr.ia}.ac.cn

Abstract

The rapid advancement of vision-language models (VLMs) in 3D domains has accelerated research in text-query-guided point cloud processing, though existing methods underperform in point-level segmentation due to inadequate 3D-text alignment that limits local feature-text context linking. To address this limitation, we propose **MR-COSMO**, a Visual-Text **M**emory **R**ecall and Direct **CrOS**s-**MO**dal Alignment Method for Query-Driven 3D Segmentation, establishing explicit alignment between 3D point clouds and text/2D image data through a dedicated direct cross-modal alignment module while implementing a visual-text memory module with specialized feature banks. This direct alignment mechanism enables precise fusion of geometric and semantic features, while the memory module employs specialized banks storing text features, visual features, and their correspondence mappings to dynamically enhance scene-specific representations via attention-based knowledge recall. Comprehensive experiments across 3D instruction, reference, and semantic segmentation benchmarks confirm state-of-the-art performance.

Extended version — https://arxiv.org/abs/2506.20991

Introduction

The emergence of large language models (LLMs) and 2D visual foundation models (VFMs) has propelled text-guided 3D segmentation to the forefront due to its significant practical implications. This technology aims to segment 3D objects or scenes using natural language inputs. Existing methods (Zhang et al. 2022; Liu et al. 2023; Qi et al. 2024; He et al. 2024; Zhen et al. 2024; Umam et al. 2024) typically employ LLMs to interpret textual inputs and leverage the CLIP model (Radford et al. 2021) to establish cross-modal associations between 2D images and text. Alternatively, these approaches utilize VFMs to process 2D visual data, relying on camera intrinsic and extrinsic parameters to establish geometric correspondences between 2D projections and 3D coordinates, thereby indirectly inferring 3D point cloud semantics through multi-view fusion.

PointCLIP (Zhang et al. 2022) extends the understanding of 3D data by utilizing CLIP's (Radford et al. 2021) framework to construct alignment between 2D images and text.

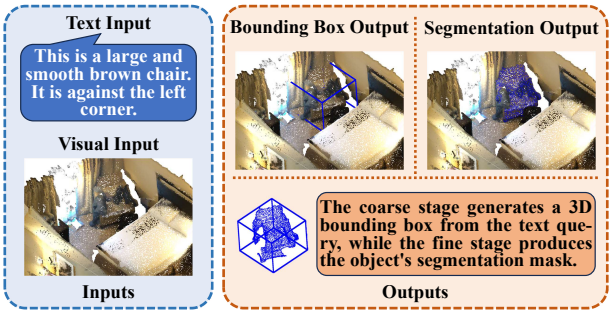


Figure 1: The inputs and outputs of the proposed coarse-to-fine query-driven 3D segmentation model.

The CLIP (Radford et al. 2021) trains an image encoder and a text encoder through contrastive learning, ensuring that matching image-text pairs converge in the embedding space while non-matching pairs diverge. Seal (Liu et al. 2023) employs VFMs for automotive point cloud segmentation, distilling semantic perceptions from VFMs to point clouds via a hyperpixel-driven contrastive learning approach on camera views. However, fine-grained segmentation requires recognizing subtle structural variations within objects, demanding a profound understanding of 3D geometries alongside the capability to capture local details and textual context. The aforementioned methods (Zhang et al. 2022; Liu et al. 2023) rely on indirect alignment strategies that use 2D images as intermediaries between 3D point clouds and other modalities. This approach is highly susceptible to errors in the computation of intrinsic/extrinsic parameters and to pixel-point misalignment artifacts. Consequently, existing methods fail to establish stable, accurate coordinate correspondence between 3D point clouds and 2D images, thus lacking the capability required for point-level fine-grained tasks.

To address these limitations, we propose MR-COSMO, a coarse-to-fine query-driven 3D segmentation model featuring a visual-text memory recall module and direct cross-modal alignment module. Figure 1 illustrates the inputs and outputs of the proposed coarse-to-fine model. Our framework first leverages cross-modal mamba-based attention to achieve simultaneous direct alignment between 3D features and both text/2D features, further constrained by contrastive learning applied to the 2D-text pairs. The aligned

*Corresponding authors

visual features are then updated via a multilayer multiscale Transformer block and processed through a detection header to generate a 3D bounding box for the input text query. Additionally, we introduce a Memory Module that stores confidence-weighted text-visual feature pairs derived from bounding box contents. This enables effective utilization of prior knowledge during subsequent scene processing, compensating for segmentation inconsistencies across scenarios. Finally, text queries and updated point features within the 3D bounding box are processed through the Memory Module, with segmentation masks generated by an iterative binary classifier. Experiments on diverse indoor and outdoor datasets demonstrate that our method has superior performance over existing methods across multiple downstream tasks. The main contributions of our work can be summarized as follows:

- We propose an innovative coarse-to-fine query-driven architecture that first generates coarse 3D object proposals and then performs query-specific fine segmentation, resolving boundary ambiguity in complex scenes.
- We introduce Direct Cross-Modal Alignment establishing 3D-text/3D-2D correspondence with Mamba-based attention, enhanced by contrastive learning constraints to overcome camera geometry dependencies.
- We develop a Memory Module that stores weighted visual-text feature pairs and dynamically recalls prior knowledge via attention mechanisms during inference, ensuring cross-scenario segmentation consistency.
- Our method demonstrates state-of-the-art performance across diverse 3D segmentation tasks, confirming robustness under varying textual queries and scene complexities through comprehensive experiments.

Related Work

Deep Learning on Point Cloud Segmentation

Recent deep learning advances have significantly progressed point cloud segmentation through novel network architectures and feature learning. By methodology and data format, mainstream approaches include: voxel-based methods, projection-based methods, 2D processing guided methods, MLP-based methods, point convolution-based methods, and attention-based methods.

Specifically, voxel-based methods (Graham, Engelcke, and Van Der Maaten 2018; Choy, Gwak, and Savarese 2019) discretize 3D space into regular grids for 3D convolution easily but face resolution-efficiency tradeoffs. Projection-based methods (Wu et al. 2018; Zhang et al. 2020) convert points to 2D representations leveraging established CNNs. 2D processing guided methods (Dai and Nießner 2018; Puy, Boulch, and Marlet 2023) adapt 2D networks/VFMs to reduce training costs at the expense of accuracy from domain gaps. MLP-based methods (Qi et al. 2017a,b) operate directly on raw point clouds, effectively enlarging receptive fields with constrained computation. Point convolution-based methods (Li et al. 2018; Thomas et al. 2019) excel at capturing fine-grained geometry via deformable kernels, yet struggles with efficiency and parameter sensitivity at scale. Recent attention-based advances (Zhao et al. 2021; Guo et al. 2021; Lai et al. 2022; Wu et al. 2022; Robert, Raguet, and Landrieu 2023; Lai et al. 2023; Wu et al. 2024b; Li et al. 2025) via Transformer and Mamba networks further improve segmentation accuracy.

Based on the current application of deep learning in point cloud segmentation, we still choose the attention network based on Transformer as the backbone.

Text-Guided Query-Driven 3D Segmentation

Text-guided query-driven 3D segmentation enables natural language-based manipulation and segmentation of 3D objects/scenes, essential for autonomous driving and embodied intelligence requiring precise scene understanding. By leveraging multimodal inputs (text/images/point clouds), existing methods (Achlioptas et al. 2020; Chen, Chang, and Nießner 2020; Huang et al. 2021; Jain et al. 2022; Zhang, Gong, and Chang 2023; Wu et al. 2023, 2024a; He and Ding 2024; Qian et al. 2024) bridge linguistic semantics with 3D geometry, enhancing accuracy in complex environments.

Specifically, text-guided 3D segmentation encompasses referring and instruction tasks. Referring segmentation outputs masks for objects specified by explicit category words. Instruction segmentation (He et al. 2024) processes functional descriptions without category words. Both require textual comprehension and world knowledge reasoning, yet accuracy lags semantic segmentation, with direct 3D-data alignment still limited.

In this regard, we design a Direct Cross-Modal Alignment module to establish direct correspondence between 3D features and other modalities, addressing errors from parametric computation and inaccurate 2D-3D mappings. Complementarily, we propose a Memory Module that leverages text-visual mapping knowledge to enhance correspondence sensitivity and segmentation accuracy.

Methodology

The framework of our MR-COSMO, featuring coarse-to-fine query-driven segmentation with Direct Cross-Modal Alignment and Memory Module, is shown in Figure 2. We process point cloud, corresponding 2D images, and text query as inputs for each individual query-driven segmentation scenario. For point clouds, we employ dual feature extraction: per-point features are extracted via MLP, while voxelized point clouds are processed through a 4-layer 3D transformer with window-shifting to obtain voxel features. For 2D images, visual features are extracted using a pretrained ResNet-50. Text features are generated with LLaMA2-7B (Touvron et al. 2023) after vectorizing queries. Finally, our Direct Cross-Modal Alignment module integrates these multimodal features.

After obtaining the aligned features, the features continue to be updated through a multi-layer multi-scale Transformer network. Subsequently, we obtain the bounding box of the target object corresponding to the text query in 3D space through a detection header. The 3D point features within the detected bounding boxes and their associated text query features are jointly input into the proposed Memory Module.

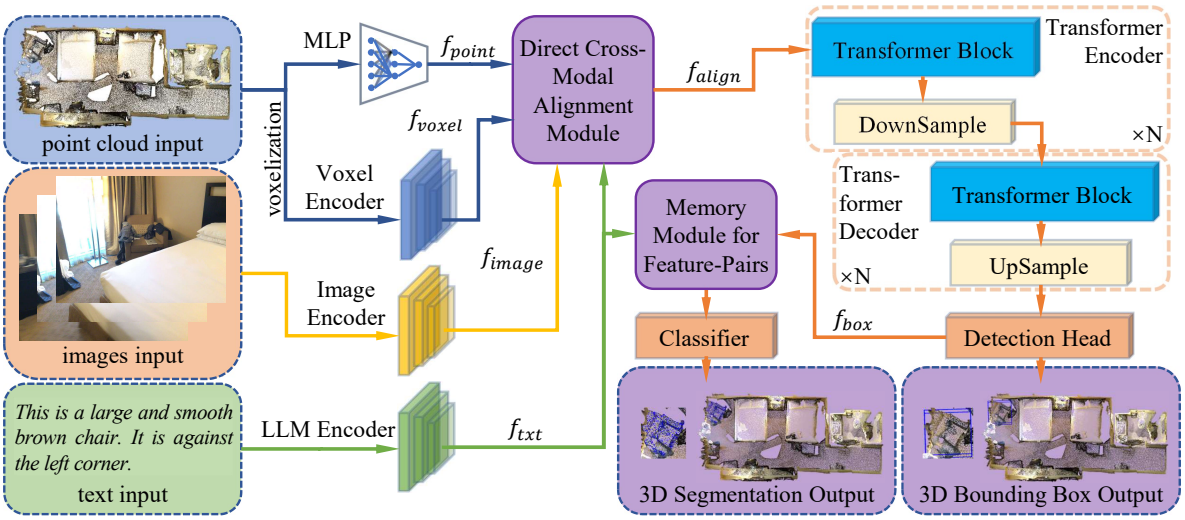


Figure 2: A general overview of the proposed network. Given point cloud, images, and text input, the point cloud is first voxelized. Then we use MLP, voxel encoder, image encoder, and LLM encoder to extract four distinct feature representations: f_{point}, f_{voxel}, f_{image}, and f_{txt}. These heterogeneous features are then unified via our novel Direct Cross-Modal Alignment (DCMA) module to generate aligned features f_{align}. Subsequently, the aligned features undergo refinement through a multi-layer Transformer encoder-decoder architecture. Following this, a detection head produces bounding box predictions and extracts point-wise features f_{box} within each detected region. The f_{box} and f_{txt} are jointly fed into the proposed Memory Module, which leverages stored cross-modal mappings as prior knowledge. Finally, an additional classifier processes the fused features to yield query-driven segmentation results.

Leveraging high-confidence text-visual feature pairs stored in the Memory Module, we iteratively train a binary classifier through an extra loss function. This enables progressive refinement from coarse bounding boxes to precise segmentation masks, establishing our coarse-to-fine framework, as illustrated in Figure 1.

Direct Cross-Modal Alignment

The proposed Direct Cross-Modal Alignment (DCMA) module, illustrated in Figure 3(a), architecturally consists of two constituent blocks: the Alignment Constrains Block and the Bidirectional Direct Alignment Block. The Alignment Constrains Block enforces contrastive learning constraints between text features and 2D image features, which are directly aligned with 3D features. The Bidirectional Direct Alignment Block implements a novel bidirectional Mamba-based cross-modal attention mechanism to establish direct alignment between 3D features and text/2D image features.

Alignment Constrains Block In order to make the subsequent 3D features have the correct correspondence with other modal features in the alignment process, we first process image features f_{imgae} and text features f_{txt} through independent encoders for feature transformation and mapping. This processing enforces matched image-text feature pairs to converge in the embedding space, as shown in Equation 1:

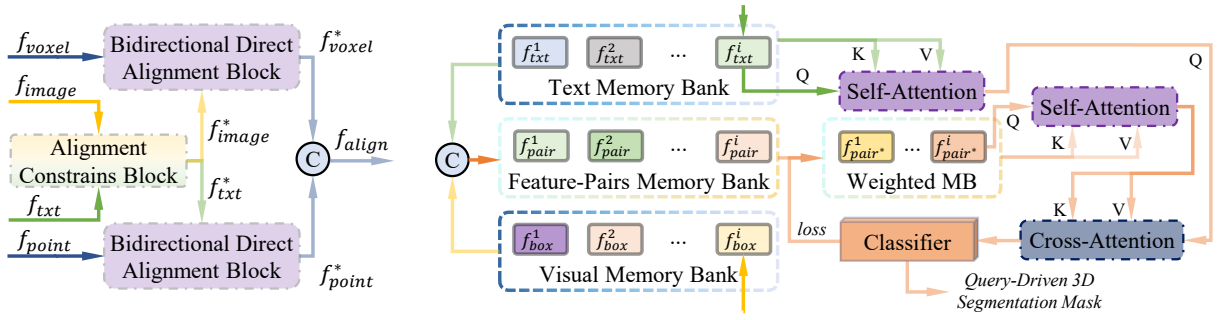
$$f^*_{image}, f^*_{txt} = \text{Encoders}_{ind}(f_{imgae}, f_{txt}). \quad (1)$$

For the aforementioned independent image and text encoders, we implement the following training procedure: For each scene in the experimental dataset, we select at most one matching image-text pair, process them through independent encoders to obtain remapped features, and compute cosine similarities between all image-text features to construct a similarity matrix. We then integrate a symmetric cross-entropy loss to maximize similarity scores for correct matching pairs (diagonal elements) while minimizing incorrect pair scores (off-diagonal elements). This contrastive strategy enforces precise alignment between 2D visual and textual representations, mapping corresponding features adjacently in high-dimensional embedding space to constrain subsequent alignment processes.

Bidirectional Direct Alignment Block After the Alignment Constraints Block, the Bidirectional Direct Alignment Block independently aligns remapped text features f^*_{txt} with point features f_{point}, and remapped image features f^*_{image} with voxel features f_{voxel}. This modality pairing offers two key benefits: aligning text directly with points avoids the pixel-to-point misalignment inherent in 2D-3D projection, while aligning image features with voxelized point clouds exploits the regular voxel structure to reduce geometric distortion.

For text-point alignment, given the remapped text features $f^*_{txt} \in \mathbb{R}^{1\times n_d}$ and the neighboring points features $f^n_{points} = [f^n_{point}; f^{n_1}_{point}; \ldots; f^{n_k}_{point}] \in \mathbb{R}^{(n_{nbr}+1)\times n_d}$ for the n-th point with k neighbors, we project both text and neighboring points features into a shared high-dimensional space of dimension D:

$$\phi_{txt} = f^*_{txt} W_{txt}, \quad (2)$$



(a) Framework of the proposed Direct Cross-Modal Alignment module.

(b) Framework of the proposed Memory Module for feature pairs.

Figure 3: Visualization of two important modules in the proposed MR-COSMO.

$$\phi_{points} = \text{MeanPool}\left(f^n_{points} W_{point}\right), \tag{3}$$

where $W_{txt} \in \mathbb{R}^{n_d \times D}$, $W_{point} \in \mathbb{R}^{n_d \times D}$, $\phi_{txt} \in \mathbb{R}^D$ is the projected text feature, $\phi_{points} \in \mathbb{R}^D$ is the aggregated point feature. And MeanPool denotes average pooling along the point dimension, we apply this operation to align the projected points features with the projected text features. We then construct a three-element sequence containing the projected text feature ϕ_{txt}, the aggregated point feature ϕ_{points}, and a copy of the text feature $\phi_{txt^{copy}}$:

$$X = \begin{bmatrix} \phi_{txt} \\ \phi_{points} \\ \phi_{txt^{copy}} \end{bmatrix} \in \mathbb{R}^{3 \times D}, \tag{4}$$

where sequence X is processed through bidirectional state space models. The forward state space model processes the sequence from top to bottom ($\phi_{txt} \rightarrow \phi_{points} \rightarrow \phi_{txt^{copy}}$):

$$h^f_t = \tilde{A}_f h^f_{t-1} + \tilde{B}_f X_t \quad \text{for} \quad t = 1, 2, 3, \tag{5}$$

$$\psi^f_t = \tilde{C}_f h^f_t + \tilde{D}_f X_t \quad \text{where} \quad h^f_0 = \mathbf{0} \in \mathbb{R}^D, \tag{6}$$

The superscripts/subscripts f and b denote forward and backward state-space processing, respectively. The backward model mirrors the forward one but operates on the reversed sequence ($\phi_{txt^{copy}} \rightarrow \phi_{points} \rightarrow \phi_{txt}$). The matrices $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D} \in \mathbb{R}^{D \times D}$ are learnable parameters. The forward and backward outputs at each position are $\psi^f_t, \psi^b_t \in \mathbb{R}^D$. Specifically, ψ^f_1 and ψ^b_3 encode pure text features after initial semantic processing; ψ^f_2 and ψ^b_2 represent text-guided point features; and ψ^f_3 and ψ^b_1 capture text refined through cross-modal interaction. The final representation is obtained by summing the terminal forward/backward outputs ($\psi^f_3 + \psi^b_1$) followed by Layer Normalization:

$$\psi^*_{point} = \text{LayerNorm}(\psi^f_3 + \psi^b_1), \tag{7}$$

and finally we get the f^*_{point} by applying an additional projection layer to project ψ^*_{point} back to the dimension $\mathbb{R}^{(n_{nbr}+1) \times n_d}$ consistent with the input.

Bidirectional direct alignment between voxel features f_{voxel} and remapped image features f^*_{image} is similar to the above process, getting voxel features f^*_{voxel} that align the image input. Subsequently, the features of the corresponding voxel for each 3D point are concatenated with the point features according to the correspondence between the 3D point and the 3D voxel to obtain the alignment features f_{align} of each 3D point.

Memory Module for Feature Pairs

Current datasets frequently exhibit imbalanced training sample distributions and inherent intra-class divergence in texture and contextual features across same-category objects. These issues result in the misclassification of objects within the same category and reduced accuracy for categories with a low number of samples. To address these challenges, we propose a Memory Module integrated into the segmentation result generation pipeline, as shown in Figure 3(b).

This module stores high-confidence (low-loss) visual-textual feature pairs to constrain the binary classifier generating segmentation results. The module receives text features f^i_{txt} and detection head-processed 3D point features f^i_{box} within bounding boxes as inputs. These features are stored in dedicated text ($\mathcal{M}_t$) and visual ($\mathcal{M}_v$) memory banks alongside prior scene features. We then concatenate corresponding elements to form feature-pairs memory bank:

$$\mathcal{M}_p = \{[f^i_{txt}; f^i_{box}] \mid i = 1, \ldots, N\}. \tag{8}$$

Each pair is assigned an initial weight $w^{(0)}_i = 1.0$. To address weighting bias, we introduce both initial and bias weighting. The initial weight is computed from mask loss:

$$w^{(\text{init})}_i = \frac{1}{\mathcal{L}_{BCE_i} + \tau}, \tag{9}$$

and all samples within the same category C are normalized whenever a new sample is added. This produces the final category-balanced weight:

$$w_i = \frac{1}{\mathcal{L}_{BCE_i} + \tau} \cdot \frac{1}{\sum_{j \in C} \frac{1}{\mathcal{L}_{BCE_j} + \tau}}, \tag{10}$$

ensuring each sample's contribution equals its initial weight divided by the category's total initial weights. The weighted feature-pairs are then stored as:

$$\widetilde{\mathcal{M}}_p = \{w_i \cdot [f^i_{txt}; f^i_{box}]\}. \tag{11}$$

During each new scene, the model retrieves stored knowledge through three attention steps: (1) text self-attention takes $f^{current}_{txt}$ as Query and $\mathcal{M}_t$ as Key/Value; (2) feature-pairs self-attention uses $[f^{current}_{txt}; f^{current}_{box}]$ as Query and $\widetilde{\mathcal{M}}_p$ as Key/Value; and (3) cross-attention takes the text self-attention output as Query and the feature-pairs output as Key/Value. The cross-attention result is passed to the binary classifier to generate masks, and its BCE loss updates the current feature-pair's weight following Equation 10, enabling dynamic weight refinement that optimizes memory influence and classifier training.

Object Function

The overall network training objective consists of two main components: (1) the combined detection loss $\mathcal{L}det$ and segmentation loss $\mathcal{L}seg$ for full-network optimization (Equation 12), where $\mathcal{L}det$ includes Smooth L1 (Equation 15) and weighted cross-entropy (Equation 16) losses, and $\mathcal{L}seg$ uses binary cross-entropy (Equation 17); and (2) the symmetric cross-entropy loss $\mathcal{L}_{DCMA}$ for training the Direct Cross-Modal Alignment module (Equation 14). Together, these form the complete objective (Equation 13):

$$\begin{aligned}\mathcal{L}_{task} &= \alpha\mathcal{L}_{det} + \beta\mathcal{L}_{seg} \\ &= \alpha\mathcal{L}_{smoothL1} + \alpha\mathcal{L}_{WCE} + \beta\mathcal{L}_{BCE},\end{aligned} \tag{12}$$

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \mathcal{L}_{DCMA}, \tag{13}$$

$$\begin{aligned}\mathcal{L}_{DCMA} &= \mathcal{L}_{SCE} \\ &= \gamma \cdot \left(-\sum_{i=1}^{N} y_i \log(p^{sce}_i)\right) + \\ &\delta \cdot \left(-\sum_{i=1}^{N} p^{sce}_i \log(y_i)\right),\end{aligned} \tag{14}$$

$$\mathcal{L}_{smoothL1}(y, \hat{y}) = \begin{cases} 0.5(\hat{y} - y)^2/\epsilon, & \text{if } |\hat{y} - y| < \epsilon \\ |\hat{y} - y| - 0.5\epsilon, & \text{otherwise} \end{cases}, \tag{15}$$

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \log(p^{wce}_{i,c}), \tag{16}$$

$$\begin{aligned}&\mathcal{L}_{BCE} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p^{bce}_i) + (1 - y_i)\log(1 - p^{bce}_i)\right],\end{aligned} \tag{17}$$

where y_i is the true label corresponding to the ith sample (where $y_i = 1$ indicates the sample belongs to a specific class, and 0 otherwise), p^{sce}_i is the predicted probability output by the model for the ith sample, y is the true value, $\hat{y}$ is the predicted value, ϵ is a smoothing threshold parameter, N is the number of samples, C is the total number of categories, $y_{i,c}$ is the true one-hot label of the ith sample, $p^{wce}_{i,c}$ is the predicted probability of the ith sample for category c, w_c is the weight of category c, and p^{bce}_i is the probability that the ith sample is predicted to be the positive class.

Method	Acc	mIoU
ReferIt3D (Achlioptas et al. 2020)	11.7	6.4
ScanRefer (Chen, Chang, and Nießner 2020)	12.0	6.9
TGNN (Huang et al. 2021)	12.9	7.1
BUTD-DETR (Jain et al. 2022)	16.3	10.9
EDA (Wu et al. 2023)	16.6	12.1
M3DRef (Zhang, Gong, and Chang 2023)	18.1	12.8
SegPoint (He et al. 2024)	31.6	<u>27.5</u>
Ours*	<u>31.9</u>	27.4
Ours	**33.8**	**28.5**

Table 1: 3D instruction segmentation results on Instruct3D (He et al. 2024). * denotes our method removing the proposed Memory Module.

Experiments

Experimental Settings

Datasets The proposed network is applicable to diverse point cloud segmentation downstream tasks, with experimental validation conducted on three key tasks: 3D instruction segmentation using Instruct3D (He et al. 2024) built upon ScanNet++ (Yeshwanth et al. 2023), 3D referring segmentation using ScanRefer (Chen, Chang, and Nießner 2020) based on ScanNet (Dai et al. 2017), and 3D semantic segmentation evaluated on both the indoor S3DIS (Armeni et al. 2016) and outdoor SemanticKITTI (Behley et al. 2019) datasets. Comprehensive dataset specifications are provided in the extended version of this paper.

Metrics Following established practices in 3D segmentation research, we adopt standard mean Intersection-over-Union (mIoU) for main evaluation. For the Instruct3D benchmark (He et al. 2024), we employ SegPoint's accuracy (Acc) metric, which defines a sample as correctly identified if IoU > 0.5 and calculates accuracy as the proportion of correctly identified samples relative to all samples. The best indicators are highlighted in **bold** and the next best indicators are <u>underlined</u>.

Implementation Details Experiments are implemented on a server equipped with four Nvidia V100 (32G $\times$ 4 GPU memory). Consistent with PTv3 (Wu et al. 2024b), we use the AdamW optimizer with a cosine scheduler during training and set 1% of the training process as a warm-up. We set the initial learning rate and weight decay to 0.005, 0.05 and 0.002, 0.005 for indoor and outdoor scenes, respectively. In addition, the number of training epochs was set to 500 and 100 for the indoor and outdoor scenes, accordingly.

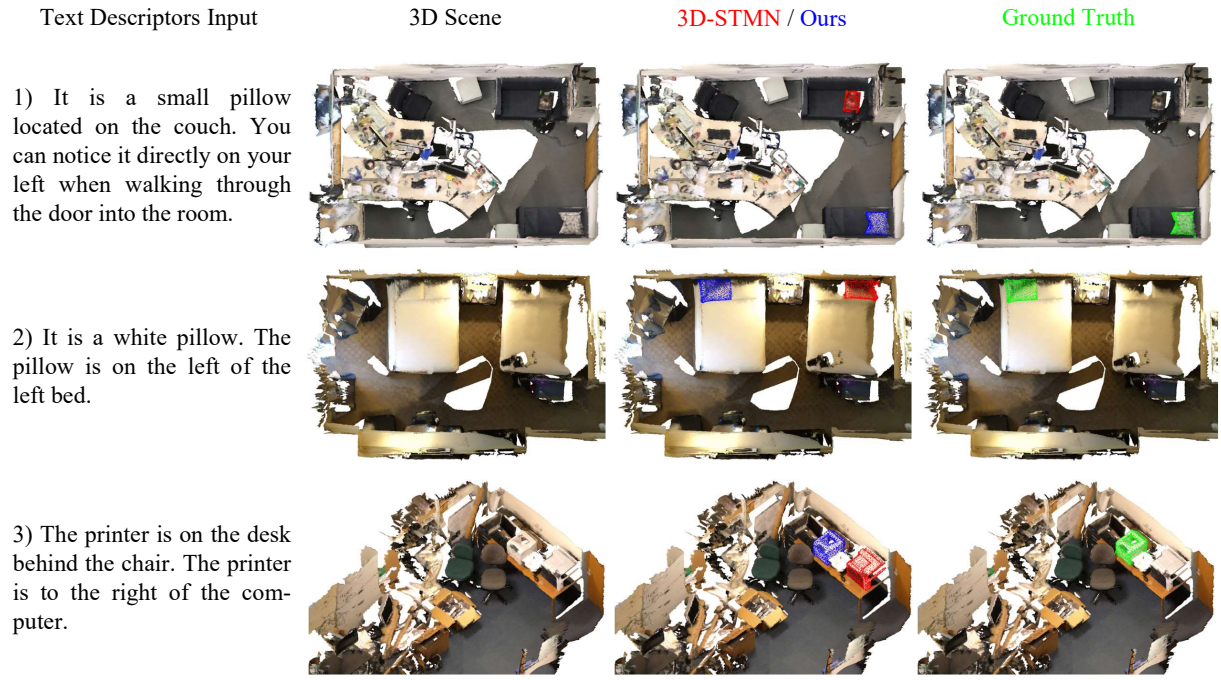


Figure 4: Qualitative results of 3D referring segmentation experiment on ScanRefer (Chen, Chang, and Nießner 2020). In the visualization results, the blue masks denote the segmentation outputs from the proposed MR-COSMO, the red masks represent the predictions of the 3D-STMN (Wu et al. 2024a), and the green masks indicate the ground truth annotations.

Method	mIoU
TGNN (Huang et al. 2021)	27.8
BUTD-DETR (Jain et al. 2022)	35.4
EDA (Wu et al. 2023)	36.2
M3DRef (Zhang, Gong, and Chang 2023)	35.7
X-RefSeg3D (Qian et al. 2024)	29.9
3D-STMN (Wu et al. 2024a)	39.5
RefMask3D (He and Ding 2024)	<u>44.8</u>
SegPoint (He et al. 2024)	41.7
Ours	**45.6**

Table 2: 3D referring segmentation results on ScanRefer (Chen, Chang, and Nießner 2020).

Experiments on 3D Instruction Segmentation

We present the performance on 3D instruction segmentation task of MR-COSMO and SOTA methods in Table 1. In the Instruct3D (He et al. 2024) experiment, MR-COSMO achieve excellent results, outperforming the SOTA method in both Acc and mIoU evaluation metrics.

Even when not incorporating the proposed Memory Module for storing high-confidence text-visual query pairs, our method has a competitive level of accuracy compared to the best existing method, SegPoint (He et al. 2024). Compare the last three rows of data in the table, our method without the Memory Module is 0.3% ahead of SegPoint (He et al. 2024) in Acc while only 0.1% behind in mIoU, and our full method is 2.2% and 1.0% ahead of SegPoint (He et al. 2024) in Acc and mIoU. These results show that our method has a strong text comprehension and reasoning ability.

Experiments on 3D Referring Segmentation

We demonstrate MR-COSMO's performance on 3D referring segmentation against SOTA methods in Table 2, achieving superior results on ScanRefer (Chen, Chang, and Nießner 2020) with at least 0.8% mIoU improvement. This demonstrates that our method not only has strong comprehension and inference ability, but also has text-visual association construction ability, which means that it can query the distribution of the corresponding objects in 3D space based on the explicit text words.

Moreover, we also conduct qualitative results of our method on the validation set of ScanRefer (Chen, Chang, and Nießner 2020), as shown in Figure 4. We select 3D-STMN (Wu et al. 2024a) as the comparison baseline because RefMask3D (He and Ding 2024) lacks pretrained models and SegPoint (He et al. 2024) has unreleased source code, making 3D-STMN the strongest fully available method for qualitative evaluation. It can be seen that our method has stronger individual discrimination and recognition ability when there are multiple similar objects of the target category in the same space. Unlike 3D-STMN's susceptibility to selecting non-target objects with complex text descriptions, our method maintains accurate target identification through the Memory Module's text-visual mapping knowledge.

Experiments on 3D Semantic Segmentation

We summarize the performance on 3D semantic segmentation task of MR-COSMO and SOTA methods in Table

Method	OA	mAcc	mIoU
PointNet (Qi et al. 2017a)	-	49.0	41.1
KPConv (Thomas et al. 2019)	-	72.8	67.1
PCT (Guo et al. 2021)	-	67.7	61.3
ST (Lai et al. 2022)	91.5	78.1	71.0
SPT (Robert, Raguet, and Landrieu 2023)	89.5	77.3	68.9
PTv2 (Wu et al. 2022)	91.6	78.0	72.6
PTv3 (Wu et al. 2024b)	-	-	73.4
PTv3+PPT (Wu et al. 2024b,c)	92.0	80.1	74.7
Ours	**92.8**	**84.3**	**75.6**

Table 3: Indoor 3D semantic segmentation results on S3DIS (Armeni et al. 2016).

Method	mIoU
SphereFormer (Lai et al. 2023)	67.8
WaffleIron (Puy, Boulch, and Marlet 2023)	68.0
2DPASS (Yan et al. 2022)	69.3
PTv2 (Wu et al. 2022)	70.3
OA-CNNs (Peng et al. 2024)	70.6
PPT+SparseUNet (Wu et al. 2024c)	71.4
PTv3+PPT (Wu et al. 2024b,c)	72.3
Ours	**73.4**

Table 4: Outdoor 3D semantic segmentation results on the validation set of SemanticKITTI (Behley et al. 2019).

3 and Table 4. Experiments on the S3DIS (Armeni et al. 2016) Area 5 (Table 3) show that our method outperforms the method (Wu et al. 2024b,c) that using multiple datasets as training data in terms of all three overall accuracy metrics. Experiments on the SemanticKITTI validation set (Table 4) show that MR-COSMO has an accuracy improvement of at least 1.1% in mIoU compared to existing methods. This can prove that allowing the 3D segmentation network to perceive the category information in the scene in advance during the feature processing stage helps its segmentation accuracy to be improved. For a fair comparison, our method is trained using only the data in the corresponding dataset in our experiments, and no data from additional datasets are added to augment the training samples. Qualitative results of our method on the SemanticKITTI (Behley et al. 2019) validation set can be found in the extended version of this paper.

Ablation Experiments

Compare the last two rows of data shown in Table 1, the method with the proposed Memory Module has an accuracy improvement of 1.9% and 1.1% in Acc and mIoU, respectively, which shows the importance of the proposed Memory Module in improving the network's ability to understand text-visual feature correspondences.

Table 5 summarizes the impact of the proposed Direct Cross-Modal Alignment and Memory Module for Feature Pairs on Instruct3D (He et al. 2024). The experimental results demonstrate that network accuracy achieves enhancements of 1.0%, 1.1%, and 2.1% under three conditions: (a) introducing the Direct Cross-Modal Alignment module, (b) deploying the Memory Module for Feature Pairs independently, and (c) combining both modules, respectively.The proposed method mitigates the negative impacts caused by error accumulation and 2D-3D coordinate mapping distortions in conventional alignment approaches through two strategic mechanisms: (1) explicitly bridging 3D features with multimodal representations via direct alignment, and (2) integrating text-2D feature correlations by contrastive learning as constraints to regulate the 3D feature alignment process. Meanwhile, the proposed Memory Module can further constrain the output process from 3D bounding boxes to segmentation results, and store high-confidence text-to-vision mapping relations to enhance the text-to-vision inference capability of the network. Combining both modules further improves accuracy in interactive segmentation tasks.

Method	mIoU	Δ
Baseline	26.4	+0.0/-2.1
w/ Direct Cross-Modal Alignment (DCMA)	27.4	+1.0/-1.1
w/ Memory Module for Feature Pairs (MMFP)	27.5	+1.1/-1.0
Only Use Voxel Encoder for Point Cloud	27.7	+1.3/-0.8
w/o Alignment Constrains Block	28.0	+1.6/-0.5
w/o Loss on Bounding Boxes	28.4	+2.0/-0.1
w/ all	28.5	+2.1/-0.0

Table 5: Ablation study of different proposed modules on Instruct3D (He et al. 2024).

In addition, we show that using both voxels and point MLP features for point cloud processing provides a 0.8% accuracy advantage over using voxels alone. We also demonstrate that applying contrastive learning for modal alignment improves the network's accuracy by 0.5%, highlighting the contribution of each module to overall performance. Finally, ablating the constraints on the detection frame results in only a 0.1% change in accuracy, which helps the network discriminate between individuals of the same class without unfairly affecting comparisons with other methods. More ablation studies on the effects of Mamba attention, backbone choices, and hyperparameter settings can be found in the extended version of this paper.

Conclusions

We propose MR-COSMO, a coarse-to-fine query-driven 3D segmentation model integrating Memory Recall and Direct Cross-Modal Alignment. This framework mitigates computational errors from intrinsic/extrinsic parameters through a Direct Cross-Modal Alignment module establishing explicit 3D feature alignment with text/image modalities. Complementarily, a Memory Module stores high-confidence text-visual feature mappings during training to leverage prior knowledge for enhancing segmentation accuracy when processing new scenes. Comprehensive experiments across diverse segmentation tasks demonstrate significant performance gains over comparative methods with consistent cross-scene generalization, confirming the method's effectiveness and robustness.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2024YFB4709100, the National Natural Science Foundation of China under Grant No. 62572468 and Beijing Natural Science Foundation under Grant No. L241012. We thank the anonymous Program Committees and Program Chairs so much for their helpful comments and suggestions.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 422–440.

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9297–9307.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 202–221.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2432–2443.

Dai, A.; and Nießner, M. 2018. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 452–468.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224–9232.

Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; and Hu, S.-M. 2021. Pct: Point cloud transformer. *Computational Visual Media*, 7: 187–199.

He, S.; and Ding, H. 2024. RefMask3D: Language-Guided Transformer for 3D Referring Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8316–8325.

He, S.; Ding, H.; Jiang, X.; and Wen, B. 2024. SegPoint: Segment Any Point Cloud via Large Language Model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 349–367.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. In *Thirty-fifth AAAI Conference on Artificial Intelligence*, 1610–1618.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 417–433.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical Transformer for LiDAR-Based 3D Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17545–17555.

Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; and Jia, J. 2022. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8500–8509.

Li, C.; Zhang, P.; Liu, B.; Wei, H.; and Wu, Y. 2025. FEAST-Mamba: FEAture and SpaTial Aware Mamba Network with Bidirectional Orthogonal Fusion for Cross-Modal Point Cloud Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4634–4642.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. Pointcnn: Convolution on x-transformed points. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 828–838.

Liu, Y.; Kong, L.; Cen, J.; Chen, R.; Zhang, W.; Pan, L.; Chen, K.; and Liu, Z. 2023. Segment Any Point Cloud Sequences by Distilling Vision Foundation Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36, 37193–37229.

Peng, B.; Wu, X.; Jiang, L.; Chen, Y.; Zhao, H.; Tian, Z.; and Jia, J. 2024. OA-CNNs: Omni-Adaptive Sparse CNNs for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21305–21315.

Puy, G.; Boulch, A.; and Marlet, R. 2023. Using a Waffle Iron for Automotive Point Cloud Semantic Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3356–3366.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5105–5114.

Qi, Z.; Fang, Y.; Sun, Z.; Wu, X.; Wu, T.; Wang, J.; Lin, D.; and Zhao, H. 2024. GPT4Point: A Unified Framework for Point-Language Understanding and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26417–26427.

Qian, Z.; Ma, Y.; Ji, J.; and Sun, X. 2024. X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured

Cross-Modal Graph Neural Networks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, 4551–4559.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Robert, D.; Raguet, H.; and Landrieu, L. 2023. Efficient 3D semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17195–17204.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. M.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A. S.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I. M.; Korenev, A. V.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M. H. M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Umam, A.; Yang, C.-K.; Chen, M.-H.; Chuang, J.-H.; and Lin, Y.-Y. 2024. PartDistill: 3D Shape Part Segmentation by Vision-Language Model Distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3470–3479.

Wu, B.; Wan, A.; Yue, X.; and Keutzer, K. 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE international conference on robotics and automation (ICRA)*, 1887–1893.

Wu, C.; Ma, Y.; Chen, Q.; Wang, H.; Luo, G.; Ji, J.; and Sun, X. 2024a. 3D-STMN: dependency-driven superpoint-text matching network for end-to-end 3D referring expression segmentation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, 5940–5948.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024b. Point Transformer V3: Simpler Faster Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4840–4851.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022. Point transformer v2: Grouped vector attention and partition-based pooling. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 33330–33342.

Wu, X.; Tian, Z.; Wen, X.; Peng, B.; Liu, X.; Yu, K.; and Zhao, H. 2024c. Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19551–19562.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2023. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19231–19242.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 677–695.

Yeshwanth, C.; Liu, Y.-C.; Nießner, M.; and Dai, A. 2023. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12–22.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. PointCLIP: Point Cloud Understanding by CLIP. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8542–8552.

Zhang, Y.; Gong, Z.; and Chang, A. X. 2023. Multi3DRefer: Grounding Text Description to Multiple 3D Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15225–15236.

Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; and Foroosh, H. 2020. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9601–9610.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; and Gan, C. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *Proceedings of the 41st International Conference on Machine Learning*, 61229–61245.