

# Masked Point-Entity Contrast for Open-Vocabulary 3D Scene Understanding

Yan Wang<sup>1\*</sup> Baoxiong Jia<sup>1\*</sup> Ziyu Zhu<sup>1,2</sup> Siyuan Huang<sup>1</sup>

<sup>1</sup>State Key Laboratory of General Artificial Intelligence, BIGAI

<sup>2</sup>Tsinghua University

<https://mpec-3d.github.io>

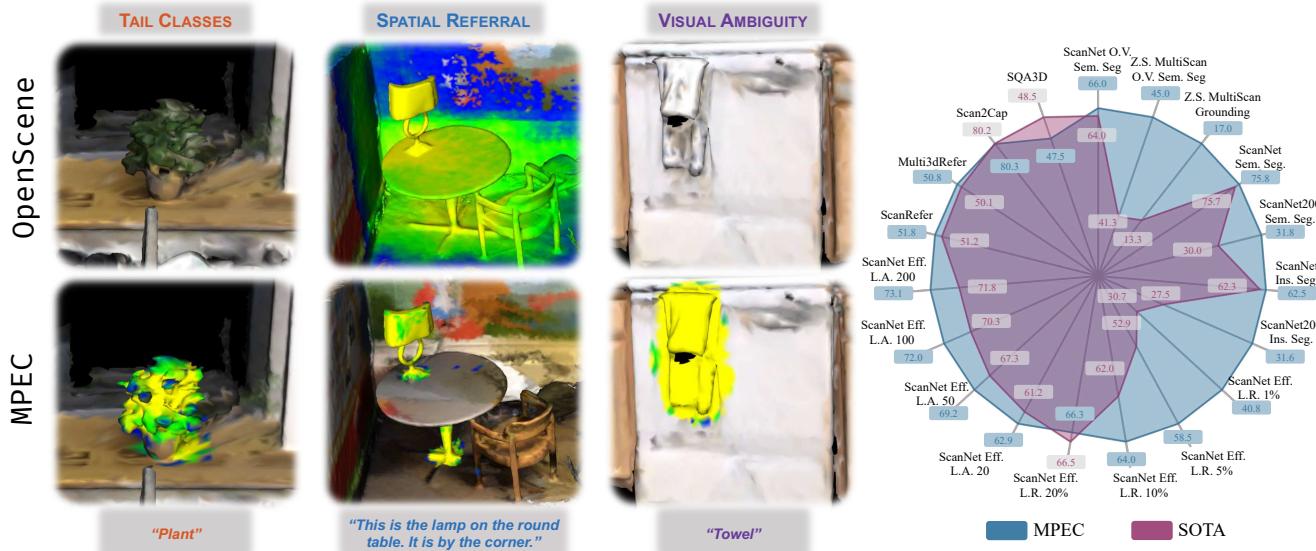


Figure 1. **Qualitative and Quantitative Analysis of MPEC Results.** MPEC achieves state-of-the-art on open-vocabulary 3D semantic segmentation. Compared with OpenScene [57], MPEC is more robust to tail classes, visual ambiguity, and detailed descriptions, e.g., spatial referral. The radar chart highlights the performance advantages of MPEC across various 3D scene understanding tasks.

## Abstract

Open-vocabulary 3D scene understanding is pivotal for enhancing physical intelligence, as it enables embodied agents to interpret and interact dynamically within real-world environments. This paper introduces **MPEC**, a novel Masked Point-Entity Contrastive learning method for open-vocabulary 3D semantic segmentation that leverages both 3D entity-language alignment and point-entity consistency across different point cloud views to foster entity-specific feature representations. MPEC improves semantic discrimination and enhances the differentiation of unique instances, achieving state-of-the-art results on ScanNet for open-vocabulary 3D semantic segmentation and demonstrating superior zero-shot scene understanding capabilities. Extensive fine-tuning experiments on 8 datasets, spanning from low-level perception to high-level reasoning tasks, showcase

the potential of learned 3D features, driving consistent performance gains across varied 3D scene understanding tasks.

## 1. Introduction

Open-vocabulary 3D scene understanding [57] is an emerging research topic with significant implications for embodied AI [2, 19, 27, 31, 55, 83, 91]. This task requires the model to act beyond recognizing only predefined 3D object categories and concepts seen during training, enabling it to identify flexible concepts based on instructions while preserving sufficient spatial or geometry information to support tasks like reconstruction [45, 46, 53, 54] and manipulation [23, 42, 91].

However, developing scene representations that possess both flexible semantic understanding capability and accurate spatial and geometric information presents non-trivial challenges. Due to the lack of available paired 3D vision-language (3D-VL) data, recent methods leverage

\* indicates equal contribution.

2D foundation models [21, 40, 41, 59] either to generate data [17, 38, 81] for 3D-language alignment training or to provide semantically aligned 2D representations from individual images which are then merged into global 3D representations [76, 82] or distilled into trainable 3D encoders [57]. However, such pipelines can struggle to preserve 3D spatial information, primarily due to the limited field of view in individual images, which hinders the capture of global spatial relationships between objects, and the challenges in maintaining multi-view consistency, especially by semantic representations rather than pixel-level information. Considering the intrinsic uniqueness of individual entities in 3D scenes, particularly for tasks requiring spatial reasoning in 3D-VL and embodied AI, we emphasize the need for representations that not only align with language but also distinguish between instances within the scene.

Previous methods for incorporating spatial and geometric information without supervision have largely focused on investigating the inherent correlations in point clouds, with much of the work centered on unsupervised and self-supervised representation learning [84], particularly via contrastive learning approaches [29, 67, 72, 74]. These methods typically create positive and negative point cloud pairs through various strategies to form the concept of neighbors [72], regions [29], and semantic prototypes [67]. However, none of these approaches are effectively suited for aligning with language in complex 3D scenes, as they tend to be either overly fragmented or coarse-grained and fail to capture the concept of objects within the scene.

In this paper, we address the aforementioned challenges by learning 3D representations that preserve both flexible semantic understanding and entity-specific geometric and spatial information. To achieve this, we propose MPEC, a simple yet effective entity-level contrastive framework for training open-vocabulary 3D scene understanding models. Our approach bridges efforts from open-vocabulary scene understanding and unsupervised scene representation learning [67, 72], while also leveraging recent advancements in 3D-VL scaling [35, 69, 80, 83]. Methodologically, we apply augmentations [72] to generate multiple views of a scene. We generate 3D object entity masks with off-the-shelf object models [62, 63], and form contrastive pairs between different views, enforcing feature similarity within the same entities across views and dissimilarity across different ones. Finally, we align 3D point features merged from different views with language generated by foundation models [35, 38, 81] on diverse scenes through contrastive alignment.

MPEC achieves state-of-the-art open-vocabulary 3D semantic segmentation results with 66.0% foreground mIoU and 81.3% foreground mAcc on ScanNet [14]. The trained 3D encoder also demonstrates strong generalization across various downstream tasks after fine-tuning, from low-level perception [22, 37] to high-level reasoning tasks [1, 6, 8, 48,

87]. Notably, MPEC especially boosts the performance of ScanNet Data Efficient Benchmark [29], *improving mIoU from 30.7% [67] to 40.8% with just 1% of the scenes trained*. We also provide an empirical analysis of the framework, breaking down each module’s contribution to the overall performance. In addition, we conduct extensive experiments to assess how different scene sources and language types affect performance. In summary, our contributions are:

- We present MPEC, a novel contrastive framework with both point-to-entity and entity-to-language alignments, achieving new state-of-the-arts in open-vocabulary 3D scene understanding tasks.
- Through extensive experiments on 7 tasks over 8 datasets, we demonstrate that MPEC achieves state-of-the-art results in both zero-shot and fine-tuned settings, particularly excelling in data-scarce scenarios.
- We analyze how model architectures and data sources impact model performance, highlighting cross-view contrast and data domain selection as key factors for contrastive learning, offering new insights for future research.

## 2. Related Work

**Open-Vocabulary 3D Scene Understanding** Open-vocabulary 3D scene understanding typically involves two signature tasks: open-vocabulary 3D semantic segmentation [7, 17, 38, 43, 57, 68, 81] and open-vocabulary 3D instance segmentation [34, 52, 63]. Both tasks aim to predict 3D segmentation masks for class labels or object instances using natural language descriptions, without requiring prior model training on specific object categories. Despite significant advancements in open-vocabulary understanding within the 2D domain [10, 21, 36, 41, 59, 71, 73, 94], achieving open-vocabulary 3D scene understanding from point clouds [17, 32, 35, 38, 57, 63, 77, 81] remains challenging due to the scarcity of large-scale training data.

Existing approaches leverage 2D foundation models [21, 40, 41, 59] to bridge the gap between language and 3D point cloud data. With off-the-shelf 3D segmentation models [47, 62] or back-projection from 2D foundation models [40, 94], existing methods learn point-language alignment for open-vocabulary semantic [17, 38, 57, 81] and instance segmentation [34, 63]. These methods align 3D point clouds and language either by directly distilling language and image features from 2D vision-language (2D-VL) models into 3D learnable backbones or by contrastively aligning point and language features, similar to approaches in 2D [59]. However, a key limitation of these models is the insufficient incorporation of 3D information during model learning, which restricts their performance on a wide range of 3D-VL tasks. For instance, in OpenScene [57], as illustrated in Fig. 1, the model struggles with 2D visual ambiguities (*e.g.*, objects with similar colors) and faces challenges in accurately identifying specific object instances. This is

sue arises primarily because the 2D-VL model used to train the 3D scene encoder was based on category-level text descriptions, which lack spatial and 3D geometric context, thereby limiting the model’s understanding of the 3D structure of objects. Therefore, we emphasize the crucial role of 3D information from the scene representation itself in open-vocabulary scene understanding tasks.

**Pre-training for 3D Point Clouds** Motivated by the successful application of the “pre-train then fine-tune” paradigm in 2D vision research, there has been growing interest in 3D point cloud representation learning [29, 33, 51, 65, 67, 70, 72, 74, 79, 84, 88], demonstrating positive impacts on downstream tasks. These approaches can be categorized into two paradigms: mask modeling and contrastive learning. Mask modeling methods [51, 65, 70, 79, 84, 88] focus on reconstructing the coordinates or features of masked points. In contrast, contrastive learning methods [29, 33, 67, 74] create positive and negative pairs within 3D point clouds for discriminative learning. To enhance the learning of contrastive-based methods, MSC [72] incorporates masked point modeling as additional reconstructive supervision. GroupContrast [67] builds on this by introducing semantic prototypes for discovering local neighborhoods. However, these methods still lack a concept of objects, making them inapplicable for aligning with language.

Meanwhile, recent research has shown a growing interest in vision-language pre-training for 3D point cloud understanding. Methods like 3D-VisTA [96], 3DVLP [86], and SynVL3D [78] employ unsupervised pre-training strategies such as masked modeling and contrastive learning [35], while other approaches utilize multi-task training for vision-language alignment [4, 9, 97]. With the emergence of 3D Large Language Model (3D-LLM) [26, 28, 39, 49, 89], recent efforts like LEO [31], MiniGPT-3D [64] and LLaVA-3D [95] have focused on aligning object-level or voxel-level features with LLMs through explicit alignment training stage [26, 44]. These studies demonstrate that improved vision-language alignment leads to enhanced performance on downstream tasks [20, 24, 30–32, 58]. However, a critical gap remains in the lack of a robust, open-vocabulary 3D encoder capable of directly processing scene-level point clouds [31, 32, 95, 97]. Our work addresses this need by introducing a powerful 3D encoder, which has the potential to further advance the capabilities of 3D-LLM.

### 3. Method

In this section, we elaborate on the proposed masked point-entity contrastive learning framework for open-vocabulary 3D scene understanding. The overall architecture is presented in Fig. 2. In essence, our model aims to learn 3D point features  $\mathbf{F}_P \in \mathbb{R}^{N \times D}$  that maintain 3D information of 3D scenes, a vision-language (VL-) adapter  $\text{Proj}_{\text{VL}}(\cdot)$  that maps

these 3D features into the semantic space which is aligned with language features  $\mathbf{F}_T$  extracted from a pre-trained text encoder. The input of our framework mainly consists of three components: 1) 3D scene point cloud  $\mathcal{P} = (\mathbf{P}, \mathbf{C})$ , where  $\mathbf{P} \in \mathbb{R}^{N \times 3}$  and  $\mathbf{C} \in \mathbb{R}^{N \times 3}$  represents spatial coordinates and colors of the points. 2)  $K$  3D entity mask proposals obtained with off-the-shelf methods which are merged into  $\mathbf{M} \in \{\text{BG}, 1, \dots, K\}^N$  where BG denotes the background. 3) Text descriptions  $\mathbf{T} = \{t_i\}_{i=1}^{N_T}$  including captions and referrals, which could be generated for each entity proposal by template-based methods or foundation models.

#### 3.1. Point-to-Entity Alignment

**Cross-view Feature Extraction** We start by extracting separate point features from different augmented views of the scene for contrastive representation learning. Specifically, given a 3D point cloud  $\mathcal{P}$ , two distinct augmented views  $\mathcal{P}_u$  and  $\mathcal{P}_v$  are generated, along with their transformed entity masks  $\mathbf{M}_u$  and  $\mathbf{M}_v$ . Following the practice of previous works [25, 72, 75], we divide the unioned set of point clouds into non-overlapping grids based on the original spatial locations. Then two sets of non-overlapping grid masks  $\mathbf{G}_u$  and  $\mathbf{G}_v$  are randomly generated and applied to the corresponding views. We use the same mask token embedding  $\mathbf{t} \in \mathbb{R}^3$  for the color of all masked points. We then feed the masked point clouds of separate views into the 3D U-Net encoder  $\mathcal{E}_{\text{pcd}}$  to extract per-point embeddings:

$$\mathbf{f}_l = \mathcal{E}_{\text{pcd}}((1 - \mathbf{G}_l)\mathcal{P}_l + \mathbf{G}_l\mathbf{T}_l), \quad l \in \{u, v\}, \quad (1)$$

where  $\mathbf{T}_u$  and  $\mathbf{T}_v$  are the expansions of the learnable embedding vector to fit the dimension with point cloud across different views, respectively.

**Point-to-Entity Contrastive Learning** After obtaining the point features from two views, we perform point-to-entity contrastive learning to incorporate entity information into our learned scene representations. Specifically, we calculate the cosine similarity matrix  $\mathbf{s}^{u \rightarrow v}$  between all pairs of matched points by computing the similarities of point  $i$  from view  $u$  to points in view  $v$  belonging to entities and background separately. Let  $d_{\cos}(\mathbf{f}_1, \mathbf{f}_2) = \frac{\mathbf{f}_1^T \cdot \mathbf{f}_2}{\|\mathbf{f}_1\| \cdot \|\mathbf{f}_2\|}$  to denote the cosine similarity function, for points  $\mathbf{E}_k^v = \{j \mid \mathbf{M}_v[j] = e_k\}$  that belong to entity  $e_k$  in view  $v$ , we have:

$$\mathbf{s}_{i,e_k}^{u \rightarrow v} = \frac{1}{|\mathbf{E}_k^v|} \sum_{j \in \mathbf{E}_k^v} \mathbf{s}_{i,j}^{u \rightarrow v} = \frac{1}{|\mathbf{E}_k^v|} \sum_{j \in \mathbf{E}_k^v} d_{\cos}(\mathbf{F}_u^i, \mathbf{F}_v^j). \quad (2)$$

Meanwhile, for the remaining  $L$  background points  $\{b_1, b_2, \dots, b_L\}$  in view  $v$ , we compute the similarity between point  $i$  to the each  $b_j$  in view  $v$  with cosine similarity  $\mathbf{s}_{i,b_j}^{u \rightarrow v} = d_{\cos}(\mathbf{F}_u^i, \mathbf{F}_v^{b_j})$ . After these calculations, we obtain the entity-aware point similarities  $\mathbf{s}_i^{u \rightarrow v}$  for point  $i$  by

$$\mathbf{s}_i^{u \rightarrow v} = \text{Concat} \left( \mathbf{s}_{i,e_1}^{u \rightarrow v}, \dots, \mathbf{s}_{i,e_K}^{u \rightarrow v}, \mathbf{s}_{i,b_1}^{u \rightarrow v}, \dots, \mathbf{s}_{i,b_L}^{u \rightarrow v} \right), \quad (3)$$

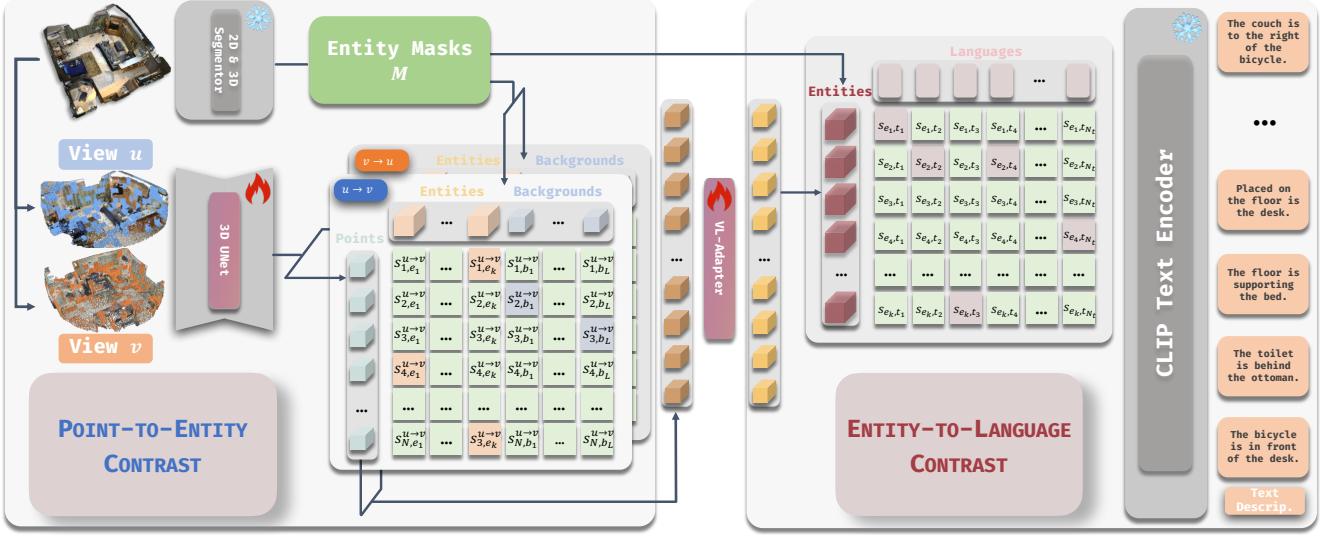


Figure 2. **The Overall Pipeline of MPEC.** Given a 3D point cloud as input, we predict entity mask proposals and generate text descriptions for each entity. Then different views of the scene are randomly masked and we replace the masked point features with learnable embeddings. A 3D UNet then extracts per-point features. Guided by the entity masks, cross-view point-to-entity contrastive learning is conducted to enforce cross-view consistency and distinguishment across different entities. Then we merge the point features from two views and align the feature dimension of points with text features extracted by CLIP with a VL-Adapter. Finally, entity-to-language contrastive learning is performed for open-vocabulary 3D scene understanding.

and add point-entity contrastive supervision depending on whether point  $i$  belongs to an entity or background in view  $u$ . If point  $i$  belongs to an entity, then we have

$$\ell_i^{u \rightarrow v} = -\log \frac{\exp(s_{i,\bar{e}}^{u \rightarrow v}/\tau)}{\sum_r \exp(s_{i,r}^{u \rightarrow v}/\tau)}, \quad (4)$$

where  $\bar{e}$  is the entity in view  $v$  that point  $i$  belongs to. If point  $i$  belongs to the background, then we use

$$\ell_i^{u \rightarrow v} = -\log \frac{\exp(s_{i,\bar{b}}^{u \rightarrow v}/\tau)}{\sum_r \exp(s_{i,r}^{u \rightarrow v}/\tau)}, \quad (5)$$

to align point  $i$  with its corresponding background point  $\bar{b}$  in view  $v$  with a temperature factor  $\tau$ . We can similarly compute the similarity from view  $v$  to  $u$  and obtain the final point-entity contrastive loss:

$$\mathcal{L}_{\text{p2e}} = \frac{1}{2} \left( \frac{1}{N} \sum_i \ell_i^{u \rightarrow v} + \frac{1}{N} \sum_j \ell_j^{v \rightarrow u} \right). \quad (6)$$

### 3.2. Entity-to-Language Alignment

To achieve open-vocabulary 3D scene understanding, we align the 3D point features with language features. First, we merge the 3D point features from both views to generate the final point features  $\mathbf{F}_P$ . Next, we feed the text descriptions of 3D entities to a frozen text encoder  $\mathcal{E}_{\text{txt}}$  to extract language features  $\mathbf{F}_T$ . Then we perform entity-to-language contrastive learning to align the 3D point features with their corresponding language features.

**Point Features Merging** In Sec. 3.1, we extract point features  $\mathbf{F}_u$  and  $\mathbf{F}_v$  for different augmented views of the input point cloud. To fully utilize the advantages of both views’ features and further align with language features, we merge  $\mathbf{F}_u$  and  $\mathbf{F}_v$  to obtain point features  $\mathbf{F}_P$  by taking the mean feature of matched points in the two views. We then feed the merged features to the VL-adapter  $\text{Proj}_{\text{VL}}(\cdot)$  to align the dimension of point and language features:

$$\mathbf{F}_P = \text{Mean}(\mathbf{F}_u, \mathbf{F}_v), \quad \mathbf{F}_{\text{VL}} = \text{Proj}_{\text{VL}}(\mathbf{F}_P). \quad (7)$$

**Language Feature Extraction** Given various text descriptions  $\mathbf{T}$ , e.g., object captions  $\mathbf{T}_{\text{cap}}$  and referrals  $\mathbf{T}_{\text{ref}}$  of different 3D entities in the scene, we random sample  $N_T$  descriptions and extract their embeddings using the pre-trained CLIP text encoder  $\mathcal{E}_{\text{txt}}$  of CLIP [59]:

$$\mathbf{F}_T = \mathcal{E}_{\text{txt}}(\mathbf{T}) \in \mathbb{R}^{N_T \times D}. \quad (8)$$

**Entity-to-Language Contrastive Learning** After obtaining the adapted 3D point features  $\mathbf{F}_{\text{VL}}$  and entity text description features  $\mathbf{F}_T$ , we conduct entity-to-language contrastive learning to align their representations. Analogous to Sec. 3.1, we employ a point-discriminative loss to enforce consistency among matched point-language pairs, while contrasting unmatched pairs. Specifically, given the language features  $\mathbf{F}_T$  and adapted vision features  $\mathbf{F}_{\text{VL}}$ , we first compute the cosine similarity matrix  $s^{t \rightarrow e}$ . For language description  $i$ , we

compute its similarity to all points in an entity  $e_k$  by

$$\mathbf{s}_{i,e_k}^{t \rightarrow e} = \frac{1}{|\mathbf{E}_k|} \sum_{j \in \mathbf{E}_k} d_{\cos}(\mathbf{F}_T^i, \mathbf{F}_{\text{VL}}^j), \quad (9)$$

where  $\mathbf{E}_k = \{j | M[j] = e_k\}$  denotes the set of points that belongs to entity  $e_k$ . We concatenate text-to-entity similarities  $\mathbf{s}_i^{t \rightarrow e}$  between text  $i$  and all entity features derived from  $\mathbf{F}_{\text{VL}}$  by

$$\mathbf{s}_i^{t \rightarrow e} = \text{Concat}(\mathbf{s}_{i,e_1}^{t \rightarrow p}, \mathbf{s}_{i,e_2}^{t \rightarrow p}, \dots, \mathbf{s}_{i,e_K}^{t \rightarrow e}). \quad (10)$$

Let  $\bar{e}$  denote the entity that aligns with text description  $i$ , we compute the text-to-entity contrastive loss with temperature factor  $\tau$  as:

$$\ell_i^{t \rightarrow e} = -\log \frac{\exp(\mathbf{s}_{i,\bar{e}}^{t \rightarrow e}/\tau)}{\sum_k \exp(\mathbf{s}_{i,k}^{t \rightarrow e}/\tau)}. \quad (11)$$

Similarly, we can compute the entity-to-text similarity  $s^{e \rightarrow t}$ . However, in contrast to the direct symmetry between point-to-entity and entity-to-point alignment in Sec. 3.1, the entity-to-text loss  $\ell^{e \rightarrow t}$  needs adjustments as different descriptions can refer to the same object entity in the scene and thus disable the application of the cross-entropy loss. Therefore, we replace the loss with a binary cross-entropy loss when applying the entity-to-language supervision:

$$\ell_k^{e \rightarrow t} = -\frac{1}{N_T} \sum_{t \in T} [\mathbb{1}(t \in \{\bar{t}\}) \log (\sigma(s_{k,t}^{e \rightarrow t})) + (1 - \mathbb{1}(t \in \{\bar{t}\})) \log (1 - \sigma(s_{k,t}^{e \rightarrow t}))], \quad (12)$$

where  $\{\bar{t}\}$  is the set of text descriptions that describes entity  $k$ ,  $\mathbb{1}(\cdot)$  is an indicator function testing if a text  $t$  is within the descriptive texts  $\{\bar{t}\}$  of the entity  $k$ . Finally, we combine the two-side contrast for entity-to-language alignment with:

$$\mathcal{L}_{\text{e2l}} = \alpha \frac{1}{N_t} \sum_{i=1}^{N_t} \ell_i^{t \rightarrow e} + \beta \frac{1}{K} \sum_{k=1}^K \ell_k^{e \rightarrow t}, \quad (13)$$

where factors  $\alpha$  and  $\beta$  serve to balance the scale of cross-entropy loss and binary cross-entropy loss.

### 3.3. Model Learning

**Training Data** We obtain training data following the data curation pipeline from SceneVerse [35]. Specifically, we use an off-the-shelf 3D instance segmentation model [97] to generate entity mask proposals on ScanNet and generate referential and descriptive text of objects with the help of scene graphs and 2D foundation models [56] for fair comparisons with existing methods. We also leverage the generated paired 3D-VL data from other real 3D scene datasets, including 3RScan [66], HM3D [60] and MultiScan [50] in SceneVerse as additional training data.

**Learning Objectives** Our training objective is fully contrastive-based. We combine both the point-to-entity alignment in Eq. (6) and the entity-to-language alignment in Eq. (13), enabling the end-to-end training of the 3D encoder  $\mathcal{E}_{\text{pcd}}$  and the VL-adapter  $\text{Proj}_{\text{VL}}(\cdot)$ , leading to a final loss function as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{p2e}} + \mathcal{L}_{\text{e2l}}. \quad (14)$$

**Implementation Details** For our default model, we adopt sparse-convolution based UNet (SPUNet) [11, 22] as the 3D encoder and CLIP [59] as the text encoder. The VL-adapter is implemented with a two-layer MLP. During training, we only update the parameters of the 3D encoder and VL-adapter and keep the text encoder frozen. We provide more training details in the *supplementary*.

## 4. Experiments

In our experiments, we focus the following questions:

- How effective are scene representations learned by MPEC for open-vocabulary scene understanding tasks?
- Does the learned model show zero-shot transfer capabilities in data-scarce scenarios?
- How do the learned semantically aligned representation compare with previous pre-training methods for diverse 3D-VL tasks, including segmentation, object referral, question answering, *etc.*?
- How does each component of MPEC contribute to model performance? Does training data domain and language type affect model performance?

In the following sections, we detail the design of experiments to address these key questions and demonstrate the effectiveness of MPEC as a general open-vocabulary pre-training pipeline as well as a plug-and-play backbone for various downstream tasks. We provide more experiment details and additional experiment results in the *supplementary*.

### 4.1. Open-vocabulary Scene Understanding

**Experiment Settings** We select open-vocabulary semantic segmentation (OV-SemSeg) as a signature task for evaluating models' open-vocabulary scene understanding capabilities. Specifically, we follow previous works [17, 38, 81] to evaluate the foreground mIoU (f-mIoU, %) and foreground mAcc (f-mAcc, %) on ScanNet by providing all class categories from ScanNet20 as input texts. We directly use the aligned 3D features  $\mathbf{F}_{\text{VL}}$  without applying the cross-view augmentation and pass the category texts into the frozen CLIP text encoder, using the best match between the two features as the prediction for the final class category.

**Results & Analyses** As shown in Tab. 1, MPEC achieves state-of-the-art performance, with 66.0% f-mIoU and 81.3% f-mAcc, outperforming previous state-of-the-art OV3D [38] by 3.0% and 6.5%, respectively.

Table 1. Open-Vocabulary 3D Semantic Segmentation on ScanNet [14]. We report foreground mIoU (f-mIoU, %) and foreground mAcc (f-mAcc, %).  $\dagger$  and  $\ddagger$  mean results reproduced by RegionPLC [81] and Uni3D [93], independently.

Method	Network	f-mIoU	f-mAcc
MaskCLIP $\dagger$ [92]	CLIP [59]	23.1	40.9
OpenScene-2D [57]	LSeg [41]	58.0	68.5
CLIP $^{2\dagger}$ [85]	CLIP [59]	-	38.5
Uni3D $\ddagger$ [93]	ViT [18]	-	45.8
OpenScene-3D $\dagger$ [57]	SPUNet16 [22]	57.2	69.9
OpenScene-3D $\dagger$ [57]	SPUNet32 [22]	57.8	70.3
PLA $\dagger$ [17]	SPUNet16 [22]	17.7	33.5
PLA $\dagger$ [17]	SPUNet32 [22]	19.1	41.5
RegionPLC [81]	SPUNet16 [22]	56.9	75.6
RegionPLC [81]	SPUNet32 [22]	59.6	77.5
OV3D [38]	SPUNet16 [22]	64.0	76.3
MPEC	SPUNet16 [22]	64.6	79.5
MPEC	SPUNet32 [22]	<b>66.0</b>	<b>81.3</b>

MPEC also shows a significant improvement ( $\sim 10\%$ ) over OpenScene, which uses pre-trained 2D-VL model as supervision for 3D representation learning. This highlights the benefit of incorporating 3D information into scene representation learning. Additionally, we report model performance using the SPUNet16 architecture for the 3D encoder, demonstrating consistent improvements over previous methods with the same backbone. These results showcase better alignment of our learned 3D scene representations to language descriptions using the MPEC framework.

**Per-category Results** Due to the page limit, we provide per-category performance analysis on ScanNet20 [14] in the *supplementary*. In conclusion, we find that MPEC achieves much higher IoU than previous open-source state-of-the-art RegionPLC [81] on many categories with similar accuracies, indicating fewer false positives and better geometric understanding.

**Qualitative Results** We provide more qualitative results in Fig. 3, including relatively **good** and **bad** cases. The basic object category understanding and more complex object referral results are shown in the top row and bottom row in Fig. 3, respectively. MPEC shows superior basic category understanding capabilities with clear object boundaries, which attribute to the fine-grained geometry understanding. Moreover, MPEC can handle visual grounding tasks that take more complex texts as input. However, when handling distractions with similar appearances and locations, MPEC still faces challenges. We attribute this to the limitations of the fixed CLIP text encoder, which struggles to handle long and detailed descriptions required for precise visual grounding. We provide further investigation on this topic in the *supplementary*.



Figure 3. Qualitative Results on ScanNet [14]. We show relatively **good** and **bad** cases in **blue** and **red**.

## 4.2. Zero-shot Transfer and Long-tail Scenarios

**Experiment Settings** To assess the generalization capability of learned representations to unseen scenes and language descriptions, we evaluate the f-mIoU and f-mAcc on OV-SemSeg in two settings:

- **Zero-shot Transfer:** For *zero-shot transfer* experiments, we evaluate MPEC on the unseen SceneVerse-val [35] and Matterport3D [5]. To clarify, the SceneVerse-val zero-shot transfer setting requires excluding MultiScan [50] scenes and languages from training while testing on MultiScan data. We adhered to this protocol, removing MultiScan data during training for fair comparisons.
- **Long-tail Scenarios:** We provide the performance on ScanNet200 [61] dataset, which is a more challenging scenario and includes a large number of test classes.

**Results & Analyses** We present experimental results on *zero-shot transfer* in Tab. 2 and Tab. 3, as well as *long-tail scenarios* in Tab. 4. Our findings are as follows:

- **Zero-shot Transfer:** For *zero-shot transfer* on SceneVerse-val [35], as shown in Tab. 2, MPEC significantly outperforms existing methods by a large margin. Specifically, for f-mIoU, we outperform the second-best OpenScene [57] by 3.7%, and for f-mAcc, we surpass the second-best RegionPLC [81] by 7.2%. We further provide results on Matterport3D [5] in Tab. 3. It's worth noting that, compared with OpenScene [57] and OV3D [38], our MPEC omits Matterport3D during training. However, MPEC achieves comparable performance with these domain-specific methods. Specifically, we achieve 47.7% f-mIoU and 69.8% f-mAcc. Compared with OV3D [38], MPEC is 2.7% lower on f-mIoU, but significantly outperforms 4.1% f-mAcc. MPEC also surpasses another *zero-shot* method RegionPLC [81] by about 20% on both f-mIoU and f-mAcc. These results show that MPEC can generalize better to unseen scenarios.
- **Long-tail Scenarios:** We provide the results on ScanNet200 [61] in Tab. 4. MPEC achieves state-of-the-art

Table 2. **Zero-Shot Open-Vocabulary 3D Semantic Segmentation on SceneVerse-val** [35]. We report f-mIoU (%) and f-mAcc (%).

Method	f-mIoU	f-mAcc
OpenScene [57]	41.3	52.5
RegionPLC [81]	39.1	56.4
<b>MPEC</b>	<b>45.0</b>	<b>63.6</b>

Table 3. **Open-Vocabulary 3D Semantic Segmentation on Mat-terport3D** [5]. We report the f-mIoU (%) and f-mAcc (%). ‡ means results reproduced by OV3D [38].

Method	Zero-Shot	f-mIoU	f-mAcc
OpenScene-3D‡ [57]	✗	49.7	64.0
OV3D [38]	✗	<b>50.4</b>	65.7
RegionPLC [81]	✓	28.9	43.8
<b>MPEC</b>	✓	47.7	<b>69.8</b>

performance, with 10.8% f-mIoU and 27.4% f-mAcc, outperforming existing models by a large margin (~10% improvement in f-mAcc). This indicates that MPEC can capture fine-grained taxonomic distinctions, ultimately benefiting complex downstream embodied tasks.

### 4.3. Fine-tuning on Downstream Tasks

**Experiment Settings** We evaluate the adaptability of our learned 3D encoder as a general 3D backbone for fine-tuning on diverse 3D tasks. Specifically, we consider the following two levels of tasks:

- **Low-level Perception:** For this group of tasks, we consider closed-set semantic and instance segmentation as representatives to compare our model against previous 3D point cloud representation learning methods. Concretely, we compare MPEC with unsupervised scene representation methods following the setting in [67] on ScanNet and ScanNet200, using mIoU and mAP@0.5 as metrics for semantic and instance segmentation respectively. We fine-tune our learned 3D scene representation  $F_P$  from the 3D encoder on the downstream tasks. We also conduct data efficiency experiments following [29].
- **High-level Reasoning:** We further test if the scene representation learned by MPEC can be directly adopted in more 3D-VL tasks including grounding, captioning, and question answering. To achieve this goal, we select a recent state-of-the-art 3D-VL model, PQ3D [97], that leverages the same backbone architecture as the ones used in MPEC for comparative studies. By replacing PQ3D’s 3D encoder with our pre-trained method, we report model performance with fine-tuning on: (i) visual grounding datasets, including ScanRefer [6], Nr3D and Sr3D [1], Multi3dRefer [87]; (ii) 3D question-answering dataset SQA3D [48]; and (iii) scene captioning dataset

Table 4. **Open-Vocabulary 3D Semantic Segmentation on ScanNet200** [61]. We report f-mIoU (%) and f-mAcc (%). ‡ means results reproduced by OV3D [38].

Method	f-mIoU	f-mAcc
PLA [17]	1.8	3.1
OpenScene-3D‡ [57]	7.3	-
RegionPLC [81]	9.1	17.3
OV3D [38]	8.7	-
<b>MPEC</b>	<b>10.8</b>	<b>27.4</b>

Scan2Cap [8]. We report model performance under the convention of each dataset. Specifically, for grounding results on ScanRefer [6] and Multi3dRefer [87], we report the grounding accuracies of predictions whose object bounding boxes overlap the ground truth with IoU  $> 0.5$  (Acc@0.5). For visual grounding on Nr3D and Sr3D of Referit3D [1], we report the grounding accuracies using ground-truth masks (Acc). For 3D dense captioning on Scan2Cap [8], we report CIDEr under IoU@0.5 (CIDEr@0.5). For situated reasoning on SQA3D [48], we report the answer accuracies (AccQA).

**Results & Analyses** We provide fine-tuning results for low-level perception tasks in Tab. 5 and results for high-level reasoning tasks in Tab. 6. Our findings on these two groups of tasks are as follows:

- **Low-level Perception:** For close-set semantic segmentation, MPEC achieves considerable performance gains over existing methods. Specifically, it surpasses previous methods on both ScanNet [14] and ScanNet200 [61], achieving mIoUs of 75.8% and 31.8%, respectively, with a notable 1.8% improvement on ScanNet200 over GC [67]. For close-set instance segmentation, MPEC performs comparably well to GC on ScanNet [14], yet significantly outperforms on ScaNet200 [61]. Specifically, our MPEC achieves mAP@0.5 of 31.6%, achieving substantial improvements over the previous state-of-the-art GC’s 27.5%. We attribute these gains to our proposed masked point-entity contrastive learning framework, which provides more discriminative features across 3D entities. Data efficiency experiments are provided in *supplementary*.
- **High-level Reasoning:** As shown in Tab. 6, when substituting the 3D backbone with MPEC for fine-tuning on high-level 3D-VL tasks, the model achieves consistent and notable improvements in most reasoning tasks. Specifically, MPEC with PQ3D task heads sets new state-of-the-art performance on 3D visual grounding datasets, including ScanRefer [6], Nr3D [1], Sr3D [1], and Multi3dRefer [87]. Additionally, our model also boosts the accuracy on SQA3D [48] from 47.1% to 47.5%. These findings highlight MPEC’s adaptability and its potential to serve as an initial backbone weight for various downstream tasks.

Table 5. **Fine-Tuning Results on Perception Tasks.** We report mIoU (%) for semantic segmentation and mAP@0.5 (%) for instance segmentation on ScanNet [14] and ScanNet200 [61].

Datasets	Semantic Segmentation (mIoU)						Instance Segmentation (mAP@0.5)					
	SC	PC [74]	CSC [29]	MSC [72]	GC [67]	MPEC	SC	PC [74]	CSC [29]	MSC [72]	GC [67]	MPEC
ScanNet [14]	72.2	74.1	73.8	75.3	75.7	<b>75.8</b>	56.9	58.0	59.4	59.6	62.3	<b>62.5</b>
ScanNet200 [61]	25.0	26.2	26.4	28.8	30.0	<b>31.8</b>	24.5	24.9	25.2	26.8	27.5	<b>31.6</b>

Table 6. **Fine-Tuning Results on Reasoning Tasks.** The grounding accuracy of ScanRefer [6] and Multi3DRefer [87] are reported under IoU@0.5. The results of Nr3D and Sr3D [1] are reported using ground-truth masks. We report CIDEr under IoU@0.5 on Scan2Cap [8] and answer accuracy on SQA3D [48].

Method	ScanRefer Acc@0.5	Nr3D Acc	Sr3D Acc	Multi3DRefer F1@0.5	Scan2Cap CIDEr@0.5	SQA3D AccQA
3DJCG	37.3	-	-	26.6	47.7	-
M3DRef-CLIP	44.7	49.4	-	38.4	-	-
3D-ViSTA	45.8	64.2	76.4	-	66.9	<b>48.5</b>
Vote2CapDetr	-	-	-	-	64.3	-
PQ3D	51.2	<b>66.7</b>	79.7	50.1	<b>80.3</b>	47.1
MPEC + PQ3D	<b>51.8</b>	<b>66.7</b>	<b>80.0</b>	<b>50.8</b>	80.2	47.5

#### 4.4. Additional Discussions

In this subsection, we conduct ablation studies on model design and data engineering.

**Model Design** In Tab. 7, we show the results with different model designs on open-vocabulary 3D semantic segmentation on ScanNet [14]. Specifically, directly applying  $\mathcal{L}_{e2e}$  with  $\mathcal{L}_{e2l}$  without employing cross mask augmentations leads to a noticeable drop on f-mIoU and f-mAcc, compared to minimizing  $\mathcal{L}_{e2l}$  alone. We hypothesize this may be attributed to the model’s tendency to overemphasize the uniqueness of each entity within a single view, thereby neglecting shared attributes across entities. By introducing masked cross-view augmentations, minimizing  $\mathcal{L}_{e2e}$  across views encourages the model to distinguish entities between views without enforcing distinctiveness within the same view. This adjustment helps preserve common attributes for different entities, such as semantic categories.

**Text Descriptions Type** The text descriptions of entities include two types: captions  $T_{cap}$  which provide detailed descriptions of an object’s visual and physical properties, and referrals  $T_{ref}$  which articulate the spatial relationships between the entity and other objects within the scene. We conduct an ablation study on each description type in Tab. 8. As shown in Tab. 8, leveraging both description types yields the best performance. This result aligns with our expectations, as combining an entity’s inherent attributes with its relationships with other entities provides a more comprehensive understanding of the 3D entity.

**Training Dataset** Our training dataset, SceneVerse [35], is composed of seven datasets, including real-world scene datasets ScanNet [14], MultiScan [50], RScan [66], ARKitScenes [3] and HM3D [60] as well as synthetic environments

Table 7. **Ablation Analysis on Model Design.** We report f-mIoU (%) and f-mAcc (%) on ScanNet [14].

$\mathcal{L}_{e2l}$	$\mathcal{L}_{e2e}$	Cross-View Aug.	f-mIoU	f-mAcc
✓			63.6	79.2
✓	✓		62.2	78.5
✓	✓	✓	<b>64.6</b>	<b>79.5</b>

Table 8. **Ablation Analysis on Text Type.** We report f-mIoU (%) and f-mAcc (%) on ScanNet [14].

$T_{cap}$	$T_{ref}$	f-mIoU	f-mAcc
✓		57.6	74.8
	✓	61.7	77.2
✓	✓	<b>64.6</b>	<b>79.5</b>

ments from Structured3D [90] and ProcTHOR [15]. For our ablation study, we discard ARKitScenes [3] and the synthetic datasets, *i.e.* Structured3D [90] and ProcTHOR [15], to analyze the influence of different real-world data sources, as shown in Tab. 9. We observe obvious performance gains when scaling up the training data.

Table 9. **Ablation Analysis on Training Dataset.** We report f-mIoU and f-mAcc on ScanNet [14] after training on different combinations of datasets.

ScanNet [14]	MultiScan [50]	RScan [66]	HM3D [60]	f-mIoU	f-mAcc
✓				56.4	74.3
✓	✓			57.7	75.1
✓		✓		59.9	76.5
✓			✓	63.3	79.3
✓	✓	✓		60.6	77.4
✓	✓		✓	63.3	79.0
✓	✓	✓	✓	62.7	78.3
✓	✓	✓	✓	<b>64.6</b>	<b>79.5</b>

## 5. Conclusion

In this paper, we propose a novel entity-level contrastive learning framework, MPEC, for open-vocabulary 3D scene understanding, which achieves state-of-the-art performance on open-vocabulary 3D semantic segmentation. Through fine-tuning on diverse downstream tasks, our model demonstrates high effectiveness and transferability, showcasing its potential as a foundational approach for 3D vision-language applications. Furthermore, our data scale ablation study reveals the importance of large-scale 3D vision-language datasets in advancing 3D foundation models, underscoring a critical area for future development.

## References

- [1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440. Springer, 2020. 2, 7, 8
- [2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1
- [3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NIPSD*, 2021. 8
- [4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, pages 16464–16473, 2022. 3
- [5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6, 7
- [6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020. 2, 7, 8
- [7] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *CVPR*, pages 7020–7030, 2023. 2
- [8] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, pages 3193–3203, 2021. 2, 7, 8
- [9] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *ICCV*, pages 18109–18119, 2023. 3
- [10] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 2
- [11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 5, 13
- [12] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. <https://github.com/Pointcept/Pointcept>, 2023. 13
- [13] Spconv Contributors. Spconv: Spatially sparse convolution library. <https://github.com/traveller59/spconv>, 2022. 13
- [14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 6, 7, 8, 13, 14, 15, 16, 17
- [15] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *NeurIPS*, pages 5982–5994. Curran Associates, Inc., 2022. 8
- [16] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 14
- [17] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *CVPR*, pages 7010–7019, 2023. 2, 5, 6, 7
- [18] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6
- [19] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, page 02783649241281508, 2023. 1
- [20] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhao Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 3
- [21] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 2
- [22] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 2, 5, 6, 13
- [23] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Concept-graphs: Open-vocabulary 3d scene graphs for perception and planning. In *ICRA*, pages 5021–5028. IEEE, 2024. 1
- [24] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3
- [25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3
- [26] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NIPS*, 36:20482–20494, 2023. 3
- [27] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NIPS*, 36:20482–20494, 2023. 1
- [28] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *CVPR*, pages 26406–26416, 2024. 3
- [29] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with

- contrastive scene contexts. In *CVPR*, pages 15587–15597, 2021. 2, 3, 7, 8, 13, 14, 15
- [30] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *NIPS*, 2024. 3
- [31] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1, 3
- [32] Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *CVPR*, 2025. 2, 3
- [33] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *ICCV*, pages 6535–6545, 2021. 3
- [34] Zhenling Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *ECCV*, pages 169–185. Springer, 2025. 2
- [35] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 2, 3, 5, 6, 7, 8, 14
- [36] Chao Jia, Yinfai Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 2
- [37] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, pages 4867–4876, 2020. 2, 13
- [38] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *CVPR*, pages 21284–21294, 2024. 2, 5, 6, 7
- [39] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. *arXiv preprint arXiv:2410.00255*, 2024. 3
- [40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2
- [41] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2, 6
- [42] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *CVPR*, 2025. 1
- [43] Ruihuang Li, Zhengqiang Zhang, Chenhang He, Zhiyuan Ma, Vishal M Patel, and Lei Zhang. Dense multimodal alignment for open-vocabulary 3d scene understanding. In *ECCV*, pages 416–434. Springer, 2024. 2
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3
- [45] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *ICLR*, 2025. 1
- [46] Ruijie Lu, Yixin Chen, Junfeng Ni, Baoxiong Jia, Yu Liu, Diwen Wan, Gang Zeng, and Siyuan Huang. Movis: Enhancing multi-object novel view synthesis for indoor scenes. In *CVPR*, 2025. 1
- [47] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boulaaras, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *CoRL*, pages 1610–1620. PMLR, 2023. 2
- [48] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sq3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 2, 7, 8
- [49] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024. 3
- [50] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In *NIPS*, 2022. 5, 6, 8
- [51] Chen Min, Liang Xiao, Dawei Zhao, Yiming Nie, and Bin Dai. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles*, 2023. 3
- [52] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, 2024. 2
- [53] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. In *NeurIPS*, 2024. 1
- [54] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *CVPR*, 2025. 1
- [55] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 1
- [56] OpenAI. Gpt-4 with vision (gpt-4v) system card, 2023. 5
- [57] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 1, 2, 6, 7, 14
- [58] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *ECCV*, pages 214–238. Springer, 2025. 3

- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 4, 5, 6, 13, 14
- [60] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 5, 8
- [61] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, pages 125–141. Springer, 2022. 6, 7, 8, 13
- [62] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pages 8216–8223. IEEE, 2023. 2
- [63] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Open-mask3d: open-vocabulary 3d instance segmentation. In *NIPS*, pages 68367–68390, 2023. 2
- [64] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *ACMMM*, pages 6617–6626, 2024. 3
- [65] Xiaoyu Tian, Haoxi Ran, Yue Wang, and Hang Zhao. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *CVPR*, pages 13570–13580, 2023. 3
- [66] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *ICCV*, 2019. 5, 8
- [67] Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, Bo-hao Peng, Hengshuang Zhao, and Jiaya Jia. Groupcontrast: Semantic-aware self-supervised representation learning for 3d understanding. In *CVPR*, pages 4917–4928, 2024. 2, 3, 7, 8, 13, 15
- [68] Pengfei Wang, Yuxi Wang, Shuai Li, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Open vocabulary 3d scene understanding via geometry guided self-distillation. In *ECCV*, pages 442–460. Springer, 2024. 2
- [69] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 2
- [70] Weijie Wei, Fatemeh Karimi Nejadasl, Theo Gevers, and Martin R Oswald. T-mae: Temporal masked autoencoders for point cloud representation learning. *arXiv preprint arXiv:2312.10217*, 2023. 3
- [71] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2
- [72] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, pages 9415–9424, 2023. 2, 3, 8, 13, 15
- [73] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, pages 3426–3436, 2024. 2
- [74] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, pages 574–591. Springer, 2020. 2, 3, 8, 13
- [75] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 3
- [76] Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. Sampro3d: Locating sam prompts in 3d for zero-shot scene segmentation. *arXiv preprint arXiv:2311.17707*, 2023. 2
- [77] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *CVPR*, pages 28274–28284, 2024. 2
- [78] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pretraining with large-scale synthetic data. *arXiv preprint arXiv:2407.06084*, 2024. 3
- [79] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *CVPR*, pages 9403–9414, 2023. 3
- [80] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyer-gar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-lmns with better grounding and less hallucination. *arXiv preprint arXiv:2406.05132*, 2024. 2
- [81] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *CVPR*, 2024. 2, 5, 6, 7, 14
- [82] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2
- [83] Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puha Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. Metascenes: Towards automated replica creation for real-world 3d scans. In *CVPR*, 2025. 1, 2
- [84] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, pages 19313–19322, 2022. 2, 3
- [85] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chao-qiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *CVPR*, pages 15244–15253, 2023. 6
- [86] Taolin Zhang, Sunan He, Tao Dai, Zhi Wang, Bin Chen, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. In *AAAI*, pages 7296–7304, 2024. 3

- [87] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, pages 15225–15236, 2023. 2, 7, 8
- [88] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *ICCV*, pages 10252–10263, 2021. 3
- [89] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 3
- [90] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, pages 519–535. Springer, 2020. 8
- [91] Peiyuan Zhi, Zhiyuan Zhang, Muzhi Han, Zeyu Zhang, Zhitian Li, Ziyuan Jiao, Baoxiong Jia, and Siyuan Huang. Closed-loop open-vocabulary mobile manipulation with gpt-4v. *arXiv preprint arXiv:2404.10220*, 2024. 1
- [92] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. 6
- [93] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 6
- [94] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 2
- [95] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 3
- [96] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pages 2911–2921, 2023. 3, 14
- [97] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. *arXiv preprint arXiv:2405.11442*, 2024. 3, 5, 7, 13

# Masked Point-Entity Contrast for Open-Vocabulary 3D Scene Understanding

## Supplementary Material

### A. Implementation Details

We provide implementation and training details of our proposed *Masked Point-Entity Contrast* (MPEC) model.

**Model Architecture** For open-vocabulary 3D semantic segmentation and zero-shot visual grounding experiments, we ensure a fair comparison with existing methods by employing SparseUNet16 and SparseUNet32 [22] as the 3D encoder and the frozen CLIP [59] as the text encoder. The vision-language adapter consists of a two-layer MLP. For downstream fine-tuning tasks, we train a MinkUNet34C [11] re-implemented with *SpConv* [13] following *Pointcept* [12] using the same point-entity contrastive supervision.

For fine-tuning on low-level perception tasks, we build our experiments using the point cloud perception codebase *Pointcept* [12]. Specifically, for all the semantic segmentation tasks, including closed-set experiments on ScanNet [14], ScanNet200 [61], we add a single linear layer as the classification head and utilize the cross-entropy loss for supervision. For instance segmentation tasks, we follow previous works [29, 67, 72, 74] to adopt PointGroup [37] as the instance segmentation head.

For fine-tuning high-level reasoning tasks, we select PQ3D [97], a state-of-the-art framework for reasoning tasks for indoor scenes as the baseline. Specifically, we replace the voxel encoder of PQ3D with our MinkUNet34C while maintaining other configurations, and fine-tune on the reasoning datasets.

**View Generation and Masking Strategy** We follow the view generation pipeline and mask strategy in MSC [72]. For a given 3D point cloud, we first create two copies and apply separate random augmentation sequences to each, generating two distinct views of the same scene. The augmentation sequence, detailed in Tab. A.1, consists of three main components: spatial augmentations, photometric augmentations and sampling augmentations.

For the masking strategy, we adopt the approach proposed in MSC [72], setting the grid size to 0.1m to partition the original coordinates into evenly spaced, non-overlapping grids. For each view, 40% of the grids are selected and the features inside are masked and replaced with learnable tokens. Importantly, the selected grids for the two views are mutually exclusive, ensuring no overlap.

Table A.1. View Generation Pipeline.

Augmentation	Value
random rotate	angle=[-1/64, 1/64], axis='x', p=1
random rotate	angle=[-1/64, 1/64], axis='y', p=1
random flip	p=0.5
random coord jitter	sigma=0.005, clip=0.02
random color brightness jitter	ratio=0.4, p=0.8
random color contrast jitter	ratio=0.4, p=0.8
random color saturation jitter	ratio=0.2, p=0.8
random color hue jitter	ratio=0.02, p=0.8
random color gaussian jitter	std=0.05, p=0.95
grid sample	grid size=0.02
random crop	ratio=0.6
center shift	n/a
color normalize	n/a

Table A.2. Fine-Tuning Setting on Low-Level Perception Tasks.

Config	Value
optimizer	SGD
scheduler	cosine decay
weight decay	1e-4
optimizer momentum	0.9
batch size	12
warmup epochs	40
epochs	800

**Training** For the point-entity contrastive learning, we utilize the AdamW optimizer with a learning rate of  $1 \times 10^{-3}$  for 500 epochs with a cosine warm-up period of 200 steps. During training, we set a batch size of 4 scenes for each GPU and sample 64 text descriptions for each scene. To balance the scale of cross-entropy loss and binary cross-entropy loss in  $\mathcal{L}_{\text{e2l}}$ , we empirically set  $\alpha$  and  $\beta$  to 1.0 and 6.0, respectively. All the contrastive learning experiments are performed on 4 NVIDIA-A100 GPUs with the longest training taking less than 4 days.

For the downstream fine-tuning experiments, we conduct all the low-level perception tasks on *Pointcept* [12] and all the high-level reasoning tasks on PQ3D [97].

The general fine-tuning setting for low-level perception tasks is shown in Tab. A.2. We adjust the learning rate based on the task. Specifically, for full-set semantic and instance segmentation fine-tuning experiments on ScanNet and ScanNet200, the learning rate is set to 0.2.

For training PQ3D on high-level reasoning tasks, we train the model on multiple 3D vision-language tasks including visual grounding, question answering, and dense captioning for 50 epochs. The model architecture uses a hidden dimension of 1024.

Table A.3. **Partial Per-Category Performance on ScanNet** [14]. We compare the IoU (%) and accuracy (%) with previous SOTA RegionPLC [81] of each category.

	chair		bookshelf		counter		toilet		sink		shower curtain		curtain	
	IoU	Acc	IoU	Acc	IoU	Acc								
RegionPLC	75.4	82.9	72.9	<b>96.3</b>	49.0	64.6	64.2	<b>98.6</b>	38.1	84.2	43.3	<b>90.7</b>	46.5	51.4
MPEC	<b>83.8</b>	<b>85.5</b>	<b>80.3</b>	92.3	<b>56.4</b>	<b>65.2</b>	<b>85.6</b>	98.2	<b>48.5</b>	<b>85.6</b>	<b>61.6</b>	87.1	<b>66.1</b>	<b>73.1</b>

sion of 768 and 4 query decoder layers. Optimization is performed using the AdamW optimizer with a learning rate of  $1 \times 10^{-4}$ , a batch size of 16, and momentum parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$ . The loss balancing weights are set to  $\lambda_{\text{gen}} = 1$  and  $\lambda_{\text{grd}} = 10$ .

## B. Additional Per-category Performance Analyses on ScanNet

We provide part of the per-category performance on ScanNet in Tab. A.3. As can be seen, though RegionPLC [81] and MPEC achieve similar accuracies on many categories, MPEC continuously outperforms RegionPLC on the IoU metric by a large margin, indicating fewer false positives and better geometric understanding. Superior results on *shower curtain* and *curtain* further highlight MPEC’s strong spatial reasoning and semantic understanding ability.

## C. Additional Experiment Results for Zero-shot 3D Visual Grounding

In Fig. 3 of the main paper, we observe that MPEC still faces challenges when dealing with complicated grounding texts qualitatively. We attribute this phenomenon to the limitations of the fixed CLIP [59] text encoder. This section provides quantitative analysis on zero-shot 3D visual grounding experiments on SceneVerse [35] to support this hypothesis.

**Experiment Settings** Following the SceneVerse-val *zero-shot* setting in [35], we remove MultiScan data during training for fair comparisons. We provide the model ground-truth object proposals and use the pooled feature  $F_{\text{VL}}$  for each object to match with the grounding text for predictions. We test the grounding accuracy of different text encoders, *i.e.*, frozen CLIP [59] and trainable BERT [16].

**Results & Analyses** We present experiments for *zero-shot* 3D visual grounding on SceneVerse-val [35] in Tab. A.4. MPEC with the frozen CLIP text encoder achieves a better overall grounding accuracy of 17% compared with existing available open-vocabulary 3D understanding models like OpenScene [57] and RegionPLC [81]. However, compared with task-specific models for 3D visual grounding, *i.e.*, 3D-ViSTA [96] and GPS [35], MPEC with the frozen (✳) CLIP

Table A.4. **Zero-Shot 3D Visual Grounding on SceneVerse-val** [35]. We report accuracy (%) on SceneVerse-val [35] and evaluate models using GT object proposals. 🔴 and ✳ indicates trainable and frozen text encoder, respectively.

Method	Text Encoder	Overall	Easy	Hard
3D-ViSTA [96]	BERT🔴	52.9	59.6	35.4
<b>GPS</b> [35]	BERT🔴	<b>59.2</b>	<b>69.4</b>	<b>44.0</b>
OpenScene [57]	CLIP✳	13.3	15.5	10.1
RegionPLC [81]	CLIP✳	10.6	11.8	8.9
MPEC	CLIP✳	17.0	23.8	6.7
MPEC	BERT🔴	42.6	56.2	22.2

text encoder is considerably lower by more than 35% (17% vs. 52%). After replacing the frozen CLIP text encoder with a trainable (🔴) BERT, the overall accuracy significantly improves from 17% to 42.6%. This underscores the limitation of the frozen LCIP text encoder, which struggles to handle long and detailed descriptions, particularly when grounding specific 3D objects in complex 3D scenes.

## D. Additional Experiment Results for Data-efficiency Fine-tuning

In this section, we provide additional fine-tuning experiment results on the ScanNet Data-Efficiency benchmark [29].

**Experiment Settings** We compare our method with previous methods on ScanNet-LR (Limited Scene Reconstruction) and ScanNet-LA (Limited Annotation) test splits. For ScanNet-LR, we use the {1%, 5%, 10%, 20%} sampled scenes provided in ScanNet-LR and use the annotations within each scene to fine-tune our pre-trained representation  $F_P$  for semantic segmentation. Similarly, For ScanNet-LA, we follow [29] and provide {20, 50, 100, 200} labeled points per scene for fine-tuning our learned representation. Notably, we train MPEC by removing ScanNet data under this setting and report the mIoU for semantic segmentation on both splits as the evaluation metric.

**Results & Analyses** As shown in Tab. A.5 and Tab. A.6, our method consistently outperforms previous methods by a large margin, particularly in scenarios with extremely limited

**Table A.5. ScanNet Limited Scene Reconstruction.** We report the mIoU (%) results on ScanNet [14] data efficient semantic segmentation benchmark with limited scene reconstruction setting.

LR	Semantic Segmentation (mIoU)					
	Pct.	SC	CSC [29]	MSC [72]	GC [67]	Ours
1%	26.1	28.9	29.2	30.7	<b>40.8</b>	
5%	47.8	49.8	50.7	52.9	<b>58.5</b>	
10%	56.7	59.4	61.0	62.0	<b>64.0</b>	
20%	62.9	64.6	64.9	<b>66.5</b>	66.3	

**Table A.6. ScanNet Limited Annotation.** We report the mIoU (%) results on ScanNet [14] data efficient semantic segmentation benchmark with limited point annotation setting.

LA	Semantic Segmentation (mIoU)					
	Pts.	SC	CSC [29]	MSC [72]	GC [67]	Ours
20	41.9	55.5	61.2	61.2	<b>62.9</b>	
50	53.9	60.5	66.8	67.3	<b>69.2</b>	
100	62.2	65.9	69.7	70.3	<b>72.0</b>	
200	65.5	68.2	70.7	71.8	<b>73.1</b>	

reconstructions ( $\sim 10\%$  improvement for 1% trained scenes). This highlights the ability of MPEC to retain language-aligned 3D feature extraction on unseen scenes and the fast adaptability of the learned representations to downstream tasks under data-scarce scenarios.

## E. More Quantitative Results

We provide more qualitative results in Fig. A.4 and Fig. A.5.



Toilet



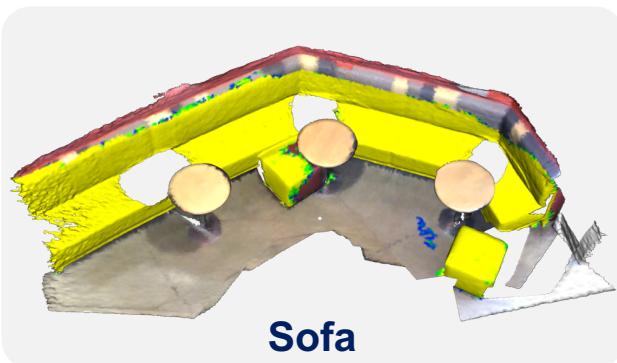
Bed



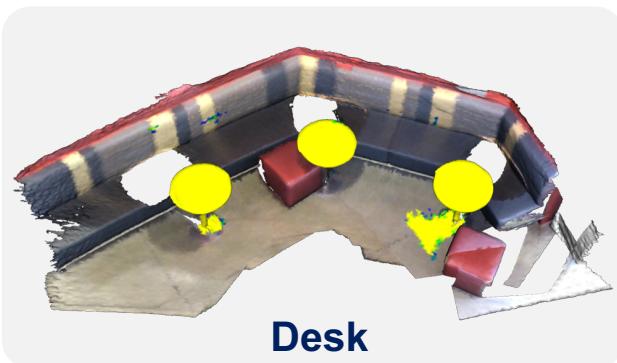
Chair



Sink



Sofa



Desk



Monitor



Desk

Figure A.4. More Qualitative Results on ScanNet [14].



Shelf



Wash Clothes



Printer



Lamp



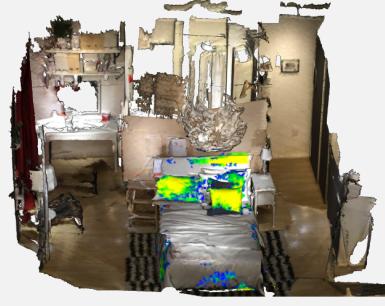
Stool



Backpack



Nightstand



Pillow

Figure A.5. More Qualitative Results on ScanNet [14].