

PLA: Language-Driven Open-Vocabulary 3D Scene Understanding

Runyu Ding^{1*†} Jihan Yang^{1*} Chuhui Xue² Wenqing Zhang² Song Bai^{2‡} Xiaojuan Qi^{1‡}
¹The University of Hong Kong ²ByteDance

Abstract

Open-vocabulary scene understanding aims to localize and recognize unseen categories beyond the annotated label space. The recent breakthrough of 2D open-vocabulary perception is largely driven by Internet-scale paired image-text data with rich vocabulary concepts. However, this success cannot be directly transferred to 3D scenarios due to the inaccessibility of large-scale 3D-text pairs. To this end, we propose to distill knowledge encoded in pre-trained vision-language (VL) foundation models through captioning multi-view images from 3D, which allows explicitly associating 3D and semantic-rich captions. Further, to foster coarse-to-fine visual-semantic representation learning from captions, we design hierarchical 3D-caption pairs, leveraging geometric constraints between 3D scenes and multi-view images. Finally, by employing contrastive learning, the model learns language-aware embeddings that connect 3D and text for open-vocabulary tasks. Our method not only remarkably outperforms baseline methods by 25.8% ~ 44.7% hIoU and 14.5% ~ 50.4% hAP₅₀ in open-vocabulary semantic and instance segmentation, but also shows robust transferability on challenging zero-shot domain transfer tasks. See the project website at <https://dingry.github.io/projects/PLA>.

1. Introduction

3D scene understanding is a fundamental perception component in real-world applications such as robot manipulation, virtual reality and human-machine interaction. Deep learning has attained remarkable success in this area [13, 39, 29]. However, deep models trained on a human-annotated dataset are only capable of understanding semantic categories in that dataset, *i.e.* closet-set prediction. As a result, they fail to recognize unseen categories in the open world (see Fig. 1). This largely restricts their applicability in real-world scenarios with unbounded categories. Besides, heavy annotation costs on 3D datasets (*e.g.* 22.3 minutes for one scene with 20 classes [7]) further make it infeasible to rely

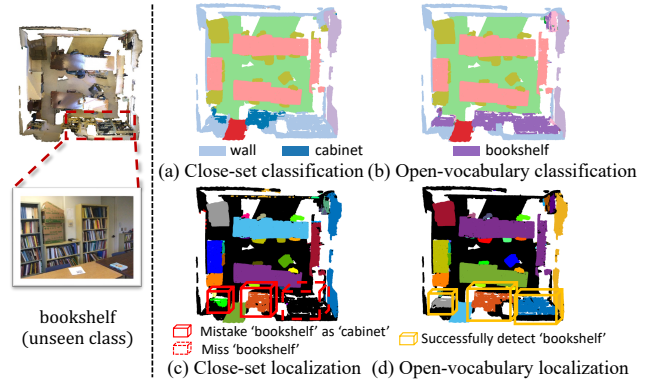


Figure 1. An example of 3D open-vocabulary scene understanding with “bookshelf” as unseen class for ScanNet [7]. The close-set model mistakes “bookshelf” as “cabinet” or simply misses “bookshelf” in (a) and (c). Our open-vocabulary model correctly localizes and recognizes “bookshelf” in (b) and (d).

on human labor to cover all real-world categories.

This motivates us to study open-vocabulary 3D scene understanding, which equips a model with the ability to localize and recognize open-set classes beyond the label space of an annotated dataset (see Fig. 1). Recently, vision-language (VL) foundation models [34, 22, 48] trained on billions of web-crawled image data with semantic-rich captions [37] are capable of learning adequate vision-language embeddings to connect text and image, which are further leveraged to solve many 2D open-vocabulary tasks including object detection [15, 36], semantic segmentation [44, 26, 52], visual question answering [32] and *etc.* Albeit significantly advancing open-vocabulary image understanding tasks, this pre-training paradigm is not directly viable in the 3D domain due to the absence of large-scale 3D-text pairs.

To this end, initial efforts [51, 20] have attempted to project 3D data into 2D modalities, such as RGB images and depth maps, enabling pre-trained VL foundation models to process the 2D data and achieve object-level open-vocabulary recognition. Nevertheless, this line of methods suffers from several major issues, making it suboptimal to handle scene-level understanding tasks (*e.g.*, instance segmentation). First, multiple RGB images and depth maps are required to represent a 3D sample, which incurs heavy computation and memory costs during training and inference. Second, the projection from 3D to 2D induces information

*Equal contribution: {ryding, jhyang}@eee.hku.hk

†Part of the work is done during an internship at ByteDance AI Lab.

‡Corresponding authors: song.site@gmail.com, xjq@eee.hku.hk

loss and prohibits direct learning from rich 3D data, leading to subpar performance. Our preliminary study shows the cutting-edge 2D open-vocabulary semantic segmentation method MaskCLIP [52] attains a mere 17.8% mIoU with a 20-fold increase in latency when applied to analyze projected 2D images from 3D ScanNet dataset.

Thus, considering the success of VL foundation models for a variety of vision-language tasks [15, 36, 44, 26, 52, 51, 20], we ask: *is it possible to elicit knowledge encoded in powerful VL foundation models to build an explicit association between 3D and language for open-vocabulary understanding?* To this end, our core idea is to exploit pre-trained VL foundation models [1, 40] to caption easily-obtained image data aligned with 3D data (*i.e.* the point set in the corresponding frustum to produce the image). Note that these images can be acquired through neural rendering [9, 47] or from the 3D data collection pipeline [7]. By doing so, we can distill semantic-rich textual descriptions to the 3D domain, which allows explicit association between 3D and vocabulary-rich text for zero-shot 3D scene understanding.

Given 3D-language association, the next question is enabling a 3D network to learn language-aware embeddings from (pseudo) captions. The key challenge stems from intricate object compositions in 3D scene-level data (see Fig. 3), making it difficult to connect objects with corresponding words in the caption. This differs from object-centric image data containing a single centered object [34]. Fortunately, the captioned multi-view images from a 3D scene are related by 3D geometry, which can be leveraged to build hierarchical point-caption pairs, including scene-, view- and entity-level captions. These multi-level point-caption pairs offer coarse-to-fine supervision signals, facilitating learning adequate visual-semantic representations from rich vocabulary by contrastive learning. Without task-specific design, our **Point-Language Association** paradigm, namely PLA, is generic for various open-vocabulary 3D scene understanding tasks, such as semantic and instance segmentation.

Experimental results for ScanNet [7] and S3IDS [2] datasets show the effectiveness of our method in in-domain open-vocabulary tasks with only category shifts, *i.e.* training and evaluation are conducted on the same dataset, surpassing baselines by 25.8% ~ 44.7% hIoU on semantic segmentation and 14.5% ~ 50.4% hAP₅₀ on instance segmentation. Besides, our model, trained on a dataset (*i.e.* ScanNet), can generalize to another dataset (*i.e.* S3IDS) with both data distribution and category shifts, manifesting its transferability. Finally, our model can benefit from more advanced foundation models that provide higher-quality caption supervision, showing its scalability and extensibility.

2. Related Work

3D scene understanding focuses on understanding the semantic meaning of objects and surrounding environment from point clouds. In this work, we focus on two fundamen-

tal scene understanding tasks: semantic and instance segmentation. *3D semantic segmentation* aims to obtain point-wise semantic predictions for point clouds. Representative works develop point-based solutions [33, 19] with elaborately designed point convolution operations [38, 43] or transformers [24] or voxel-based [13, 6] methods with 3D sparse convolutions [14] to produce point-wise segmentation results. *3D instance segmentation* further targets distinguishing different object instances based on semantic segmentation. Existing approaches either adopt a top-down solution [46, 45] via predicting 3D bounding box followed by mask refinement, or a bottom-up [23, 39] approach through grouping points. However, existing methods cannot recognize open-set novel categories, which we aim to address.

Zero-shot and open-vocabulary understanding aims to recognize novel classes that are not annotated in training data. Early approaches mainly follow zero-shot settings that can be coarsely grouped into discriminative methods [41, 3] and generative methods [4, 16]. 3DGenZ [28] extends [4] to the 3D scenario for zero-shot semantic segmentation. Going beyond zero-shot learning, the more general open-vocabulary setting assumes a large vocabulary corpus is accessible during training [50]. Existing *2D open-vocabulary learning* works either exploit massive annotated image-text pairs to provide weak supervision for expanding vocabulary size [50, 54] or leverage pre-trained VL models from large-scale image-caption pairs, such as CLIP [34], to address open-vocabulary recognition where knowledge distillation [36, 15, 49] and prompt learning [12, 11] are studied.

In comparison, *3D open-vocabulary learning* is still in its infancy with only a few explorations focusing on object classification [51, 20]. They attempt to project object-level 3D point clouds to multi-view 2D images and depth maps to adopt the pre-trained VL model to generate open-vocabulary predictions, which, however, suffer from heavy computation and poor performance if applied to 3D scene understanding tasks. In this work, we propose a language-driven 3D open-vocabulary framework that directly associates 3D with text descriptions leveraging multi-view images and VL foundation models. It can be generally applied to various scene understanding tasks and is efficient with only the 3D network employed in training and inference.

3. Method

3.1. Preliminary

3D open-vocabulary scene understanding aims to localize and recognize unseen categories without corresponding human annotation as supervision. Formally, annotations on semantic and instance levels $\mathcal{Y} = \{\mathbf{y}^{\text{sem}}, \mathbf{y}^{\text{ins}}\}$ are divided into base \mathcal{C}^B and novel \mathcal{C}^N categories. In the training stage, the 3D model can access all point clouds $\mathcal{P} = \{\mathbf{p}\}$ but only annotations for base classes \mathcal{Y}^B , unaware of both annotations \mathcal{Y}^N and category names concerning novel classes \mathcal{C}^N .

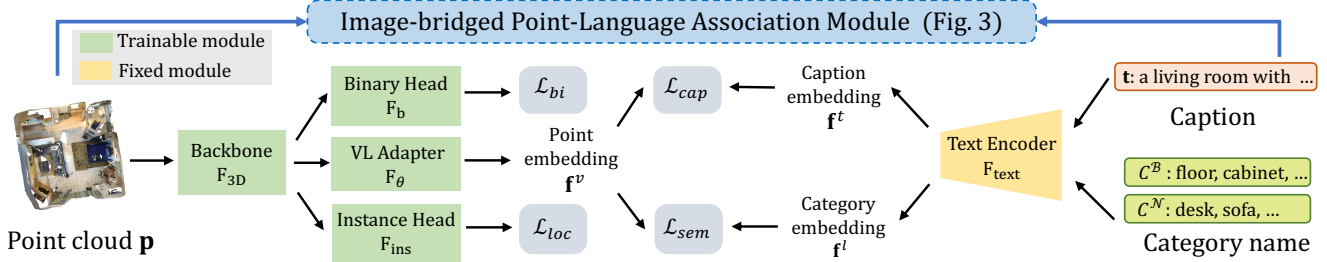


Figure 2. Our language-driven 3D scene understanding paradigm. Different from the close-set network, the learnable semantic head is replaced by category embeddings encoded by a text encoder from category names. Binary head is to rectify semantic scores with base and novel probability as conditions. Instance head is tailored to instance segmentation. Most importantly, to endow the model with rich semantic space to improve open-vocabulary capability, we supervise point embeddings with caption embeddings based on point-language association (see Fig. 3 for details). Best viewed in color.

However, during inference, the 3D model needs to localize objects and classify points belonging to both base and novel $\mathcal{C}^B \cup \mathcal{C}^N$ categories.

As for a typical scene understanding network, it consists of a 3D encoder F_{3D} , a dense semantic classification head F_{sem} and an instance localization head F_{loc} (see Suppl. for details). Its inference pipeline can be demonstrated below,

$$\mathbf{f}^p = F_{3D}(\mathbf{p}), \quad \mathbf{s} = \sigma \circ F_{sem}(\mathbf{f}^p), \quad \mathbf{z} = F_{loc}(\mathbf{f}^p, \mathbf{s}), \quad (1)$$

where \mathbf{p} is the input point cloud, \mathbf{f}^p is point-wise visual feature, \mathbf{s} is semantic score, \mathbf{z} is the instance proposal output and σ is the softmax function. With these network predictions, we can then calculate semantic classification loss \mathcal{L}_{sem} with semantic label \mathbf{y}^{sem} , and localization loss \mathcal{L}_{loc} with instance label \mathbf{y}^{ins} similar to [23, 39] as Eq (2). Notice that \mathbf{y}^{sem} and \mathbf{y}^{ins} only relate to base categories \mathcal{C}^B .

$$\mathcal{L}_{sem} = \text{Loss}(\mathbf{s}, \mathbf{y}^{sem}), \quad \mathcal{L}_{loc} = \text{Loss}(\mathbf{z}, \mathbf{y}^{ins}). \quad (2)$$

3.2. Open-Vocabulary Setups

Though we can train a scene understanding model with loss functions in Eq. (2), it is actually a close-set model with a close-set classifier F_{sem} , incapable of recognizing unseen categories. In this regard, we introduce the text-embedded semantic classifier to obtain an open-vocabulary model and propose a binary calibration module to correct the bias toward base categories for open-vocabulary inference.

3.2.1 Text-Embedded Semantic Classifier

First, as shown in Fig. 2, to make the model become an open-vocabulary learner, we replace its learnable semantic classifier F_{sem} with category embeddings \mathbf{f}^l and a learnable vision-language adapter F_θ to match the dimension between 3D features \mathbf{f}^p and \mathbf{f}^l as follows,

$$\mathbf{f}^v = F_\theta(\mathbf{f}^p), \quad \mathbf{s} = \sigma(\mathbf{f}^l \cdot \mathbf{f}^v), \quad (3)$$

where \mathbf{f}^v is the projected features with the VL adapter F_θ , $\mathbf{f}^l = [\mathbf{f}_1^l, \mathbf{f}_2^l, \dots, \mathbf{f}_k^l]$ is a series of category embeddings obtained by encoding category names \mathcal{C} with a frozen text encoder F_{text} such as BERT [10] or CLIP [34] (see Fig. 2). The prediction is made by calculating the cosine similarity among projected point features \mathbf{f}^v and categories \mathbf{f}^l and

then selecting the most similar category. Notice that \mathbf{f}^l only contains embeddings belonging to base categories \mathcal{C}^B during training, but embeddings related to both base and novel classes $\mathcal{C}^B \cup \mathcal{C}^N$ are used during open-vocabulary inference. With category embeddings \mathbf{f}^l as a classifier, the model can support open-vocabulary inference with any desired categories. The above design generally follows LSeg [26] and is named LSeg-3D as a baseline.

3.2.2 Semantic Calibration with Binary Head

Although the model has already possessed open-vocabulary capability, we empirically find that it can hardly make any correct predictions on novel classes but mistakes them for base classes. As the model is only trained to recognize base categories, it inevitably produces over-confident predictions on base classes regardless of their correctness, also known as the calibration problem [17]. To this end, we propose a binary calibration module to rectify semantic scores with the probability of a point belonging to base or novel classes.

Specifically, as shown in Fig. 2, the binary head F_b is employed to distinguish annotated (*i.e.* base) and unannotated (*i.e.* novel) points. During training, F_b is optimized with:

$$\mathbf{s}^b = F_b(\mathbf{f}^p), \quad \mathcal{L}_{bi} = \text{BCELoss}(\mathbf{s}^b, \mathbf{y}^b), \quad (4)$$

where $\text{BCELoss}(\cdot, \cdot)$ is the binary cross-entropy loss, \mathbf{y}^b is the binary label and \mathbf{s}^b is the predicted binary score indicating the probability that a point belongs to novel categories. In the inference stage, we then exploit the binary probability \mathbf{s}^b to correct the over-confident semantic score \mathbf{s} as follows,

$$\mathbf{s} = \mathbf{s}_B \cdot (1 - \mathbf{s}^b) + \mathbf{s}_N \cdot \mathbf{s}^b, \quad (5)$$

where \mathbf{s}_B is the semantic score computed solely on base classes with novel class scores set to zero. Similarly, \mathbf{s}_N is computed only for novel classes, setting base class scores to zero. We empirically show that the probability calibration largely improves the performance of both base and novel categories (see Sec. 5), demonstrating that our design effectively corrects over-confident semantic predictions.

3.3. Image-Bridged Point-Language Association

With the text-embedded classifier and the binary semantic calibration module, we obtain a deep model with

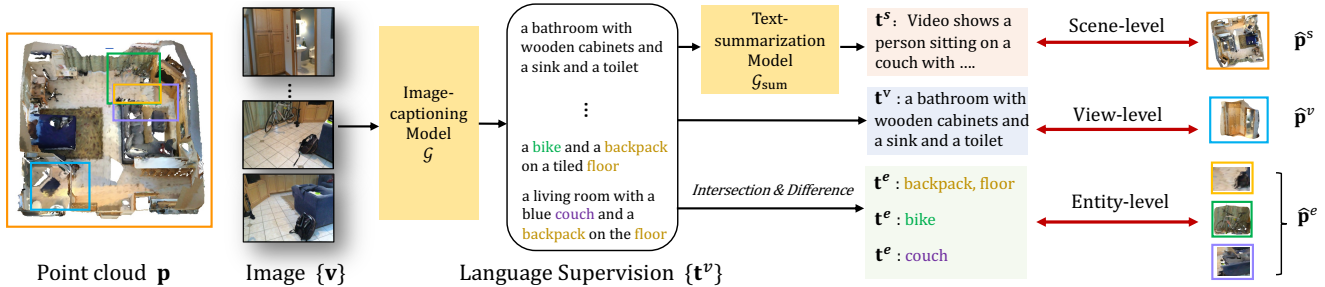


Figure 3. Image-bridged point-language association. We present hierarchical scene-level, view-level and entity-level point-language association manners to assign partial point set with caption supervision through multi-view RGB images and VL foundation models.

open-vocabulary capability. Nevertheless, its performance on novel categories is very close to random guesses as shown in Table 5. Recent success of open-vocabulary works [26, 36, 15] in 2D vision community shows the effectiveness of introducing language supervision to guide vision backbones. Language supervision can not only enable the vision backbone to access abundant semantic concepts with a large vocabulary size but also assist in mapping vision and language features into a common space to facilitate multi-modality downstream tasks. However, Internet-scale paired point-text data are not as readily available as image-text pairs on social media, which largely hinders the development of language-driven 3D understanding.

To address this challenge, we propose PLA, an image-bridged point-language association module to provide language supervision for 3D scene perception without human annotation (see Fig. 2 & Fig. 3). Our core idea is to use multi-view images of a 3D scene as a bridge to access knowledge encoded in VL foundation models. As shown in Fig. 3, a text description is first generated by a powerful image-captioning model taking images of 3D scenes as input, and then associated with a set of points in the 3D scene using the projection matrix between images and 3D scenes. We elaborate on our captioning procedure as well as the designed hierarchical point-caption association as follows.

3.3.1 Caption Multi-View Images

As image captioning is a fundamental task in VL research area [18], various foundation models [40, 1, 30] trained with massive samples are readily available for solving this task. Specifically, taking the j^{th} image of the i^{th} scene \mathbf{v}_{ij} as input, the pre-trained image-captioning model \mathcal{G} can generate its corresponding language description \mathbf{t}_{ij}^v as follows,

$$\mathbf{t}_{ij}^v = \mathcal{G}(\mathbf{v}_{ij}). \quad (6)$$

Surprisingly, though \mathcal{G} has not been specifically trained on the 3D scene understanding dataset, the entities in generated captions already cover the whole semantic label space of the popular 3D scene understanding dataset ScanNet [7]. In addition, the caption \mathbf{t} provides fairly accurate and comprehensive descriptions for room types, semantic categories with color and texture attributes, and even spatial relations

(see language supervision $\{\mathbf{t}^v\}$ examples in Fig. 3 and more examples in Suppl.).

3.3.2 Associate Point Cloud with Language

Given the image-caption pairs, the next step is to connect a point set $\hat{\mathbf{p}}$ to language \mathbf{t} with images \mathbf{v} as bridge as follows:

$$\text{Explore } \langle \hat{\mathbf{p}}, \mathbf{t} \rangle \text{ with } \langle \hat{\mathbf{p}}, \mathbf{v} \rangle \text{ and } \langle \mathbf{v}, \mathbf{t} \rangle. \quad (7)$$

Here, we propose three association fashions on point sets with different spatial scales.

Scene-Level Point-Caption Association. The simplest and coarsest association manner is to link language supervision to all points in a given 3D point cloud scene $\hat{\mathbf{p}}^s = \mathbf{p}$. As illustrated in Fig. 3, we take all 2D image captions \mathbf{t}_{ij}^v of a given scene \mathbf{p}_j to obtain a scene-level caption \mathbf{t}_j^s via a text summarizer [25] \mathcal{G}_{sum} as follows:

$$\mathbf{t}_j^s = \mathcal{G}_{\text{sum}}(\{\mathbf{t}_{1j}^v, \mathbf{t}_{2j}^v, \dots, \mathbf{t}_{n_j j}^v\}), \quad (8)$$

where n_j is the number of images of scene \mathbf{p}_j . By forcing each scene \mathbf{p} to learn from the corresponding scene descriptions \mathbf{t}^s , abundant vocabulary and visual-semantic relationships are introduced to improve the language understanding capability of a 3D network. Despite the simplicity of scene-level caption, we empirically find that it can lift the model’s open-vocabulary capability by a large margin (see Sec. 5).

View-Level Point-Caption Association. Albeit effective, scene-level caption only provides a single caption for all points in a scene, which overlooks the relation of language to local 3D point clouds, rendering it sub-optimal for scene understanding tasks. In this regard, we further propose a view-level point-caption association that leverages the geometrical relationship between image and points to assign each image caption \mathbf{t}^v with a point set inside the 3D view frustum $\hat{\mathbf{p}}^v$ of the given image \mathbf{v} (see blue box in Fig. 3). Specifically, to obtain the view-level point set $\hat{\mathbf{p}}^v$, we first back-project the RGB image \mathbf{v} to 3D space using the depth information \mathbf{d} to get its corresponding point set $\hat{\mathbf{p}}$:

$$[\hat{\mathbf{p}} \mid \mathbf{1}] = \mathbf{T}^{-1} [\mathbf{v} \mid \mathbf{d}], \quad (9)$$

where $[\cdot \mid \cdot]$ denotes block matrix, $\mathbf{T} \in \mathbb{R}^{3 \times 4}$ is the projection matrix comprising of camera intrinsic matrix and rigid transformations obtained by sensor configurations or mature SLAM approaches [8]. As back-projected points $\hat{\mathbf{p}}$ and

points in 3D scene \mathbf{p} may be only partially overlapped, we then compute their overlapped regions to get the view-level point set $\hat{\mathbf{p}}^v$ as follows,

$$\hat{\mathbf{p}}^v = V^{-1}(R(V(\hat{\mathbf{p}}), V(\mathbf{p}))), \quad (10)$$

where V and V^{-1} are the voxelization and reverse-voxelization processes, and R denotes the radius-based nearest-neighbor search [53]. Such a view-based association enables the model to learn with region-level language description, which largely strengthens the model’s recognition and localization ability on unseen categories.

Entity-Level Point-Caption Association. Although view-level caption can already associate each image-caption \mathbf{t}^v with a concrete partial point set in a 3D scene, such an association still constructs on a large 3D area (*i.e.* around 25K points) with multiple semantic objects/categories as shown in Fig. 3. This is not friendly for the 3D network to learn fine-grained point-wise semantic attributes and instance-wise position information from caption supervision. In this regard, we further propose a fine-grained point-language association that owns the potential to build entity-level point-caption pairs, *i.e.* object instances with a caption.

Specifically, as illustrated in Fig. 3, we leverage the differences and intersections of adjacent view-level point sets $\hat{\mathbf{p}}^v$ and their corresponding view-caption \mathbf{t}^v to obtain the entity-level associated points $\hat{\mathbf{p}}^e$ and caption \mathbf{t}^e . First, we calculate entity-level caption \mathbf{t}^e as below:

$$w_i = E(\mathbf{t}_i^v), \quad (11)$$

$$w_{i \setminus j} = w_i \setminus w_j, \quad w_{j \setminus i} = w_j \setminus w_i, \quad w_{i \cap j} = w_i \cap w_j, \quad (12)$$

$$\mathbf{t}^e = \text{Concate}(w^e), \quad (13)$$

where E denotes extracting a set of entity words w from caption \mathbf{t}^v , \setminus and \cap represent the set difference and intersection separately, and Concate denotes the concatenation of all words with spaces to form an entity-level caption \mathbf{t}^e . Similarly, we can easily calculate entity-level point sets and associate them to previously obtained entity-level captions to form point-caption pairs as below:

$$\hat{\mathbf{p}}_{i \setminus j}^e = (\hat{\mathbf{p}}_i^v \setminus \hat{\mathbf{p}}_j^v), \quad \hat{\mathbf{p}}_{j \setminus i}^e = (\hat{\mathbf{p}}_j^v \setminus \hat{\mathbf{p}}_i^v), \quad \hat{\mathbf{p}}_{i \cap j}^e = \hat{\mathbf{p}}_i^v \cap \hat{\mathbf{p}}_j^v, \quad (14)$$

$$\langle \hat{\mathbf{p}}_{i \setminus j}^e, \mathbf{t}_{i \setminus j}^e \rangle, \langle \hat{\mathbf{p}}_{j \setminus i}^e, \mathbf{t}_{j \setminus i}^e \rangle, \langle \hat{\mathbf{p}}_{i \cap j}^e, \mathbf{t}_{i \cap j}^e \rangle. \quad (15)$$

With entity-level $(\hat{\mathbf{p}}^e, \mathbf{t}^e)$ pairs, we further filter them to ensure each entity-level points set $\hat{\mathbf{p}}^e$ relates to at least one entity and focuses on a small enough 3D space as follows,

$$\gamma < |\hat{\mathbf{p}}^e| < \delta \cdot \min(|\hat{\mathbf{p}}_i^v|, |\hat{\mathbf{p}}_j^v|) \quad \text{and} \quad |\mathbf{t}^e| > 0, \quad (16)$$

where γ is a scalar to define minimal number of points, δ is a ratio to control the maximum size of $\hat{\mathbf{p}}^e$, and caption \mathbf{t}^e is not empty. Such a constraint helps focus on a fine-grained 3D space with fewer entities in each caption supervision.

Comparison Among Different Point-Caption Association Manners. The above-proposed three coarse-to-fine point-caption association manners actually hold different merits and drawbacks. As shown in Table 1, the scene-level association has the simplest implementation but obtains the

	scene-level	view-level	entity-level
complexity	simplest	middle	hardest
# captions	1,201	24,902	6,163
# points for each caption	145,171	24,294	3,933

Table 1. Comparison among point-caption association manners.

coarsest correspondence between captions and points (*i.e.* each caption corresponds to over 140K points); the view-level association provides point-language mapping relation at a finer level, enjoying a larger semantic label space (*i.e.* over 20× more captions) and a more localized point set (*i.e.* around 6× fewer corresponding points per caption) than scene caption; the entity-level association owns the most fine-grained correspondence relation, matching each caption to only 4K points on average, and thus can further benefit dense prediction and instance localization in downstream tasks. We empirically show that the fine-grained association and the semantic-rich label space are two important factors for open-vocabulary perception tasks (see Sec. 5).

3.4. Contrastive Point-Language Training

With obtained point-caption pairs $(\hat{\mathbf{p}}, \mathbf{t})$, we are ready to guide the 3D network F_{3D} to learn from vocabulary-rich language supervisions. Here, we introduce a general point-language feature contrastive learning that can be applied to all kinds of coarse-to-fine point-caption pairs.

Specifically, we first obtain caption embeddings \mathbf{f}^t with a pre-trained text encoder F_{text} . As for the associated partial point set $\hat{\mathbf{p}}$, we select its corresponding point-wise features from adapted features \mathbf{f}^v and leverage global average pooling to obtain its feature vector $\mathbf{f}^{\hat{\mathbf{p}}}$ as follows,

$$\mathbf{f}^t = F_{\text{text}}(\mathbf{t}), \quad \mathbf{f}^{\hat{\mathbf{p}}} = \text{Pool}(\hat{\mathbf{p}}, \mathbf{f}^v). \quad (17)$$

We then adopt contrastive loss as [50] to pull corresponding point-caption feature embeddings closer and push away unrelated point-caption features as follows,

$$\mathcal{L}_{\text{cap}} = -\frac{1}{n_t} \sum_{i=1}^{n_t} \log \frac{\exp(\mathbf{f}_i^{\hat{\mathbf{p}}} \cdot \mathbf{f}_i^t / \tau)}{\sum_{j=1}^{n_t} \exp(\mathbf{f}_i^{\hat{\mathbf{p}}} \cdot \mathbf{f}_j^t / \tau)}, \quad (18)$$

where n_t is the number of point-caption pairs in any given association fashion and τ is a learnable temperature to modulate the logits as CLIP [34]. It is also worth noting that we remove duplicate captions in a batch to avoid noisy optimization during contrastive learning. With Eq. (17) and Eq. (18), we can easily compute caption losses on scene-level $\mathcal{L}_{\text{cap}}^s$, view-level $\mathcal{L}_{\text{cap}}^v$ and entity-level $\mathcal{L}_{\text{cap}}^e$. Our final caption loss is a weighted combination as follows,

$$\mathcal{L}_{\text{cap}}^{\text{all}} = \alpha_1 * \mathcal{L}_{\text{cap}}^s + \alpha_2 * \mathcal{L}_{\text{cap}}^v + \alpha_3 * \mathcal{L}_{\text{cap}}^e, \quad (19)$$

where α_1 , α_2 and α_3 are trade-off factors. As shown in Fig. 2, the overall training objective can be written as

$$\mathcal{L} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{cap}}^{\text{all}} + \mathcal{L}_{\text{bi}}. \quad (20)$$

Method	C^N prior	ScanNet									S3DIS					
		B15/N4			B12/N7			B10/N9			B8/N4			B6/N6		
		hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N
LSeg-3D [26]	×	00.0	64.4	00.0	00.9	55.7	00.1	01.8	68.4	00.9	00.1	49.0	00.1	00.0	30.1	00.0
3DGenZ [28]	✓	20.6	56.0	12.6	19.8	35.5	13.3	12.0	63.6	06.6	08.8	50.3	04.8	09.4	20.3	06.1
3DTZSL [5]	✓	10.5	36.7	06.1	03.8	36.6	02.0	07.8	55.5	04.2	08.4	43.1	04.7	03.5	28.2	01.9
PLA (w/o Cap.)	×	39.7	68.3	28.0	24.5	70.0	14.8	25.7	75.6	15.5	13.0	58.0	07.4	12.2	54.5	06.8
PLA	×	65.3	68.3	62.4	55.3	69.5	45.9	53.1	76.2	40.8	34.6	59.0	24.5	38.5	55.5	29.4
PLA (w/ self-train)	✓	70.3	68.9	71.7	61.1	70.4	54.0	59.2	76.9	48.2	36.1	59.7	26.0	46.7	58.9	38.7
Fully-Sup.	✓	73.3	68.4	79.1	70.6	70.0	71.8	69.9	75.8	64.9	67.5	61.4	75.0	65.4	59.9	72.0

Table 2. Results for open-vocabulary 3D semantic segmentation on ScanNet and S3DIS in terms of hIoU, mIoU^B and mIoU^N. C^N prior denotes whether novel category names C^N need to be known during training. PLA (w/o Cap.) denotes training without point-caption pairs as supervision. Best open-vocabulary results are highlighted in **bold**.

Method	C^N prior	ScanNet									S3DIS					
		B13/N4			B10/N7			B8/N9			B8/N4			B6/N6		
		hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N
LSeg-3D [26]	×	05.1	57.9	02.6	02.0	50.7	01.0	02.4	59.4	01.2	00.5	58.3	00.3	01.1	41.4	00.5
PLA (w/o Cap.)	×	21.0	59.6	12.6	11.1	56.2	06.2	15.9	63.2	09.1	01.8	59.3	00.9	01.3	49.2	01.2
PLA	×	55.5	58.5	52.9	31.2	54.6	21.9	35.9	63.1	25.1	15.0	59.0	08.6	16.0	46.9	09.8
PLA (w/ self-train)	✓	58.6	58.0	59.2	41.4	56.9	32.6	42.1	61.1	32.1	26.7	60.3	17.2	23.4	45.6	15.8
Fully-Sup.	✓	64.5	59.4	70.5	62.5	57.6	62.0	62.0	65.1	62.0	57.6	60.8	54.6	57.4	50.0	67.5

Table 3. Results for open-vocabulary 3D instance segmentation on ScanNet and S3DIS in terms of hAP₅₀, mAP₅₀^B and mAP₅₀^N.

4. Experiments

4.1. Basic Setups

Datasets and Perception Tasks. To validate the effectiveness of our point-language association paradigm, we conduct experiments on two datasets: ScanNet [7] densely annotated in 20 classes and S3DIS [2] with 13 classes on both semantic and instance segmentation tasks.

Category Partitions. Without standard open-vocabulary partitions on these two datasets, we build an open-vocabulary benchmark with multiple base/novel partitions. To circumvent model confusion, we disregard the “otherfurniture” class in ScanNet and the “clutter” class in S3DIS as they lack exact semantic meanings and can include any semantic categories. As for ScanNet, we randomly partition the rest 19 classes into 3 base/novel partitions for semantic segmentation, *i.e.* B15/N4, B12/N7 and B10/N9, where B15/N4 indicates 15 base and 4 novel categories. We also follow SoftGroup [39] to exclude two background classes and thus obtain B13/N4, B10/N7, and B8/N9 partitions for instance segmentation on ScanNet. As for S3DIS, we randomly shuffle the rest 12 classes into 2 base/novel splits, *i.e.* B8/N4, B6/N6 for both semantic and instance segmentation. Specific category splits are presented in the Suppl..

Metrics. We employ widely adopted mean intersection over union (mIoU) and mean average precision under 50% IoU threshold (mAP₅₀) as evaluation metrics for semantic and instance segmentation, respectively. These metrics are calculated on base and novel classes separately with superscripts of B and N (*e.g.* mIoU^B). Further, we use harmonic mean IoU (hIoU) and AP₅₀ (hAP₅₀) as major indicators following popular zero-shot learning works [41, 44] to consider category partition between base and novel.

Architectures and Baseline Methods. We adopt the popular and high-performance sparse convolutional UNet [13, 6] as 3D encoder F_{3D} , the text encoder of CLIP as F_{text} , two fully-connected layers with batch normalization [21] and ReLU [31] as VL adapter F_{θ} , an UNet decoder as binary head F_b . Also, we utilize the state-of-the-art instance segmentation network SoftGroup [39] for instance head F_{ins} .

As for baseline methods, other than the above-mentioned **LSeg-3D** in Sec.3.2.1, we also re-produce two 3D zero-shot learning methods **3DGenZ** [28] and **3DTZSL** [5] with task-tailored modifications. The implementation details are provided in the Suppl..

4.2. Main Results

3D Semantic Segmentation. As shown in Table 2, compared to LSeg-3D [26] baseline, our method obtains around 51.3% ~ 65.3% and 34.5% ~ 38.5% hIoU improvements among different partitions on ScanNet and S3DIS respectively, demonstrating its superior open-vocabulary capability. Even compared to previous zero-shot methods 3DGenZ [28] and 3DTZSL [5] that know novel category names during training, our method still obtains 35.5% ~ 54.8% improvements in terms of hIoU among various partitions on ScanNet. Especially, our PLA trained model largely surpasses its no caption supervision counterparts (*i.e.* PLA (w/o Cap.)) by 25.6% ~ 30.8% hIoU and 21.6% ~ 26.3% hIoU on ScanNet and S3DIS, respectively. It is noteworthy that the improvement from our method is consistent on different base/novel partitions and datasets, further illustrating its robustness and effectiveness.

3D Instance Segmentation. As demonstrated in Table 3, our method remarkably surpasses baseline methods by 29.2% ~ 50.4% hAP₅₀ and 14.5% ~ 14.9% hAP₅₀ among different base/novel partitions on ScanNet and S3DIS, re-

spectively. Such outstanding performance indicates our contrastive point-language training helps the 3D backbone learn not only semantic attributes but also instance localization information from captions. Notice that the improvement for S3DIS is slighter than ScanNet on both semantic segmentation and instance segmentation. This is actually caused by S3DIS’s small number of training samples (only 271 scenes) and much fewer point-caption pairs owing to fewer overlapped regions between images and 3D scenes.

Self-Bootstrap with Novel Category Prior. As some existing zero-shot methods (*i.e.* 3DGenZ [28] and 3DTZSL [5]) can access novel category names but no human-annotation during training, here we also provide a simple variant to leverage such novel category prior in self-training fashion [42]. As shown in Table 2 and 3, PLA (w/ self-train) obtains around 2% ~ 12% gains among semantic and instance segmentation on two datasets. This demonstrates that our model can further self-bootstrap its zero-shot capability and extend its vocabulary size without any human annotation.

4.3. Zero-shot Domain Transfer

Our method already shows excellent potential in solving in-domain open-vocabulary scene understanding tasks with category shifts. However, transferable open-vocabulary learners across different domains/datasets also merit exploration, as they face both category and data distribution shifts. In this regard, we conduct zero-shot domain transfer experiments that train the model on ScanNet’s base classes and test it on all S3DIS classes without fine-tuning. Notably, S3DIS has 4 categories not present in ScanNet. As shown in Table 4, our PLA consistently outperforms LSeg-3D [26] by 7.7% ~ 18.3% mIoU for semantic segmentation and 5.0% ~ 9.5% mAP₅₀ for instance segmentation. Such outstanding improvements substantiate our model’s generality for both category shift and data distribution shift. Note that we do not use the binary head for domain transfer here, as the base/novel partition is dataset-specific. We leave calibrating base and novel semantic predictions in out-of-domain open-vocabulary scenarios to future work.

5. Ablation Studies

In this section, we examine key components of our framework through in-depth ablation studies. Experiments are conducted on ScanNet B15/N4 partition by default. The default setting is marked in gray.

ScanNet partition	S3DIS Semantic (mIoU)		S3DIS Instance (mAP ₅₀)	
	LSeg-3D	PLA	LSeg-3D	PLA
B19/N0	42.5	50.2 (+7.7)	37.5	43.6 (+6.1)
B15/N4	30.2	48.5 (+18.3)	31.2	40.7 (+9.5)
B12/N7	26.1	38.3 (+12.2)	28.2	35.1 (+6.9)
B10/N9	34.5	48.1 (+13.6)	33.8	38.8 (+5.0)

Table 4. Zero-shot domain transfer results for semantic segmentation and instance segmentation on ScanNet → S3DIS.

Component Analysis. We investigate the effectiveness of

our proposed binary calibration module and three coarse-to-fine point-caption supervision here. As shown in Table 5, adopting binary head for semantic calibration greatly surpasses baseline LSeg-3D by 39.8% hIoU on semantic segmentation and 15.9% hAP₅₀ on instance segmentation. Such performance lifts on both base and novel classes verify that it correctly rectifies semantic scores.

As for point-caption association manners, they all substantially improve results by a large margin of 14.8% ~ 23.8% hIoU and 31.8% ~ 35.6% hAP₅₀ on semantic and instance segmentation, respectively. Among three association fashions, entity-level caption supervision performs the best, demonstrating that fine-grained language-point correspondence is one of the most vital considerations for constructing point-caption pairs. Notice that when we combine different types of captions, the model will not always obtain improvements in all scenarios, potentially caused by the difficulty of simultaneously optimizing multiple caption losses with various granularities on some tasks.

Components				hIoU / mIoU ^B / mIoU ^N	hAP ₅₀ / mAP ₅₀ ^B / mAP ₅₀ ^N
Binary	Cap ^s	Cap ^v	Cap ^e		
				00.0 / 64.4 / 00.0	05.1 / 57.9 / 02.6
✓				39.8 / 68.5 / 28.1	21.0 / 59.6 / 12.8
✓	✓			54.6 / 67.9 / 45.7	52.8 / 57.8 / 36.6
✓		✓		61.3 / 68.5 / 55.5	55.9 / 58.9 / 53.3
✓			✓	63.6 / 67.8 / 60.0	56.6 / 59.0 / 54.4
✓	✓	✓		61.9 / 68.1 / 56.8	54.9 / 59.5 / 51.0
✓		✓	✓	65.3 / 68.3 / 62.4	55.5 / 58.5 / 52.9
✓	✓	✓	✓	64.6 / 69.0 / 60.8	54.5 / 58.2 / 51.4

Table 5. Component analysis on ScanNet. Binary denotes binary head calibration. Cap^s, Cap^v and Cap^e denotes scene-level, view-level and entity-level caption supervision, respectively.

Caption Composition Analysis. As a caption can composite entities (*e.g.* sofa), their relationships (*e.g.* spatial relation) and attributes (*e.g.* color and texture), we investigate which types of words mainly contribute to the open-vocabulary capability. As shown in Table 6, when only keeping entity phrases in the caption, (a) variant even outperforms the full caption variant. In addition, if we only keep entities that exactly match category names in captions, obtained (b) variant suffers over 13% mIoU degradation on novel categories, showing that diverse entity words to expand semantic space is a crucial factor for captions. Furthermore, although the (c) variant introduces both correct base and novel label names in the caption, it still obtains slightly inferior performance to our foundation-model-generated caption, illustrating existing foundation models are powerful enough to provide promising supervision.

Caption Composition	hIoU / mIoU ^B / mIoU ^N
(a) keep only entities	65.7 / 69.0 / 62.7
(b) keep only label names	57.6 / 68.5 / 49.6
(c) ground-truth label names	64.8 / 68.1 / 61.9
(d) full caption	65.3 / 68.3 / 62.4

Table 6. Ablation of caption composition.

Text Encoder Selection. Here, we compare different text

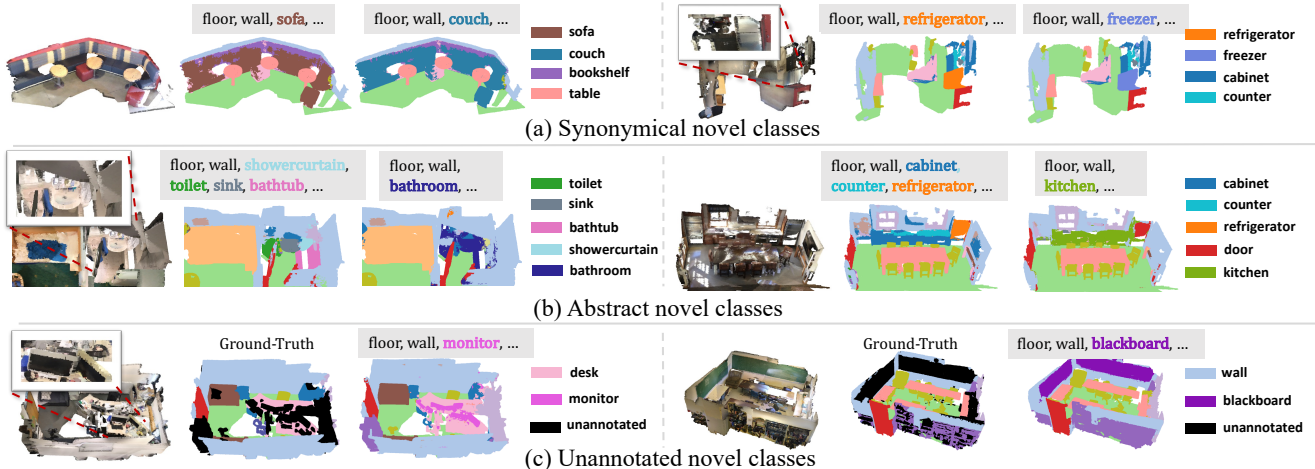


Figure 4. Qualitative results of recognizing out-of-vocabulary classes. (a) demonstrates the results of recognizing synonymical classes. (b) shows the segmentation results on abstract concepts. (c) presents the results of segmenting unannotated categories in the dataset.

encoders F_{text} for extracting caption and category embeddings. As shown in Table 7, the vision-language pre-trained text encoder of CLIP [34] shows over 7% higher mIoU^N than BERT [10] and GPT2 [35] that are only pre-trained on language modality. This demonstrates that the vision-aware text encoder can provide better language embedding for 3D-language tasks since 3D also leverages texture, shape and RGB information as images for recognition.

Text Encoder	BERT [10]	GPT2 [35]	CLIP [34]
$\text{hIoU} / \text{mIoU}^B / \text{mIoU}^N$	61.2 / 68.7 / 55.2	61.0 / 69.1 / 54.6	65.3 / 68.3 / 62.4

Table 7. Ablation of text encoder.

Foundation Model for Image Captioning. By default, we employ one of the most popular open-source image captioning models, GPT-ViT2 [1], on the HuggingFace platform to generate captions in main experiments. However, as shown in Table 8, the recent state-of-the-art foundation model OFA [40] can consistently surpass GPT-ViT2 on three partitions, which reflects the potential of our method to be further boosted with stronger foundation models.

model	$\text{hIoU} / \text{mIoU}^B / \text{mIoU}^N$		
	B15/N4	B12/N7	B10/N9
ViT-GPT2 [1]	65.3 / 68.3 / 62.4	55.3 / 69.5 / 45.9	53.1 / 76.2 / 40.8
OFA [40]	65.6 / 68.3 / 63.1	57.5 / 69.8 / 48.9	56.6 / 75.9 / 45.1

Table 8. Ablation of VL foundation model for image captioning.

6. Qualitative Analysis

To more straightforwardly illustrate the open-vocabulary ability of our method, we present some interesting qualitative results in terms of recognizing synonymical classes, abstract classes and even unannotated classes.

Synonymical Novel Classes. Here, we substitute class names with related but new words for inference. As illustrated in Fig. S9 (a), when we replace “sofa” with “couch” or “refrigerator” with “freezer”, the model still attains a high-quality segmentation mask. This demonstrates our model is robust to recognize synonymical concepts.

Abstract Novel Classes. Apart from object entities, we find the model is able to understand more abstract concepts such as room types. As shown in Fig. S9 (b), by removing “shower curtain”, “toilet”, “sink” and “bathtub” in input categories and adding “bathroom”, the predicted “bathroom” roughly covers the real bathroom region. The right example shows the model can also understand ‘kitchen’ regions. It indicates our model is capable to recognize out-of-vocabulary and abstract concepts beyond concrete semantic objects.

Unannotated Novel Classes. As current 3D datasets fail to annotate all classes due to insufferable annotation costs, our model owns the potential to recognize those unannotated classes with high-quality predictions, facilitating open-world applications. As shown in Fig. S9 (c), the model successfully identifies “monitor” and “blackboard” that are not included in the dataset annotations with accurate masks.

7. Conclusion

We propose PLA, a general and effective language-driven 3D scene understanding framework that enables the 3D model to localize and recognize novel categories. By leveraging images as a bridge, we construct hierarchical point-language pairs harvesting powerful 2D VL foundation models and geometric constraints between 3D scenes and 2D images. We employ contrastive learning to pull features of such associated pairs closer, introducing rich semantic concepts into the 3D network. Extensive experimental results show the superiority of our method on not only in-domain open-vocabulary semantic and instance segmentation, but also challenging out-of-domain zero-shot transfer.

Acknowledgement. This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant no. 17202422), and RGC matching fund scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded

by The Hong Kong Jockey Club Charities Trust.

References

- [1] Vit-gpt2 image captioning. <https://huggingface.co/nlpconnect/vit-gpt2-image-captioning/discussions>.
- [2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016.
- [3] Donghyeon Baek, Youngmin Oh, and Bumsuh Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9536–9545, 2021.
- [4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [5] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 923–933, 2020.
- [6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.
- [7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [8] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017.
- [9] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.
- [12] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. 2022.
- [13] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.
- [14] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.
- [15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.
- [16] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020.
- [17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- [18] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- [19] Qianguai Huang, Weiyue Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2635, 2018.
- [20] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv preprint arXiv:2210.01055*, 2022.
- [21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [22] Chao Jia, Yinfei 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- [23] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [24] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022.
- [25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

- [26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [28] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pages 992–1002. IEEE, 2021.
- [29] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021.
- [30] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- [31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.
- [32] AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. *arXiv preprint arXiv:2205.00949*, 2022.
- [33] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.
- [34] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [36] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022.
- [37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [38] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE International Conference on Computer Vision*, 2019.
- [39] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022.
- [40] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.
- [41] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.
- [42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.
- [43] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021.
- [44] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021.
- [45] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019.
- [46] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019.
- [47] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.
- [48] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021.
- [49] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022.
- [50] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.
- [51] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.
- [52] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022.
- [53] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

- [54] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

Outline

In this supplementary file, we provide more experimental results and details not elaborated on in our main paper due to page length limits:

- Sec. **S1**: Details of our open-vocabulary scene understanding benchmark.
- Sec. **S2**: Limitation analysis of PointCLIP for scene understanding tasks.
- Sec. **S3**: Additional experimental results on re-partition results, per-class results, error bar results, fully-supervised results with caption supervision and combination of caption supervisions.
- Sec. **S4**: Examples of image-caption pairs and hierarchical point-caption pairs.
- Sec. **S5**: Qualitative results of open-vocabulary scene understanding.
- Sec. **S6**: Limitation and open problems.

S1. Implementation Details

Here, we present the implementation details of dataset category partition, network modifications, baseline setups, hyper-parameter configurations and usage of images.

S1.1. Dataset Category Partition

As mentioned in Sec. 4.1 of the main paper, we build a 3D open-vocabulary benchmark on ScanNet [7] and S3DIS [2] with multiple base/novel partitions. ScanNet [7] consists of 1,613 scenes (1,201 scenes for training, 312 scenes for validation and 100 for testing) densely annotated in 20 classes. We discard the ‘otherfurniture’ class and partition the rest 19 classes into three partitions for semantic segmentation as shown in Table S9. Note that the B15/N4 partition adheres to the 3DGenZ [28] partitioning scheme. As for instance segmentation, we follow SoftGroup [39] to ignore two background classes (*i.e.* wall and floor) and obtain corresponding partitions (see Table S10).

S3DIS [2] contains 271 scans across 6 building areas along with 13 categories. Following previous work [33], we treat the 5th area as the validation split and other areas as the training split. We discard the ‘clutter’ class and partition the rest 12 classes into two partitions for both semantic segmentation and instance segmentation as demonstrated in Table S11.

S1.2. Network Modifications

In this section, we elaborate on how to extend a close-set network to an open-vocabulary learner for semantic segmentation and instance segmentation. We employ sparse-convolution-based UNet [13] with a base hidden dimension of 16 as our backbone F_{3D} .

First, as illustrated in Fig. S5 (a), the close-set network contains a learnable semantic head F_{sem} that classifies a fixed number of categories. As discussed in Sec. 3.2 in the main paper, to obtain an open-vocabulary model, we replace the semantic head F_{sem} with a vision-language (VL) adapter F_θ and the category embedding f^l encoded by a fixed text encoder F_{text} . Note that the category embedding f^l can be treated as replacing the weights of the classifier. The category embedding f^l encodes semantic attributes of base classes in the training stage and encodes any desired categories during inference to achieve open-vocabulary semantic segmentation.

Further, as we follow SoftGroup [39] to develop instance head F_{ins} , we modify the close-set designs in SoftGroup to obtain an open-vocabulary instance head. First, as shown in Fig. S6, the seg head and the score head that produce per-class confidence in the vector form are modified to class-agnostic modules that produce a single scalar for each generated instance proposal. In this way, we can train these two heads without needing to know novel categories. Second, the learnable cls head that predicts the classification scores of generated proposals is replaced by the proposal-level pooling of semantic scores s , which can be extended to arbitrary categories. Finally, the class statistics, such as the average number of points in an instance mask for each class, which assists proposal grouping, are removed to avoid leakage of novel class information. We empirically show that those modifications cause little degradation of fully-supervised performance by 1.1% mAP₅₀, as demonstrated in Table S12. Note that we train the model from scratch rather than fine-tuning a supervised pretrained model, as SoftGroup does, to prevent leakage of novel classes during training. Additionally, we use a smaller hidden dimension size for the UNet backbone. Consequently, our reproduced performance differs from that in the original paper.

S1.3. Baseline Setups

As mentioned in Sec. 4.1 of the main paper, we follow LSeg [26] to implement LSeg-3D as a baseline with UNet [13, 6] backbone, vision-language adapter implemented by MLP and the CLIP [34] ViT-B/16 text encoder. For the other two 3D zero-shot methods, 3DGenZ [28] and 3DTZSL [5], we reproduce them with the same network and CLIP text embedding for fair comparisons. Specifically, for 3DGenZ [28], instead of training on samples that only contain base classes, we train it on the whole training dataset with points belonging to novel classes ignored during optimization. Besides, we remove calibrated stacking that aims to alleviate bias towards seen classes since it brings extremely minor performance gains in our implementations. As for 3DTZSL [5] designed for object classification, we extend it to segmentation via learning with triplet loss on the point level instead of the sample level. We implement its projection net with 2 fully-connected layers and the Tanh

Partition	Base Categories	Novel Categories
B15/N4	wall, floor, cabinet, bed, chair, table, door, window, picture, counter, curtain, refrigerator, showercurtain, sink, bathtub	sofa, bookshelf, desk, toilet
B12/N7	wall, floor, cabinet, sofa, door, window, counter, desk, curtain, refrigerator, showercurtain, toilet	bed, chair, table, bookshelf, picture, sink, bathtub
B10/N9	wall, floor, cabinet, bed, chair, sofa, table, door, window, curtain	bookshelf, picture, counter, desk, refrigerator, showercurtain, toilet, sink, bathtub

Table S9. Category partitions for open-vocabulary semantic segmentation on ScanNet.

Partition	Base Categories	Novel Categories
B13/N4	cabinet, bed, chair, table, door, window, picture, counter, curtain, refrigerator, showercurtain, sink, bathtub	sofa, bookshelf, desk, toilet
B10/N7	cabinet, sofa, door, window, counter, desk, curtain, refrigerator, showercurtain, toilet	bed, chair, table, bookshelf, picture, sink, bathtub
B8/N9	cabinet, bed, chair, sofa, table, door, window, curtain	bookshelf, picture, counter, desk, refrigerator, showercurtain, toilet, sink, bathtub

Table S10. Category partitions for open-vocabulary instance segmentation on ScanNet.

activation function, the same as its paper claimed.

S1.4. Hyper-Parameter Configurations

We train 19,216 iterations on ScanNet and 4,080 iterations on S3DIS for semantic segmentation. For instance segmentation, we train 24,020 iterations on ScanNet and 9,160 iterations on S3DIS. The learning rate is initialized as 0.004 with cosine decay. We adopt the AdamW [27] optimizer and run all experiments with 32 batch size on 8 NVIDIA V100 or NVIDIA A100.

For entity-level captions, we filter out some $\langle \hat{\mathbf{p}}^e, \mathbf{t}^e \rangle$ pairs to guarantee the point set $\hat{\mathbf{p}}^e$ is small enough containing only a few entities. Specifically, we set the minimal points γ as 100 and the ratio that controls the maximum number of points δ as 0.3. As for the caption loss, we set α_1 , α_2 and α_3 as 0, 0.05 and 0.05 for scene-level $\mathcal{L}_{\text{cap}}^s$, view-level $\mathcal{L}_{\text{cap}}^v$ and entity-level loss $\mathcal{L}_{\text{cap}}^e$ for ScanNet, respectively. For S3DIS, we set α_1 , α_2 , and α_3 as 0, 0.08, and 0.02 separately.

S1.5. Usage of Images

For ScanNet, we use a 25,000-frame subset[§] from ScanNet images for captioning. For S3DIS, as each scene contains a widely varying number of images, we subsample its images to caption at most 50 images per scene. It is worth noting that some S3DIS scenes lack corresponding images; we consequently cannot provide language supervision for those scenes without images during training.

[§]https://kaldir.vc.in.tum.de/scannet_benchmark/documentation

S2. Analysis of PointCLIP for Scene Understanding

In recent years, 2D open-vocabulary understanding [15, 36, 44, 26] achieves unprecedented success driven by transferable vision-language models such as CLIP [34] trained on large-scale image-caption pairs. Inspired by that success, PointCLIP [20] has made the first attempt to transfer the knowledge of CLIP into the 3D domain for zero-shot and few-shot object classification tasks. PointCLIP projects 3D point clouds into 2D multi-view depth maps and leverages CLIP to process multi-view depth images to obtain predictions. Finally, the predictions are assembled into 3D predictions. Though some progress has been made in object-level understanding, our experimental results show that PointCLIP is not suitable for scene-level understanding tasks with poor performance and heavy inference overheads.

Task-specific modifications. To extend PointCLIP for 3D scene understanding, we make the following modifications. First, we follow the state-of-the-art 2D open-vocabulary semantic segmentation method MaskCLIP [52] to modify the attentive pooling layer of CLIP’s vision encoder for obtaining pixel-wise dense predictions. Second, instead of using self-rendered images, we utilize collected depth images captured by depth sensors since they are realistic with more accurate depth values. We also explore utilizing collected RGB images to avoid modal gaps caused by using depth images. Finally, to assemble multi-view 2D results into 3D, other than voting to get object-wise predictions, we back-project all multi-view image predictions into 3D space via 3D geometry and assign predictions to each point of 3D scenes by searching nearest neighbors in back-projected 3D point clouds.

Partition	Base Categories	Novel Categories
B8/N4	ceiling, floor, wall, beam, column, door, chair, board	window, table, sofa, bookcase
B6/N6	ceiling, wall, beam, column, chair, bookcase	floor, window, door, table, sofa, board

Table S11. Category partitions for open-vocabulary semantic and instance segmentation on S3DIS.

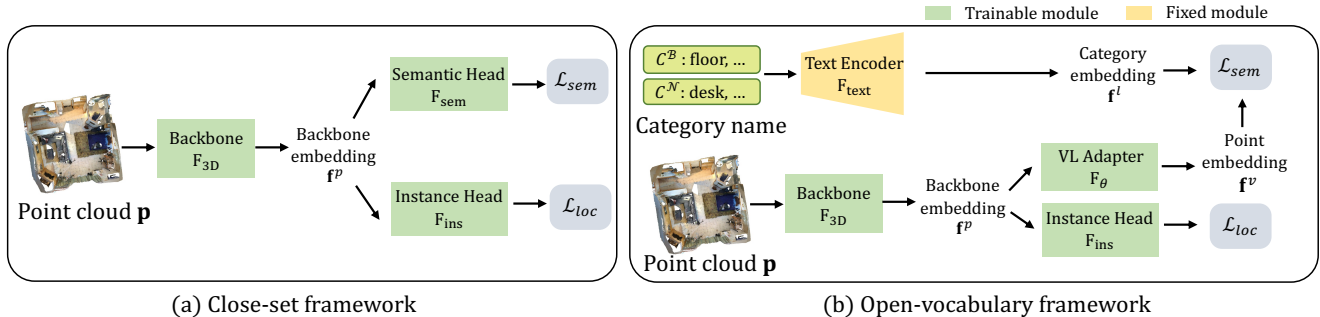


Figure S5. Comparison between close-set scene understanding framework and open-vocabulary scene understanding framework.

Components			mAP ₅₀
per-class seg head and score head	cls head	class statistics	
✓	✓	✓	61.8
	✓	✓	62.0
		✓	61.1
			60.7

Table S12. Fully-supervised instance segmentation results of different SoftGroup variants upon ScanNet in terms of mAP₅₀.

Results. As shown in Table S13, with depth images as

Input	2D mIoU	3D mIoU	latency (ms)
depth images	02.2	01.7	1667
RGB images	17.8	17.2	1667

Table S13. Results of zero-shot 3D semantic segmentation using PointCLIP on ScanNet.

input, the modified PointCLIP obtains only 2.2% mIoU on 2D semantic segmentation with 5,436 validation samples of ScanNet. The assembled 3D prediction only attains 1.7% mIoU on 312 samples, which is very close to random guesses. When alternated to use RGB images as input, the performance lifts to 17.8% mIoU on 2D and 17.2% mIoU on 3D, demonstrating that using RGB images can avoid annoying modal gaps. However, the performance is still moderate, which suggests this projection-based stream of work is sub-optimal for tackling 3D scene understanding tasks. Though further fine-tuning on seen categories might benefit model performance, this line of research has a key limitation: by projecting 3D data to 2D, it suffers from informa-

tion loss and makes the model unable to directly learn from information-rich 3D data.

In addition, to assess the model efficiency, we use latency to measure the execution speed of model inference on a single GeForce RTX 2080Ti. As shown in Table S13, PointCLIP takes an average of 1667ms to process images of one 3D scene, which is rather costly, not to mention the post-processing time for back-projection and results ensemble. Instead, our 3D network only costs 83ms to process one 3D sample, which is 20 times more efficient than PointCLIP.

In sum, the poor zero-shot performance, information loss from projection, and heavy computation costs render this line of work not suitable for 3D scene understanding and prevent us from exploring further on this stream of work.

S3. Additional Experimental Results

S3.1. Re-partition Experiments

Splits	hIoU / mIoU ^B / mIoU ^N	
	LSeg-3D [26]	Ours
random-sample 1	00.0 / 61.7 / 00.0	65.3 / 68.3 / 62.4
random-sample 2	00.0 / 48.5 / 00.0	53.1 / 70.1 / 42.7
random-sample 3	00.3 / 66.1 / 00.2	60.9 / 69.2 / 54.5
frequency-sample	00.0 / 68.7 / 00.0	62.6 / 69.0 / 57.3

Table S14. Results of re-sampled base and novel categories.

To ensure the reliability of results, we randomly re-sample base and novel categories three times and sample it based on class frequency for the B15/N4 ScanNet semantic segmentation task. As shown in Table S14, our method consistently exceeds LSeg-3D baseline among four different splits by a large margin of 53.1% ~ 65.3% hIoU, which

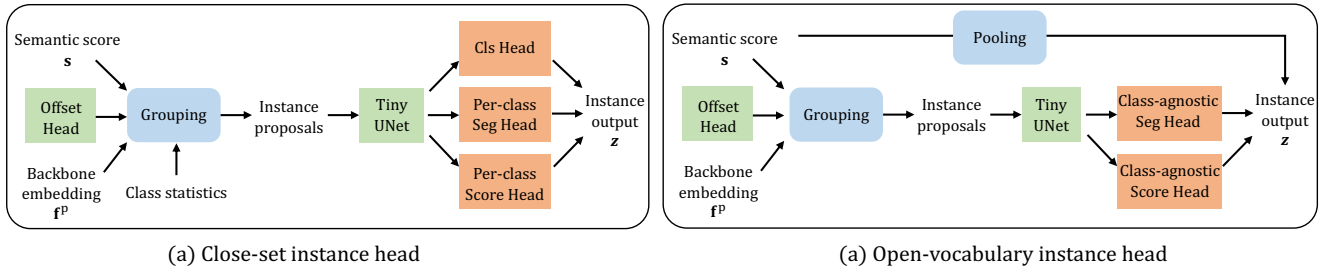


Figure S6. Comparison between close-set instance head and open-vocabulary instance head.

reveals the robustness of our methods in handling different novel classes.

S3.2. Per-class Results

We present per-category performances of our open-vocabulary 3D scene understanding framework on semantic and instance segmentation. As shown in Table S15 and Table S16, novel classes generally perform worse than base classes without annotation supervision. With the space of novel categories enlarged (e.g. from B15/N4 to B12/N7 partition), the performance on novel classes degrades (e.g. ‘bookshelf’ obtains 7.4% mIoU drop from B15/N4 to B12/N7 partition on semantic segmentation) due to the insufficient seen-category data to tune the model.

S3.3. Error Bar

Here, to show the robustness of our experimental results, we repeat the experiments on open-vocabulary semantic and instance segmentation three times and report their average along with standard deviation. As shown in Table S17 and Table S18, the results on base classes are slightly more stable than novel classes with lower standard deviations, which demonstrates the higher confidence uncertainty of novel class predictions. Besides, results on ScanNet are more stable than S3DIS, which indicates that the sample size and diversity contribute a lot to the performance stability.

S3.4. Equipping Fully-Supervised Model with Point-Caption Supervision.

As demonstrated in Table S19, fully-supervised models equipped with caption supervision loss perform similarly to those without it, as they already have access to annotations for all categories. In this scenario, our language supervision neither hinders nor enhances fully-supervised performance, validating our fairness in using the fully-supervised model for comparison in the main paper.

S3.5. Combination of Caption Supervisions.

The combination of three captions, including the scene-level caption, can result in a 0.6% increase in hIoU, as

shown in Table S20. However, finding such a right balance between these captions requires sophisticated loss trade-off techniques that are not universally applicable across different datasets. Therefore, the scene-level caption is not used in our paper for the sake of generalization. Further studies on effectively combining caption supervisions would be a future investigation.

S4. Caption Examples

In this section, we present examples of image-caption pairs obtained by vision-language (VL) foundation models and examples of hierarchical associated point-caption pairs. As illustrated in Fig. S7, image captions describe main entities of images along with room types (e.g. kitchen), texture (e.g. leather), color (e.g. green) or spatial relationships (e.g. on top of), conveying rich semantic clues with large vocabulary size. Moreover, uncommon classes such as ‘buddha statue’ are also correctly detected, reflecting the generalizability of existing VL foundation models and semantic comprehensiveness of generated captions.

With obtained image-caption pairs, we are capable to associate 3D points and captions hierarchically leveraging geometric constraints between 3D point clouds and multi-view images. As shown in Fig. S8 (a), the scene-level caption describes each area/room (e.g. kitchen, living room) in the whole scene with abundant vocabulary, providing semantic-rich language supervision. View-level caption in Fig. S8 (b) focuses on single view frustums of the 3D point cloud, capturing more local details with elaborate text descriptions, which enables the model to learn region-wise vision-semantic relationships. Additionally, as shown in Fig. S8 (c), the entity-level caption covers only a few entities in small 3D point sets with concrete words, providing more fine-grained supervisions to learn object-level understanding and localization.

S5. Qualitative Results

Here, we provide some qualitative results on open-vocabulary semantic segmentation and instance segmentation as illustrated in Fig. S9. Compared to the LSeg-

Task	Partition	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	shower c.	toilet	sink	bathtub
Sem.	B15/N4	84.6	95.0	64.9	81.1	87.9	75.9	72.2	61.9	62.1	69.5	30.9	60.1	46.5	70.7	50.5	66.1	56.8	59.0	81.7
	B12/N7	84.7	95.1	65.3	57.8	44.2	75.9	34.5	62.5	62.3	62.1	20.5	57.8	61.4	72.4	47.9	64.9	85.9	28.4	69.6
	B10/N9	83.8	95.2	64.3	80.9	88.0	78.5	73.2	60.6	61.5	68.6	17.7	23.4	51.3	70.6	25.7	38.2	51.3	27.3	61.7
Inst.	B13/N4	–	–	50.5	77.0	82.9	43.4	75.4	49.0	46.0	43.7	46.5	33.7	23.2	54.1	49.6	56.0	97.8	47.5	85.8
	B10/N7	–	–	53.7	62.7	11.2	70.5	27.2	47.7	45.7	30.0	01.5	39.9	40.8	50.6	68.6	84.6	92.9	24.6	00.0
	B8/N9	–	–	45.1	77.4	82.2	84.2	74.2	48.9	51.0	30.0	00.5	02.1	16.8	44.9	28.3	35.1	94.3	16.6	00.0

Table S15. Per-class results of 3D open-vocabulary scene understanding on ScanNet. Performance on novel class are marked in blue.

Task	Partition	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board
Sem.	B8/N4	93.9	97.8	82.9	00.0	17.2	15.6	53.7	35.8	86.3	05.3	37.3	43.3
	B6/N6	93.7	79.1	80.1	00.1	28.5	24.1	08.4	37.6	87.0	54.0	24.0	06.9
Inst.	B8/N4	89.5	100.0	50.8	00.0	35.3	36.2	60.5	00.1	84.6	01.9	00.8	59.4
	B6/N6	89.5	60.2	17.9	00.0	41.5	10.2	02.1	00.6	86.2	45.1	00.1	02.2

Table S16. Per-class results of 3D open-vocabulary scene understanding on S3DIS.

3D baseline that always confuses unseen classes as seen classes, our framework successfully recognizes novel categories with accurate semantic masks, which shows our point-caption association injects rich semantic concepts into the 3D network. Additionally, the instance prediction masks of our framework are also accurate, while the LSeg-3D baseline misses novel objects or predicts incomplete object masks. It reflects the strong generalized localization ability of our framework.

S6. Limitation and Open Problems

Although our language-driven open-vocabulary 3D scene understanding framework introduces rich semantic concepts for learning adequate visual-semantic relationships, it still suffers from limitations in the following aspects. First comes the calibration problem that the model tends to produce over-confident predictions on base classes, which lies in both semantic and instance segmentation tasks. For semantic segmentation, though the binary head is developed to calibrate semantic scores for in-domain open-vocabulary scene understanding, it fails to rectify predictions for out-of-domain transfer tasks. Trained on the dataset-specific base/novel partition, the binary head is hard to generalize to other datasets with data distribution shifts, which encourages us to design more transferable score calibration modules in the future. As for the instance segmentation task, though we largely address the localization problem for novel classes through fine-grained point-caption pairs, the calibration problem also exists in the proposal grouping process, where objects of novel classes cannot

group well and probably obtain incomplete instance masks. We also leave it as a challenge that needs to be resolved further.

The second problem is that S3DIS achieves slightly worse open-vocabulary performance than ScanNet, largely due to its limited sample size and diversity, as well as much fewer point-caption associations. Inspired by our zero-shot transfer results, we believe it is an appealing alternative to pre-train on a large dataset with rich semantic information and then fine-tune it on the small-scale dataset, which we leave for future study.

Round	ScanNet									S3DIS					
	B15/N4			B12/N7			B10/N9			B8/N4			B6/N6		
	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N	hIoU	mIoU ^B	mIoU ^N
1	66.3	68.4	64.2	54.3	69.5	44.6	52.8	76.2	40.6	33.2	58.2	23.3	39.4	57.2	30.0
2	65.2	68.6	62.2	54.8	69.7	45.2	53.3	75.6	40.9	37.0	59.5	26.9	39.5	55.1	30.8
3	64.5	67.8	60.8	59.7	69.2	48.0	53.2	76.6	40.8	33.7	59.4	23.5	36.5	54.3	27.5
Average	65.3	68.3	62.4	55.3	69.5	45.9	53.1	76.2	40.8	34.6	59.0	24.5	38.5	55.5	29.4
Std	00.9	00.4	01.7	01.3	00.2	01.8	00.3	00.5	00.2	02.1	00.7	02.0	01.7	01.5	01.7

Table S17. Repeat results for open-vocabulary 3D semantic segmentation on ScanNet and S3DIS in terms of hIoU, mIoU^B and mIoU^N.

Round	ScanNet									S3DIS					
	B13/N4			B10/N7			B8/N9			B8/N4			B6/N6		
	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N	hAP ₅₀	mAP ₅₀ ^B	mAP ₅₀ ^N
1	54.9	58.1	52.0	33.1	52.5	24.1	34.5	62.1	23.9	19.3	59.2	11.5	10.9	49.2	06.1
2	56.7	57.9	55.5	28.4	55.1	19.1	37.5	63.8	26.5	9.2	57.4	05.0	19.8	46.7	12.6
3	55.0	59.5	51.1	32.1	56.3	22.5	35.7	63.5	24.8	16.8	60.0	09.7	17.4	44.9	10.8
Average	55.5	58.5	52.9	31.2	54.6	21.9	35.9	63.1	25.1	15.0	59.0	08.6	16.0	46.9	09.8
Std	01.0	00.9	02.3	02.5	01.9	02.6	01.5	00.9	01.3	04.3	01.1	02.7	04.6	02.2	03.4

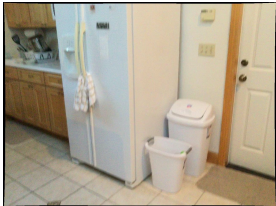
Table S18. Repeat results for open-vocabulary 3D instance segmentation on ScanNet and S3DIS in terms of hAP₅₀, mAP₅₀^B and mAP₅₀^N.

Method	mIoU	mIoU ^B / mIoU ^N		
		B15/N4	B12/N7	B10/N9
Fully-Sup.	70.62	68.4 / 79.1	70.0 / 71.8	75.8 / 64.9
Fully-Sup. + Caption	70.82	68.7 / 78.9	70.3 / 71.7	76.7 / 64.6

Table S19. Fully-supervised results equipped with point-caption supervision.

α_1 (scene)	α_2 (view)	α_3 (entity)	hIoU / mIoU ^B / mIoU ^N
0.000	0.050	0.050	65.3 / 68.3 / 62.4
0.033	0.033	0.033	64.6 / 69.0 / 60.8
0.010	0.045	0.045	65.9 / 68.2 / 63.8

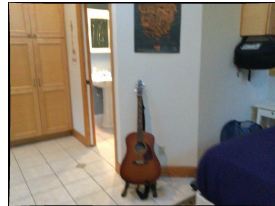
Table S20. Ablation for caption loss weights on ScanNet B15/N4.



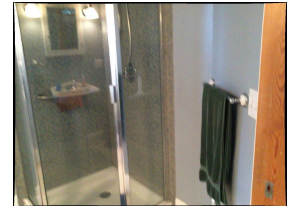
a kitchen with a refrigerator and a trash can



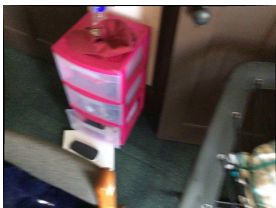
a living room with a couch and a bar



a guitar sitting on the floor in a room



a bathroom with a shower and a green towel



a pink plastic container with a bunch of boxes on the floor



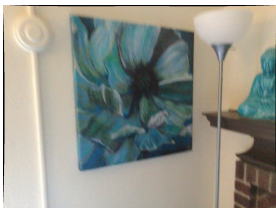
a toaster oven sitting on top of a kitchen counter



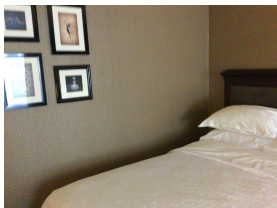
three leather chairs and a stool in a living room



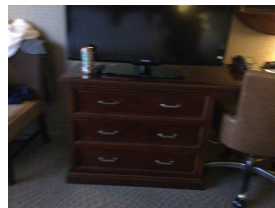
the back of a computer screen on a table



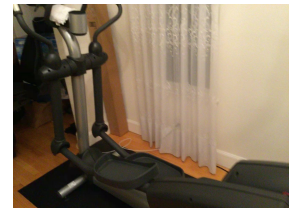
a painting of a flower next to a lamp and a buddha statue



a bedroom with a bed and pictures on the wall



a dresser with drawers and a tv on top of it



a treadmill in the corner of a room

Figure S7. Examples of image-caption pairs by image-captioning model ViT-GPT2 [1].



Video shows a person sitting on a couch with their feet on a rug. A guitar is sitting in a room next to a bed. A toaster oven is sitting on top of a kitchen counter. A bike is parked in a living room with a tiled floor.



A living room is clean and ready for the flooring to be installed. A bed with a gold blanket and a laptop on top of it. A bag of clothes sitting on a chair in a living room. A treadmill in the corner of a room. an exercise bike in a room with a white curtain.

(a) scene-level caption



a kitchen with a refrigerator and a trash can



a bedroom with a bed and pictures on the wall



a dresser with drawers and a tv on top of it



a toaster oven sitting on top of a kitchen counter

(b) view-level caption



table couch living



chair couch



hotel lamp bed



tv

(c) entity-level caption

Figure S8. Examples of hierarchical point-caption pairs from ScanNet [7]

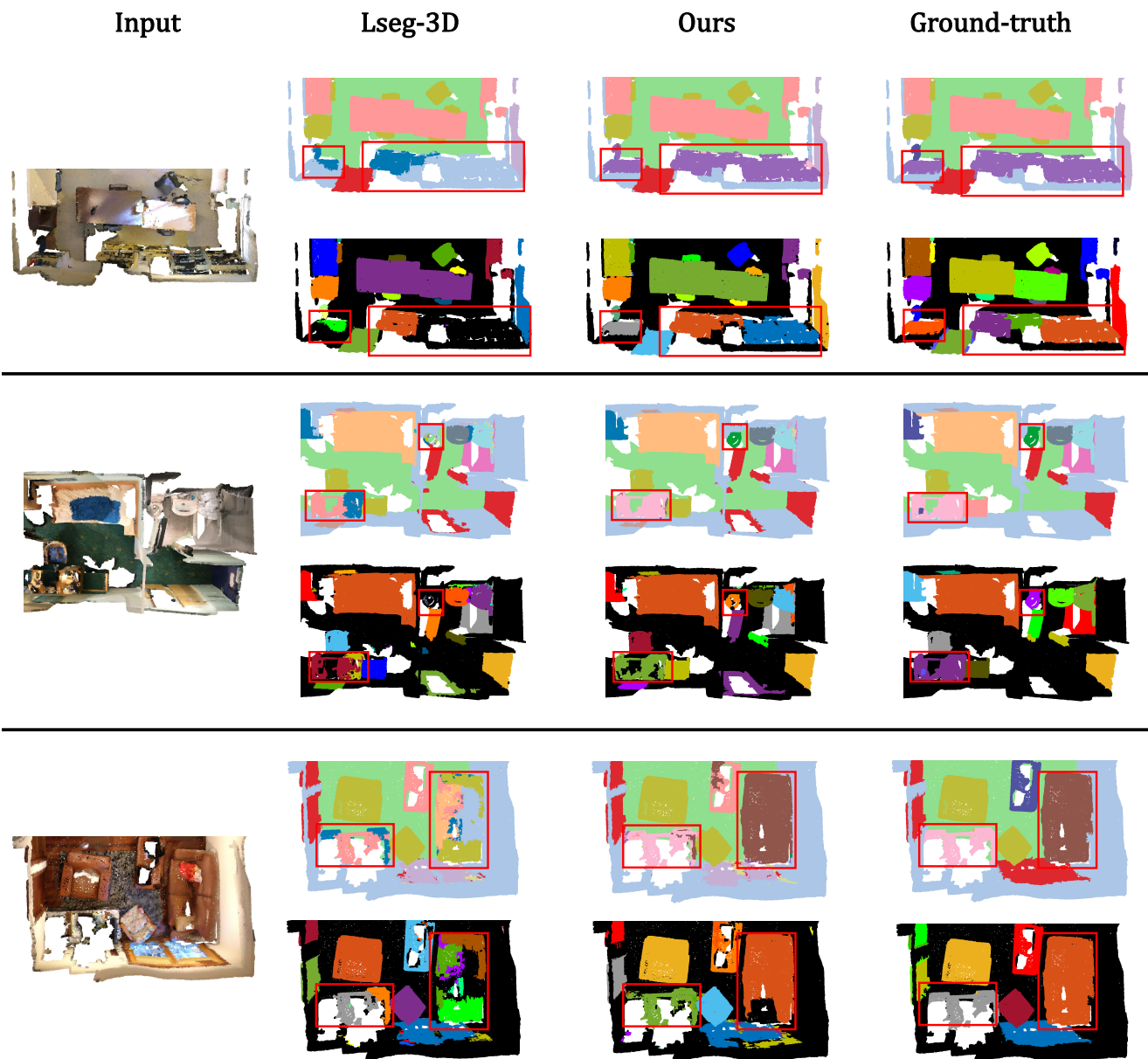


Figure S9. Qualitative results of open-vocabulary semantic segmentation and instance segmentation. In each example, the first row illustrates the semantic masks and the second row shows the instance masks. Novel classes are highlighted in red bounding boxes.