# Cross-Modal Match for Language Conditioned 3D Object Grounding

**Yachao Zhang**[1], **Runze Hu**[2], **Ronghui Li**[1], **Yanyun Qu**[3], **Yuan Xie**[4], **Xiu Li**[1]*

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2] School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China
[3]School of Informatics, Xiamen University, Xiamen, 361000, China
[4]School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China
{yachaozhang, li.xiu}@sz.tsinghua.edu.cn

## Abstract

Language conditioned 3D object grounding aims to find the object within the 3D scene mentioned by natural language descriptions, which mainly depends on the matching between visual and natural language. Considerable improvement in grounding performance is achieved by improving the multi-modal fusion mechanism or bridging the gap between detection and matching. However, several mismatches are ignored, i.e., mismatch in local visual representation and global sentence representation, and mismatch in visual space and corresponding label word space. In this paper, we propose cross-modal match for 3D grounding from mitigating these mismatches perspective. Specifically, to match local visual features with the global description sentence, we propose BEV (Bird's-eye-view) based global information embedding module. It projects multiple object proposal features into the BEV and the relations of different objects are accessed by the visual transformer which can model both positions and features with long-range dependencies. To circumvent the mismatch in feature spaces of different modalities, we propose cross-modal consistency learning. It performs cross-modal consistency constraints to convert the visual feature space into the label word feature space resulting in easier matching. Besides, we introduce label distillation loss and global distillation loss to drive these matches learning in a distillation way. We evaluate our method in mainstream evaluation settings on three datasets, and the results demonstrate the effectiveness of the proposed method.

## Introduction

3D scene understanding based on point cloud has attracted a lot of attention and achieved great success along with the development of deep learning (Guo et al. 2020; Yan et al. 2020). Most of the existing 3D scene understanding methods focus on visual modality, *i.e.,* point cloud modality (Graham, Engelcke, and Van Der Maaten 2018; Wang et al. 2019; Zhang et al. 2021b), image (Long, Shelhamer, and Darrell 2015; Olaf Ronneberger 2015) or the fusion of them (Jaritz et al. 2020; Peng et al. 2021; Zhang et al. 2022).

Recently, language conditioned 3D grounding aims to discover and locate an object in the 3D scene referred to by the natural language sentence. It can enhance the interaction between humans and machines by enabling more natural and intuitive communication using natural language.

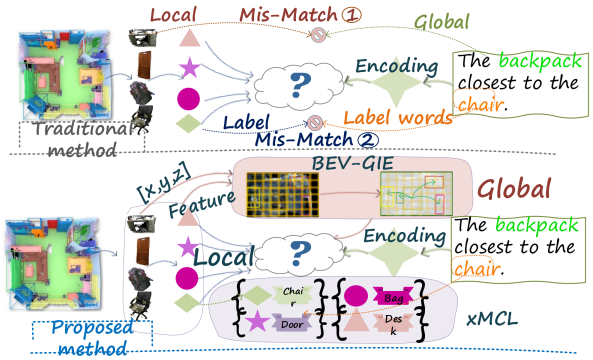


Figure 1: An overview of the traditional two-stage method (the top part) and our xM_Match (the bottom part). {Desk} denotes the cross-modal consistency constraint (xMCL).

Existing 3D grounding methods can be divided into two types: one-stage method and two-stage method. The former focuses on bridging the gap between detection and matching in the 3D visual grounding task, and thus achieving the target localization at a single stage. However, these methods suffer from the high training complexity compared to their two-stage counterpart, because the multi-modal feature extraction (language sentences and the entire 3D scene) and target object regression are implemented simultaneously. Distinctly, the two-stage method primarily attempts to explore relations between proposals and referred language sentences to distinguish the target object. For this type of method, there are mismatches between visual and linguistic modalities, mainly in two aspects. 1) Mismatch in local view and global view. Since the visual features are extracted from each proposal individually (local view representation) (Yang et al. 2021; Chen et al. 2022b; Bakr, Alsaedy, and Elhoseiny 2022), the visual features lack the interaction between every object while the representation of linguistic description is based on the global view of the entire scene. Even some fancy techniques, *i.e.,* graph neural network (Huang et al. 2021), and attention mechanisms (Zhao et al. 2021), focus on modeling the relations of two modalities to promote the

*Corresponding Author.

matching, mismatch in global and local representation is still inevitable. 2) Mismatch in different feature spaces. In previous methods (Achlioptas et al. 2020; Luo et al. 2022; Chen et al. 2022b), the different modal features are extracted independently without explicitly modeling modal consistency, resulting in the mismatch of visual feature and label word feature spaces.

In this paper, we focus on how to mitigate these mismatches and propose a cross-modal match method dubbed as xM_Match. In contrast to the existing two-stage approach (top part of Fig. 1), we introduce two new components, *i.e.*, BEV-based global information embedding (BEV-GIE) and cross-modal consistency learning (xMCL) shown as the bottom part of Fig. 1. Specifically, BEV-GIE can provide global information for object proposal features from the perspective of a lightweight manner. It projects multiple object proposals of a scene into a unified view and the relations of different objects are accessed by the visual transformer that specializes in modeling both positions and features with long-range dependencies. Therefore, the global information of the 3D scene is introduced, improving the matching with the global natural language description. To further alleviate the mismatch feature space between visual and label words, we introduce xMCL, which performs both point cloud recognition and cross-modal consistency constraints to help align visual feature space to the feature space of object label words. Additionally, we introduce two distillation losses to facilitate cross-modal match learning.

To summarize, the following are the main contributions:

- We present xM_Match for language conditioned 3D object grounding from a novel perspective of promoting cross-modal matching between visual data and natural language.
- To make the local visual representation match the global description of natural language, we present BEV-based global information embedding module to supplement the global information for visual features extracted from separated object proposals.
- To enhance multi-modal interaction, we introduce cross-modal consistency learning to align visual feature space to the label word feature space.
- Extensive experimental results demonstrate that xM_Match achieves state-of-the-art performance and outperforms the most of competitors on three datasets.

## Related Work

### Scene Understanding Based on Point Cloud

Recently, deep learning on point clouds has become even thriving, which has been successfully used to solve various 3D vision problems, including 3D shape classification (Qi et al. 2017; Zhang, Hua, and Yeung 2019; Li et al. 2019), 3D object detection and tracking (Qi et al. 2019; Yang, Luo, and Urtasun 2018; Jiang et al. 2020), and 3D point cloud segmentation (Zhang et al. 2021a,b; Landrieu and Simonovsky 2018; Hu et al. 2020). According to the data type of input for neural networks, existing 3D scene understanding methods can be divided into multi-view based methods, volumetric-based methods, and point-based methods. Point-based methods directly work on raw point clouds which can avoid explicit information loss (Guo et al. 2020). Qi *et al.* (Qi et al. 2017) proposed a hierarchical network, named PointNet++, by capturing fine geometric structures from the neighborhood of each point and is widely used in various tasks citehu2021simulation as the backbone for 3D point cloud feature extraction. Based on the convenience of point-based methods and fair comparison, we also choose PointNet++ to serve our visual feature extraction in the 3D grounding task.

### 3D Visual Grounding

Multi-modality (including 2D images, 3D point clouds, and language) brings important cues to improve 3D scene perception for the agent (Wu et al. 2023). 3D grounding task requires a model to find the object mentioned by natural language in a wild point cloud. However, 3D grounding is still in its infancy due to the unique challenges confronted by the processing of point clouds, language, and their matching. With some methods being proposed to deal with the above challenge, these methods can be roughly divided into two types: one-stage method and two-stage method.

In one-stage methods (Luo et al. 2022; Wu et al. 2023) linguistic features are densely fused with every point or sampled point of the entire scene to generate multi-modal feature maps for regressing the bounding box. However, they are usually computationally massive, due to the time-consumed extracting features for the entire point cloud and modeling the relationship of all candidate points and linguistic features. The two-stage method is a mainstream method, which transforms the regression problem into a matching problem (Feng et al. 2021; He et al. 2021; Roh et al. 2022; Yang et al. 2021; Yuan et al. 2021), where a detection-then-matching strategy is introduced. These methods mainly focus on better modeling the relationship among objects and language to locate the target object. Some fancy techniques, *i.e.,* graph neural network (Huang et al. 2021), and attention mechanisms (Zhao et al. 2021), are proposed to improve the matching performance. To avoid the huge computational consumption of large-scale point cloud scene feature extraction, these methods use feature extraction independently on the proposal sub-point cloud and then compute matching scores with linguistic features. However, the global information of the scene is not well preserved, which obviously mismatches the reference sentences covering the whole scene. In this paper, we supplement the global information of visual features with the help of BEV.

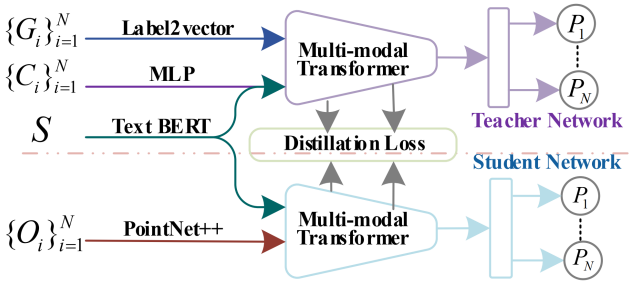


Figure 2: ViL3DRel framework.

# Proposed Method

## Problem Definition and Preliminary

**Problem Definition.** Given one 3D scene represented by a point cloud $P_s \in \mathbb{R}^{K\times 6}$ with $K$ points, the goal of the 3D grounding task is to discover and locate an object in this scene referred by the natural language sentence $S$. Each object can be represented as a subset of $P_s$, denoted as $O_i \subset P_s, O_i \in \mathbb{R}^{K_i\times 6}$ which contains $K_i$ points and each point represented by 3-d coordinates (XYZ) and 3-d color values (RGB). We follow the two-stage method, where a list of object proposals $\{O_1, \cdots, O_N\}$ is obtained via 3D object detector (Jiang et al. 2020) or ground-truth annotations (depending on the evaluation setup), and then the 3D grounding model $\mathcal{F}_\Theta(\cdot)$ outputs the matching one referred by $S$ among $N$ object proposals. Compactly, it can be formulated as:

$$O_T = \mathcal{F}_\Theta(\{O_1, \cdots, O_N\}, S), \tag{1}$$

where $O_T$ is the bounding box $B_T \in \mathbb{R}^6$ of target object.

**Preliminary.** ViL3DRel (Chen et al. 2022b) utilizes a cross-modal transformer to explore the relations of natural language and proposal embedding, and a knowledge distillation strategy is given. The framework is shown in Fig. 2. The distillation loss can be formulated as:

$$\begin{aligned}&\mathcal{L}_{local} = \\ &\mathcal{D}[\mathcal{F}^T_{\Theta_T}(f^{rgb}(C_i), f^{gt}(G_i)|S)||\mathcal{F}^S_{\Theta_S}(f^{PN2}(O_i)|S)],\end{aligned} \tag{2}$$

where $\mathcal{F}^{T/S}_{\Theta_{T/S}}(\cdot)$ denotes the Teacher/Student network, and $C_i$ and $G_i$ are the color of point cloud and object label word, respectively. $\mathcal{D}(\cdot||\cdot)$ denotes the distillation loss in ViL3DRel. $f^{rgb}(C_i)$, $f^{gt}(G_i)$, and $f^{PN2}(O_i)$ are multi-modal data encoding networks.

## Overview of xM_Match

The language conditioned 3D grounding requires bridging the gap between the linguistic and visual modalities. We find that the mismatches of cross-modal are mainly manifested in two aspects: i) Mismatch in local visual representation and global sentence representation. Natural language describes the target object according to the global information of the scene, and these two-stage methods independently encode every object proposal where the global information hidden between objects has been discarded. ii) Mismatch in visual space and corresponding label word space. 3D point cloud and object label words are encoded independently using the pre-trained networks resulting in different feature spaces, which can not be favorable for feature interaction.

Therefore, we propose xM_Match, and the overall framework is shown in Fig. 3. Specifically, we use ViL3DRel as the baseline, and first introduce BEV-GIE module which introduces global information about the 3D scene to alleviate the mismatch of local visual feature and global language feature from a lightweight perspective. Then, we introduce cross-modal consistency learning (xMCL) to supervise visual feature learning label-word-related vectors to alleviate the lack of label word encoding in the student network. It promotes multi-modal feature space alignment and facilitates subsequent matching. Additionally, we impose global distillation loss and label distillation loss to promote the learning of the above modules.

## Cross-modal Consistency Learning

With the help of natural language representation of object label words, the 3D grounding performance of the teacher network can be greatly improved. However, label information is not available during model inference, so it is impossible to obtain this representation. The visual feature space is mismatched with object label word encoding and knowledge transfer is limited from the teacher network to the student network. Our goal is that the visual feature can express the 3D scene and also encode the object label word.

We hold that a memory bank of object label word features for every category can reach the above goal. During student network training, we can query the object label words feature from this bank by similarity retrieval. As different modalities belong to different feature spaces, the key problem that needs to be solved in detail is how to make visual coding effectively retrieve memorized features. Therefore, the consistency constraint is introduced.

**Consistency constraint.** We introduce two heads for the point cloud encoder. One is used for object recognition and the other for mapping the 3D visual feature space of the proposal to the label word feature space. To ensure the alignment of the two feature spaces, we directly align the features between the encoding point cloud feature and the label word feature by minimizing:

$$\mathcal{L}_{xmmse} = \frac{1}{N}\sum_{i=1}^{N}||H_1(f^{PN2}(O_i)) - f^{gt}(G_i)||_2, \tag{3}$$

where $H_1(\cdot)$ is the mapping head, containing two linear layers.

The correlation level consistency can enhance the consistency of the two modal feature spaces. Therefore, we construct two correlation graphs according to the discrepancy between different objects for two modalities and constrain their consistency, where the label word graph $\mathcal{G}^{gt}_{ij}$ is formulated as:

$$\mathcal{G}^{gt}_{ij} = \frac{f^{gt}(G_i)\cdot f^{gt}(G_j)}{||f^{gt}(G_i)|| \times ||f^{gt}(G_j)||}. \tag{4}$$

We utilize the same way to get the visual feature graph $\mathcal{G}^{visual}_{ij}$ by replacing the $f^{gt}(O^{gt}_j)$ with $H_2(f^{PN2}(O_i))$. The correlation level consistency $\mathcal{L}_{xmgraph}$ is formulated as:

$$\mathcal{L}_{xmgraph} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}||\mathcal{G}^{gt}_{ij} - \mathcal{G}^{visual}_{ij}||_2. \tag{5}$$

The overall objective of cross-modal consistency constraint is $\mathcal{L}_{xm}$ formulated as:

$$\mathcal{L}_{xm} = \mathcal{L}_{xmgraph} + \mathcal{L}_{xmmse}. \tag{6}$$

**Label word-like encoding.** The memory bank $M \in \mathbb{R}^{C\times dim}$ of the label words is initialized with the label-to-vector same as the baseline in the teacher network and without updating. $C$ is the number of categories. To ensure that the input of the student network is as consistent as possible with the teacher network, the mapping head features ($\hat{e}^{gt}_{O_i} = H_1(f^{PN2}(O_i)), \in \mathbb{R}^{1\times dim}$) of the student network

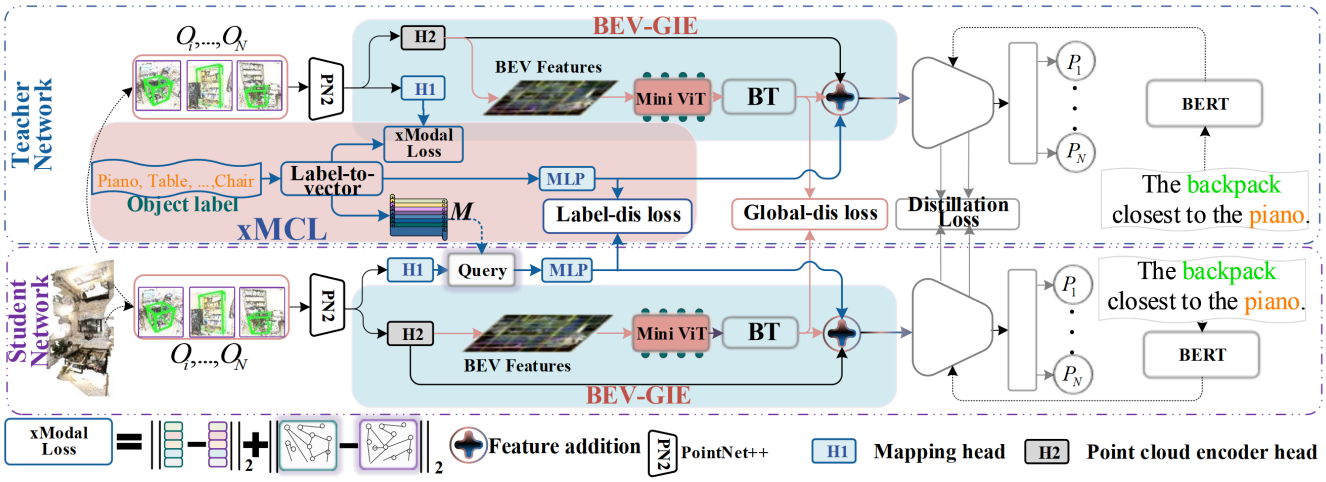


Figure 3: xM_Match framework. BT and BERT are the global feature backtrack and the pre-trained text embedding model.

are used to query the features of the corresponding category labels from the memory bank, where an attention-based query strategy is proposed. Firstly, we can obtain the attention map ($A$) by:

$$A = \frac{\hat{e}^{gt}_{O_i} \cdot M^T}{||\hat{e}^{gt}_{O_i}|| \times ||M^T||}, \tag{7}$$

where $A \in \mathbb{R}^{1\times C}$ denotes the similarity between an object and all category codes. Superscript $^T$ represents matrix transpose. The label word-like encoding of the student network can be updated by:

$$\hat{e}^{gt}_{O_i} := A \cdot M. \tag{8}$$

To learn label word-like encoding better for student network, a label distillation loss $\mathcal{L}_{gt}$ is introduced to constrain the consistency with the teacher network, which can be formulated as:

$$\mathcal{L}_{gt} = ||MLP(\hat{e}^{gt}_{O_i}) - MLP(e^{gt}_{O_i})||_2, \tag{9}$$

where $MLP(\cdot)$ denotes the linear layers followed by label word feature and label word-like encoding.

## BEV-based Global Information Embedding

To describe an object in one scene, natural language is usually based on the global information, *i.e.,* the object size, distance, relationships and spatial location of every object. The two-stage methods perform a detection-then-matching strategy (Yang et al. 2021; Chen, Chang, and Nießner 2020; Chen et al. 2022b) for every sub-point cloud $O_i$ independently or only using the center coordinate of every object to model the global relations. This global information based solely on the center position information is insufficient.

A direct solution is to encode the entire scene, which typically contains millions of points. Directly modeling global features undoubtedly increases the complexity of point cloud representation. Considering a lightweight way to embed global information, the BEV map is emerging,

| R | Hor. | Between | Allocentric | Support | Vertical |
|---|---|---|---|---|---|
| P | 81% | 9% | 4% | 2% | 4% |

Table 1: Statistics of Sr3D. R and P denote relationships and proportion.

wherein a mini transformer is used to simultaneously characterize the positional relationship between BEV planes (positional encoding) and the correlation learning between multiple objects (multi-head attention mechanism). The module is identical in both branches, we only detail the BEV-GIE module in the teacher network branch for a concise presentation as follows.

**BEV projection.** In real scenarios, more horizontal relations are used to describe the target object. For example, there are over 80% of objects are described by horizontal relationship (Hor.) in Sr3D (Achlioptas et al. 2020) (refer to Tab. 1). Therefore, we choose projection-based BEV to model global horizontal relations.

For a scene point cloud $P_s = \{O_1, ..., O_N\}$, we extract the features by $f^{PN2}(O_i)$ for every object proposals same as baseline. For multi-scale PointNet++, we use inverse distance weighted average based on 3 nearest neighbors to up-sample the third scale to the same scale as the second scale and concatenate two scale features. For an object $O_i$, the concatenated feature of one point $p$ can be denoted as $f^{O_i}_p, p = \{1, ..., K_i\}$. We aggregate the point features in all objects of the entire scene: $f^{P_s} = \{f^{O_i}_p\} \mid i = \{1, ..., N\}, p = \{1, ..., K_i\}$.

We quantize it along the $x$-axis and $y$-axis to generate pillar voxels evenly. The points are assigned to these voxels according to their coordinates. The feature of a voxel is obtained by $max$-pooling ($MAX(\cdot)$) of points inside it. For example, the feature in the $i, j$-th grid cell is:

$$\begin{aligned} f^{bev}_{i,j} = &MAX(\{f^{P_s} \mid (i-1)w < x_p < iw, \\ &(j-1)w < y_p < jw\}), p = \{1, ..., KN\}, \end{aligned} \tag{10}$$

| Methods | Nr3D | | | | | Sr3D | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Easy | Hard | V-Dep. | V-Indep. | Overall | Easy | Hard | V-Dep. | V-Indep. |
| ReferIt3D | 35.6 | 43.6 | 27.9 | 32.5 | 37.1 | 40.8 | 44.7 | 31.5 | 39.2 | 40.8 |
| ScanRefer | 34.2 | 41.0 | 23.5 | 29.9 | 35.4 | - | - | - | - | - |
| TGNN | 37.3 | 44.2 | 30.6 | 35.8 | 38.0 | - | - | - | - | - |
| InstanceRefer | 38.8 | 46.0 | 31.8 | 34.5 | 41.9 | 48.0 | 51.1 | 40.5 | 45.4 | 48.1 |
| FFL-3DOG | 41.7 | 48.2 | 35.0 | 37.1 | 44.7 | - | - | - | - | - |
| 3DVG | 40.8 | 48.5 | 34.8 | 34.8 | 43.7 | 51.4 | 54.2 | 44.9 | 44.6 | 51.7 |
| TransRefer3D | 42.1 | 48.5 | 36.0 | 36.5 | 44.9 | 57.4 | 60.5 | 50.2 | 49.9 | 57.7 |
| LanguageRefer | 43.9 | 51.0 | 36.6 | 41.7 | 45.0 | 56.0 | 58.9 | 49.3 | 49.2 | 56.3 |
| SAT | 49.2 | 56.3 | 42.4 | 46.9 | 50.4 | 57.9 | 61.2 | 50.0 | 49.2 | 58.3 |
| 3D-SPS | 51.5 | 58.1 | 45.1 | 48.0 | 53.2 | 62.6 | 56.2 | 65.4 | 49.2 | 63.2 |
| Multi-view | 55.1 | 61.3 | 49.1 | 54.3 | 55.4 | 64.5 | 66.9 | 58.8 | 58.4 | 64.7 |
| ViL3DRel | 64.4 | 70.2 | 57.4 | 62.0 | 64.5 | 72.8 | 74.9 | 67.9 | 63.8 | 73.2 |
| LAR | 48.9 | 58.4 | 42.3 | 47.4 | 52.1 | 59.4 | 63.0 | 51.2 | 50.0 | 59.1 |
| **xM_Match** | **66.2** | **72.8** | **59.9** | **63.8** | **67.5** | **74.6** | **75.9** | **71.3** | **65.0** | **74.7** |

Table 2: Grounding accuracy (%) on Nr3D and Sr3D datasets with ground-truth object proposals. "V-Dep." and "V-Indep." represent view-dependent setting and view-independent setting, respectively.

where $f_{i,j}^{bev} \in \mathcal{R}^{1\times dim}$. The size of a grid cell is $w \times w$. $x_p$/$y_p$ is the $x$/$y$ coordinate of 3D point $p$, *i.e.*, its locations in the BEV space. Finally, the BEV feature map of $P_s$ is:

$$f^{bev} = \left\{ f_{i,j}^{bev} \mid i \in \{1, 2 \ldots, W\}, j \in \{1, 2 \ldots, L\} \right\}, \quad (11)$$

where $f^{bev} \in \mathcal{R}^{W\times L\times dim}$. $W$ and $L$ denote the number of grid cells along the $x$-axis and $y$-axis, respectively. Meanwhile, we record the proposal identity of the point in each cell, denoted as $idx^{proposal}$.

**Global information embedding.** To embed the global information, an encoder that can well embody both scene embedding and spatial position is required. Vision Transformer (ViT) contains the self-attention mechanism and positional encoding, which allows it to capture long-range dependencies between image patches (Hu et al. 2022). These components enable ViT to learn global BEV map representations. Considering that we have already obtained the features of each cell, we introduce a lightweight version with 3 layers of ViT denoted as $ViT(\cdot)$ to maintain the efficiency.

We use cells of the BEV map as patches and flatten them into a sequence of tokens. These tokens coupled with pixel position embedding of the BEV map are processed by multiple transformer layers that allow for global interactions between all the BEV cells through self-attention mechanisms, formulated as:

$$f^{ViT} = ViT(f^{bev}). \quad (12)$$

**Global feature backtrack.** To attach global information for every object proposal, we need to fuse the patch feature containing global information with the local object proposal. We utilize the index of BEV projection to backtrack the patch feature vectors to the local point features and perform an averaging operation on the points within the proposal, denoted as: $f^{global} = Q^{scatter\_mean}(f^{ViT}, idx^{proposal})$.

Finally, we combine this patch feature vector with each object proposal feature through a summation operation:

$$f^{O_i} = H_2(f^{PN2}(O_i)) + f^{global}. \quad (13)$$

For the student branch, we get the global features $\hat{f}^{global}$ and $\hat{f}^{O_i}$ same as the teacher branch. To train the student network, we introduce global distillation loss $\mathcal{L}_{global}$ and can be formulated as:

$$\mathcal{L}_{global} = \frac{1}{dim} \sum_{d=1}^{dim} MSE(f_d^{global} - \hat{f}_d^{global}). \quad (14)$$

## Overall Objective Function

We follow 3D object grounding loss $\mathcal{L}_{og}$, sentence classification loss $\mathcal{L}_{sent}$ used in the previous works (Achlioptas et al. 2020; Chen, Chang, and Nießner 2020). Based on distillation loss $\mathcal{L}_{local} = \mathcal{L}_{atten} + \mathcal{L}_{hidden}$, we add global distillation loss $\mathcal{L}_{global}$ and label distillation loss $\mathcal{L}_{gt}$. For object classification losses, based on two object classification losses $\mathcal{L}_{obj}^{u}$ and $\mathcal{L}_{obj}^{m}$ in previous works (Achlioptas et al. 2020; Chen, Chang, and Nießner 2020; Chen et al. 2022b), we introduce the cross-modal loss $\mathcal{L}_{xm}$. Therefore, the overall training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{og} + \mathcal{L}_{sent} + \mathcal{L}_{obj}^{u} + \mathcal{L}_{obj}^{m} + \mathcal{L}_{xm} + \mathcal{L}_{local} + \lambda_a \mathcal{L}_{gt} + \lambda_b \mathcal{L}_{global}, \quad (15)$$

where $\lambda_a$ and $\lambda_b$ are used to trade off the proposed two distillation losses.

# Experiments and Results

## Datasets

We leverage three recently released datasets, *i.e.,* Nr3D (Achlioptas et al. 2020), Sr3D (Achlioptas et al. 2020) and ScanRefer (Chen, Chang, and Nießner 2020) built on the 3D scenes of ScanNet (Dai et al. 2017) to evaluate performance. We follow the official split for training and validation.

**Additional split validation subsets.** For Nr3D and Sr3D datasets, two splits during evaluation are introduced. 1) According to the number of distractors (more distractors indicate more difficulty), the sentences are split into an "easy"

subset (less than or equal to 2 distractors) and a "hard" subset (more than 2 distractors) in evaluation. 2) According to whether the sentence requires a specific viewpoint to ground the referred object, the dataset can also be partitioned into "view-dependent" and "view-independent" subsets.

**Evaluation Metrics.** For Nr3D and Sr3D datasets, we choose the default ground-truth object proposals evaluation setting. The metric is the accuracy (%) of selecting the target bounding box among the proposals. For ScanRefer, we utilize the ground-truth object proposals and proposals obtained by a 3D detector. acc@0.25 and acc@0.5 are used to evaluate the performance of detected proposals. The acc@0.25/0.5 represents the percentage of correctly predicted bounding boxes whose IoU is larger than 0.25/0.5 with the ground truth.

## Implementation Details

For fair comparisons, we use the same point cloud backbone (PointNet++ (Qi et al. 2017)) and text encoding module (BERT (Kenton and Toutanova 2019)) same with ViL3DRel (Chen et al. 2022b). For PointNet++ (Qi et al. 2017), we set $K_i = 1024$, which denotes sampling $1,024$ points for all the objects. We set batch size as $128$, and learning rate as $0.0005$ with a warm-up of $5,000$ iterations and cosine decay scheduling. Our model is trained $100$ epochs using Adam optimizer. We directly set $\alpha_a = 1$ and $\alpha_b = 1$. We set the grid $w$ of BEV as $0.5m$. We implement our model by using PyTorch based on Python 3.8. It is trained and evaluated on one NVIDIA RTX 3090 GPU with 24GB RAM.

## Comparison to State-of-the-art Methods

**Compared methods.** We choose methods directly related to ours for comparison, containing ReferIt3D (Achlioptas et al. 2020), ScanRefer (Chen, Chang, and Nießner 2020), InstanceRefer (Yuan et al. 2021), FFL-3DOG (Feng et al. 2021), 3DVG (Zhao et al. 2021),TransRefer3D (Roh et al. 2022), LanguageRefer (He et al. 2021), SAT (Yang et al. 2021), 3D-SPS (Luo et al. 2022), Multi-view (Huang et al. 2022), ViL3DRel (Chen et al. 2022b), LAR (Bakr, Alsaedy, and Elhoseiny 2022), TGNN (Huang et al. 2021), SAT (Yang et al. 2021), BUTD-DETR (Jain et al. 2022), D3Net (Chen et al. 2022a), EDA (Wu et al. 2023).

**Evaluation on ground-truth object proposals.** Overall, compared with recent 3D grounding methods, xM_Match only uses 3D point cloud and achieves the best performance on all settings against the state-of-the-art methods even some of them using 2D image assistance.

From Tab. 2, xM_Match gains the improvements of 1.8% and 1.8% in terms of overall to baseline (ViL3DRel) on real-world Nr3D and synthetic Sr3D datasets (Achlioptas et al. 2020). The performance improvement is more evident in terms of view-dependent and hard settings, which are two more challenging subsets. The improvements are 2.5% (Hard) and 1.8% (V-Dep.) on Nr3D, 3.4% (Hard) and 1.2% (V-Dep.) on Sr3D. Because ours introduces global information embedding and multi-modal consistency constraints, which improves the matching between language descriptions and visual features, resulting in better matching performance. Furthermore, we report the results on ScanRefer dataset with ground-truth object proposals in Tab. 3. It can be observed that we achieve consistent improvements.

We also give the qualitative results of xM_Match and baseline on Nr3D datasets. From Fig. 4, our method accurately gives the target objects for the two challenging cases. For example, the sample of the second row in Fig. 4 is difficult because the positions are very near to the two objects with the same category. It requires global information which contains the correlation between the target object and other referred objects. For the failure case in the third row of Fig. 4, due to the lack of clear view guidance in the corresponding sentence and the symmetrical distractors in this scene, the failure case is mainly caused by inaccurate language description. Therefore, we can conclude that 3D grounding can benefit from our xM_Match.

| Methods | Modality | Det-pro. | | GT-pro. |
|---|---|---|---|---|
| | | 0.25 | 0.5 | Overall |
| ScanRefer | 3D+2D | 41.2 | 27.4 | - |
| ReferIt3D | 3D | 26.4 | 16.9 | 46.9 |
| TGNN | 3D | 37.4 | 29.7 | - |
| InstanceRefer | 3D | 40.2 | 32.9 | - |
| SAT | 3D+2D | 44.5 | 30.1 | 53.8 |
| FFL-3DOG | 3D | 41.3 | 34.0 | - |
| 3DVG | 3D+2D | 47.6 | 34.7 | - |
| 3D-SPS | 3D+2D | 48.8 | 37.0 | - |
| 3DJCG | 3D+2D | 49.6 | 37.3 | - |
| BUTD-DETR | 3D | 50.4 | 38.6 | - |
| D3Net | 3D+2D | - | 37.9 | - |
| ViL3DRel | 3D | 47.9 | 37.7 | 59.8 |
| **xM_Match** | 3D | **51.8** | **39.3** | **60.6** |
| EDA | 3D | 54.6 | 42.3 | - |
| **EDA+xM_Match** | 3D | **54.9** | **42.7** | - |

Table 3: Grounding accuracy (%) on ScanRefer (Chen, Chang, and Nießner 2020) datasets with detector proposals and ground-truth object proposals.

**Evaluation on detected object proposals.** The results of using detected proposals on ScanRefer dataset are reported in Tab. 3. Compared to the baseline (ViL3DRel), ours gains clear improvements. EDA (Wu et al. 2023) utilizes RoBERTa (Liu et al. 2019) which is a better text encoding model than BERT, and achieves better performance than ours with ViL3DRel backbone. We also modify our BEV-GIE and cross-modal learning and added them to EDA, abbreviated as **EDA+xM_Match**. It can be seen that EDA can benefit from our method. These results demonstrate the effectiveness of the proposed method.

## Ablation Study

**Effectiveness of BEV-based global information embedding.** In our approach, we construct BEV-based global information embedding (BEV-GIE) to model the global information. To demonstrate its benefits, we conduct an ablation study by removing BEV-GIE from our method, denoted as w/o BEV-GIE. From the comparison of #1 and #2 of Tab.

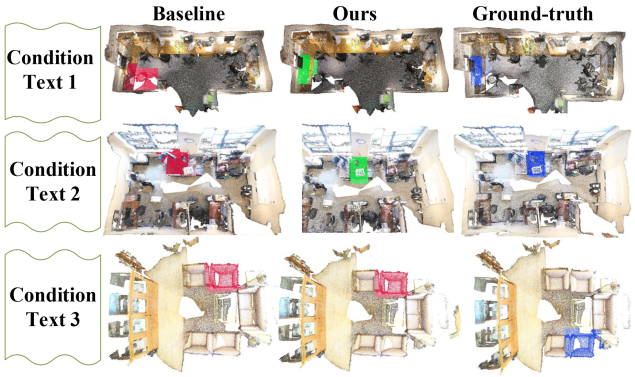


Figure 4: Qualitative results on Nr3D. The correctly predicted, incorrectly predicted, and ground-truth target objects are marked in green, red, and blue, respectively. Condition 1-3 are "The desk beneath the clock on the right hand side of the two desks grouped together", "The single computer tower under the desk in the corner", and "Chair on the left side closest to couch".

| Settings | Nr3D | | | Sr3D | | |
|---|---|---|---|---|---|---|
| | O | E | H | O | E | H |
| xM_Match | 66.2 | 72.8 | 59.9 | 74.6 | 75.9 | 71.3 |
| w/o BEV-GIE | 65.0 | 72.5 | 57.8 | 73.3 | 74.6 | 70.2 |
| w/o xMCL | 65.3 | 72.4 | 58.5 | 73.2 | 75.6 | 67.4 |
| w/o $\mathcal{L}_{gt}$ | 65.8 | 73.0 | 58.8 | 74.2 | 76.6 | 69.4 |
| w/o $\mathcal{L}_{global}$ | 65.4 | 72.7 | 58.4 | 74.0 | 76.3 | 68.6 |

Table 4: Ablation study of different components. O, E, and H denote as Overall, Easy, and Hard, respectively.

4, the model of w/o BEV-GIE gets performance degradation by 1.2% and 1.3% on Sr3D and Nr3D in terms of overall, respectively. For the Hard subsets of two datasets, the performance degradation is more significant. Because, in this setting, a scene contains more than two distractors (objects of the same category as the target object), and the lack of global information has a greater impact on performance. These results show that global information can provide additional supplementary information.

**Effectiveness of cross-modal consistency learning.** We study the performance of the model without cross-modal consistency learning (w/o xMCL), whose results are reported in #3. From the comparison of #1 and #3 in Tab. 4, it drops 0.9% and 1.4% in terms of overall accuracy on the two datasets. The consistent decreasing trends are also revealed for both Easy and Hard subset settings. Therefore, xMCL can also promote the learning of multi-modal learning.

**Effectiveness of introduced distillation losses.** To evaluate whether label distillation loss and global distillation loss can promote cross modal learning, we remove any of them and report their performance in #4 and #5 respectively. From the comparison of #4, #5 and #1, These distillations can further improve the performance.

| | Settings | Overall | Easy | Hard |
|---|---|---|---|---|
| #1 | BEV-based fusion | 74.6 | 75.9 | 71.3 |
| #2 | Plain concatenation | 71.6 | 73.2 | 67.7 |
| #3 | Attention-based query | 74.6 | 75.9 | 71.3 |
| #4 | Direct feature using | 73.6 | 74.9 | 70.6 |

Table 5: Comparison with alternative components on Sr3D.

Overall, according to the comparison of #1 with #2∼#5, the network can benefit from the key components, which demonstrates the effectiveness of xM_Match.

## Model Analysis

**BEV-based fusion *vs.* Plain concatenation**. To better demonstrate the effectiveness of the global and local fusion based on BEV map and transformer (BEV-based fusion), we study the performance of a general way to fuse global and local features, *i.e.,* directly features concatenation dubbed plain concatenation. The results are shown in Tab. 5. The plain concatenation achieves 71.6% on Sr3D. The BEV-based design gets better performance by a significant margin (3.0%) against the plain concatenation. Therefore, our method works by BEV-GIE design instead of as a feature concatenation strategy.

**Attention-based query *vs.* Direct feature using.** We directly use the features of mapping head (Directly feature using) instead of our attention-based query. The results are shown in #4 of Tab. 5. From the comparison of #3 in Tab. 4 and the #4 in Tab. 5, It can be observed that direct feature using strategy has a slight positive effect, while the attention-based query strategy improves significantly (1.7% improvement than w/o $\mathcal{L}_{xm}$ ). These results demonstrate that exploiting the matching of multi-modal feature space is rewarding for 3D grounding.

## Conclusion

We propose xM_Match, a novel language conditioned 3D object grounding method with explicit global information embedding and multi-modal consistency constraints. In contrast to existing two-stage 3D grounding methods, we co-encode multiple independently encoded object proposals into a horizontal view. This can address the mismatch of local visual representation and global sentence representation. Besides, we solve the feature space mismatch of visual space and corresponding label word space by cross-modal consistency constraint. In addition, we introduce two distillation losses to drive teacher-student network learning. According to the extensive experiments on three datasets, xM_Match gets better performance under both real-world and synthetic settings against state-of-the-art methods.

## Acknowledgements

This work was supported in part by the China Postdoctoral Science Foundation (No.2023M731957), in part by the National Natural Science Foundation of China under Grant 62306165, in part by the Shenzhen Key Laboratory

of next generation interactive media innovative technology (No.ZDSYS20210623092001004).

## References


Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 422–440.

Bakr, E.; Alsaedy, Y.; and Elhoseiny, M. 2022. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *NeurIPS*, 37146–37158.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in RGB-D scans using natural language. In *ECCV*, 202–221.

Chen, D. Z.; Wu, Q.; Nießner, M.; and Chang, A. X. 2022a. D3Net: A unified speaker-listener architecture for 3D dense captioning and visual grounding. In *ECCV*, 487–505.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022b. Language conditioned spatial relation reasoning for 3D object grounding. In *NeurIPS*, 1–14.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 5828–5839.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3D visual graph network for object grounding in point cloud. In *ICCV*, 3722–3731.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 9224–9232.

Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; and Bennamoun, M. 2020. Deep learning for 3d point clouds: A survey. *T-PAMI*, 43(12): 4338–4364.

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *ACMMM*, 2344–2352.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, 11108–11117.

Hu, R.; Monebhurrun, V.; Himeno, R.; Yokota, H.; and Costen, F. 2022. An uncertainty analysis on finite difference time-domain computations with artificial neural networks: improving accuracy while maintaining low computational costs. *IEEE Antennas and Propagation Magazine*, 65(1): 60–70.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-guided graph neural networks for referring 3D instance segmentation. In *AAAI*, 1610–1618.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022. Multi-view transformer for 3d visual grounding. In *CVPR*, 15524–15533.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, 417–433.

Jaritz, M.; Vu, T.-H.; Charette, R. d.; Wirbel, E.; and Pérez, P. 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 12605–12614.

Jiang, L.; Zhao, H.; Shi, S.; Liu, S.; Fu, C.-W.; and Jia, J. 2020. PointGroup: Dual-set point grouping for 3D instance segmentation. In *CVPR*, 4867–4876.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Landrieu, L.; and Simonovsky, M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *ICCV*, 4558–4567.

Li, G.; Muller, M.; Thabet, A.; and Ghanem, B. 2019. Deepgcns: Can GCNs go as deep as CNNs? In *ICCV*, 9267–9276.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3D-SPS: Single-stage 3d visual grounding via referred point progressive selection. In *CVPR*, 16454–16463.

Olaf Ronneberger, T. B., Philipp Fischer. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Peng, D.; Lei, Y.; Li, W.; Zhang, P.; and Guo, Y. 2021. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3D semantic segmentation. In *ICCV*, 7108–7117.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 9277–9286.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 5099–5108.

Roh, J.; Desingh, K.; Farhadi, A.; and Fox, D. 2022. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, 1046–1056. PMLR.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph CNN for learning on point clouds. *ACM Transactions On Graphics*, 38(5): 1–12.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2023. EDA: Explicit text-decoupling and dense alignment for 3D visual and language learning. In *CVPR*, 19231–19242.

Yan, X.; Zheng, C.; Li, Z.; Wang, S.; and Cui, S. 2020. PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, 5589–5598.

Yang, B.; Luo, W.; and Urtasun, R. 2018. Pixor: Real-time 3d object detection from point clouds. In *CVPR*, 7652–7660.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, 1856–1866.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*, 1791–1800.

Zhang, Y.; Li, M.; Xie, Y.; Li, C.; Wang, C.; Zhang, Z.; and Qu, Y. 2022. Self-supervised Exclusive Learning for 3D Segmentation with Cross-Modal Unsupervised Domain Adaptation. In *ACMMM*, 3338–3346.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021a. Weakly supervised semantic segmentation for large-scale point cloud. In *AAAI*, 3421–3429.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021b. Perturbed Self-Distillation: Weakly supervised large-scale point cloud semantic segmentation. In *ICCV*, 15520–15528.

Zhang, Z.; Hua, B.-S.; and Yeung, S.-K. 2019. ShellNet: Efficient point cloud convolutional neural Nnetworks using concentric shells statistics. In *ICCV*, 1607–1616.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2928–2937.