Context-aware Alignment and Mutual Masking for 3D-Language Pre-training

Zhao Jin[1] Munawar Hayat[2] Yuwei Yang[1] Yulan Guo[3] Yinjie Lei[1,✉]
[1]Sichuan University [2]Monash University [3]Sun Yat-sen University
jinzhao@stu.scu.edu.cn munawar.hayat@monash.edu yuwei@stu.scu.edu.cn
guoyulan@sysu.edu.cn yinjie@scu.edu.cn

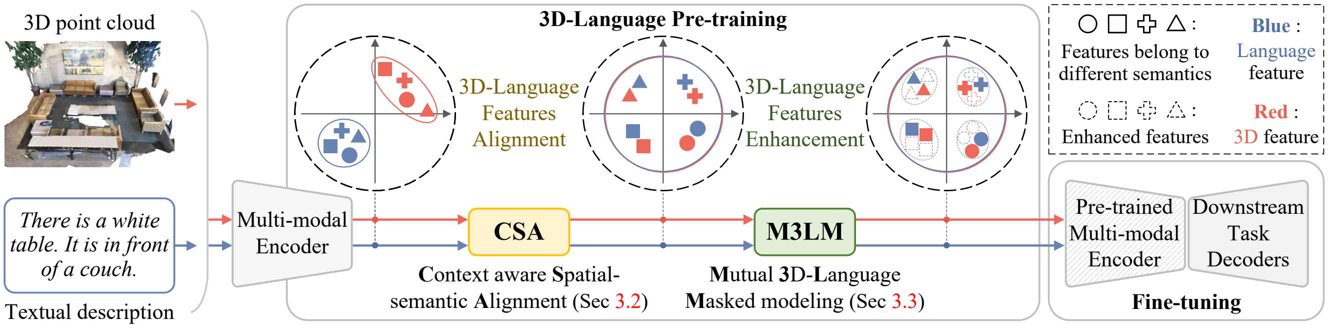

3D point cloud
There is a white table. It is in front of a couch.
Textual description
Multi-modal Encoder
3D-Language Pre-training
3D-Language Features Alignment
3D-Language Features Enhancement
CSA
M3LM
Context aware Spatial-semantic Alignment (Sec 3.2)
Mutual 3D-Language Masked modeling (Sec 3.3)
○ □ ✣ △ : Features belong to different semantics
Blue : Language feature
Enhanced features
Red : 3D feature
Pre-trained Multi-modal Encoder
Downstream Task Decoders
Fine-tuning

Figure 1. Illustration of our 3D-Language Pre-training framework. We first semantically align 3D-language features (Sec. 3.2), and then further enhance their granularity using mutual masked learning (Sec. 3.3). Our learned multi-modal features generalize well across various downstream tasks, including 3D visual grounding, 3D dense captioning and 3D question answering.

Abstract

3D visual language reasoning plays an important role in effective human-computer interaction. The current approaches for 3D visual reasoning are task-specific, and lack pre-training methods to learn generic representations that can transfer across various tasks. Despite the encouraging progress in vision-language pre-training for image-text data, 3D-language pre-training is still an open issue due to limited 3D-language paired data, highly sparse and irregular structure of point clouds and ambiguities in spatial relations of 3D objects with viewpoint changes. In this paper, we present a generic 3D-language pre-training approach, that tackles multiple facets of 3D-language reasoning by learning universal representations. Our learning objective constitutes two main parts. 1) Context aware spatial-semantic alignment to establish fine-grained correspondence between point clouds and texts. It reduces relational ambiguities by aligning 3D spatial relationships with textual semantic context. 2) Mutual 3D-Language Masked modeling to enable cross-modality information exchange. Instead of reconstructing sparse 3D points for which language can hardly provide cues, we propose masked proposal reasoning to learn semantic class and mask-invariant representations. Our proposed 3D-language pre-training method achieves promising results once adapted to various downstream tasks, including 3D visual grounding, 3D dense captioning and 3D question answering. Our codes are available at https://github.com/leolyj/3D-VLP

1. Introduction

3D Vision and Language (3D V+L) reasoning aims to jointly understand 3D point clouds and their textual descriptions. It lies at the intersection of 3D visual understanding and natural language processing, and plays an important role in applications *e.g.*, Metaverse, AR/VR and autonomous robots. 3D V+L reasoning has recently gained significant research interest, with multiple works tackling 3D visual grounding [1, 8, 33, 59], 3D dense captioning [11, 23, 58] and 3D question answering [3, 51, 54].

Despite promising progress made towards solving 3D visual reasoning tasks, the existing approaches are highly specialized and task specific. This is in contrast to multi-modal reasoning from RGB images, where the dominant approach is to pre-train a generic model on large scale image-text paired data, and then adapt this model for multiple downstream tasks. The pre-training step enables learning highly transferable and generic cross-modality representations via

[✉] Corresponding Author: Yinjie Lei (yinjie@scu.edu.cn) techniques such as *image-text feature alignment* [21,41] and *masked signal reconstruction* [10,28]. For RGB images, the transfer learning from pre-trained Vision-Language models achieves impressive results on numerous downstream tasks (*e.g*., image-text retrieval [28, 49], visual question answering [46,47] and image captioning [17,48]). However, due to unique challenges posed by irregular and unstructured point cloud data, 3D-language pre-training to learn a unified representation space, that can be transferred across tasks, has not yet been investigated in the existing literature.

As point clouds have different characteristics from 2D images, 3D-Language Pre-training (3D-LP) poses multiple unique challenges: **1)** Available 3D-language samples are limited. Compared to image-text samples that can be web crawled, the existing pairwise point cloud and language samples are much scarce. **2)** Point clouds are naturally unstructured. Unlike 2D images having pixels densely arranged in regular grids, point clouds are highly sparse and irregularly distributed. **3)** The spatial relations between 3D objects are complex, as they are not restricted to a 2D plane, and introduce ambiguities with viewpoint changes.

In this paper, we propose a 3D-language pre-training approach that aims to establish fine-grained interactions between point clouds and their textual descriptions, thus learning universal multi-modal features for various 3D V+L tasks, as illustrated in Fig 1. First, to bridge the distribution discrepancy between 3D geometric features and their text semantics, we propose a Context aware Spatial-semantic Alignment (CSA) strategy (Sec. 3.2). Different from the global contrastive learning in image-text, we align point cloud and language features from semantic and contextual perspectives separately, so that the spatial context between 3D objects and the semantic context in the language are simultaneously considered to overcome relational ambiguity. We further introduce Mutual 3D-Language Masked modeling (M3LM) (Sec. 3.3) that reconstructs the masked parts and enable meaningful cross-modal information exchange to enhance the feature of both modality. Due to the irregular structure and variable (unfixed) number of 3D points, existing masking methods that reconstruct raw input signal are not suitable to learn effective representation for point clouds. We propose to reconstruct the semantic class and high-level features of masked 3D objects by taking complementary information from language, which gives the model more meaningful objective than merely reconstructing the *xyz* of points. In our approach, we predict the semantic class of masked 3D objects and reconstruct momentum-distilled encoded features for the unmasked input. We jointly train the 3D-language model with our proposed multi-task learning objectives to learn and semantically align multi-modal features that generalize well across tasks. Through experiments on various downstream 3D V+L tasks, we demonstrate the versatility of our proposed 3D-language pre-training for three different tasks on ScanRefer [8], Scan2Cap [11] and ScanQA [3] benchmark datasets. Our main contributions are:

- We propose a pre-training method to learn transferable 3D-language representations to solve 3D visual grounding, 3D dense captioning and 3D question answering from a unified perspective.
- In order to jointly train point cloud and language encoders, we propose context-aware 3D-language alignment and mutual masked modeling strategies, which ensure that the learned multi-modal features are semantically-aligned and complement each other.
- We consistently surpass existing task-specific methods on ScanRefer [8] (+2.1 Acc@0.5), Scan2Cap [11] (+5.5 CIDEr@0.5) and ScanQA [3] (+1.3 EM@1) benchmark datasets, achieving new state-of-the-arts.

2. Related Work

2.1. 3D Vision and Language Reasoning

3D Visual Grounding. As one of the earliest investigated 3D V+L task, visual grounding on 3D point clouds has attracted a wide research interest [1, 6, 8, 19, 20, 30, 33, 53, 59, 61]. [8] is the first work to introduce the 3D visual grounding task with ScanRefer dataset, aiming to localize objects in the scene using language. ReferIt3D [1] also released two datasets similar to ScanRefer, *i.e*., Nr3D and Sr3D, in which object bounding box labels are given and localization is no more required. Most existing methods solve 3D visual grounding by extracting proposals that are then matched with textual descriptions. *Detection-based* methods [6, 19, 53, 61] apply 3D detectors (*e.g*., VoteNet [39]) to encode object proposals and match with language through fusion transformer. *Segmentation-based* methods [18, 59] extract object instances by segmentation backbone (*e.g*., PointGroup [22]), and learn the relationships between instance candidates and referring text. Recently [33] proposes a *single-stage* method 3D-SPS, which progressively fuses word and point features to select text-relevant keypoints to predict the referring bounding boxes. Different from above, [20, 57] propose to solve 3D visual grounding by aligning all mentioned object phrases with corresponding bounding boxes. However, they require extra object name information for the implementation of their method.

3D Dense Captioning. The captioning task is studied after 3D visual grounding, with the release of Scan2Cap dataset [11], where a model learns contextual relationships between object proposals with a graph module and decodes descriptive tokens for each object. [23] improves the graph module by progressively encoding multi-order relations. To better utilize 2D knowledge, [58] transfers color and texture from 2D images to 3D proposals in a teacher-student paradigm.

3D Question Answering. Recently, 3D Question Answering (3D-QA) on point clouds has been explored in [3,34,51,54]. On the ScanQA dataset [3], the 3D-QA task is defined as predicting the answer text as well as the question-related 3D bounding box. [3] proposes a transformer-based network to fuse 3D proposals with the question text embeddings and then generate answer from the fused features.

While the above 3D vision-language tasks have witnessed impressive research progress, the developed approaches are highly specialized and only work for the considered task. Developing a general-purpose solution capable of tackling multiple facets of 3D visual-language reasoning in a unified manner still remains unresolved. [6] shows that 3D visual grounding and dense captioning are mutually beneficial to each other and can be solved jointly. In this work, we go beyond and develop a generic approach, where a 3D vision-language model is trained to capture the intrinsic correlations between point clouds and language, to learn representations that can transfer across multiple tasks.

2.2. Vision-Language Pre-training

Learning a unified embedding space via Vision-Language Pre-training (VLP) [10,25,28,32,43] has recently gained significant attention. ALIGN [21] and CLIP [41] apply contrastive objectives on massive web-crawled image-text pairs to obtain a joint embedding space for vision and language. [27] propose to align paired image and text features with contrastive loss and then fuse them for cross-modal reasoning. Following this approach, [52] proposes a triple contrastive learning approach to additionally model intra-modal relationships, and [13] learns a codebook to bridge the modality gap for better alignment. For RGB images, these pre-trained models show remarkable performance once adapted to various downstream V+L tasks.

Despite astounding progress witnessed in VLP for RGB images [10,25,28,32,43], learning a unified cross-modality embedding space for 3D-language remains an open research problem, largely because of the unique challenges posed by the lack of paired 3D-language data, sparse and unordered nature of the 3D point clouds, and complex relationships between objects in 3D where the textual description may change with viewpoints. [60] transfers knowledge from CLIP [41] to 3D, however, since CLIP uses global contrastive learning, [60] can only do object-level reasoning. We directly pre-train on point clouds and texts to achieve more fine-grained scene-level understanding.

2.3. Mask-based Representation Learning

Self-supervised learning by reconstructing a masked portion of the input has emerged as a powerful tool across multiple domains. In the field of natural language processing (NLP), BERT [24] and its variants [31] have achieved remarkable generalization ability to a variety of tasks through masked language modeling (MLM). Once extended to visual understanding, masked auto-encoder (MAE) [15, 35, 36] and BERT-like masked image modeling (MIM) [5, 50, 55] have shown their promises. Further, conditional MLM [10, 28] and masked multi-modal modeling [7, 14, 16, 26, 42, 62] have developed masking based reconstruction schemes to learn multi-modal features in VLP. However, multi-modal masked reasoning is only explored in the context of RGB images, and not for point clouds. In this work, we investigate multi-modal masked learning on language and point clouds to mine mutual complementary information between 3D geometry and semantics.

3. Methodology

3.1. Model Architecture

As shown in Fig. 2, our model for 3D-language pre-training mainly consists of three parts: point cloud encoder, language encoder and cross-modal fusion decoder.

Point cloud encoder. The point cloud encoder is used to extract object proposal features. We first use VoteNet [39] with PointNet++ [40] backbone to encode and extract M object proposals from the input point cloud $\boldsymbol{C} \in \mathbb{R}^{N\times(3+D)}$, which represents N points with their 3D coordinates and D-dim auxiliary feature (color, normal vectors *etc.*) The encoded object proposal features $\mathbf{P} \in \mathbb{R}^{M\times D_p}$ will be used to predict M 3D bounding-boxes by a detection head. Motivated by [6], we subsequently feed $\mathbf{P} \in \mathbb{R}^{M\times D_p}$ into a transformer-based relation module to enhance the contextual relationships between objects. The resulting enhanced proposal features are $\mathbf{F}_p \in \mathbb{R}^{M\times D_p}$.

Language encoder. We encode the input textual description $T = \{w_i\}_{i=1}^{L}$ with GloVe [38] to obtain word representation $\mathbf{W} \in \mathbb{R}^{L\times 300}$, where L is the number of words in a sentence. $\mathbf{W}$ is then fed into a GRU cell to model word contextual relationships. We can obtain word features $\mathbf{F}_w = \{\mathbf{w}_i\}_{i=1}^{L} \in \mathbb{R}^{L\times D_w}$ and sentence feature $\mathbf{s} \in \mathbb{R}^{D_w}$. $\mathbf{s}$ is then used to predict the target object class of T.

Cross-modal fusion decoder. The encoded language and object features are fused through H layers of connected cross-attention [44] blocks, discussed in Sec. 3.3.

3.2. Context aware Spatial-semantic Alignment

Since the 3D object features and language features are extracted separately, they reside in their own spaces and directly fusing them leads to ambiguities. Different from traditional image-text contrastive learning which only achieves global feature alignment, 3D point clouds require more fine-grained cross-modality interaction, due to complex spatial relations between objects in a 3D scene. To this end, we propose a context aware spatial-semantic alignment (CSA) strategy that aligns 3D object and language features from semantic and contextual perspectives, as shown in Fig. 3.

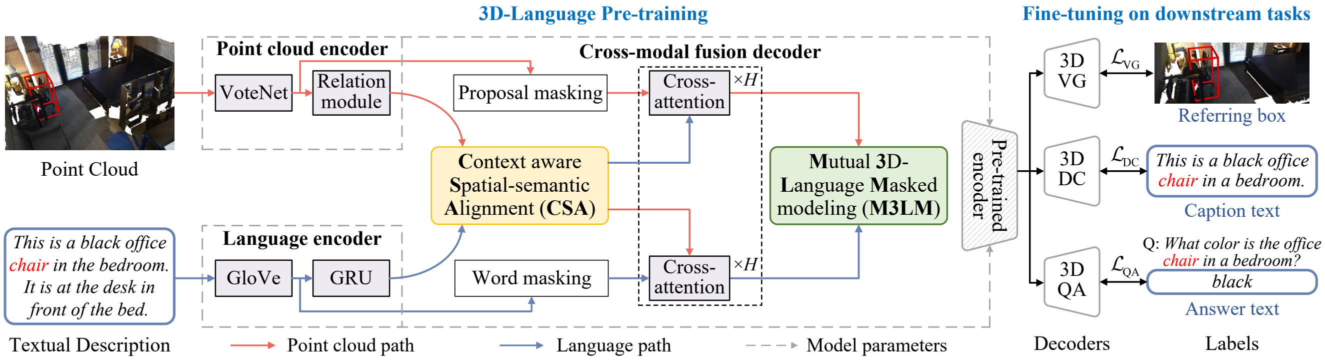

3D-Language Pre-training
Fine-tuning on downstream tasks
Point cloud encoder
VoteNet
Relation module
Cross-modal fusion decoder
Proposal masking
Cross-attention
×H
Context aware Spatial-semantic Alignment (CSA)
Mutual 3D-Language Masked modeling (M3LM)
Pre-trained encoder
Point Cloud
Language encoder
GloVe
GRU
Word masking
Cross-attention
×H
This is a black office chair in the bedroom. It is at the desk in front of the bed.
Textual Description
Point cloud path
Language path
Model parameters
3D VG
$\mathcal{L}_{\mathrm{VG}}$
Referring box
3D DC
$\mathcal{L}_{\mathrm{DC}}$
This is a black office chair in a bedroom.
Caption text
3D QA
$\mathcal{L}_{\mathrm{QA}}$
Q: What color is the office chair in a bedroom?
black
Answer text
Decoders
Labels

Figure 2. The overview of our method. Given input point cloud with textual description, we first encode them with a point cloud encoder and language encoder separately. Then we jointly pre-train multi-modal encoders with context aware spatial-semantic alignment (see Fig. 3) and mutual 3D-language masked modeling (see Fig. 4). The pre-trained encoder will be adapted to multiple downstream 3D V+L tasks for fine-tuning, *i.e*., 3D Visual Grounding (3DVG), 3D Dense Captioning (3DDC) and 3D Question Answering (3DQA).

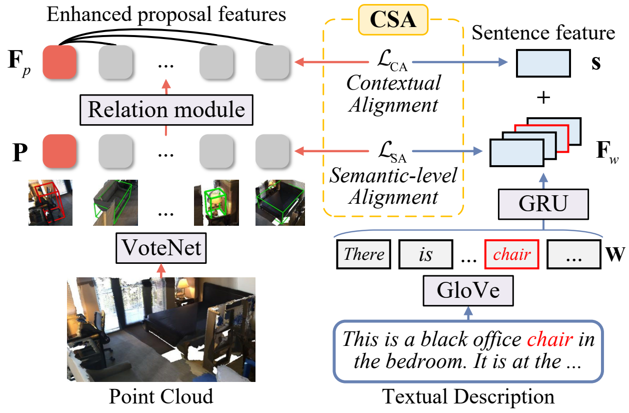


Figure 3. Illustration of CSA (Sec. 3.2) module that aligns 3D and language features from both semantic and contextual perspectives.

Semantic-level Alignment (SA). We align the object proposal features $\mathbf{P}$ and word features $\mathbf{F}_w$ belonging to the same semantic category. For instance, given a sentence "*the chair in front of table*", and paired point cloud as input, and object proposal feature $\mathbf{p}$ classified as "Chair". Then we align $\mathbf{p}$ with the word feature $\mathbf{w}$ located in the position of word "*chair*". We apply a contrastive loss on $\mathbf{P}$ and $\mathbf{F}_w$ to achieve the semantic-level cross-modality alignment:

$$\mathcal{L}_{\mathrm{SA}} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{L_k}\left[\log\frac{\exp\left(\mathbf{p}^k\cdot\mathbf{w}_i^k/\tau\right)}{\sum_{n=1}^{N}\sum_{j=1}^{L_k}\exp\left(\mathbf{p}^k\cdot\mathbf{w}_j^n/\tau\right)}\right], \tag{1}$$

where N and L_k are the training batch size and the total number of words in batch k. $\mathbf{p}^k$ and $\mathbf{w}_i^k$ are from a paired input and they belong to the same semantic class. τ is the temperature parameter. At this stage, we align the object proposal feature before attending to its surrounding objects, which can be considered as the non-contextual alignment. Next, we introduce our context-level alignment strategy.

Contextual Alignment (CA). During feature encoding, a relation module and a GRU cell are respectively used to model the contextual relationship within each modality. Even though the features of individual objects are semantically aligned with linguistic features, 3D contextual features may still introduce ambiguities with linguistic features after relational learning due to complex spatial relations between 3D objects. Since the contextual words in the sentence describe the relation of the referring objects, they can provide cues for the point cloud encoder to learn better spatial context. So we further align the enhanced proposal features $\mathbf{F}_p$ and sentence features $\mathbf{s}$ to achieve contextual alignment:

$$\mathcal{L}_{\mathrm{CA}} = -\frac{1}{N}\sum_{k=1}^{N}\left[\log\frac{\exp\left(\mathbf{f}_p^k\cdot\mathbf{s}^k/\tau\right)}{\sum_{n=1}^{N}\exp\left(\mathbf{f}_p^k\cdot\mathbf{s}^n/\tau\right)}\right], \tag{2}$$

where $\mathbf{s}^k$ is the sentence feature in batch k, and $\mathbf{f}_p^k$ is the referred object feature from $\mathbf{F}_p$. Our final loss for the CSA module is given by $\mathcal{L}_{\mathrm{CSA}} = \mathcal{L}_{\mathrm{SA}} + \mathcal{L}_{\mathrm{CA}}$.

3.3. Mutual 3D-Language Masked Modelling

After aligning object features with corresponding linguistic features, they are fused and then used for joint V+L reasoning. To enable meaningful bidirectional interaction between language and point clouds, as shown in Fig. 4, we propose mutual 3D-language masked modeling (M3LM), that masks both 3D proposals and text to jointly learn mutual complementary spatial and semantic information.

3D-Language Mutual masking. We jointly mask the predicted 3D proposals $\mathbf{P}$ and the word representations $\mathbf{W}$ before cross-modal fusion. Specifically, we randomly mask out 75% of the 3D proposals and feed the remaining visible proposal features $\hat{\mathbf{P}}$ to the transformer encoder of the relation module. After relational learning, the encoded visible object features are given by $\hat{\mathbf{F}}_p$. For the language, we randomly mask out 20% of words and replace the masked part

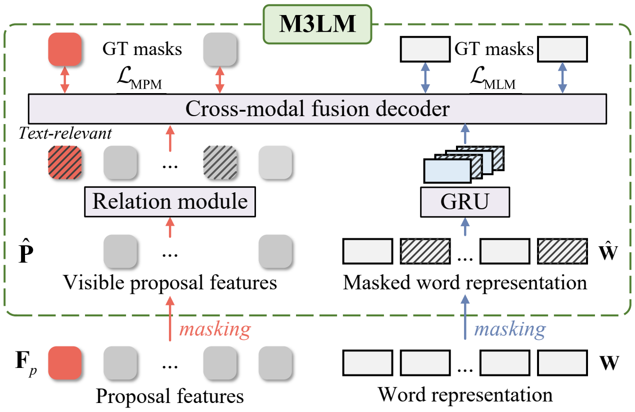


Figure 4. Illustration of M3LM (Sec. 3.3) module. We jointly mask on proposal features and word representation to enable learning cross-modal complementary information.

with the "unk" token to obtain $\hat{\mathbf{W}}$. Then after GRU cell, they are encoded as masked word features $\hat{\mathbf{F}}_w$.

Masked Proposal Modeling (MPM). Before feeding $\hat{\mathbf{F}}_p$ to the fusion decoder, we concatenate mask tokens $\mathbf{M}$ with the visible object features to obtain the full token sets, similar to MAE [15]. The full token sets $\hat{\mathbf{S}}_p = \text{concat}(\hat{\mathbf{F}}_p, \mathbf{M})$ for decoder are also added with positional embeddings $\mathbf{e}_p$, to incorporate the locations of the proposals and masks in the 3D scene. We calculate positional embedding for each token by applying a linear layer on its 27-dim predicted 3D box center and corner coordinates. This preserves the spatial relationship between the mask and visible objects and enables masked reasoning based on language.

In the cross-modal fusion decoder, the cross-attention between masked proposal features and word features is calculated to obtain the fused feature $\hat{\mathbf{F}}_{pw}$ for all proposals:

$$\hat{\mathbf{F}}_{pw} = \text{CrossAtt}\left(\hat{\mathbf{S}}_p + \mathbf{e}_p, \mathbf{F}_w, \mathbf{F}_w\right), \tag{3}$$

where $\mathbf{F}_w$ denotes the word features for input text. After the fusion process, $\hat{\mathbf{F}}_{pw}$ is used to predict the semantic class of the masked objects. However, we notice *information bias* exists between input point cloud and text *i.e.*, each of input text only describes a certain spatial relation about the referring object in the scene. While predicting masked objects, the textual description can hardly provide complementary information if the location of mask is away from the descriptive region. As the correspondence between input text and the referring object is given, we propose to only calculate loss for the *text-relevant* masks $\mathbf{M}_t$. Specifically, for each mask, if the Intersection-over-Union (IoU) between its predicted box and ground-truth box of the text is higher than 0.25, it is considered as text-relevant. In addition to the semantic class reasoning for text-relevant objects, we also encourage all the masks to learn intrinsic representation that are robust to masking. To this end, we force the masked proposals after relational learning and cross-modal fusion to be consistent with them in the unmasked branch. Specifically, the unmasked branch is momentum-updated by the exponential moving average (EMA) on masked branch without requiring gradient updates, and the full sets of proposals are fed to it. Then an $\mathcal{L}_1$ loss is defined between the output features of masked branch and unmasked branch. The MPM loss includes the Cross-Entropy loss for the text-relevant mask predictions $f_p(\mathbf{M}_t)$ and $\mathcal{L}_1$ loss for all masks:

$$\mathcal{L}_{\text{MPM}} = \text{CE}\left(f_p(\mathbf{M}_t), \mathbf{M}_{gt}\right) + \mathcal{L}_1\left(\hat{\mathbf{F}}_{pw}, \dot{\mathbf{F}}_{pw}\right), \tag{4}$$

where $\mathbf{M}_{gt}$ are the ground-truth classes for masked objects, and $\dot{\mathbf{F}}_{pw}$ denotes fused features from the unmasked branch.

Masked Language Modeling (MLM). We also predict the masked words conditioned on corresponding point cloud and unmasked part of the sentence. The masked word features $\hat{\mathbf{F}}_w$ are fused with proposal features $\mathbf{F}_p$ by:

$$\hat{\mathbf{F}}_{wp} = \text{CrossAtt}\left(\hat{\mathbf{F}}_w, \mathbf{F}_p, \mathbf{F}_p\right), \tag{5}$$

where $\hat{\mathbf{F}}_{wp}$ denotes the word features after fusing with proposal feature and is used to predict the vocabulary probability $f_w(\hat{\mathbf{F}}_w)$. The MLM loss is calculated as:

$$\mathcal{L}_{\text{MLM}} = \text{CE}\left(f_w(\hat{\mathbf{F}}_w), y_w\right), \tag{6}$$

where y_w are ground-truth language token labels for masks. The total loss of mutual 3D-language masked modeling is given by $\mathcal{L}_{\text{M3LM}} = \mathcal{L}_{\text{MPM}} + \mathcal{L}_{\text{MLM}}$.

3.4. Training Strategy

We jointly train the point cloud encoder with a detection loss $\mathcal{L}_{\text{det}}$ as in [6]. For the language encoder, to predict object class, the language classification loss $\mathcal{L}_{\text{lang}}$ is used for supervision. We also include the object-language matching task to predict which one of the detected bounding boxes matches the textual description, by using a cross-entropy loss $\mathcal{L}_{\text{match}}$. The overall pre-training loss is the weighted sum of all these losses: $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{det}} + 0.3\mathcal{L}_{\text{lang}} + 0.3\mathcal{L}_{\text{match}} + 5\mathcal{L}_{\text{CSA}} + 0.2\mathcal{L}_{\text{M3LM}}$, where the weights are empirically set to balance the order of magnitude. During finetuning stage, we disable $\mathcal{L}_{\text{M3LM}}$ and add the task-specific loss with rest of the $\mathcal{L}_{\text{pre}}$. Note that the cross-modal fusion decoder is only used for M3LM during pre-training and is replaced with a light-weight structure as per the final task.

4. Experiments

4.1. Settings and Implementation Details

3D Visual Grounding. We use ScanRefer [8] dataset for visual grounding evaluations. It includes $51,583$ descriptions of $11,046$ objects, which are from 800 scenes of

Table 1. Comparison of 3D visual grounding results on ScanRefer [8] dataset. We report accuracy (Acc) under 0.25 and 0.5 IoU, which is calculated between predicted and ground-truth 3D bounding boxes. Compared with highly specialized methods, our approach achieves consistent gains across different experimental settings. The best and second best results are in **bold** and underlined.

Method	Publication	Data	Unique		Multiple		Overall	
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Validation Set								
ScanRefer [8]	ECCV2020	3D only	67.64	46.19	32.06	21.26	38.97	26.10
InstanceRefer [59]	ICCV2021	3D only	77.45	**66.83**	31.27	24.77	40.23	32.93
3DVG-Transformer [61]	ICCV2021	3D only	77.16	58.47	38.38	28.70	45.90	34.47
3DJCG [6]	CVPR2022	3D only	78.75	61.30	40.13	30.08	47.62	36.14
3D-SPS [33]	CVPR2022	3D only	**81.63**	64.77	39.48	29.61	47.65	36.43
Ours		3D only	79.35	62.60	**42.54**	**32.18**	**49.68**	**38.08**
ScanRefer [8]	ECCV2020	2D+3D	76.33	53.51	32.73	21.11	41.19	27.40
InstanceRefer [59]	ICCV2021	2D+3D	77.45	66.83	31.27	24.77	40.23	32.93
3DVG-Transformer [61]	ICCV2021	2D+3D	81.93	60.64	39.30	28.42	47.57	34.67
3DJCG [6]	CVPR2022	2D+3D	83.47	64.34	41.39	30.82	49.56	37.33
3D-SPS [33]	CVPR2022	2D+3D	84.12	66.72	40.32	29.82	48.82	36.98
D3Net [9]	ECCV2022	2D+3D	-	**70.35**	-	30.05	-	37.87
Ours		2D+3D	**84.23**	64.61	**43.51**	**33.41**	**51.41**	**39.46**
Online Benchmark								
ScanRefer [8]	ECCV2020	2D+3D	68.59	43.53	34.88	20.97	42.44	26.03
3DVG-Transformer [61]	ICCV2021	2D+3D	75.76	55.15	42.24	29.33	49.76	35.12
InstanceRefer [59]	ICCV2021	3D only	77.82	66.69	34.57	26.88	44.27	35.80
BUTD-DETR [20]	ECCV2022	3D only	78.48	54.99	39.34	24.80	48.11	31.57
3DJCG [6]	CVPR2022	2D+3D	76.75	60.59	43.89	31.17	51.26	37.76
D3Net [9]	ECCV2022	2D+3D	79.23	**68.43**	39.05	30.74	48.06	39.19
Ours		2D+3D	**81.37**	62.41	**45.44**	**33.17**	**53.49**	**39.72**

ScanNet [12]. To jointly evaluate the performance of detection and grounding, we report the grounding accuracy with different IoU scores s between predicted and ground truth bounding boxes, denoted as Acc@sIoU. Following previous works [6, 8, 33, 59, 61], we use Acc@0.25IoU and Acc@0.5IoU as the main evaluation metrics for 3D visual grounding. The accuracy is reported on "unique" and "multiple" categories. If there is only a single object of a class in the scene, it is regarded as "unique", otherwise "multiple". We also report overall accuracy of all samples.

3D Dense Captioning. We follow Scan2Cap [11] for the evaluations on dense captioning. Scan2Cap dataset is based on ScanRefer [8], with descriptions longer than 30 tokens truncated, and two special tokens [SOS] and [EOS] added to the start and end of each description. We use the metrics CIDEr [45], BLEU [37], METEOR [4] and ROUGE-L [29], which are briefly denoted by C, B-4, M and R, respectively. Similar to grounding, the above metrics m for captioning are also calculated with different IoU scores as $m@k\text{IoU} = \frac{1}{N}\sum_{i=1}^{N} m_i u_i$, where $u_i \in \{0, 1\}$ is set to 1 if the IoU score for the i-th detected box is greater than k, and 0 otherwise.

3D Question Answering. ScanQA [3] is a recently proposed dataset for 3D question answering on point clouds. It collects $41,363$ questions and $58,191$ free-form answers for the ScanNet [12]. We follow [3] and adopt exact matches EM@K as the evaluation metric, which means the percentage of predictions in which the top-K predicted answers match with any one of the ground-truth answers. Since some of the questions can be answered with multiple expressions, we also report captioning metrics CIDEr [45], BLEU [37], METEOR [4], ROUGE-L [29] and SPICE [2].

Implementation Details. All experiments are conducted using a single NVIDIA A10 24GB GPU. We pre-train our model on ScanRefer [8] dataset for 150 epochs, and then fine-tune it for 100, 50 and 50 epochs, respectively, for 3D visual grounding, 3D dense captioning and 3D question answering. We follow [6, 8, 11, 33] to evaluate grounding and captioning using "3D only" and "2D + 3D" inputs, where "3D only" means the input includes "*xyz* + *normals* + *RGB*" and "2D + 3D" means "*xyz* + *normals* + *multiviews*", where "*multiviews*" are the 128-dim image features extracted as in [8]. See Appendix-A for other details.

4.2. Experimental Results

3D Visual Grounding. Tab. 1 compares different approaches on 3D visual grounding. We observe that the overall accuracies (*i.e*., Acc@0.25 and Acc@0.5) of our method are higher than existing methods. Compared with 3DJCG [6], we achieve ∼2.2% and 2.0% improvement on the ScanRefer online benchmark in terms of overall Acc@0.25 and Acc@0.5 respectively. Consistent gains are seen on the validation set. Notably, our gains are more pronounced for "multiple" and "overall" than "unique", suggesting that our approach understands complex 3D scenes better, benefiting from the well-explored cross-modal contextual and mutual complementary information during the pre-training stage.

Table 2. Comparison of 3D dense captioning results on Scan2Cap [11] dataset. All the listed methods adopt VoteNet [39] as the 3D detector. We report CIDEr (C) [45], BLEU-4 (B-4) [37], METEOR (M) [4] and ROUGE-L (R) [29] under 0.25 and 0.5 IoU of predicted bounding box. The best and second best results are in **bold** and underlined.

Method	Publication	Input	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5
Scan2Cap [11]	CVPR2021	3D only	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57
X-Trans2Cap [58]	CVPR2022	3D only	58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97
MORE [23]	ECCV2022	3D only	58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33
3DJCG [6]	CVPR2022	3D only	60.86	39.67	27.45	**59.02**	47.68	31.53	24.28	51.08
Ours		3D only	**64.09**	**39.84**	**27.65**	58.78	**50.02**	**31.87**	**24.53**	**51.17**
Scan2Cap [11]	CVPR2021	2D + 3D	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.48
X-Trans2Cap [58]	CVPR2022	2D + 3D	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
MORE [23]	ECCV2022	2D + 3D	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
D3Net(CIDEr+lis.) [9]	ECCV2022	2D + 3D	-	-	-	-	47.32	24.76	21.66	43.62
3DJCG [6]	CVPR2022	2D + 3D	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80
Ours		2D + 3D	**70.73**	**41.03**	**28.14**	**59.72**	**54.94**	**32.31**	**24.83**	**51.51**

Table 3. Comparison of 3D question answering results on ScanQA [3] dataset. We report EM@1 and EM@10 with additional captioning metrics to jointly reflect the accuracy of the predicted answer. The best and second best results are in **bold** and underlined.

Method	Split	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
VoteNet [39] + MCAN [56]	Validation set	17.33	45.54	28.09	16.72	10.75	6.24	29.84	11.41	54.68	10.65
ScanRefer [8] + MCAN [56]		18.59	46.76	26.93	16.59	11.59	7.87	30.03	11.52	55.41	11.28
ScanQA [3]		20.28	50.01	29.47	19.84	14.65	9.55	32.37	12.60	61.66	11.86
Ours		**21.65**	**50.46**	**30.53**	**21.33**	**16.67**	**11.15**	**34.51**	**13.53**	**66.97**	**14.18**
VoteNet [39] + MCAN [56]	Test w/ objects	19.71	50.76	29.46	17.23	10.33	6.08	30.97	12.07	58.23	10.44
ScanRefer [8] + MCAN [56]		20.56	52.35	27.85	17.27	11.88	7.46	30.68	11.97	57.36	10.58
ScanQA [3]		23.45	**56.51**	31.56	21.39	15.87	**12.04**	34.34	13.55	67.29	11.99
Ours		**24.58**	55.97	**33.15**	**22.65**	**16.38**	11.23	**35.97**	**14.16**	**70.18**	**12.71**
VoteNet [39] + MCAN [56]	Test w/o objects	18.15	48.56	29.63	17.80	11.57	7.10	29.12	11.68	53.34	10.36
ScanRefer [8] + MCAN [56]		19.04	49.70	26.98	16.17	11.28	7.82	28.61	11.38	53.41	10.63
ScanQA [3]		20.90	**54.11**	30.68	21.20	15.81	10.75	31.09	12.59	60.24	11.29
Ours		**21.56**	53.89	**31.48**	**23.56**	**19.62**	**15.84**	**31.79**	**13.13**	**63.40**	**12.53**

On the "unique" category, the methods [9, 59] using PointGroup [22] achieve higher Acc@0.5, as it predicts better bounding boxes. However, our method achieves the best performance of all VoteNet-based approaches.

3D Dense Captioning. For 3D dense captioning, we add a captioning head following 3DJCG [6] and fine-tune the network with additional captioning loss. From results in Tab. 2, our method outperforms other approaches that adopt VoteNet as the detector. Notably, we get significant improvements of CIDEr [45] for 3D dense captioning. We achieve 6.0% and 5.4% improvement in C@0.25 and C@0.5 compared with 3DJCG [6], suggesting that our generated descriptions are more in line with human consensus. It indicates that compared with the joint training strategy of 3DJCG, our proposed pre-training strategy helps the model learn generic 3D-language features, resulting in more accurate and consensual textual descriptions for 3D point clouds.

3D Question Answering. To evaluate our pre-trained model on 3D question answering, we follow [3] to add a 3D-language fusion module with an answer classification head to replace the cross-modal fusion module. In Tab. 3, the quantitative comparison shows our method surpasses existing models on both EM and captioning metrics. Note that the textual input of 3D-QA is a question, which is distinct from the referring expression of grounding. Our method still benefits the 3D question-answering task as the learned multi-modal features after pre-training are semantically-aligned and enhanced in granularity.

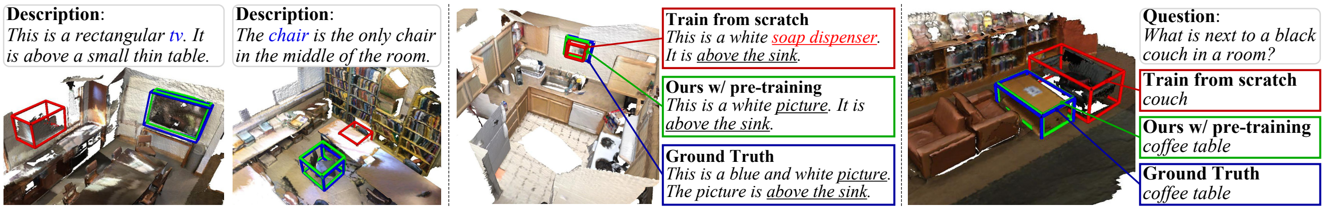
4.3. Ablation Study

Effectiveness of pre-training tasks. We analyze contributions of different pre-training objectives in Tab. 4. Here, "train from scratch" means directly training on the downstream task without pre-training. We gradually add our objectives *i.e.* semantic-level alignment (SA), contextual alignment (CA), masked proposal modeling (MPM) and masked language modeling (MLM). We enumerate the main metrics for 3D visual grounding, 3D dense captioning and 3D question answering. Tab. 4 shows the proposed pre-training objectives progressively improve downstream performance. Notably, MPM and SA achieve the most significant gains on Acc@0.25 (+0.94) and Acc@0.5 (+1.09) respectively, which indicate the effectiveness of our alignment and masked modeling in 3D-LP. Fig. 5 qualitatively shows our pre-training yields more accurate localization, caption and answer results than training from scratch.

Reconstruction objective in MPM. In the MPM of mutual 3D-language masked modeling (Sec. 3.3), we predict the

Table 4. Ablation on our proposed pre-training objectives. We compare training from scratch by progressively adding different pre-training objectives to evaluate their effect. The performances of each method on three downstream tasks are shown.

Method	CSA		M3LM		3D Visal Grounding		3D Dense Captioning				3D-QA	
	SA	CA	MLM	MPM	Acc@0.25	Acc@0.5	C@0.5	B-4@0.5	M@0.5	R@0.5	EM@1	EM@10
Train from scratch					50.04	37.01	49.17	30.50	24.06	50.07	20.32	49.41
+ SA	✓				50.19	38.10	51.26	30.50	24.48	50.75	20.86	50.29
+ CA		✓			50.22	37.63	51.59	30.23	24.32	50.29	20.78	50.01
+ CSA	✓	✓			50.33	38.20	52.11	30.42	24.53	50.19	21.01	50.34
+ CSA + MLM	✓	✓	✓		50.81	38.91	51.25	31.13	24.42	50.96	21.28	50.46
+ CSA + MPM	✓	✓		✓	51.27	39.12	53.59	30.97	24.54	51.08	21.24	50.49
+ CSA + M3LM	✓	✓	✓	✓	**51.41**	**39.46**	**54.94**	**32.31**	**24.83**	**51.51**	**21.65**	**50.54**


Description: *This is a rectangular tv. It is above a small thin table.*

Description: *The chair is the only chair in the middle of the room.*

Train from scratch
This is a white soap dispenser. It is above the sink.

Ours w/ pre-training
This is a white picture. It is above the sink.

Ground Truth
This is a blue and white picture. The picture is above the sink.

Question: *What is next to a black couch in a room?*

Train from scratch
couch

Ours w/ pre-training
coffee table

Ground Truth
coffee table

(a) 3D visual grounding on ScanRefer [8] (b) 3D dense captioning on Scan2Cap [11] (c) 3D question answering on ScanQA [3]

Figure 5. Visual comparison on three downstream 3D V+L tasks. Blue, red and green represent the ground-truth label, results of training from scratch and ours with complete pre-training objectives, respectively.

Table 5. Ablation on reconstruction objectives for MPM.

Reconstruction objective	Acc@0.5
None	38.20
Raw points (*xyz*)	38.18
Raw points (*xyz* + *RGB*)	38.35
Proposal class	38.52
Text-relevant proposal class	38.85
Text-relevant proposal class + masked feature	39.12

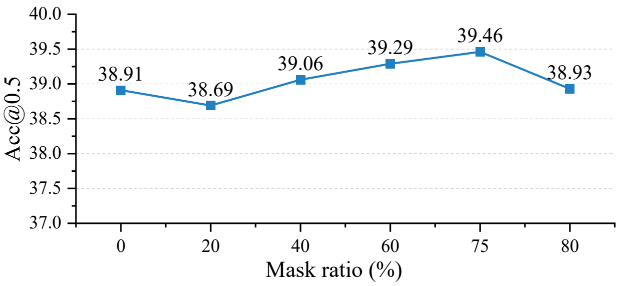


Figure 6. Performance curve with different proposal mask ratio.

semantic class of text-relevant masked proposals and constrain the consistency of fused feature after masking. And it is different from the common approaches (*e.g.*, MAE [15]) that reconstruct raw pixel value of input image. We conduct experiments to compare the effects of different reconstruction objectives for 3D point clouds. As shown in Tab. 5, reconstruction of raw points (*i.e. xyz* or *xyz* + *RGB*, see Appendix-A for implementation details) leads to unsatisfactory results, while reconstruction of semantic class is beneficial, indicating our MPM is more suitable for unstructured point clouds. With the additional text-relevant filtering strategy and momentum-distilled masked feature reconstruction, our method achieves the best performance.

Influence of mask ratio. The ratio of masked tokens in full token sets is an important parameter in mask-based representation learning. Fig. 6 illustrates the trend in Acc@0.5 with different proposal mask ratio R_p. We found the fine-tuning performance benefits from relative high proposal mask ratio, and the best result is obtained in $R_p = 75\%$. We analyse it is because the large information redundancy in predicted proposals since an object can be encoded by multiple proposals. However, when R_p is too large ($> 75\%$), the performance drops as the visible proposals are too few.

5. Conclusion

In this work, we present a 3D point cloud and language pre-training framework to learn universal representation that transfers well across multiple 3D V+L tasks. Our method jointly trains point cloud and language encoders with context-aware alignment and mutual masked modeling to achieve fine-grained 3D-language interaction. With extensive experiments, we verify the effectiveness of our proposed method. And the learned multi-modal 3D-language features, once evaluated on multiple 3D V+L benchmarks, consistently surpass existing task-specific methods. In the future work, with the availability of large-scale 3D-language data, we will explore the scalability and and cross-domain generalization of our approach.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No.62276176), was also partially supported by the National Natural Science Foundation of China (No.U20A20185, 61972435).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, pages 422–440, 2020. 1, 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398, 2016. 6

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 1, 2, 3, 6, 7

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop*, pages 65–72, 2005. 6, 7

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[6] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3DJCG: A unified framework for joint dense captioning and visual grounding on 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 2, 3, 5, 6, 7

[7] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision (ECCV)*, pages 202–221, 2020. 1, 2, 5, 6, 7

[9] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3Net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in RGB-D scans. *arXiv preprint arXiv:2112.01551*, 2021. 6, 7

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120, 2020. 2, 3

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021. 1, 2, 6, 7

[12] 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 (CVPR)*, pages 5828–5839, 2017. 6

[13] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15651–15660, 2022. 3

[14] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3, 5, 8

[16] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Chen Wu, Xiujun Shu, and Bo Ren. VLMAE: Vision-language masked autoencoder. *arXiv preprint arXiv:2208.09374*, 2022. 3

[17] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989, 2022. 2

[18] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3D instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021. 2

[19] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15524–15533, 2022. 2

[20] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Looking outside the box to ground language in 3D scenes. *arXiv preprint arXiv:2112.08879*, 2021. 2, 6

[21] 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 (ICML)*, pages 4904–4916, 2021. 2, 3

[22] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2020. 2, 7

[23] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. MORE: Multi-order relation mining for dense captioning in 3D scenes. *arXiv preprint arXiv:2203.05203*, 2022. 1, 2, 7

[24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, 2019. 3

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021. 3

[26] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022. 3

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 3

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, pages 121–137, 2020. 2, 3

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop*, pages 74–81, 2004. 6, 7

[30] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-It-in-RGBD: A bottom-up approach for 3D visual grounding in RGBD images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6032–6041, 2021. 2

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[33] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3D-SPS: Single-stage 3D visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16454–16463, 2022. 1, 2, 6

[34] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. *arXiv preprint arXiv:2210.07474*, 2022. 3

[35] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-MAE: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 3

[36] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022. 3

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 6, 7

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3

[39] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9277–9286, 2019. 2, 3, 7

[40] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017. 3

[41] 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 (ICML)*, pages 8748–8763, 2021. 2, 3

[42] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, 2022. 3

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[45] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6, 7

[46] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[47] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 2

[48] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[49] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. COOKIE: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2208–2217, 2021. 2

[50] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 3

[51] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. CLEVR3D: Compositional language and elementary visual reasoning for question answering in 3D real-world scenes. *arXiv preprint arXiv:2112.11691*, 2021. 1, 3

[52] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15671–15680, 2022. 3

[53] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2D semantics assisted training for 3D visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1856–1866, 2021. 2

[54] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3D question answering. *arXiv preprint arXiv:2112.08359*, 2021. 1, 3

[55] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, 2022. 3

[56] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 7

[57] Zhihao Yuan, Xu Yan, Zhuo Li, Xuhao Li, Yao Guo, Shuguang Cui, and Zhen Li. Toward explainable and fine-grained 3D grounding through referring textual phrases. *arXiv preprint arXiv:2207.01821*, 2022. 2

[58] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8563–8573, 2022. 1, 2, 7

[59] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1791–1800, 2021. 1, 2, 6, 7

[60] 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 (cvpr)*, pages 8552–8562, 2022. 3

[61] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021. 2, 6

[62] Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. MAMO: Masked multimodal modeling for fine-grained vision-language representation learning. *arXiv preprint arXiv:2210.04183*, 2022. 3