# 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment

Ziyu Zhu[1,2*] Xiaojian Ma[2] Yixin Chen[2] Zhidong Deng[1✉] Siyuan Huang[2✉] Qing Li[2✉]
[1]Tsinghua University [2]National Key Laboratory of General Artificial Intelligence, BIGAI, China
3d-vista.github.io

## Abstract

*3D vision-language grounding (3D-VL) is an emerging field that aims to connect the 3D physical world with natural language, which is crucial for achieving embodied intelligence. Current 3D-VL models rely heavily on sophisticated modules, auxiliary losses, and optimization tricks, which calls for a simple and unified model. In this paper, we propose 3D-VisTA, a pre-trained Transformer for 3D Vision and Text Alignment that can be easily adapted to various downstream tasks. 3D-VisTA simply utilizes self-attention layers for both single-modal modeling and multi-modal fusion without any sophisticated task-specific design. To further enhance its performance on 3D-VL tasks, we construct ScanScribe, the first large-scale 3D scene-text pairs dataset for 3D-VL pre-training. ScanScribe contains 2,995 RGB-D scans for 1,185 unique indoor scenes originating from ScanNet and 3R-Scan datasets, along with paired 278K scene descriptions generated from existing 3D-VL tasks, templates, and GPT-3. 3D-VisTA is pre-trained on ScanScribe via masked language/object modeling and scene-text matching. It achieves state-of-the-art results on various 3D-VL tasks, ranging from visual grounding and dense captioning to question answering and situated reasoning. Moreover, 3D-VisTA demonstrates superior data efficiency, obtaining strong performance even with limited annotations during downstream task fine-tuning.*

## 1. Introduction

Aligning the 3D physical world with natural language is a crucial step towards embodied artificial intelligence [18, 26, 37], where intelligent agents can understand and further execute human instructions in the real world [5, 29]. Recently, 3D vision-language (3D-VL) tasks have attracted growing interest [19], including 3D visual grounding [8, 1], dense captioning [11], grammar learning [23], question answering [3, 56], and situated reasoning [36].

However, most of the models developed for 3D-VL only focus on one or two of these 3D-VL tasks and employ task-specific designs [7, 3, 36, 35, 10]. For instance, 3D-SPS [35] and BUTD-DETR [27] progressively discover the target object by attending VL features and detecting objects in each layer. 3DVG [55], MVT [24], and ViL3DRel [10] improve 3D visual grounding by explicitly infusing spatial relation information into the model design. 3DJCG [7] jointly learns 3D dense captioning and visual grounding via a shared 3D object proposal module [16] with two separate task-specific heads [7]. Additionally, training these models often requires manually specified auxiliary losses (*e.g.*, 3D object detection/classification and text classification [35, 24, 7, 3, 36]) or optimization tricks (*e.g.*, knowledge distillation [4, 53] ). The lack of a simple and unified approach creates a significant gap in developing a general-purpose 3D-VL model.

To fill such gap, we introduce **3D-VisTA**, a Transformer-based model for 3D Vision and Text Alignment that can be easily adapted to various downstream tasks. Unlike previous models that design sophisticated task-specific modules,

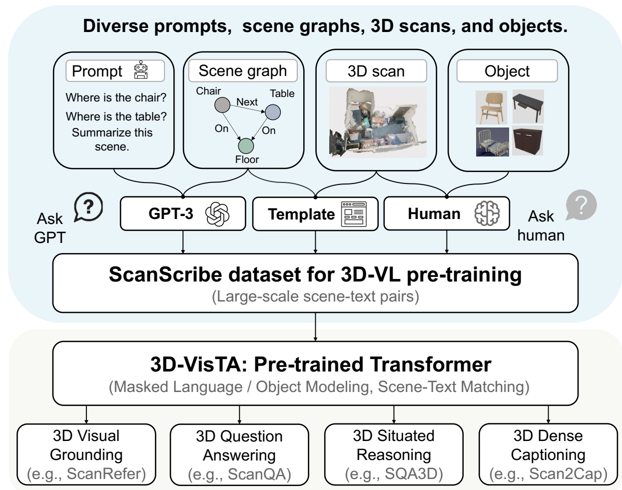



Figure 1: Overall framework of our 3D-VisTA pipeline. We collect diverse prompts, scene graphs, 3D scans, and objects to construct ScanScribe dataset. Through self-supervised pre-training, 3D-VisTA supports various downstream tasks including 3D visual grounding, dense captioning, question answering, and situated reasoning.

*Work done as an intern at BIGAI. ✉ Corresponding author.

we simply utilize a vanilla self-attention transformer [46] for both single-modal modeling and multi-modal fusion in the 3D-VisTA. As a general approach to further enhance 3D spatial comprehension [10, 55, 7], we explicitly encode the pairwise spatial relations between objects into the self-attention weights for 3D object modeling.

Inspired by the success of large-scale pre-training in NLP [15, 41, 42, 6, 52, 31], CV [22, 17, 21, 25, 38], and 2D-VL [30, 2, 34, 40], we propose to pre-train 3D-VisTA on 3D scene-text data, aiming for better performances on 3D-VL tasks. To this end, we construct ScanScribe, the first large-scale 3D scene-text pairs dataset for 3D-VL pre-training. We first collect RGB-D scans of indoor scenes from ScanNet [12] and 3R-Scan [48] datasets. We also randomly replace some objects in the scene with objects from the Objaverse 3D object database [13] based on their categories, in order to increase object diversity. To obtain the text, we transform the text from existing datasets based on ScanNet into scene descriptions, including the question-answer pairs from ScanQA [3] and the referring expressions from ScanRefer [8] and ReferIt3D [1]. We further leverage the scene graph annotations [51] of scans from 3R-Scan, and adopt both templates and GPT-3 [6] to generate scene descriptions from their scene graphs. In total, ScanScribe contains 278K 3D scene-text pairs for 2,995 RGB-D scans of 1,185 indoor scenes, with 56.1K unique object instances.

We pre-train 3D-VisTA on the proposed ScanScribe dataset. Our pre-training tasks include masked language modeling, masked object modeling, and scene-text matching. Notably, similar objectives are widely adopted in 2D-VL yet rarely explored in the 3D-VL domain. The proposed pre-training procedure effectively learns the alignment between 3D point clouds and texts, which eliminates the need for auxiliary losses and optimization tricks in downstream task fine-tuning. On six challenging 3D-VL tasks, ranging from visual grounding (*i.e.*, ScanRefer [8], Nr3D/Sr3D [1]) and dense captioning (*i.e.*, Scan2Cap [11]) to question answering (*i.e.*, ScanQA [3]) and situated reasoning (*i.e.*, SQA3D [36]), fine-tuned 3D-VisTA raises the SOTA results on ScanRefer by 8.1% (acc@0.5), on Sr3D by 3.6%, on Scan2Cap by 10.1%(C@0.25), on ScanQA by 3.5%/2.1% (EM@1), and on SQA3D by 1.9%. Moreover, 3D-VisTA demonstrates superior data efficiency, obtaining strong results with only 30% of the annotations for these downstream tasks.

Our main contributions can be summarized as follows:

- We propose 3D-VisTA, a simple and unified Transformer for aligning 3D vision and text. The proposed Transformer simply utilizes the self-attention mechanism, without any complex task-specific design.
- We construct ScanScribe, a large-scale 3D-VL pre-training dataset that contains 278K 3D scene-text pairs for 2,995 RGB-D scans of 1,185 unique indoor scenes.
- We introduce a self-supervised pre-training scheme for 3D-VL, with masked language/object modeling and scene-text matching. It effectively learns the 3D point cloud and text alignment and further simplifies and improves downstream task fine-tuning.
- We fine-tune 3D-VisTA and achieve state-of-the-art performances on various 3D-VL tasks, ranging from visual grounding and dense captioning to question answering and situated reasoning. 3D-VisTA also demonstrates superior data efficiency, obtaining strong results even with limited annotations.

## 2. Related Work

**3D Vision-language Learning.** Recently, there has been growing interest in 3D vision-language (3D-VL) learning. Unlike traditional scene understanding, 3D-VL tasks connect the physical world to natural language, which is crucial for achieving embodied intelligence [18]. In this emerging area, Chen *et al*. [8] and Achlioptas *et al*. [1] concurrently introduce ScanRefer and ReferIt3D datasets for benchmarking natural language grounding to 3D object properties and relations. Besides 3D visual grounding, Azuma *et al*. [3] develop a 3D question-answering dataset named ScanQA that requires a model to answer a question about objects and their relations given a 3D scene. More recently, Ma *et al*. [36] propose a situated reasoning task called SQA3D for embodied scene understanding in 3D scenes.

Several models have been proposed for these benchmarks [8, 1, 35, 27, 55, 24, 10, 20, 43]. Notably, 3D-SPS [35] and BUTD-DETR [27] progressively discover the target object by leveraging cross attention mechanism and language guidance. 3DVG [55], MVT [24], and ViL3DRel [10] tackle 3D visual grounding by explicitly infusing spatial relation information into their models. Although these works have achieved impressive results in bridging 3D vision and language, they still rely heavily on task-specific knowledge in model design [55, 24, 10] and sophisticated optimization techniques [10, 27, 35]. In contrast, the proposed 3D-VisTA unifies visual grounding, question-answering, and situated reasoning through a simple Transformer-based architecture. Training 3D-VisTA is also straightforward, without requiring any auxiliary losses or sophisticated optimization techniques. Refer to Table 1 for a detailed comparison between 3D-VisTA and other 3D-VL models w.r.t. task, auxiliary Loss, and architecture.

**Large-scale Pre-training.** In recent years, large-scale pre-training has become a cornerstone of natural language processing (NLP), computer vision (CV), and 2D vision-and-language (2D-VL) domains. The introduction of the transformer-based architecture [47], especially BERT [15] and GPT [41, 42, 6], has led to significant improvements in various NLP tasks. The success of these models has led to the development of more advanced pre-training techniques such as XLNet [52] and RoBERTa [31]. These models have

Table 1: The comparison between 3D-VisTA and other models w.r.t. tasks, auxiliary losses, and task-specific architectures."VG" stands for visual grounding, "QA" for question answering, "SR" for situation reasoning, "DC" for dense captioning. "DET" stands for object detection loss, "KD" for knowledge distillation loss, "O-CLS" for object classification loss, and "T-CLS" for text classification loss. "CA" stands for cross attention, "2D" for 2D features, "MV" for multi-view features, and "LC" for language-conditioned modules.

| **Method** | **Task** | **Auxiliary loss** | | | | **Architecture** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | DET | KD | O-CLS | T-CLS | CA | 2D | MV | LC |
| MVT [24] | VG | | | ✓ | ✓ | ✓ | | ✓ | |
| 3D JCG [7] | VG, DC | ✓ | | ✓ | ✓ | ✓ | ✓ | | |
| ViL3DRel [10] | VG | | ✓ | ✓ | ✓ | ✓ | | | ✓ |
| ScanQA [3] | QA | ✓ | | ✓ | | ✓ | ✓ | | |
| SQA3D [36] | SR | ✓ | | | | ✓ | | | |
| 3D-VisTA (ours) | VG,QA,SR,DC | × | × | × | × | × | × | × | × |

Table 2: The comparison between ScanScribe and other 3D-VL datasets. "VG" stands for Visual Grounding, "QA" for Question Answering, "SR" for Situated Reasoning, and "PT" for Pre-training. "Vocab." denotes the text vocabulary size.

| **Dataset** | **Task** | **Size** | **Vocab.** |
|---|---|---|---|
| Nr3D [1] | VG | 30.0K | 2,986 |
| Sr3D [1] | VG | 90.5K | 158 |
| ScanRefer [8] | VG | 36.7K | 4,197 |
| ScanQA [3] | QA | 26.5K | 3,357 |
| SQA3D [36] | SR | 33.4K | 4,535 |
| ScanScribe | PT | 278.0K | 8,197 |

achieved state-of-the-art performance on a wide range of NLP tasks, including text classification, question answering, and language generation. The most successful pre-training approach in CV is the ImageNet [14] pre-training, which has been used as a starting point for a wide range of downstream tasks such as object detection and image segmentation. Recently, the introduction of transformer-based models such as ViT [17] and Swin Transformer [32] has led to significant improvements in various CV tasks. The field of 2D-VL has also seen significant advancements due to pre-training techniques. In particular, the introduction of the ViLBERT [34] and LXMERT [45] models has led to state-of-the-art performance on tasks such as visual question answering and image captioning. More recently, the development of CLIP [40], ALIGN [50], and Flamingo [2] has shown that large-scale pre-training on image-text pairs leads to better cross-modal understanding and the emerge of in-context learning in a zero-shot or few-shot manner.

Although large-scale pre-training has become a crucial technique in NLP, CV, and 2D-VL, it has rarely been explored in 3D-VL. [7, 9] explore multi-task learning of visual grounding and dense captioning, and then further fine-tune their models on each task. The exploration of 3D-VL pre-training may be hindered by the lack of a large-scale pre-training dataset. Therefore, we construct ScanScribe, the first large-scale 3D scene-text pairs dataset for 3D-VL pre-training. As shown in Table 2, ScanScribe is much larger than existing 3D-VL datasets and also has more diverse text. Pre-training 3D-VisTA on ScanScribe has led to significant improvements on 3D-VL tasks, so we believe ScanScribe can fuel the exploration of 3D-VL pre-training in the future.

## 3. 3D-VisTA

In this section, we introduce 3D-VisTA, a simple and unified Transformer for aligning 3D scenes and text. As illustrated by Fig. 2, 3D-VisTA takes a pair of scene point cloud and sentence as input. It first encodes the sentence via a text encoding module and processes the point cloud via a scene encoding module. Then the text and 3D object tokens are fused by a multi-modal fusion module to capture the correspondence between 3D objects and text. 3D-VisTA is pre-trained using self-supervised learning and can be easily fine-tuned to various downstream tasks. Next, we describe each module in detail.

### 3.1. Text Encoding

We adopt a four-layer Transformer to encode the sentence $S$ into a sequence of text tokens $\{w_{\text{cls}}, w_1, w_2, \cdots, w_M\}$, where $w_{cls}$ is a special classification token (`[CLS]`) and $M$ is the sentence length. This text encoding module is initialized by the first four layers of a pre-trained BERT [15].

### 3.2. Scene Encoding

Given the point cloud of a 3D scene, we first use segmentation masks to break down the scene into a bag of objects. The segmentation masks can be either obtained from ground truth or instance segmentation models [16, 28, 44]. For each object, we sample 1024 points and normalize their coordinates into a unit ball. Then the object point cloud is fed into PointNet++ [39] to obtain its point features and semantic class. We compose the point features $f_i$, the semantic class embedding $c_i$, and the location $l_i$ (*i.e.*, 3D position, length, width, height) as the representation of the object token $i$:

$$o_i = f_i + W_c c_i + W_l l_i, i = 1, 2, ..., N, \tag{1}$$

where $W_c$ and $W_l$ are additional projection matrices to map $c_i$ and $l_i$ into the same dimension as $f_i$.

To further provide a contextual representation of objects, we capture the object-to-object interactions by infusing object tokens into a four-layer Transformer. Motivated by previous works [55, 24, 10], we explicitly encode the pairwise spatial relations of objects into the Transformer (*Spatial transformer* in Fig. 2). More specifically, we follow [10] to define the pairwise spatial features for the object pair $i, j$:

$$s_{ij} = [d_{ij}, \sin(\theta_h), \cos(\theta_h), \sin(\theta_v), \cos(\theta_v)],$$

where $d_{ij}$ is the Euclidean distance and $\theta_h, \theta_v$ are the horizontal and vertical angles of the line connecting the centers

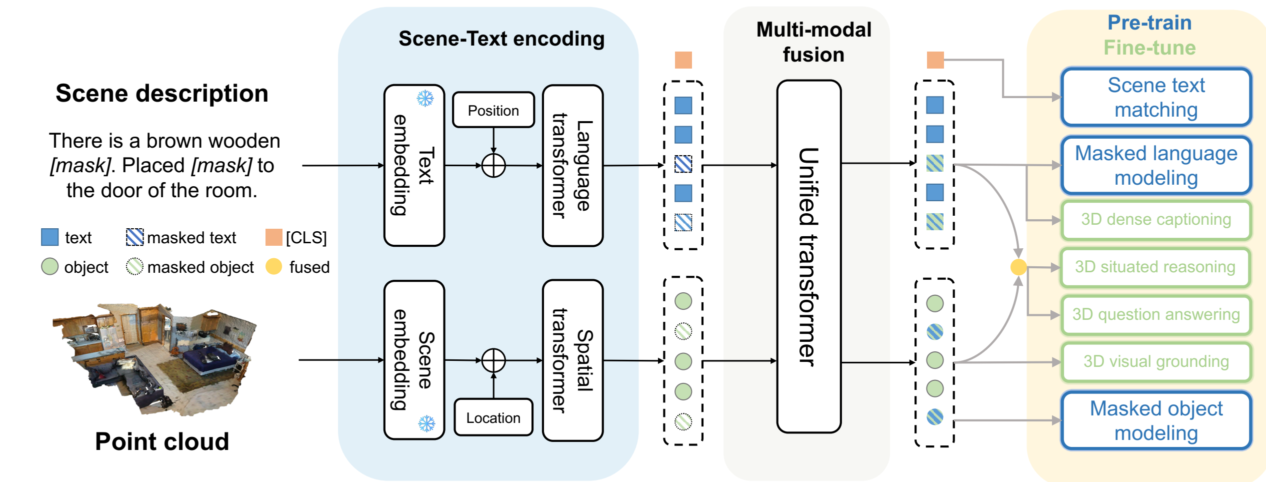



Figure 2: The model architecture of our 3D-VisTA, which includes text encoding, scene encoding, and multi-modal fusion modules. 3D-VisTA is pre-trained by self-supervised learning objectives, which include masked language modeling, masked object modeling, and scene-text matching. Pre-trained 3D-VisTA can be easily adapted to various downstream tasks by adding lightweight task heads without task-specific design like auxiliary losses and optimization tricks.

of objects $i, j$. The pairwise spatial features $S = [s_{ij}] \in \mathbb{R}^{N \times N \times 5}$ are used to modulate the attention weights of the self-attention layers in the Transformer:

$$\mathrm{Attn}(Q, K, V, S) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_h}} + \log \sigma(Sw)\right) V,$$

where $w \in \mathbb{R}^5$ is used to map the spatial features to the attention scores and $\sigma$ is the sigmoid function.

### 3.3. Multi-modal Fusion

We simply concatenate the text and the 3D object tokens and send them to a $L$-layer Transformer (*Unified transformer* in Fig. 2) for multi-modal fusion. Learnable type embeddings are added to the tokens to differentiate text and 3D objects. We denote the output of the multi-modal fusion module as $\{\mathbf{w}_{\text{cls}}, \mathbf{w}_{1:M}, \mathbf{o}_{1:N}\}$ for `[CLS]`, text tokens, and 3D object tokens, respectively.

### 3.4. Self-supervised Pre-training

To learn the 3D scene and text alignment in a self-supervised manner, we pre-train 3D-VisTA on 3D scene-text pairs via the following proxy tasks:

**Masked Language Modeling (MLM).** We follow the BERT pre-training [15] to perform MLM: (1) 15% of the text tokens are randomly chosen; (2) 80% of the time: replace these tokens with `[MASK]`; (2) 10% of the time: replace these tokens with some random text tokens; (3) 10% of the time: these tokens remain unchanged. The model is trained to predict the masked text tokens given the remaining text and 3D object tokens:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{o})\sim D} \log P_\theta \left(\mathbf{w_m} \mid \mathbf{w}_{\backslash \mathbf{m}}, \mathbf{o}\right). \quad (2)$$

**Masked Object Modeling (MOM).** Similar to MLM, we mask out 10% of 3D object tokens. However, we mask a 3D object token by only replacing its point features and semantic embedding (*i.e.*, "$f_i + W_c c_i$" in Eq. (1)) with a learnable mask embedding but keep its positional information (*i.e.*, "$W_l l_i$" in Eq. (1)) unchanged. The model is trained to utilize the position clue of the masked object to predict its semantic class $c$ given the remaining 3D objects and text:

$$\mathcal{L}_{\text{MOM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{o})\sim D} \log P_\theta \left(c(\mathbf{o_m}) \mid \mathbf{o}_{\backslash \mathbf{m}}, \mathbf{w}\right). \quad (3)$$

**Scene-Text Matching (STM).** While masked language and object modeling enable local text-object alignment in a fine-grained granularity, we also perform scene-text matching to enhance the global fusion of scene and text, which we find very beneficial for downstream question-answering tasks. More specifically, we extract the output corresponds to `[CLS]` as the global representation of the input scene-text pair, and feed it into a two-layer MLP to predict if the scene and the text are matched:

$$\mathcal{L}_{\text{STM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{o})\sim D} \log P_\theta \left(y \mid \mathbf{w}, \mathbf{o}\right). \quad (4)$$

In practice, 30% of the samples in a training batch are negative pairs, created by replacing the scene point cloud or text with a randomly selected sample.

**Final loss.** Our final pre-training objective is obtained by simply adding the losses of the proxy tasks above:

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MOM}} + \mathcal{L}_{\text{STM}} \quad (5)$$

Notably, the proposed pre-training scheme is self-supervised and task-agnostic, unlike the supervised multi-task learning used in previous work [7] that requires task supervision.

Table 3: The composition of ScanScribe. *We only use Objaverse to provide candidate object replacement for the 3D scenes in other two datasets; thus no scene-text pair is generated.

| Source | 3D | | | Text | | | Scene-Text |
|---|---|---|---|---|---|---|---|
| | Scan | Scene | Object | Human | Template | GPT-3 | Pairs |
| ScanNet | 1,513 | 707 | 36.2K | 93.2K | 90.5K | - | 183.7K |
| 3R-Scan | 1,482 | 478 | 13.6K | - | 89.6K | 4.7K | 94.3K |
| Objaverse* | - | - | 6.3K | - | - | - | - |
| ScanScribe | 2,995 | 1,185 | 56.1K | 93.2K | 180.1K | 4.7K | 278.0K |

### 3.5. Downstream Task Finetuning

The pre-trained 3D-VisTA can be easily adapted to various 3D-VL tasks by adding lightweight task heads. More specifically, we fine-tune 3D-VisTA on the following tasks:

**3D Visual Grounding** tasks a model to locate a target object in a 3D scene from a referring expression. To find the referred object, we apply a two-layer MLP to each object token $\mathbf{o}_i$, and obtain the probability of the object being referred to. The model is fine-tuned using the cross-entropy loss.

**3D Dense Captioning** is introduced by [11] to test a model's ability of detecting and describing objects in a 3D scene. Following [30], we take $\mathbf{w}_{1:M}$ and predict text tokens autoregressively to generate a sentence. The model is fine-tuned using cross-entropy loss.

**3D Question Answering** requires a model to answer an object-related question given a 3D scene. Following [3], we feed the text tokens $\mathbf{w}_{1:M}$ and the object tokens $\mathbf{o}_{1:N}$ into a modular co-attention network (MCAN) [54] to produce answers. The model is fine-tuned using the QA loss and the object localization loss.

**3D Situated Reasoning** is recently proposed by [36] to benchmark the 3D scene understanding of embodied agents. To adapt 3D-VisTA to this task, we concatenate the situation description and the question into a single input sentence. The answer classification is similar to the 3D question answering task. The model is fine-tuned using the answer loss.

In general, we find adapting 3D-VisTA to these downstream tasks much simpler than previous methods [8, 24, 10, 3, 36], as 3D-VisTA is simply fine-tuned using the task loss only, without the need for any auxiliary losses (*e.g.*, sentence/object classification loss [8, 3]) or optimization tricks (*e.g.*, multi-view aggregation [24] and knowledge distillation [10]). This makes 3D-VisTA a more unified and general-purpose 3D-VL model.

## 4. ScanScribe

In recent years, large-scale pre-training has been widely used to improve the performance on downstream tasks in CV [49], NLP [15], and 2D-VL [30, 45]. However, large-scale pre-training has barely been touched in the 3D-VL domain, possibly due to the lack of pre-training datasets for 3D-VL. To facilitate the exploration of 3D-VL pre-training, we build a large-scale 3D scene-text pairs dataset, named ScanScribe. As illustrated in Table 3, the construction of 3D scene-text pairs in ScanScribe comprises two parts:

**3D scenes.** We collect RGB-D scans of indoor scenes from ScanNet [12] and 3R-Scan [48]. To increase the diversity of 3D objects in these scenes, 10% of the object instances in each scene are randomly replaced by objects from the Objaverse 3D object database[13] based on their categories. For each ScanNet and 3R-Scan object category, we download about 40 object instances from Objaverse as candidate object replacements. As a result, we collect 2,995 RGB-D scans of 1,185 indoor scenes, with 56.1K unique object instances.

**Text.** For the scans from ScanNet, we transform the text from existing datasets based on ScanNet into scene descriptions, including the question-answer pairs from ScanQA [3] and the referring expressions from ScanRefer [8] and ReferIt3D [1]. For the scans from 3R-Scan, we adopt both templates and GPT-3 [6] to generate scene descriptions based on their scene graph annotations [51]. Specifically, for each `object`, we first extract all the ⟨`object`, `relation`, `neighbor`⟩ triplets from the scene graph. We then use the template "This is a `object`, a `neighbor` is `relation` to `object`" to generate the descriptions. Note that we only choose objects with fewer than 7 neighbors in a template-based generation. We further explore using GPT-3 to generate the descriptions with the following prompt "`object` is `relation` to `neighbor` *...(repeat until all the neighbors have been used)*. Where is `object`? or Summarize the scene." Ultimately, 278K scene descriptions are generated for the collected 3D scenes.

## 5. Experiments

### 5.1. Experimental Settings

**Implementation Details.** The pre-training runs for 30 epochs with a batch size of 128. We use the AdamW [33] optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$. The learning rate is set to $1e^{-4}$, with a warmup of 3,000 steps, and cosine decay. During pre-training, we use ground-truth segmentation masks to generate object-level point clouds.During fine-tuning, we use ground-truth masks or Mask3d [44], which depends on the task setting. On the ScanRefer dataset, we also incorporate PointGroup [28] for comparison with previous approaches. In ablation studies, we use ground-truth masks in all tasks for simplicity. Both pre-training and fine-tuning are conducted on a single NVIDIA A100 80GB GPU.

**3D Visual Grounding.** We evaluate our model on three datasets for this task: ScanRefer [8], Nr3D, and Sr3D [1]. For Nr3D/Sr3D, we follow ReferIt3D [1] to use ground-truth object masks and report the results as the grounding accuracy, *i.e.*, whether the model correctly selects the referred object among ground-truth object proposals. For ScanRefer, we follow [8] to use detector-generated object proposals and report the results as Acc@$k$($k \in \{0.25, 0.5\}$), *i.e.*, the frac-

Table 4: Grounding accuracy (%) on Nr3D and Sr3D with ground-truth object proposals. Δ denotes the performance difference between 3D-VisTA and 3D-VisTA (scratch). 3D-VisTA achieves competitive results with SOTA on Nr3D and outperforms SOTA on Sr3D.

| Method | Nr3D | | | | | Sr3D | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Easy | Hard | View Dep | View Indep | Overall | Easy | Hard | View Dep | View Indep |
| 3DVG-Trans [55] | 40.8 | 48.5 | 34.8 | 34.8 | 43.7 | 51.4 | 54.2 | 44.9 | 44.6 | 51.7 |
| TransRefer3D [20] | 48.0 | 56.7 | 39.6 | 42.5 | 50.7 | 57.4 | 60.5 | 50.2 | 49.9 | 57.7 |
| LAR [4] | 48.9 | 58.4 | 42.3 | 47.4 | 52.1 | 59.4 | 63.0 | 51.2 | 50.0 | 59.1 |
| SAT [53] | 56.5 | 64.9 | 48.4 | 54.4 | 57.6 | 57.9 | 61.2 | 50.0 | 49.2 | 58.3 |
| 3D-SPS [35] | 51.5 | 58.1 | 45.1 | 48.0 | 53.2 | 62.6 | 56.2 | 65.4 | 49.2 | 63.2 |
| MVT [24] | 59.5 | 67.4 | 52.7 | 59.1 | 60.3 | 64.5 | 66.9 | 58.8 | 58.4 | 64.7 |
| ViL3DRel [10] | **64.4** | 70.2 | **57.4** | **62.0** | 64.5 | 72.8 | 74.9 | 67.9 | **63.8** | 73.2 |
| 3D-VisTA (scratch) | 57.5 | 65.9 | 49.4 | 53.7 | 59.4 | 69.6 | 72.1 | 63.6 | 57.9 | 70.1 |
| 3D-VisTA | 64.2 | **72.1** | 56.7 | 61.5 | **65.1** | **76.4** | **78.8** | **71.3** | 58.9 | **77.3** |
| Δ | **6.7 ↑** | **6.2 ↑** | **7.3 ↑** | **7.8 ↑** | **5.7 ↑** | **6.8 ↑** | **6.7 ↑** | **7.7 ↑** | **1.0 ↑** | **7.2 ↑** |

Table 5: Grounding accuracy (%) on ScanRefer with detected object proposals. "Det." represents the 3D object detection module used in the model. "VN" stands for VoteNet [16], "PG" for PointGroup [28], and M3D for Mask3D [44], while "Opt." denotes jointly optimizing the object detector on ScanRefer. Mask3D significantly improves the grounding accuracy by providing more accurate object proposals.

| Method | Det. | Unique | | Multiple | | Overall | |
|---|---|---|---|---|---|---|---|
| | | acc@0.25 | acc@0.5 | acc@0.25 | acc@0.5 | acc@0.25 | acc@0.5 |
| 3DVG-Trans [55] | Opt. | 81.9 | 60.6 | 39.3 | 28.4 | 47.6 | 34.7 |
| 3D-SPS [35] | Opt. | **84.1** | 66.7 | 40.3 | 29.8 | 48.8 | 37.0 |
| 3DJCG [7] | Opt. | 83.5 | 64.3 | 41.4 | 30.8 | 49.6 | 37.3 |
| SAT [53] | VN | 73.2 | 50.8 | 37.6 | 25.2 | 44.5 | 30.1 |
| MVT [24] | PG | 77.7 | 66.5 | 31.9 | 25.3 | 40.8 | 33.3 |
| ViL3DRel [10] | PG | 81.6 | 68.6 | 40.3 | 30.7 | 47.9 | 37.7 |
| 3D-VisTA (scratch) | PG | 76.0 | 66.9 | 33.3 | 27.0 | 41.2 | 34.4 |
| 3D-VisTA | PG | 77.0 | 67.9 | 37.9 | 30.4 | 45.2 | 37.3 |
| 3D-VisTA (scratch) | M3D | 77.4 | 70.9 | 38.7 | 34.8 | 45.9 | 41.5 |
| 3D-VisTA | M3D | 81.6 | **75.1** | **43.7** | **39.1** | **50.6** | **45.8** |
| Δ | M3D | **4.2 ↑** | **4.2 ↑** | **5.0 ↑** | **4.3 ↑** | **4.7 ↑** | **4.3 ↑** |

Table 6: Captioning results on Scan2Cap dataset. "C" stands for "CIDEr", "B-4" for "BLEU-4", "M" for "METEOR", and "R" for "ROUGE", respectively. "@0.25" and "@0.5" represent the overlap ratios between the predicted boxes and ground truth boxes.

| Method | @0.25 | | | | @0.5 | | | |
|---|---|---|---|---|---|---|---|---|
| | C | B-4 | M | R | C | B-4 | M | R |
| Scan2Cap [11] | 53.7 | 34.3 | 26.1 | 55.0 | 35.2 | 22.4 | 21.4 | 43.5 |
| 3DJCG [7] | 60.9 | **39.7** | 27.5 | **59.0** | 47.7 | 31.5 | 24.3 | 51.8 |
| 3D-VisTA (scratch) | 66.8 | 36.6 | 28.0 | 58.4 | 61.6 | 34.1 | 26.8 | **55.0** |
| 3D-VisTA | **71.0** | 36.5 | **28.4** | 57.6 | **66.9** | **34.0** | **27.1** | 54.3 |
| Δ | **4.2 ↑** | **0.1 ↓** | **0.4 ↑** | **0.8 ↓** | **5.3 ↑** | **0.1 ↓** | **0.3 ↑** | **0.7 ↓** |

tion of referring queries whose predicted box overlaps the ground truth with IoU $> k$.

**3D Dense Captioning** We evaluate our model on the Scan2cap dataset [11] and report the text similarity metrics under different box overlap ratios.

**3D Question Answering.** We evaluate our model on the ScanQA dataset [3] and use exact matches (EM@1 and EM@10) as the evaluation metric. We also report several sentence evaluation metrics, including BLEU-4, ROUGE, METEOR, and CIDEr. Both test sets (w/ or w/o objects) of ScanQA are used in our evaluation.

**3D Situated Reasoning** We evaluate our model on the SQA3D dataset [36] and report the answer accuracy under different types of questions as the evaluation metric.

## 5.2. Downstream Task Results

In this section, we discuss the experimental results of the downstream tasks and compare the proposed 3D-VisTA model with the state-of-the-art (SOTA) methods. Results are presented in Tables 4 to 8 and Fig. 3 and the main observations from these results are as follows:

1. **Even trained from scratch, 3D-VisTA achieves competitive performances with SOTA methods.** Specifically, 3D-VisTA (scratch) obtains an overall accuracy of 57.5% and 69.6% on Nr3D and Sr3D, which outperforms most previous models; it gets an EM@1 accuracy of 25.2% on ScanQA, which is 1.7% higher than SOTA. Of note, 3D-VisTA is trained on these datasets simply using the task losses, without any auxiliary losses or optimization tricks,

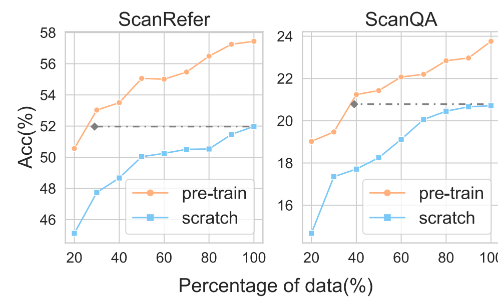
Table 7: Answer accuracy on ScanQA using object proposals from Mask3D. Each entry denotes “test w/ object” / “test w/o object”.

| Method | EM@1 | EM@10 | BLEU-4 | ROUGE | METEOR | CIDEr |
|---|---|---|---|---|---|---|
| Image+MCAN [3] | 22.3 / 20.8 | 53.1 / 51.2 | 14.3 / 9.7 | 31.3 / 29.2 | 12.1 / 11.5 | 60.4 / 55.6 |
| ScanRefer+MCAN [3] | 20.6 / 19.0 | 52.4 / 49.7 | 7.5 / 7.8 | 30.7 / 28.6 | 12.0 / 11.4 | 57.4 / 53.4 |
| ScanQA [3] | 23.5 / 20.9 | 56.5 / **54.1** | 12.0 / 10.8 | 34.3 / 31.1 | 13.6 / 12.6 | 67.3 / 60.2 |
| 3D-VisTA (scratch) | 25.2 / 20.4 | 55.2 / 51.5 | 10.5 / 8.7 | 35.5 / 29.6 | 13.8 / 11.6 | 68.6 / 55.7 |
| 3D-VisTA | **27.0 / 23.0** | **57.9** / 53.5 | **16.0 / 11.9** | **38.6 / 32.8** | **15.2 / 12.9** | **76.6 / 62.6** |
| Δ | **1.8** ↑ / **2.6** ↑ | **2.7** ↑ / **2.0** ↑ | **5.5** ↑ / **3.2** ↑ | **3.1** ↑ / **3.2** ↑ | **1.4** ↑ / **1.3** ↑ | **8.0** ↑ / **6.9** ↑ |

Table 8: Answer accuracy on SQA3D using object proposals from Mask3D. Pre-training improves the results of most question types.

| Method | Test set | | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | What | Is | How | Can | Which | Other | |
| GPT-3 [36] | **39.7** | 46.0 | 40.5 | 45.6 | 36.1 | 38.4 | 41.0 |
| ClipBERT [36] | 30.2 | 60.1 | 38.7 | 63.3 | 42.5 | 42.7 | 43.3 |
| SQA3D(w/o s) [36] | 28.6 | 65.0 | 47.3 | 66.3 | 43.9 | 42.9 | 45.3 |
| SQA3D [36] | 31.6 | 63.8 | 46.0 | 69.5 | 43.9 | 45.3 | 46.6 |
| 3D-VisTA (scratch) | 32.1 | 62.9 | **47.7** | 60.7 | 45.9 | **48.9** | 46.7 |
| 3D-VisTA | 34.8 | **63.3** | 45.4 | **69.8** | **47.2** | 48.1 | **48.5** |
| Δ | **2.7** ↑ | **0.4** ↑ | **2.3** ↓ | **9.1** ↑ | **1.3** ↑ | **0.8** ↓ | **1.8** ↑ |

Figure 3: The performance of finetuning 3D-VisTA using various amounts of training data.




indicating that 3D-VisTA is a very simple yet effective architecture for 3D-VL tasks.

2. **Pre-training on ScanScribe significantly improves the performance of 3D-VisTA.** Overall, the pre-training improves the accuracy on Nr3D/Sr3D by 6.7%/6.8%, the acc@0.25/0.5 on ScanRefer by 4.7%/4.3%, the EM@1 on ScanQA by 1.8%/2.6%, the C@0.25 on Scan2Cap by 4.2%, and the average accuracy on SQA3D by 1.8%. These large improvements consolidate the efficacy of ScanScribe for the 3D-VL pre-training.

3. **The pre-trained 3D-VisTA outperforms SOTA by a large margin.** 3D-VisTA outperforms ViL3DRel [10] on Sr3D by 3.6% and on ScanRefer by 2.7%/8.1% (acc@0.25/0.5), beats ScanQA [3] by 3.5%/2.1 (EM@1), Scan2Cap SOTA by 10.1%/19.2% (C@0.25/0.5), SQA3D [36] by 1.9% (Avg.). 3D-VisTA sets a new record for these 3D-VL tasks and may inspire future research on 3D-VL pre-training.

4. **Finetuning 3D-VisTA on downstream tasks with limited annotations achieves strong results.** As shown in Fig. 3, being fine-tuned using 30% and 40% of the annotations on ScanRefer and ScanQA, the pre-trained 3D-VisTA can achieve better performance than the one trained from scratch with full data. We hypothesize that 3D-VisTA has successfully captured the alignment between 3D objects and text via pre-training and is thus able to readily adapt to downstream tasks of various formats. It also reveals the potential of 3D-VisTA to learn unseen tasks in a zero-shot or few-shot manner, which has emerged in NLP [6] and 2D-VL [2] via large-scale pre-training.

### 5.3. Ablation Studies

In this section, we conduct ablation studies to analyze the impact of several important hyperparameters, including Transformer depth, pre-training objectives, and data amount.

**Transformer Depth.** Since the model size is a key factor in the pre-training of NLP and 2D-VL, we study the effect of the transformer depth by varying the number of layers in the multimodal fusion module. As shown in Table 9a, using 4 layers achieves the best performance and simply adding more layers does not help. This observation is somewhat contradictory to the ones from NLP and 2D-VL. It points out that although ScanScribe is much larger than existing 3D-VL datasets, it is still far from enough to unleash the full potential of pre-training in the 3D-VL domain.

**Pre-training Objectives.** Table 9b presents the ablation study for the pre-training objectives. The MLM objective alone slightly benefits question answering (QA), but brings a negative impact on visual grounding (VG). Adding MOM and STM boosts the performance of both QA and VG, which highlights the importance of MOM and STM for aligning 3D vision and text. Overall, using all three objectives together leads to the best performance for both tasks, with STM and MOM providing the greatest improvements in accuracy.

**Pre-training Data.** Table 9c presents the results using various configurations of pre-training data. We can see that simply using the ScanNet data for pre-training, which is from the same domain as downstream tasks, leads to a significant improvement in VG and QA. This validates the effectiveness

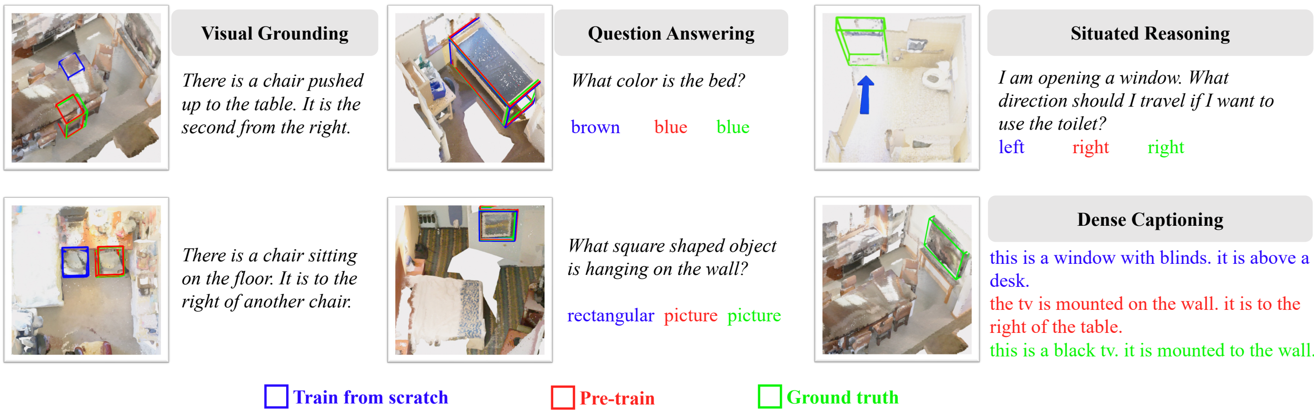



Figure 4: Qualitative results for various tasks. *Italic text* stand for the inputs, blue boxes or text for the predictions from 3D-VisTA trained from scratch, red for the predictions from pre-trained 3D-VisTA, and green for the ground truth, respectively. The results show that pre-training improves the understanding of spatial relations, visual concepts, and situations.

Table 9: Ablation studies of 3D-VisTA w.r.t. Transformer depth, pre-training objectives, and pre-training data. We report the grounding accuracy on ScanRefer for Visual Grounding (VG) and the EM@1 accuracy on ScanQA for Question Answering (QA).

(a) Transformer Depth

| # layer | VG | QA |
|---|---|---|
| 2 | 55.8 | 23.7 |
| 4 | 57.4 | 23.8 |
| 6 | 56.6 | 22.8 |
| 8 | 56.3 | 22.7 |

(b) Pre-training Objectives

| MLM | MOM | STM | VG | QA |
|---|---|---|---|---|
| × | × | × | 52.0 | 20.7 |
| ✓ | × | × | 51.5 | 21.3 |
| ✓ | ✓ | × | 57.1 | 22.5 |
| ✓ | ✓ | ✓ | 57.4 | 23.8 |

(c) Pre-training Data

| ScanNet | 3R-Scan | Objaverse | VG | QA |
|---|---|---|---|---|
| × | × | × | 52.0 | 20.7 |
| ✓ | × | × | 54.6 | 22.6 |
| ✓ | ✓ | × | 56.5 | 23.5 |
| ✓ | ✓ | ✓ | 57.4 | 23.8 |

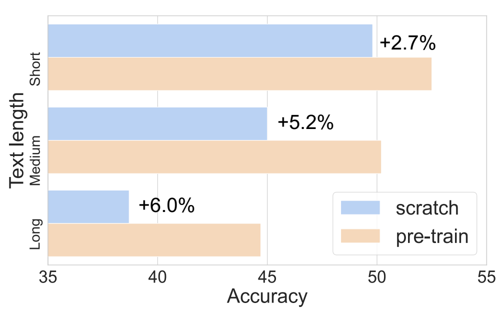



Figure 5: The performance gap between scratch and pre-training over different sentence lengths ($\leq 15, \leq 30, > 30$) in ScanRefer.

of pre-training, even in the case of no additional 3D data than downstream tasks. Adding 3R-Scan and Objaverse increases the amount and the diversity of 3D data, which further boosts the accuracy of both VG and QA. Overall, the best performance for both tasks is achieved when all three data sources are used. This points out a promising path for improving 3D-VL tasks — collecting more data for pre-training.

### 5.4. Qualitative Studies and Additional Results

In this section, we perform additional studies to better understand how pre-training helps. As shown in Fig. 4, pre-training improves the spatial understanding of 3D-VisTA for visual grounding, so it can better align with human prior viewpoint and reason over spatial relations. This is very helpful when the model needs to distinguish the target object from multiple instances of the same class. Pre-training also helps with a better understanding of visual concepts like colors and shapes, and situations for question answering and situated reasoning. Besides, pre-training enhances the capability of aligning long text with 3D scenes, as evidenced by the larger improvement over longer queries in Fig. 5.

## 6. Conclusion

This paper proposes 3D-VisTA, a simple yet effective architecture for 3D-VL tasks. The model simply uses self-attention layers and can be easily adapted to various downstream tasks, without requiring any auxiliary loss or optimization trick. We also introduce ScanScribe, the first large-scale 3D scene-text pairs dataset for 3D-VL pre-training. The pre-trained 3D-VisTA achieves state-of-the-art results on a variety of 3D-VL tasks with superior data efficiency, paving the path to future foundation models for 3D-VL tasks.

**Future Works.** Currently, 3D-VisTA uses an offline 3D object detection module, which may be a bottleneck for further improvement. Jointly optimizing the object detection module in the pre-training phase is an interesting future direction. Besides, the data amount in ScanScribe is still insufficient for large-scale 3D-VL pre-training, so scaling up the pre-training dataset as well as the model size is a promising direction to further improve the 3D-VL learning.

**Acknowledgements.** The authors would like to thank Hongming Xu at BIGAI for the help on Mask3D. This work is supported in part by the National Key R&D Program of China (2022ZD0114900) and the National Science Foundation of China (NSFC) under Grant No. 62176134.

## 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. Springer, 2020. 1, 2, 3, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 3, 7

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 1, 2, 3, 5, 6, 7

[4] Eslam Mohamed Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 6

[5] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *6th Annual Conference on Robot Learning*, 2022. 1

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 2, 5, 7

[7] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 1, 2, 3, 4, 6

[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. Springer, 2020. 1, 2, 3, 5

[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. 3

[10] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 5, 6, 7

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021. 1, 2, 5, 6

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 2, 5

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 2, 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 2, 3, 4, 5

[16] Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 202–210. Springer, 2019. 1, 3, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[18] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 1, 2

[19] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Conference on Robot Learning*, 2022. 1

[20] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 2, 6

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[23] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[24] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15524–15533, 2022. 1, 2, 3, 5, 6

[25] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[26] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*, pages 9118–9147. PMLR, 2022. 1

[27] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision (ECCV)*, pages 417–433. Springer, 2022. 1, 2

[28] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2020. 3, 5, 6

[29] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022. 1

[30] 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. Springer, 2020. 2, 5

[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. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019. 5

[34] 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 (NeurIPS)*, 32, 2019. 2, 3

[35] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16454–16463, 2022. 1, 2, 6

[36] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3, 5, 6, 7

[37] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 1

[38] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[40] 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. PMLR, 2021. 2, 3

[41] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. 2

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[43] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 2

[44] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 3, 5, 6

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 3, 5

[46] 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. 2

[47] 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. 2

[48] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *International Conference on Computer Vision (ICCV)*, pages 7658–7667, 2019. 2, 5

[49] 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. 5

[50] Yiren Wang, Shiyang Huang, Tianyu Gao, Xu Zhang, Xu Han, and Zhangyang Wang. Align: Adaptive fine-tuning for long-tailed instance generation via contrastive pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 3

[51] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2021. 2, 5

[52] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:5754–5764, 2019. 2

[53] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *International Conference on Computer Vision (ICCV)*, pages 1856–1866, 2021. 1, 6

[54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. 5

[55] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021. 1, 2, 3, 6

[56] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie Zhao, Lipeng Wang, and Xibo Fan. Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1