SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding

Baoxiong Jia*, Yixin Chen*, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang

* indicates equal contribution
State Key Laboratory of General Artificial Intelligence, BIGAI
https://scene-verse.github.io

Abstract. 3D vision-language (3D-VL) grounding, which aims to align language with 3D physical environments, stands as a cornerstone in developing embodied agents. In comparison to recent advancements in the 2D domain, grounding language in 3D scenes faces two significant challenges: (i) the scarcity of paired 3D-VL data to support grounded learning of 3D scenes, especially considering complexities within diverse object configurations, rich attributes, and intricate relationships; and (ii) the absence of a unified learning framework to distill knowledge from grounded 3D data. In this work, we aim to address these major challenges in 3D-VL by examining the potential of systematically upscaling 3D-VL learning in indoor scenes. We introduce the first **million-scale** 3D-VL dataset, SceneVerse, encompassing 68K indoor scenes and comprising 2.5M vision-language pairs collected from both human annotations and our scalable scene-graph-based generation approach. We demonstrate that this scaling allows for a unified pre-training framework, Grounded Pre-training for Scenes (GPS), for 3D-VL learning. Through extensive experiments, we showcase the effectiveness of GPS by achieving state-of-the-art performance on existing 3D visual grounding and question-answering benchmarks. We also show that the data scaling effect is not limited to GPS, but is generally beneficial for models on tasks like 3D semantic segmentation. The vast potential of SceneVerse and GPS is unveiled through zero-shot transfer experiments in challenging 3D-VL tasks.

Keywords: 3D Vision-Language · Data Scaling · Grounded Scene Understanding

1 Introduction

The foundation of human cognitive development lies in the grounding of language within the physical world [53, 81, 108]. Recent progress in Large Language Models (LLMs) [10, 11, 83] has markedly promoted the alignment between vision and language [3, 59, 75] utilizing billion-scale vision-language datasets [79, 107]. However, with these advancements predominantly focusing on the 2D domain, the grounded understanding of 3D physical environments remains in an incipient stage [1, 5, 16]. Recognizing the pivotal role of grounded 3D experiences in

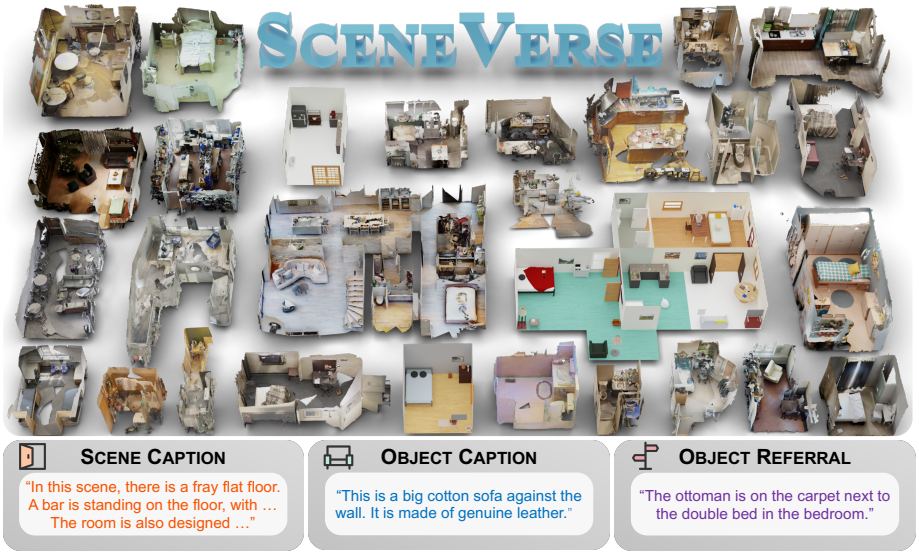

Fig. 1: Overview of SceneVerse. A million-scale 3D vision-language dataset that comprises over 68K various 3D indoor scenes and 2.5M aligned scene-language pairs in the form of scene caption, object caption, and object referral.

shaping human cognition [7, 8], there is a compelling need to focus on exploring vision-language learning in the context of 3D scenes.

Seeking insights from success in 2D vision-language (2D-VL) learning, a major factor to the success was the notable scale-up of paired vision-language data [15, 52, 79]. However, applying this experience directly from 2D to 3D is fraught with challenges. Primarily, 3D data collection heavily relies on the scanning device, making it inherently much more complex and expensive than gathering 2D images. Despite steady efforts to increase the volume of 3D scene data [9, 24, 66, 98], most datasets remain limited to thousands of scenes, substantially lagging behind the scale of existing 2D datasets. This gap is further widened by the inherent complexities of 3D scenes, which feature a multitude of object instances with diverse attributes, varying arrangements, and intricate inter-object relationships. These unique aspects of 3D scenes not only make the accurate description of objects and their relations more challenging but also considerably increase the number of language descriptions required for thorough scene depiction. Consequently, this presents a significant challenge in gathering sufficient and high-quality paired scene-language data for grounded scene understanding.

To confront these challenges, we propose SceneVerse, the first **million-scale** dataset aimed at advancing 3D vision-language (3D-VL) learning for grounded scene understanding. At the scene level, we unify 3D scene data from existing datasets [9, 24, 66, 76, 85], aligning scenes and annotations from various capturing sources, and supplement the collection with synthetic scenes [28, 106]. This compilation represents the most extensive 3D scene data gathered to date, amounting to 68K scenes. For language, we first present 97K newly-annotated referring expressions, the most extensive thus far. We additionally propose an

automated generation pipeline utilizing 3D scene graphs [4,86] and LLMs to create comprehensive, high-quality scene-language pairs. This refined collection, totaling 2.5M scene-language pairs, provides detailed and comprehensive descriptions of both object-level and scene-level descriptions within the 3D scene.

We thoroughly investigate the potential offered by SceneVerse with large-scale pre-training, introducing Grounded Pre-training for Scenes (GPS), a novel and unified pre-training framework designed for scene and object-level alignment without auxiliary losses. Through multi-level contrastive alignment, we achieve significant performance boosts on 3D-VL tasks, such as grounding and question answering, setting new state-of-the-art results via a simple and effective pre-training process. We unveil the vast possibilities offered by SceneVerse and GPS in 3D-VL tasks in a zero-shot transfer setting. Additionally, we show that the scaling effect in SceneVerse is not limited to GPS, but generally benefits models in tasks like 3D semantic segmentation. At last, we offer deeper insights into the data-scaling in SceneVerse through extensive ablative experiments, pointing out future directions. Our main contributions are as follows:

1. We introduce SceneVerse, the first million-scale 3D-VL dataset for grounded scene understanding. SceneVerse encompasses 68K 3D scenes coupled with 2.5M scene-language pairs, sourced through a combination of human annotation and automated generation methods. This represents a significant improvement in terms of data diversity and scale compared to prior datasets.
2. We propose GPS, a transformer-based model trained with multi-level scene-text alignment that achieves state-of-the-art results on existing 3D-VL grounding and question-answering benchmarks by pre-training on SceneVerse.
3. We demonstrate that with the data scale-up and model design, our pre-trained GPS exhibit emerging zero-shot generalization capabilities in grounded scene understanding. We also show that this scaling effect is not limited to GPS, but is generally beneficial for models on tasks like semantic segmentation.

2 Related Work

Datasets for Grounded 3D Understanding Obtaining aligned 3D-language data is a inherently difficult task. In 3D object modeling, pioneering works like ShapeNet [14] sourced 3D assets from online repositories, leading to a proliferation of high-quality 3D object datasets [23, 69, 91]. Notably, recent developments include internet-scale data collection with Objaverse [26, 27], accompanied by the integration of object-level captions [93] for 3D-language alignment. Models trained on these datasets demonstrate an enhanced understanding of objects, evident in classification [60], generation [61], and captioning tasks [63].

In contrast, developing datasets for grounded 3D scene understanding is even more challenging due to the extensive requirements for scene acquisition and annotation. Existing works curate RGB-D and scanned indoor scene datasets [9,13, 24,66,76,85] and synthetic scenes [28,50,95,106] used for benchmarking tasks like 3D object detection and segmentation [31,46,67,80,84]. These semantically labeled scenes are subsequently used in fine-grained scene grounding tasks like object

Table 1: Comparison of SceneVerse with existing 3DVL Datasets. SceneVerse expands the data scale of prior work by order of magnitude. "VG" stands for Visual Grounding, "QA" for Question Answering, "PT" for Pre-training and "MT" for Multi-tasking. "Anno." denotes language from human annotations and "Syn." for template-based or LLM generated descriptions.

Dataset	3D		Task	Obj.	Scene	Obj.	Quality	New		Existing		Total
	Scene	Obj.		Caption	Caption	Referral	Check	Anno.	Syn.	Anno.	Syn.	
ScanRefer [16]	|	|	VG	✗	✗	✓	✓	52K	-	-	-	52K
ReferIt3D [1]			VG	✗	✗	✓	✓	42K	200K	-	-	242K
ScanQA [5]	1.5K	33K	QA	-	-	-	✓	27K	-	-	-	27K
SQA3D [65]			QA	-	-	-	✓	33K	-	-	-	33K
Multi3DRefer [104]	|	|	VG	✗	✗	✓	✓	-	10K	52K	-	62K
Cap3D [63]	-	666K	VG	✗	✓	✗	✗	58K	666K	-	-	724K
ScanScribe [109]	3K	56K	PT	✗	✗	✓	✗	-	90K	94K	94K	278K
3D-LLM [40]	1.5K	186K	MT	✓	✓	✓	✗	-	659K	-	-	659K
EmbodiedScan [87]	5K	890K	VG	✗	✗	✓	✗	-	970K	-	-	970K
LEO [41]	3K	56K	MT	✓	✓	✓	✓	-	188K	235K	90K	513K
SceneVerse	**68K**	**1.5M**	VG	✓	✓	✓	✓	**96K**	**2.1M**	**94K**	**200K**	**2.5M**

referral [1,16,104], captioning [17,19,22,99], vision-language-navigation [39,64,73, 88] and reasoning [5,40,65]. Recent works exploit the representation of 3D scene graph (3DSG) [4,20,78,86], which concisely describes scenes with hierarchical structures. This representation is notably advantageous for planning [2,77] and captioning [35], owing to its compatibility with LLMs for flexible description generation [40,41]. Nonetheless, as shown in Tab. 1, most datasets are constrained in both scene and language scales, underscoring the need for scaling up fine-grained and aligned scene-language data to enhance grounded scene understanding.

Vision-Language Learning Recent years have witnessed tremendous progress in 2D-VL [3, 21, 25, 57, 59, 75], empowered by transformer-based pre-training models [11, 29, 72] and large-scale image-language datasets [15, 79]. A central theme across 2D-VL domains is the effectiveness of data scaling [49], as demonstrated by improved alignment and expanded capabilities in open-vocabulary understanding [33, 51, 54, 58] through a contrastive pre-training pipeline [75].

However, in grounded scene understanding, the primary challenge for models has been the limited availability of paired 3D scene-language data, which restricts the application of insights drawn from 2D-VL. Current models for 3D scene grounding [6,18,37,43,45,62,92,97,105] heavily rely on task-specific knowledge in both model and loss designs or advanced optimization strategies [109]. To bridge this gap, there has been a growing emphasis on employing pre-trained 2D-VL models for 3D-VL [36, 38, 74, 82, 93, 102, 103]. Yet, these models mostly draw information available from 2D-VL models (*e.g.*, object attribute, affordance, *etc.*), falling short on capturing crucial 3D information like object spatial relationships which are necessary for more fine-grained tasks such as grounded human-scene [20, 44, 47, 48, 89, 90] and robot-scene interactions modeling [34, 55, 68, 70]. This urges the need for a multi-level alignment between language and 3D scenes, particularly regarding 3D-specific information. Considering the nascent stage of existing 3D pre-training methods [30, 94, 109, 110], we believe SceneVerse and GPS have the potential to spearhead new avenues in 3D-VL research.

3 SceneVerse

SceneVerse is designed for grounded scene understanding with 3D scenes curated from diverse existing datasets of both real and synthetic environments. Regarding language, we employ both human annotation and a novel automated generation pipeline to collect comprehensive and high-quality language for both object-level and scene-level descriptions. We provide details regarding data collection in the following sections.

3.1 Scene Curation

To address the scarcity of available 3D scene data, we construct SceneVerse by unifying 3D scene data from various existing datasets. We use real-world scene datasets, including ScanNet [24], ARKitScenes [9], HM3D [76], 3RScan [85] and MultiScan [66], alongside synthetic environments from Structured3D [106] and ProcTHOR [28]. The inclusion of these synthetic datasets is mainly motivated by their potential as scalable data sources for 3D-VL alignment. To ensure cohesion across various sources, we conduct preprocessing steps such as room segmentation, point subsampling, axis alignment, normalization, and semantic label alignment. Each scan is represented by a point cloud $\mathrm{P} \in \mathbb{R}^{N \times 8}$, wherein each point is defined by its 3D coordinates, RGB color, instance id, and semantic label. In total, we curate $68,406$ 3D scenes in SceneVerse.

3.2 Referral Annotation by Humans

In the curated scenes of SceneVerse, we present the most comprehensive set of human-annotated, context-rich object referrals to date, serving as a valuable benchmark for assessing grounded scene understanding capabilities. The human annotations contain $96,863$ descriptions in ARKitScenes [9], HM3D [76] and MultiScan [66]. During the annotation process, one human annotator was assigned to write at least 20 words to distinctly refer to a single 3D object within a 3D scene. Each referral text then undergoes independent verification by two additional reviewers, both mandated to accurately locate the referenced object based on the 3D scene and the annotated referral text. Any object referrals that do not pass the verification by either reviewer are flagged for re-annotation.

3.3 3D Scene Graph Construction

Our 3D scene graph is defined as a hierarchical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v \in \mathcal{V}$ represents one distinct 3D object instance, parameterized by its centroid $\boldsymbol{p}_i \in \mathbb{R}^3$ and bounding box size of $\boldsymbol{b}_i = (b_x, b_y, b_z) \in \mathbb{R}^3$. The edges $\mathcal{E}$ represent spatial relationships between nodes. To construct the scene graph $\mathcal{G}$, we instantiate the nodes with the instance annotation from the point clouds and assign object classes with their corresponding semantic labels. Following prior work [1, 86], we consider the **Vertical proximity**, **Horizontal proximity** and **Multi-object Relationships** as spatial relations. For a more detailed description of the scene graph construction and relationship determination, please refer to *supplementary*.

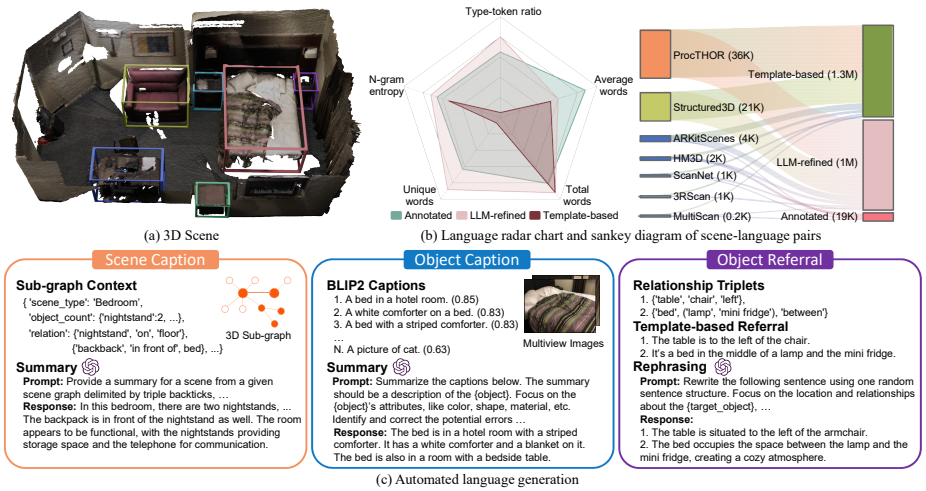

Fig. 2: SceneVerse collection and statistics. Given a 3D scene (a), our automated pipeline (c) generates three types of description including scene caption, object caption and object referral. (b) SceneVerse data comparison and composition.

3.4 Language Generation with LLMs

The scene-language pairs in SceneVerse aim to capture varying aspects of the 3D scene, including detailed object attributes in object captioning, spatial relationships between objects in object referral, and global scene descriptions in scene captioning. Based on 3D scene graphs, we utilize both templates and LLMs to automatically generate descriptions on these three granularities.

Object Captioning Object captions aim to provide detailed descriptions of an object's visual and physical properties, facilitating object-level grounding with its distinctive features. Given the multi-view images, we utilize the point cloud of the object $v \in \mathcal{V}$ to identify its occurrence in the images through rendering. The images are then cropped with the rendered bounding boxes and processed through BLIP2 [56] to generate initial object captions. We select the top 10 sentences with the highest CLIP [75] similarity score and minimal occlusion and utilize an LLM to obtain a refined and coherent summary of the captions. The detailed object captioning pipeline is illustrated in *supplementary*.

Object Referral Object relationship captions refer to objects by articulating their spatial relationships in the scene. Spatial relationship triplets (v_i, v_j, e_{ij}) are first extracted from the constructed 3D scene graph. We design various templates to generate descriptions for each relationship type, assigning the entities in the form of (target-object, spatial-relation, anchor-object(s)). This results in examples like "the chair is next to the li2023behaviorarmchair", "facing the sofa, there is a suitcase far to the right of the shoes", and "the fridge is between cabinet and sofa". Our designed templates span passive and active tenses, as well as inversion clauses, contributing to the richness of the generated text. To enhance the descriptions' naturalness, we employ LLM for sentence rephrasing. Statistics for the descriptions before and after rephrasing are presented in Fig. 2 (b).

Scene Captioning The scene-level captions emphasize global information, portraying the key objects in the scene along with their attributes and functionalities. We use the constructed 3D scene graph and prompt LLMs to generate these captions. We random sample a subset of edges and nodes from the scene graph each time as the scene context to enhance the diversity of scene captions. The object counts are also provided as LLM prompts, together with the room type and object attributes if such annotations are available in the dataset.

3.5 Data Quality and Statistics

Data Quality The 96K human-annotated set of SceneVerse is collected through AMT, where 82 humans are employed for annotation and 18 for verification. All final annotations passed the reference verification, with a re-annotation rate of 4.8%. For our automatic language generation pipeline, we conduct extensive prompt tuning and iterate with human feedback for LLMs on object captioning, summary, and rephrasing. To verify the efficacy of the pipeline, we conduct a quality check where 12K generated object-level descriptions are randomly selected for human verification. Results demonstrate a 96.93% pass rate, surpassing that in ReferIt3D [1] with 86.1% pass rate on 2K samples.

Statistics In total, SceneVerse comprises a total of 68,406 room-level 3D scans, with the source composition shown in Fig. 2 (b). The dataset contains 1.5M object instances ranging in 2290 object categories. Our generated 3D scene graph comprises 21 types of relationships following prior work [1, 86]. For the language descriptions, we generate 1M template-based texts and 1M sentences rephrased by Llama [83] and GPT-3.5 [71]. As can be seen from the radar chart and examples in Fig. 2, the diversity of the LLM-refined data, particularly in sentence length and variety, closely aligns with the characteristics of annotated descriptions, surpassing the template-based data. Together with existing sources (294K) and our newly annotated set (96K), SceneVerse contains 2.5M scene-language pairs in total. All the rephrasing and summary prompts, along with the complete set of relationships, are detailed in *supplementary*.

4 Grounded Pre-training for Scenes

In this section, we introduce GPS, an efficient transformer-based model trained with multi-level contrastive losses for aligning 3D scenes and texts. As shown in Fig. 3, we echo the language descriptions collected at different granularities to form contrastive objectives at both object-level, referral-object-level, and scene-level in GPS. We describe the design of each level in the following sections.

4.1 Object-level Grounding

Given a 3D scene point cloud $\mathcal{S}$, we use an off-the-shelf 3D object segmentation model to decompose it into a bag of N objects $\mathcal{S} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_n\}_{i=1}^{N}$. We extract

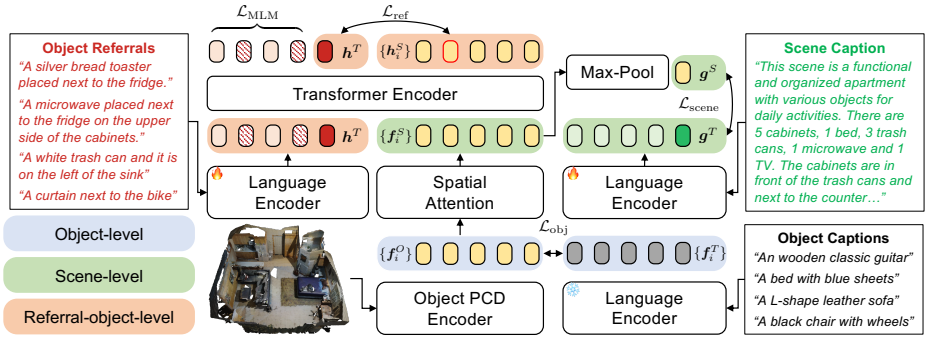

Fig. 3: Overview of GPS model. We use contrastive alignment at three levels $\mathcal{L}_{\text{obj}}$, $\mathcal{L}_{\text{scene}}$, and $\mathcal{L}_{\text{ref}}$ and a masked language modeling objective $\mathcal{L}_{\text{MLM}}$ for model learning.

object features $\{\boldsymbol{f}_i^O\}$ with an object point cloud encoder and text features $\{\boldsymbol{f}_i^T\}$ by feeding object-captions $\{\boldsymbol{T}_i^{\text{obj}}\}$ into a frozen language model. Following [93], we perform cross-modal alignment on the object features and text features via:

$$\mathcal{L}_{\text{obj}} = -\frac{1}{2}\sum_{(p,q)}\left(\log\frac{\exp\left(D^{\text{obj}}(p,q)\right)}{\sum_r \exp\left(D^{\text{obj}}(p,r)\right)} + \log\frac{\exp\left(D^{\text{obj}}(p,q)\right)}{\sum_r \exp\left(D^{\text{obj}}(r,q)\right)}\right), \tag{1}$$

where $D^{\text{obj}}(p,q) = (\boldsymbol{f}_p^O \boldsymbol{f}_q^T/\tau)$ denotes the dot product between object and text features and (p,q) denotes a pair of aligned object-text pair in the training batch and r iterates over all object-text pairs in the training batch. Similar to CLIP [75], we use a learnable temperature parameter τ to facilitate model learning.

4.2 Scene-level Grounding

With aligned object features, we encode the scene by incorporating object spatial locations into the extracted object features. Specifically, we use a spatial transformer model to encode extracted object features $\{\boldsymbol{f}_i^O\}$ with their spatial location features $\{\boldsymbol{l}_i\}$ following [18, 109]:

$$\boldsymbol{f}^S = \text{SpatialAttn}(\{\boldsymbol{f}_i^O\}, \{\boldsymbol{l}_i\})$$

where $\{\boldsymbol{f}_i^S\}$ denotes the feature of object $\boldsymbol{o}_i$ after encoding with spatial location features. To perform scene-level alignment, we operate on these scene-level object features $\{\boldsymbol{f}_i^S\}$ and align it with the scene caption $\boldsymbol{T}^{\text{scene}}$. Specifically, we feed the object features into a projection layer and use max-pooling over all object features to obtain the scene feature $\boldsymbol{g}^S$. Similar to object-level grounding, we pass the scene caption through a tunable language model to obtain text feature $\boldsymbol{g}^T$ and perform scene-level contrastive alignment through:

$$\mathcal{L}_{\text{scene}} = -\frac{1}{2}\sum_{(p,q)}\left(\log\frac{\exp\left(D^{\text{scene}}(p,q)\right)}{\sum_r \exp\left(D^{\text{scene}}(p,r)\right)} + \log\frac{\exp\left(D^{\text{scene}}(p,q)\right)}{\sum_r \exp\left(D^{\text{scene}}(r,q)\right)}\right), \tag{2}$$

where $D^{\text{scene}}(p, q) = (\boldsymbol{g}_p^S \boldsymbol{g}_q^T / \tau)$ denotes the dot product between scene feature $\boldsymbol{g}_p^S$ and scene caption feature $\boldsymbol{g}_q^T$ for each pair of aligned scene-text pairs in the training batch and r iterates over all scene-text pairs in the training batch.

4.3 Referral-object-level Grounding

To model the relationships revealed in referring expressions, we employ a self-attention-based reasoning transformer for grounding object referrals in scenes. This transformer takes in scene-object features $\{\boldsymbol{f}_i^S\}$ and an object referral $\boldsymbol{T}^{\text{ref}}$ and performs self-attention to learn relationships between text descriptions and object relationships. We use the same tunable language encoder as in scene-level grounding for extracting per-object referral features. We pass this text feature together with scene-object features into the self-attention transformer to obtain the aligned object features $\boldsymbol{h}_i^S$ and the sentence-level referral feature $\boldsymbol{h}^T$. We then perform the referral-object-level contrastive alignment following:

$$\mathcal{L}_{\text{ref}} = -\log \frac{\exp\left(\bar{\boldsymbol{h}}^S \boldsymbol{h}^T / \tau\right)}{\sum_p \exp\left(\boldsymbol{h}_p^S \boldsymbol{h}^T / \tau\right)}, \tag{3}$$

where $\bar{\boldsymbol{h}}^S$ denotes the feature of the referred object, p iterates over all objects within the same scene. Notably, in contrast to inter-scene contrast that was done in object- and scene-level alignment, we force the selection of positive pairs to be within the same scene to provide intra-scene contrast for fine-grained object grounding. This mimics the success of intra-image and inter-image contrasts commonly used for region-word alignment in 2D-VL models [101].

To learn the multi-level alignment between 3D scenes and language, we first train the point cloud encoder with an object-level grounding objective to obtain a good feature initialization for grounding objects in scenes. During the scene grounding stage, we train our inter- and intra-scene objectives together with a masked language modeling loss $\mathcal{L}_{\text{MLM}}$ over the inputted object-referral texts to tune the parameters within the language encoder and self-attention transformer. Above all, the learning of GPS could be summarized as optimizing:

$$\mathcal{L} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{ref}} + \mathcal{L}_{\text{MLM}}.$$

5 Experiments

In this section, we present experimental results addressing the following questions:

1. How effective is the data scaling in SceneVerse for 3D visual grounding? Does the scale-up benefit common 3D-VL tasks (*e.g.*, 3D question answering, open-vocabulary 3D semantic segmentation) and pre-training-based models?
2. How well is the GPS pre-training pipeline for 3D-VL tasks? Does it exhibit similar properties of 2D-VL models in 3D-VL tasks?
3. What is offered by SceneVerse and GPS and what is missing?

In the following sections, we describe in detail the model performance regarding these key topics. Due to the page limit, we direct readers to the *supplementary* for implementation details, qualitative results, and more experimental analyses.

Table 2: 3D visual grounding results on Nr3D, Sr3D, and ScanRefer. We use "*pre-train*" for our model trained on SceneVerse w/o additional fine-tuning, and "*fine-tune*" for its data-specific fine-tuned version. Best results are highlighted in **bold**.

Method	Nr3D					Sr3D					ScanRefer Acc@0.5		
	Overall	Easy	Hard	V-Dep.	V-Indep.	Overall	Easy	Hard	V-Dep.	V-Indep.	Overall	Unique	Multiple
3DVG-Trans [105]	40.8	48.5	34.8	34.8	43.7	51.4	54.2	44.9	44.6	51.7	34.7	60.6	28.4
TGNN [42]	37.3	44.2	30.6	35.8	38.0	45.0	48.5	36.9	45.8	45.0	29.7	56.8	23.2
TransRefer3D [37]	48.0	56.7	39.6	42.5	50.7	57.4	60.5	50.2	49.9	57.7	-	-	-
InstanceRefer [100]	38.8	46.0	31.8	34.5	41.9	48.0	51.1	40.5	45.8	48.1	32.9	66.8	24.7
FFL-3DOG [32]	41.7	48.2	35.0	37.1	44.7	-	-	-	-	-	34.0	67.9	25.7
LAR [6]	48.9	58.4	42.3	47.4	52.1	59.4	63.0	51.2	50.0	59.1	-	-	-
SAT [97]	56.5	64.9	48.4	54.4	57.6	57.9	61.2	50.0	49.2	58.3	30.1	50.8	25.2
3D-SPS [62]	51.5	58.1	45.1	48.0	53.2	62.6	56.2	65.4	49.2	63.2	37.0	66.7	29.8
3DJCG [12]	-	-	-	-	-	-	-	-	-	-	37.3	64.3	30.8
BUTD-DETR [45]	54.6	60.7	48.4	46.0	58.0	67.0	68.6	63.2	53.0	67.6	39.8	66.3	35.1
MVT [43]	59.5	67.4	52.7	59.1	60.3	64.5	66.9	58.8	58.4	64.7	33.3	66.5	25.3
ViL3DRel [18]	64.4	70.2	57.4	**62.0**	64.5	72.8	74.9	67.9	63.8	73.2	37.7	68.6	30.7
EDA [92]	52.1	58.2	46.1	50.2	53.1	68.1	70.3	62.9	54.1	68.7	42.3	68.6	37.6
3D-VisTA (*scratch*) [109]	57.5	65.9	49.4	53.7	59.4	69.6	72.1	63.6	57.9	70.1	41.5	70.9	34.8
3D-VisTA [109]	64.2	72.1	56.7	61.5	65.1	76.4	78.8	71.3	58.9	77.3	45.8	75.1	39.1
Ours (*scratch*)	58.7	67.0	50.9	55.8	59.8	68.4	70.5	63.4	53.1	69.0	40.4	71.3	34.7
Ours (*pre-train*)	55.2	62.8	48.0	45.5	58.8	74.1	76.4	68.5	54.1	75.0	47.1	77.4	41.6
Ours (*fine-tuned*)	**64.9**	**72.5**	**57.8**	56.9	**67.9**	**77.5**	**80.1**	**71.6**	**62.8**	**78.2**	**48.1**	**77.9**	**42.7**

5.1 3D Visual Grounding

Settings We evaluate our model on three commonly-used datasets for 3D visual grounding: ScanRefer [16], Nr3D, and Sr3D [1]. For Nr3D and Sr3D, we follow Achlioptas *et al*. [1] and report the grounding accuracies of models using ground-truth object masks. For ScanRefer, we follow Zhu *et al*. [109] and use Mask3D [80] to generate object proposals. Results are reported as Acc@0.5 to evaluate the correctness of predictions whose object bounding boxes overlap the ground truth with IoU > 0.5. For comparisons, we compare with existing baselines by providing the results of pre-trained GPS and dataset-specific fine-tuned GPS. Please see more details in the *supplementary*.

Results and Analyses As shown in Tab. 2, GPS trained on SceneVerse achieves state-of-the-art results on all existing 3D-VL grounding benchmarks. Initially, when GPS is trained directly on the training sets of benchmark datasets, labeled as Ours (*scratch*), it underperforms compared to existing models that employ more complex structures or loss designs. This result underscores the data-intensive nature of the contrastive alignment paradigm. However, when presented with extensive training data in SceneVerse, the results of our model without additional fine-tuning, *i.e*., Ours (*pre-train*), significantly improves and already achieves state-of-the-art results on benchmarks like ScanRefer. Moreover, the dataset-specific fine-tuned model, *i.e*., Ours (*fine-tuned*), *consistently outperforms existing baselines with only a simple projection MLP* added on top of the pre-trained model, jointly optimized during fine-tuning without any other auxiliary architecture or loss objective. These results underscore the strong potential of both the SceneVerse and GPS for 3D-VL tasks.

Table 3: Zero-shot transfer on existing benchmarks. "SR" stands for ScanRefer.

Method	Nr3D	Sr3D	SR@0.25	SR@0.5
3D-VisTA (*scratch*)	57.5	69.6	45.9	41.5
3D-VisTA (*zero-shot*)	35.2	31.2	33.2	29.6
3D-VisTA (*zero-shot text*)	43.1	36.1	41.1	36.4
Ours (*scratch*)	58.7	68.4	44.5	40.4
Ours (*zero-shot*)	32.4	33.3	35.2	31.1
Ours (*zero-shot text*)	41.9	38.1	40.7	35.8

Table 4: Zero-shot transfer on SceneVerse-val. Evaluation uses GT object proposals following Nr3D/Sr3D.

Method	Overall	Easy	Hard	V-Dep.	V-Indep.
3D-VisTA (*scratch*)	40.7	53.1	21.6	37.3	44.3
3D-VisTA (*zero-shot*)	52.9	59.6	35.4	53.7	52.2
3D-VisTA (*zero-shot text*)	58.1	70.0	39.6	52.5	64.1
Ours (*scratch*)	38.5	50.2	20.8	33.7	43.9
Ours (*zero-shot*)	59.2	69.4	44.0	53.1	66.3
Ours (*zero-shot text*)	60.6	70.9	45.1	54.8	67.3

5.2 Zero-Shot Transfer

Settings To better evaluate the effectiveness of both the SceneVerse data and the GPS model, we further perform zero-shot transfer experiments to test the models' capability in 4 benchmarks, ScanRefer, Sr3D, Nr3D, and SceneVerse-val. We create SceneVerse-val using $8.5K$ annotated object referrals of 271 scenes in MultiScan, and randomly split the scenes following a 4:1 train / test split for creating the held-out test set. We mainly consider 2 specific transfer settings in our experiments: (i) *zero-shot*: models trained by removing all the scenes from the target dataset, tested on held-out unseen scenes, and (ii) *zero-shot text*: Models trained on data that include the training set of scenes from the target dataset, yet tested exclusively with unseen scene-text distribution. Specifically, for the *zero-shot text* setting, we use the generated texts in SceneVerse as fine-tuning sources for the *zero-shot* model. We mainly compare our model against a recent pre-training-based model 3D-VisTA. See more details on experimental setting and implementation in the *supplementary*.

Results and Analyses We present the results of zero-shot transfer experiments in Tab. 3 and Tab. 4 with the following key observations:

- Our GPS model demonstrates superior generalization to unseen scenes compared to the 3D-VisTA model. In zero-shot transfer scenarios, our model consistently outperforms 3D-VisTA across established benchmarks and SceneVerse-val. This indicates the effectiveness of contrastive alignment over traditional classification objectives, aligning with the advancements seen in 2D-VL models for open-vocabulary grounding and transfer capabilities
- SceneVerse dataset substantially enhances 3D-VL grounding capabilities through zero-shot transfer, especially when provided with relatively limited training data, *i.e.*, SceneVerse-val. As demonstrated in Tab. 4, there is a significantly improved performance when comparing models trained on SceneVerse in a zero-shot manner to those trained from scratch. This indicates that SceneVerse can effectively capture knowledge for general 3D scene grounding. Consequently, *this underscores its potential as a go-to pre-training dataset for 3D-VL tasks.*
- The impact of our extensive collection and scalable generation of scene-text pairs is further evidenced by the results in the *zero-shot text* setting. Notably, as shown in Tab. 3, these automatically generated scene-text pairs supply ample knowledge for comprehending the scene distribution. This contributes significantly to the substantial improvement over the *zero-shot* performance.

Table 5: 3D question answering results on ScanQA and SQA3D. We report EM@1 score on ScanQA and SQA3D evaluation sets.

Model	ScanQA			SQA3D
	val	w/obj	w/o obj	
ScanRefer+MCAN [5]	18.6	20.6	19.0	-
ScanQA [5]	20.3	23.5	20.9	46.6
SQA3D [65]	-	-	-	47.2
3D-VisTA [109]	22.4	**27.0**	23.0	48.5
3D-LLM [40]	20.5	19.1	-	-
Ours	**22.7**	25.0	**23.5**	**49.9**

Table 6: Exisiting 3D backbones pre-trained on SceneVerse for open-vocabulary 3D semantic segmentation on ScanNet. "SPUNet" denotes SparseUNet proposed in [94].

Model	Network	mIoU	Δ	mAcc	Δ
OpenScene [74]	SPUNet16	57.2	-	69.9	-
PLA [30]	SPUNet16	17.7	-	33.5	-
RegionPLC [94]	SPUNet16	56.9	-	75.6	-
RegionPLC+SceneVerse	SPUNet16	**58.2**	+1.7%	**77.3**	+2.2%
OpenScene [74]	SPUNet32	57.8	-	70.3	-
PLA [30]	SPUNet32	19.1	-	41.5	-
RegionPLC [94]	SPUNet32	59.6	-	77.5	-
RegionPLC+SceneVerse	SPUNet32	**61.0**	+2.3%	**79.7**	+2.8%

5.3 Additional 3D-VL Tasks

Settings We evaluate the effectiveness of GPS and SceneVerse on additional 3D-VL tasks: (i) 3D question answering (3D-QA) on ScanQA [5] and SQA3D [65], and (ii) open-vocabulary 3D semantic segmentation (OV-Seg) on ScanNet.

- In the 3D-QA task, we follow Zhu *et al.* [109] and evaluate models over the exact match metic (EM@1) on the validation and test sets of ScanQA, as well as the test set of SQA3D. We pre-train GPS on SceneVerse and fine-tune the model on the 3D-QA dataset to compare with state-of-the-art models.
- In the OV-Seg task, as GPS builds upon an object-centric design and thus is not directly applicable to semantic segmentation, we consider testing the effectiveness of SceneVerse on improving existing 3D models. Specifically, we follow the open-vocabulary semantic segmentation settings proposed by Yang *et al.* [94] and report the mIoU and mAcc score. We compare with existing works by pre-training the RegionPLC [94] model on SceneVerse.

Results and Analyses We present the results of 3D-QA experiments in Tab. 5 and the results of OV-Seg experiments in Tab. 6. The analyses are as follows:

- As shown in Tab. 5, our model achieves state-of-the-art results on both benchmarks, outperforming recent strong pre-training-based baselines like 3D-VisTA and 3D-LLM. As SceneVerse currently contains only descriptions of objects and scenes, we believe involving more types of language descriptions (*e.g.*, question-answer pairs, dialogues) is a promising direction for further improving model performance on these downstream tasks.
- As shown in Tab. 6, we observe consistent performance improvement of existing 3D backbone models on this task when pre-trained with SceneVerse data. This result validates that the collected data in SceneVerse can effectively boost the performance of existing models on scene understanding tasks. We further provide results of state-of-the-art 3D models that are pre-trained on SceneVerse on the close-vocabulary 3D semantic segmentation task in the *supplementary*.

5.4 Ablative Studies and Discussion

In this section, we present further discussions on both the data collected in SceneVerse and the GPS model design. We aim to elucidate the effects of

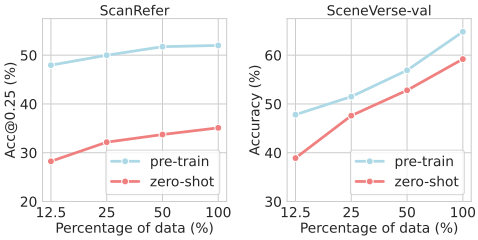

Fig. 4: Model performance *v.s.* data scale. Plots show that models consistently improve in both the pre-train and zero-shot transfer settings on ScanRefer and SceneVerse-val with data scaling-up.

Table 8: Cross domain transfer results of models pre-trained on real and synthetic datasets. "S3D" stands for Structured3D.

Real	Synthetic	SceneVerse-val	S3D	ProcTHOR
All	✗	64.8	37.1	43.4
✗	S3D	7.0	85.1	16.1
✗	ProcTHOR	4.2	16.3	91.0

Table 7: Ablation on text data source used in model pre-training. All models are tested on ScanRefer with no additional finetuning.

Template	LLM	Anno.	Acc@0.25	Acc@0.5
✗	✗	✗	43.5	38.4
✓	✗	✗	50.9	46.1
✓	✓	✗	51.1	46.3
✓	✓	✓	52.0	47.1

Table 9: Ablation on model design on SceneVerse-val. We use "Obj-lvl", "Scene-lvl" to denote object and scene alignment loss, and "MLM" for the mask language modeling loss.

Obj-lvl.	MLM	Scene-lvl.	Overall	Easy	Hard
✗	✗	✗	64.8	75.4	48.7
✓	✗	✗	65.2	77.1	47.4
✓	✓	✗	62.4	73.4	45.8
✓	✓	✓	**66.9**	**77.8**	**50.3**

data scaling and show more clearly its effectiveness in 3D scene understanding. Regarding the experimental settings and more results discussion, refer to the *supplementary*. The following points are specifically discussed in this section:

How important is data-scaling? We conduct ablation studies over the amount of data used while pre-training GPS. We consider the model trained with $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ of SceneVerse to show the effectiveness of data-scaling on model performance in the pre-train and zero-shot transfer settings in ScanRefer and SceneVerse-val. As shown in Fig. 4, we observe consistent performance improvement over the increase of data scale for both settings. We provide additional experiments in the *supplementary* to show that such scaling effect is not only beneficial for 3D-VL grounding but also for other 3D tasks like semantic segmentation [80, 96].

How is the generated data compared with human-annotated data? We assess the performance of models trained using various scene-text sources, specifically focusing on their performance in the ScanRefer dataset without additional fine-tuning. As shown in Tab. 7, models trained with our template-based generated texts and Large Language Model (LLM)-refined texts show significant improvements over models trained solely on ScanRefer. More importantly, these variants of our model already achieve state-of-the-art results compared with previous baselines. This indicates the effectiveness of our text-generation pipeline. Finally, we observe that adding human-annotated data is still beneficial for model performance. However, the improvement is relatively marginal over models trained on our generated data.

What is the role of the synthetic scenes in this scale-up process? With synthetic data providing large-scale and diverse scene data for 3D-VL tasks,

we evaluate the models' domain transfer (Sim2Real) capability. Specifically, we compare models trained on all real scenes in SceneVerse against models trained exclusively on two synthetic subsets of SceneVerse, *i.e.*, Structured3D and ProcTHOR. As shown in Tab. 8, models trained on synthetic subsets demonstrate remarkable performance on their corresponding test sets while suffering when transferred to real or other synthetic scenes. In contrast, the model trained on real scene-text pairs exhibits less performance drop when generalizing to synthetic scenes. This result affirms the domain gap between real and synthetic scenes in 3D-VL grounding and shows that a simple scale-up in the number of scenes is insufficient when naturalness can not be guaranteed. Considering the scalability of our language generation pipeline and the scaling effect shown in our experiments, the rate-determining step for further scaling-up 3D-VL comes to the collection of diverse, high-quality, and realistic scenes that capture natural 3D scene distributions.

How important is the design of each module in GPS? We provide ablative analyses of our multi-level contrastive alignment design in Tab. 9. We mainly consider removing objectives in our model to reveal the effectiveness of each level of alignment. We choose the referral-object-level alignment objective as the default setting and consider removing: (i) object-level alignment objective, (ii) masked language modeling objective, and (iii) scene-level alignment objective. When removing the object-level alignment objective, we learn the object point cloud encoder with the referral-object-level alignment and without pre-training. As shown in Tab. 9, we test different models on the SceneVerse-val without additional fine-tuning. Results show that the scene-level alignment objective is crucial for referral object grounding in SceneVerse-val with the $\sim$5% performance drop. Similar observations could be made for the model trained without object-level alignment ($\sim$2% drop) and masked language modeling objective ($\sim$1.5% drop). These results affirm the effectiveness of our model design.

6 Conclusion

In this work, we scale up 3D-VL for grounded scene understanding. We present SceneVerse, a million-scale 3D-VL dataset covering various scenes and multi-level scene descriptions sourced from both human annotation and our proposed scene-text generation approach. Utilizing SceneVerse, we propose Grounded Pre-training for Scenes (GPS), a model trained with multi-level scene-language contrastive alignment. Through extensive experiments, we show that GPS achieves state-of-the-art results on common 3D-VL tasks including grounding and question answering. We further conduct zero-shot transfer experiments to show the improved generalization performances of GPS trained on SceneVerse compared with previous baselines. We also demonstrate that the scaling effect of SceneVerse is generally beneficial for existing 3D models on 3D-VL tasks like semantic segmentation. We hope our efforts and successful scale-up attempts in SceneVerse could pave the way for new research paradigms in 3D-VL.

Acknowledgement The authors would like to thank Yaowei Zhang (BIGAI) for his help on online visualization and other colleagues from BIGAI General Vision Lab for fruitful discussions. The authors would also like to thank the anonymous reviewers for their constructive feedback.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: Proceedings of European Conference on Computer Vision (ECCV) (2020)
2. Agia, C., Jatavallabhula, K.M., Khodeir, M., Miksik, O., Vineet, V., Mukadam, M., Paull, L., Shkurti, F.: Taskography: Evaluating robot task planning over large 3d scene graphs. In: Proceedings of Conference on Robot Learning (CoRL) (2022)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
4. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of International Conference on Computer Vision (ICCV) (2019)
5. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
6. Bakr, E., Alsaedy, Y., Elhoseiny, M.: Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
7. Barsalou, L.W.: Perceptual symbol systems. Behavioral and brain sciences **22**(4), 577–660 (1999)
8. Barsalou, L.W.: Grounded cognition. Annu. Rev. Psychol. **59**, 617–645 (2008)
9. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In: Proceedings of Advances in Neural Information Processing Systems Datasets and Benchmarks (NeurIPS Datasets and Benchmarks Track) (2021)
10. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
11. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
12. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
13. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. Proceedings of International Conference on 3D Vision (3DV) (2017)
14. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)

15. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
16. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: Proceedings of European Conference on Computer Vision (ECCV) (2020)
17. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3net: a speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. In: Proceedings of European Conference on Computer Vision (ECCV) (2022)
18. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Language conditioned spatial relation reasoning for 3d object grounding. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
19. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
20. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: Proceedings of International Conference on Computer Vision (ICCV) (2019)
21. Chen, Y., Li, Q., Kong, D., Kei, Y.L., Zhu, S.C., Gao, T., Zhu, Y., Huang, S.: Yourefit: Embodied reference understanding with language and gesture. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
22. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
23. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
24. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
25. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
26. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
27. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
28. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Procthor: Large-scale embodied ai using procedural generation. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
29. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2018)

30. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
31. Ding, Z., Han, X., Niethammer, M.: Votenet: A deep learning label fusion method for multi-atlas segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2019)
32. Feng, M., Li, Z., Li, Q., Zhang, L., Zhang, X., Zhu, G., Zhang, H., Wang, Y., Mian, A.: Free-form description guided 3d visual graph network for object grounding in point cloud. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
33. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: Proceedings of European Conference on Computer Vision (ECCV) (2022)
34. Gong, R., Huang, J., Zhao, Y., Geng, H., Gao, X., Wu, Q., Ai, W., Zhou, Z., Terzopoulos, D., Zhu, S.C., Jia, B., Huang, S.: Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
35. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. arXiv preprint arXiv:2309.16650 (2023)
36. Ha, H., Song, S.: Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In: Proceedings of Conference on Robot Learning (CoRL) (2022)
37. He, D., Zhao, Y., Luo, J., Hui, T., Huang, S., Zhang, A., Liu, S.: Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In: Proceedings of ACM International Conference on Multimedia (MM) (2021)
38. Hegde, D., Valanarasu, J.M.J., Patel, V.: Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
39. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
40. Hong, Y., Lin, C., Du, Y., Chen, Z., Tenenbaum, J.B., Gan, C.: 3d concept learning and reasoning from multi-view images. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
41. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. arXiv preprint arXiv:2311.12871 (2023)
42. Huang, P.H., Lee, H.H., Chen, H.T., Liu, T.L.: Text-guided graph neural networks for referring 3d instance segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI) (2021)
43. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
44. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
45. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: Proceedings of European Conference on Computer Vision (ECCV) (2022)

46. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
47. Jiang, N., Liu, T., Cao, Z., Cui, J., Zhang, Z., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
48. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
49. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
50. Khanna, M., Mao, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
51. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
52. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: International Journal of Computer Vision (IJCV) (2017)
53. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
54. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: Proceedings of International Conference on Learning Representations (ICLR) (2022)
55. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Proceedings of Conference on Robot Learning (CoRL) (2023)
56. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning (ICML) (2023)
57. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of International Conference on Machine Learning (ICML) (2022)
58. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
59. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
60. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. arXiv preprint arXiv:2305.10764 (2023)
61. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)

62. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
63. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
64. Ma, C.Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., Xiong, C.: Self-monitoring navigation agent via auxiliary progress estimation. In: Proceedings of International Conference on Learning Representations (ICLR) (2019)
65. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. In: Proceedings of International Conference on Learning Representations (ICLR) (2023)
66. Mao, Y., Zhang, Y., Jiang, H., Chang, A., Savva, M.: Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
67. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
68. Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J.L., Singh, R., Guo, Y., Mazhar, H., et al.: Orbit: A unified simulation framework for interactive robot learning environments. Robotics and Automation Letters (RA-L) (2023)
69. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
70. Mu, T., Ling, Z., Xiang, F., Yang, D., Li, X., Tao, S., Huang, Z., Jia, Z., Su, H.: Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2021)
71. OpenAI: Introducing chatgpt. `https://openai.com/blog/chatgpt` (2022)
72. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
73. Pashevich, A., Schmid, C., Sun, C.: Episodic transformer for vision-and-language navigation. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
74. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
75. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning (ICML) (2021)
76. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In: Proceedings of Advances in Neural Information Processing Systems Datasets and Benchmarks (NeurIPS Datasets and Benchmarks Track) (2021)
77. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In: Proceedings of Conference on Robot Learning (CoRL) (2023)

78. Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carlone, L.: Kimera: From slam to spatial perception with 3d dynamic scene graphs. International Journal of Robotics Research (IJRR) (2021)
79. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
80. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: Proceedings of International Conference on Robotics and Automation (ICRA) (2023)
81. Smith, L., Gasser, M.: The development of embodied cognition: Six lessons from babies. Artificial life **11**(1-2), 13–29 (2005)
82. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
83. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
84. Vu, T., Kim, K., Luu, T.M., Nguyen, T., Yoo, C.D.: Softgroup for 3d instance segmentation on point clouds. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
85. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of International Conference on Computer Vision (ICCV) (2019)
86. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
87. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
88. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
89. Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: Move as you say interact as you can: Language-guided human motion generation with scene affordance. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
90. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
91. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
92. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

93. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
94. Yang, J., Ding, R., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. arXiv preprint arXiv:2304.00962 (2023)
95. Yang, Y., Jia, B., Zhi, P., Huang, S.: Physcene: Physically interactable 3d scene synthesis for embodied ai. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
96. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. arXiv preprint arXiv:2304.06906 (2023)
97. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
98. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
99. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., Li, Z.: X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
100. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
101. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2022)
102. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
103. Zhang, R., Wang, L., Qiao, Y., Gao, P., Li, H.: Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
104. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3d objects. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
105. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
106. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Proceedings of European Conference on Computer Vision (ECCV) (2020)
107. Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W.Y., Choi, Y.: Multimodal c4: An open, billion-scale corpus of images interleaved with text. arXiv preprint arXiv:2304.06939 (2023)

108. Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y.N., et al.: Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. Engineering **6**(3), 310–345 (2020)
109. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of International Conference on Computer Vision (ICCV) (2023)
110. Zhu, Z., Zhang, Z., Ma, X., Niu, X., Chen, Y., Jia, B., Deng, Z., Huang, S., Li, Q.: Unifying 3d vision-language understanding via promptable queries. In: Proceedings of European Conference on Computer Vision (ECCV) (2024)