TOD3Cap: Towards 3D Dense Captioning in Outdoor Scenes

Bu Jin[1,2], Yupeng Zheng[1,2⋆], Pengfei Li[3], Weize Li[3], Yuhang Zheng[4], Sujie Hu[3], Xinyu Liu[5], Jinwei Zhu[3], Zhijie Yan[3], Haiyang Sun[2], Kun Zhan[2], Peng Jia[2], Xiaoxiao Long[6], Yilun Chen[3], and Hao Zhao[3]

[1]CASIA [2]Li Auto [3]AIR, Tsinghua University [4]Beihang University [5]HKUST [6]HKU
jinbu18@mails.ucas.ac.cn, zhengyupeng2022@ia.ac.cn

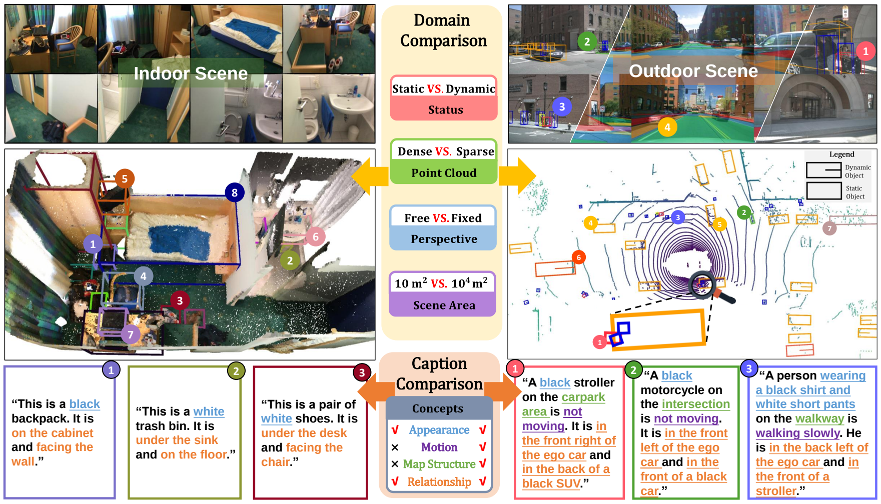
Abstract. 3D dense captioning stands as a cornerstone in achieving a comprehensive understanding of 3D scenes through natural language. It has recently witnessed remarkable achievements, particularly in indoor settings. However, the exploration of 3D dense captioning in outdoor scenes is hindered by two major challenges: 1) the **domain gap** between indoor and outdoor scenes, such as dynamics and sparse visual inputs, makes it difficult to adapt existing indoor methods directly; 2) the **lack of data** with comprehensive box-caption pair annotations specifically tailored for outdoor scenes. To this end, we introduce the new task of outdoor 3D dense captioning. As input, we assume a LiDAR point cloud and a set of RGB images captured by the panoramic camera rig. The expected output is a set of object boxes with captions. To tackle this task, we propose the ***TOD^{3}Cap* network**, which leverages the BEV representation to generate object box proposals and integrates Relation Q-Former with LLaMA-Adapter to generate rich captions for these objects. We also introduce the ***TOD^{3}Cap* dataset**, the first million-scale dataset to our knowledge for 3D dense captioning in outdoor scenes, which contains 2.3M descriptions of 64.3K outdoor objects from 850 scenes in nuScenes. Notably, our *TOD^{3}Cap* network can effectively localize and caption 3D objects in outdoor scenes, which outperforms baseline methods by a significant margin (+9.6 CiDEr@0.5IoU). Code, dataset and models are publicly available at https://github.com/jxbbb/TOD3Cap.

Keywords: 3D dense captioning · 3D scene understanding · 3D vision language · Dataset

1 Introduction

Recently, the community has witnessed significant progress in 3D dense captioning. By explicitly formulating the understanding of 3D scenes with natural language, it exhibits diverse applications in cross-modal retrieval [8, 23], robotic navigation [21, 41, 47, 55], interactive AR/VR [35] and autonomous driving [26, 37, 42, 43]. In this challenging setting, an algorithm is required to localize all of the objects in a 3D scene and caption their diverse attributes.

[⋆] indicates the corresponding author.



1 "This is a black backpack. It is on the cabinet and facing the wall."

2 "This is a white trash bin. It is under the sink and on the floor."

3 "This is a pair of white shoes. It is under the desk and facing the chair."

1 "A black stroller on the carpark area is not moving. It is in the front right of the ego car and in the back of a black SUV."

2 "A black motorcycle on the intersection is not moving. It is in the front left of the ego car and in the front of a black car."

3 "A person wearing a black shirt and white short pants on the walkway is walking slowly. He is in the back left of the ego car and in the front of a stroller."

Fig. 1: We introduce the task of 3D dense captioning in outdoor scenes (right). Given point clouds (right middle) and multi-view RGB inputs (right top), we predict box-caption pairs of all objects in a 3D outdoor scene. There are several fundamental domain gaps (middle column) between indoor and outdoor scenes, including Status, Point Cloud, Perspective, and Scene Area, bringing new challenges specific to outdoor scenes. Meanwhile, our outdoor 3D dense captioning (right bottom) contains more comprehensive concepts than indoor scenes (left bottom).

Some previous works [4,7,11,12,25,46,51] have explored the 3D dense captioning task and achieved promising results. However, these methods primarily focus on 3D dense captioning in indoor scenes, while outdoor 3D dense captioning has been rarely explored. Besides, taking a closer look at the significant differences between indoor and outdoor scenes (shown in Fig. 1), we argue it is sub-optimal to directly adapt these indoor methods to outdoor scenes, because:

- **Dynamic, not static.** Outdoor scenes are typically dynamic, necessitating the detection and tracking of objects with temporally changing status.
- **Sparse LiDAR point clouds.** The utilization of sparse point clouds collected through LiDAR for outdoor scenes presents significant challenges in shape understanding. What's worse, the sparsity level is spatially varying.
- **Fixed camera perspective.** While indoor scene scanning allows free camera trajectories (e.g., around an object of interest), outdoor scenes typically feature a fixed 6-camera rig, presenting a higher degree of self-occlusion.
- **Larger areas.** Outdoor scenes usually cover a significantly larger area.

These domain gaps pose significant challenges for successful 3D dense captioning in outdoor scenes. In this paper, we first formalize the new task of outdoor 3D dense captioning. It takes LiDAR point clouds and panoramic RGB images as inputs and the expected output is a set of object boxes with captions. Then, we

Table 1: Overview of existing 3D captioning datasets. The App., Mot., Env., Rel. denote Appearance, Motion, Environment and Relationship, respectively.

Dataset	Domain	Dense Capt.	App.	Mot.	Env.	Rel.	♯Scenes	♯ Frames	♯Sentences
Objaverse [17]	Object	✗	✓	✗	✗	✗	-	N/A	800K
SceneVerse [24]	Indoor	✓	✓	✗	✓	✓	68K	N/A	2.5M
SceneFun3D [18]	Indoor	✗	✗	✓	✗	✓	710	N/A	14.8K
ScanRefer [6]	Indoor	✓	✓	✗	✗	✓	800	N/A	51.5K
ReferIt3D [1]	Indoor	✓	✓	✗	✗	✓	800	N/A	41.5K
Multi3DRefer [53]	Indoor	✓	✓	✗	✗	✓	800	N/A	61.9K
nuCaption [49]	**Outdoor**	✗	✗	✓	✓	✓	265	34.1K	420K
Rank2Tell [39]	**Outdoor**	✗	✓	✓	✓	✓	116	5.8K	-
***TOD^{3}Cap* (Ours)**	**Outdoor**	✓	✓	✓	✓	✓	**850**	**34.1K**	**2.3M**

propose a transformer-based architecture, named *TOD^{3}Cap* network, to address this task. Specifically, we first create a unified BEV map from 3D LiDAR point clouds and 2D multi-view images. Then we use a query-based detection head to generate object proposals. We also employ a Relation Q-Former to capture the relationship between object proposals and the scene context. The object proposal features are finally fed into a vision-language model to generate dense captions. Thanks to the usage of adapter [52], *TOD^{3}Cap* network does not require retraining of the language model and thus we can leverage the commonsense in language foundation models pre-trained on a large corpus of data.

Apart from the fact that indoor-outdoor domain gaps render indoor architectures unsuitable for outdoor scenes, successfully addressing outdoor 3D dense captioning also suffers from the data hungriness [50] issue, i.e., the lack of aligned box-caption pairs for outdoor scenes. To facilitate future research in outdoor 3D dense captioning, we collect the *TOD^{3}Cap* dataset, which provides box-wise natural language captions for LiDAR point cloud and panoramic RGB images from nuScenes [3]. In total, we acquire 2.3M captions of 63.4k outdoor instances. To the best of our knowledge, our *TOD^{3}Cap* dataset is the first 3D dense captioning effort of million-scale sentences, the largest one to date for outdoor scenes. To summarize, our contributions are as follows:

- We introduce the outdoor 3D dense captioning task to densely detect and describe 3D objects, using LiDAR point clouds along with a set of panoramic RGB images as inputs. Its unique challenges are highlighted in Fig. 1.
- We provide the *TOD^{3}Cap* dataset containing 2.3M descriptions of 63.4k instances in outdoor scenes and adapt existing state-of-the-art approaches on our proposed *TOD^{3}Cap* dataset for benchmarking.
- We show that our method outperforms the baselines adapted from representative indoor methods by a significant margin (+**9.6 CiDEr@0.5IoU**).

2 Related Work

3D Dense Captioning. Recently, the community has witnessed significant progress in 3D dense captioning [4, 7, 11, 12, 25, 46, 51]. There are mainly two paradigms in previous research: "detect-then-describe" [4, 7, 12, 25, 46, 51] and "set-to-set" [11]. The "detect-then-describe" paradigm first utilizes a detector to generate proposals and then employs a generator to generate captions. For example, Scan2Cap [12] utilizes a VoteNet [36] to localize the objects in the scene, a graph-based relation module to model object relations and a decoder to generate sentences. [7,13] delve deeper to demonstrate the mutually reinforcing effect of dense captioning and visual grounding tasks. Another approach to address the problem is the "set-to-set" paradigm, like Vote2Cap-DETR [10] and its subsequent work [11]. These methods treat the 3D dense captioning as a set-to-set problem and utilize the one-stage architecture to address it. Additionally, several works [9,19,49,56] focus on large-scale pretraining by multitask settings to solve the 3D dense captioning task. However, these methods are mainly focused on indoor scenarios and are difficult to adapt directly to outdoor scenes. In contrast, our proposed *TOD*^{3}*Cap* network is aimed at outdoor 3D dense captioning.

3D Captioning Datasets. Obtaining 3D language descriptions that are both object-centric and context-aware is a difficult job. Most commonly used datasets for 3D dense captioning are ScanRefer [6] and ReferIt3D (Nr3D) [1], based on the richly-annotated 3D indoor dataset - Scannet [15]. Notably, although recent developments like Objaverse [16, 17] have attempted large-scale object captioning for 3D-language alignment, they lack scene context information. Recently proposed indoor scene datasets like SceneVerse [24], SceneFun3D [18], and Multi3DRefer [53] focus on large-scale scene-graph captioning, object part-level captioning, and multi-object relationship captioning, respectively. However, existing datasets are mostly based on indoor scenes, which fail to cover unique scientific challenges of outdoor scenes as shown in Fig. 1. nuCaption [49] and Rank2Tell [39] are designed for outdoor scenes, but they focus only on event-centric scene captioning instead of dense captioning. By contrast, our proposed *TOD*^{3}*Cap* dataset provides dense object-centric language descriptions in outdoor scenes. We show the statistical comparison of our dataset with existing 3D captioning datasets in Tab. 1, highlighting its unique value.

BEV-based 3D Perception. In recent years, there has been a rapid development and an increasing interest in BEV-based 3D perception techniques [22, 27, 28, 40], because BEV representation has proven to be highly beneficial for outdoor perception tasks such as 3D object detection and tracking. The Lift-Splat-Shoot [34] and its subsequent research [20, 27] project image features into BEV pillar using predicted depth probabilities. BEVFormer [28] utilizes a spatial cross attention to aggregate 2D image features into the BEV space and employs a temporal self attention to fuse temporal feature to model object motion. BEV-Fusion [30] combines point cloud features from LiDAR and image features to enhance the geometric information in the BEV space. Inspired by them, our method fuses features from LiDAR and multi-view images and utilizes temporal fusion for obtaining richer contextual information and modeling object motion, which helps to address the challenges of outdoor dense captioning.

3 *TOD^{3}Cap* Dataset

To facilitate research on outdoor 3D dense captioning task, we introduce *TOD^{3}Cap*, a million-scale multi-modal dataset that extends the nuScenes [3] with dense captioning annotations. We introduce the data collection pipeline in Sec. 3.1, and show the overall statistics of our proposed dataset in Sec. 3.2.

3.1 Data Collection

In this section, we introduce the data collection pipeline of the proposed dataset. We leverage a popular and large outdoor dataset nuScenes [3] encompassing 850 scenes for 3.4k frames. Each frame comprises 6 images taken from 6 cameras and point clouds from one LiDAR. The original dataset provides 3D bounding box annotations of 23 classes. We extend it to 3D dense captioning by annotating the appearance, motion, environment and relationship for all of the objects.

Collection Principle. When describing an object in outdoor scenes, humans consider a series of questions [14]: “What is it and what does it look like?”, “What is it doing?”, “Where is it?”, “What is around it?”, which we refer to as their appearance, motion, environment and relationship, respectively.

Appearance: The ability to describe what an object looks like is a hallmark of human intelligence. To answer the question, human annotators should recognize both the *category* of the object and its *visual attribute* (color, material, etc). For example, there is a person wearing blue shirts and black jeans.

Motion: Different from the static indoor scenes, outdoor scenes are generally dynamic. In our annotation, we focus on the *movement* of the object. For example, a cat is moving away quickly or a dog is approaching slowly.

Environment: For outdoor scenes, an object's relative position in its environment is critical. So we ask the annotators to position the object roughly with its *environment*. For example, there is a car in the parking lot.

Relationship: Humans tend to find a *reference* to describe an object, like “the motorcycle next to the white truck” or “the stroller in the back left of the ego car”. Following [1], we use the following compositional template for relation:

$$\text{<target-object> <spatial-relation> <anchor-object>}, \tag{1}$$

where the target object represents the object to be described and the anchor object represents the anchor to describe the target.

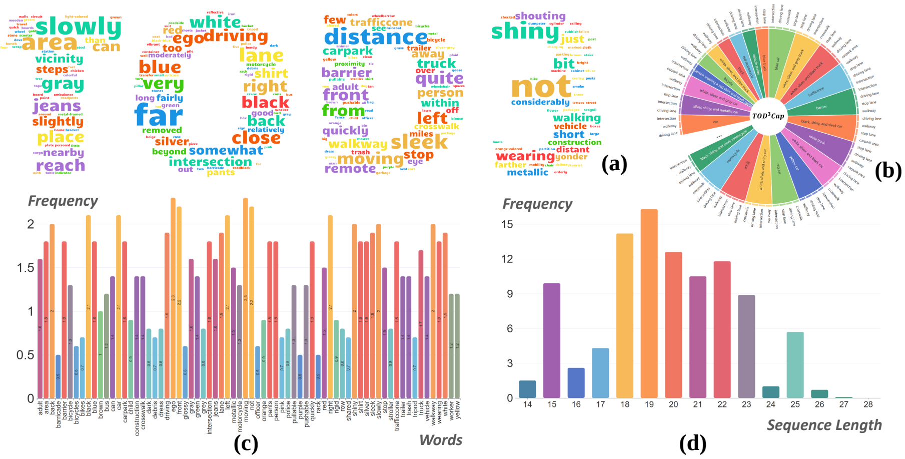


Fig. 2: The statistical properties of *TOD*^{3}*Cap* dataset. (a) The word cloud visualization of *TOD*^{3}*Cap*. (b) The visualization of object-environment relationship. (c) Statistics (in percentage) of the top 70 frequent words. (d) Sentence length Distribution.

Annotation and Verification. With multiple annotation and validation steps, expert annotators make high-quality annotations for the object-level captions.

Notably, considering the significant success of large foundation models in language auto-labeling, we deploy a semi-automatic pipeline. Specifically, we first project the pre-labeled 3D bounding box to 2D. The 2D bounding box is then used to crop the camera image to an image patch that primarily consists of one object, which is then passed as input to a pre-trained captioning model (i.e., LLaMA-Adapter [52]) to generate the captions for each object. Afterwards, we employ human annotators to perform strict correction and refinement of the generated sentences. After labeling all of the four parts of the caption, we utilize GPT-4 [32] to summarize them. Subsequently, the human workers are employed to check the correctness, fluency and readability of the captions. The annotation will not be reserved until three annotators reach an agreement. We elaborate the details of the annotating process in the Appendix.

3.2 Data Statistics

In total, we employ ten expert human annotators to work for about 2000 hours. The total number of language descriptions is about 2.3M, with an average of 67.4 descriptions per frame and 2705.9 descriptions per scene. We showcase the properties of our dataset in Fig. 2. The descriptions cover over 500 types of outdoor objects with a total vocabulary of about 2k words. We find that the appearance of the object is generally more diverse than other attributes. The proportion of vocabulary for the appearance, motion, environment, and relationship is 69.7%, 2.6%, 7.1%, and 20.6%. Moreover, we find that humans use more words to describe the relations of objects. The average words of different parts are 3.7, 2.0, 2.9 and 11.2. Since our captions are very diverse and complex, successful dense captioning involves understanding the object-centric properties, object dynamics, object-object interactions, and object-environment relationships. More details about the dataset are provided in the appendix.

4 *TOD^{3}Cap* Network

To deal with the challenging outdoor 3D dense captioning problem, we propose a new end-to-end method named *TOD^{3}Cap* network. An overview of *TOD^{3}Cap* network architecture is shown in Fig. 3.

Firstly, BEV features are extracted from 3D LiDAR point cloud and 2D multi-view images, followed by a query-based detection head that generates a set of 3D object proposals from the BEV features (see Sec. 4.1). **Secondly**, to capture the relationships between object proposals and scene context, we utilize a Relation Q-Former where the objects interact with other objects and the surrounding environment to get the context-aware features (see Sec. 4.2). **Finally**, with an Adapter [52], the object proposal features are processed to be prompts for the language model to generate dense captions. This formulation does not require a re-training process of the language model and thus we can leverage the commonsense of large foundation models pre-trained on a large corpus of data (see Sec. 4.3).

4.1 BEV-based Detector

Given multi-view camera images $I = \{I_i\}_{i=1}^{N} \in \mathbb{R}^{N \times H_c \times W_c \times 3}$ and LiDAR point clouds $L \in \mathbb{R}^{N_p \times 3}$, we first transform them into the unified BEV features $F_b \in \mathbb{R}^{H_b \times W_b \times C}$ and generate object proposals.

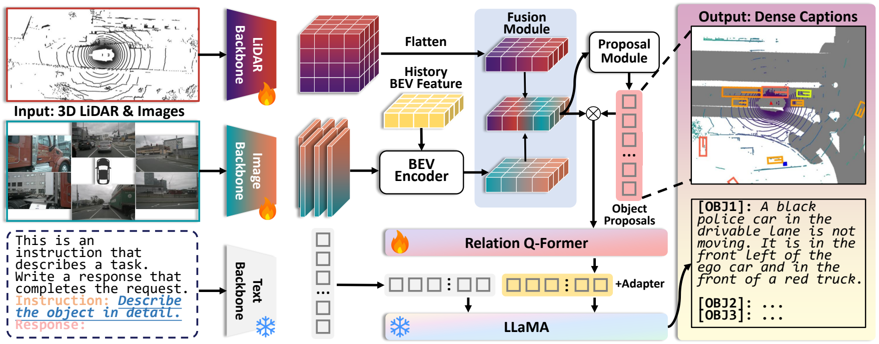
For multi-view images I, following [28], a spatial-temporal BEV encoder is used to lift image features to BEV space and effectively fuse the history BEV features to model dynamics. Specifically, we first extract multi-view image features from I with an image backbone. A set of learnable BEV queries $Q_c \in \mathbb{R}^{H_b \times W_b \times C}$ specific to camera are then updated by interacting with these features via spatial cross-attention layers [28] to capture the spatial information, resulting in F_c:

$$F_c = \text{Spatial-Cross-Attention}(Q_c, \text{Backbone}(I)).$$

To model temporal dependency and capture dynamic features, if the preserved BEV features F_c^p of the previous timestamp exist, the BEV queries Q_c will first interact with F_c^p through temporal self-attention layers, resulting in Q_c':

$$Q_c' = \text{Temporal-Self-Attention}(Q_c, F_c^p).$$

For the initial timestamp, the BEV queries Q_c are duplicated and fed into the temporal self-attention layers. The resulted Q_c' are then taken as the input of the spatial cross-attention layers as a substitute for Q_c.



```
This is an
instruction that
describes a task.
Write a response that
completes the request.
Instruction: Describe
the object in detail.
Response:
```

```
[OBJ1]: A black
police car in the
drivable lane is not
moving. It is in the
front left of the
ego car and in the
front of a red truck.

[OBJ2]: ...
[OBJ3]: ...
```

Fig. 3: Architecture of our proposed *TOD*^{3} *Cap* network. Firstly, BEV features are extracted from 3D LiDAR point cloud and 2D multi-view images, followed by a detection head that generates a set of 3D object proposals from the BEV features. Secondly, to capture the relationship information, we utilize a Relation Q-Former where the objects interact with other objects and the surrounding environment to get the context-aware features. Finally, with an Adapter, the features are processed to be prompts for the language model to generate dense captions. This formulation does not require a re-training process of the language model.

For multi-modal input, we utilize BEVFusion [30] to obtain unified BEV representation. Specifically, a LiDAR backbone is first employed to extract voxelized LiDAR features. Then, the features are flattened along the height dimension, leading to the BEV features $F_l \in \mathbb{R}^{H_b \times W_b \times C}$. Finally, the BEV features of the two different modalities are fused together with a convolutional fusion module to acquire the unified BEV features F_b.

Subsequently, we exploit a query-based object proposal generation module that takes the BEV features F_b as input to generate the object box proposals $\hat{B} = \{\hat{B}_i\}_{i=1}^{K} \in \mathbb{R}^{K \times D}$, where K is the preset number of object queries and D corresponds to the dimension of proposal feature. The process of proposal generation aligns with that in traditional detection head like DETR [5].

4.2 Relation Q-Former

After obtaining the BEV features F_b and object proposals $\hat{B}$, a relation query transformer (Relation Q-Former) is designed to extract context-aware features for each object. Specifically, we first create object queries by encoding the object proposals $\hat{B}$ with a learnable MLP, resulting in object features with the same feature dimension as F_b. These features are then concatenated and fed into the Relation Q-Former, which comprises several self-attention layers for feature interaction. As shown later, this module improves performance significantly.

$$Q_B = \text{Relation Q-Former}(\text{MLP}(\hat{B}), F_b).$$

The resulting object queries Q_B are taken as input to a captioning decoder for natural language generation, which will be elaborated in the next section.

4.3 Captioning Decoder

Inspired by the recent advancements of LLMs in contextual reasoning, we employ a frozen LLM as our language generator, which takes object queries Q_B as input and output descriptions for each object. To ensure the dimension consistency between Q_B and the hidden layers of the LLM, we first use an MLP to transform the dimension of Q_B, resulting in Q'_B. We further employ an Adapter [52] to align the object proposal representation with the feature space of the pre-trained language model, which bridges the modality gap. The adapted object features serve as prompts $\mathcal{V}$ for the LLM to generate corresponding captions.

$$\begin{aligned} Q'_B = \mathrm{MLP}(Q_B), \quad \mathcal{V} = \mathrm{Adapter}(Q'_B), \\ \hat{\mathcal{C}} = & \mathrm{LLM}(\mathcal{T}, \mathcal{V}), \end{aligned}$$

where $\mathcal{T}$ is the system text prompt (as shown in the left-bottom corner of Fig. 3) and $\hat{\mathcal{C}} = \{\hat{w}_i\}_{i=1}^{M}$ is the resulting caption, which consists of M words.

During training process, we take the standard cross-entropy loss as the captioning loss $\mathcal{L}_{\text{cap}}$ and train the model in the *teacher-forcing*[1] manner:

$$\mathcal{L}_{\text{cap}} = \sum_{i=1}^{M} \mathcal{L}_{\text{cap}}(w_i) = -\sum_{i=1}^{M} \log \hat{p}\left(w_i \mid w_{[1:i-1]}, \mathcal{T}, \mathcal{V}, \theta_{\text{LLM}}\right), \tag{2}$$

where $\mathcal{C} = \{w_i\}_{i=1}^{M}$ is the ground truth caption, θ_{LLM} represent the weights of the LLM and $\hat{p}$ is the predicted probability. Note that θ_{LLM} are frozen to reduce the computation cost and mitigate the catastrophic forgetting problem of LLM.

Moreover, considering the memory burden and optimization difficulty when generating hundreds of sentences during training, we do not feed all the object queries into the captioning decoder at once. Instead, we filter the queries by a 3D hungarian assigner [48] to get those matched with the ground truth and then randomly sample a subset during training. During inference, we apply non-maximum suppression (NMS) to suppress overlapping proposals.

4.4 Loss Function

We utilize L_1 loss as $\mathcal{L}_{\text{obj}}$ to supervise 3D bounding box regression for object proposal generation and use $\mathcal{L}_{\text{cap}}$ for captioning. Then the overall loss for dense captioning is calculated as the weighted combination:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{obj}} + \beta \mathcal{L}_{\text{cap}}. \tag{3}$$

where hyper-parameters α and β are set to $\alpha = 10$ and $\beta = 1$ in our experiments.

[1] It means using ground truth words as the conditioning during training, which differs from the auto-regressive testing setting that uses predicted words for conditioning.

5 Experiments

We conduct a comprehensive evaluation of adapted state-of-the-art baseline methods and ours on *TOD*^{3}*Cap* dataset. In Sec. 5.1, we describe the evaluation metrics, the implementation details of our model and adapted baselines. In Sec. 5.2, we compare adapted indoor baselines with our proposed method on the introduced dataset. Finally in Sec. 5.3, we conduct a comprehensive ablation study to validate the effectiveness of *TOD*^{3}*Cap* network design.

5.1 Experimental Setup

Dataset and Metrics. We inherit the official nuScenes split setting for *TOD*^{3}*Cap*, where the train/val scenes are 700 and 150, respectively. The reported results are calculated on the val split for all following experiments. The $m@kIoU$ metric [12] is leveraged for the evaluation of the 3D outdoor dense captioning task. Specifically, we denote each ground truth box-caption pair as $(B_i, \mathcal{C}_i)$, where B_i and $\mathcal{C}_i$ are the bounding box label and the ground truth caption for the i-th object. The predicted box-caption pair matched with the ground truth is denoted as $(\hat{B}_i, \hat{\mathcal{C}}_i)$. For all $(B_i, \mathcal{C}_i)$ and $(\hat{B}_i, \hat{\mathcal{C}}_i)$, the m@kIoU is defined as:

$$m@kIoU = \frac{1}{N_{\text{gt}}} \sum_{i=1}^{N_{\text{gt}}} m\left(\hat{\mathcal{C}}_i, \mathcal{C}_i\right) \cdot \mathbb{I}\left\{\text{IoU}\left(\hat{B}_i, B_i\right) \geq k\right\}, \tag{4}$$

where N_{gt} is the number of the ground truth objects and m represents the standard image captioning metrics, including CIDEr [44], BLEU [33], METEOR [2], Rouge [29], abbreviated as C, B, M, R, respectively.

Table 2: Quantitative results on *TOD*^{3}*Cap* dataset. The "*" represents that we replace the scene encoder with BEV encoder for adaptation. All of the methods are trained to full convergence on the *TOD*^{3}*Cap* dataset for fair comparison. Our *TOD*^{3}*Cap* network outperforms other methods with a clear margin, using various inputs.

Method	Input	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5
TOD^3*Cap* (Ours)	2D	96.2	45.0	34.2	67.4	94.1	47.6	33.3	65.4
Scan2Cap* [12]	3D	50.6	34.3	25.2	57.9	43.3	31.3	22.8	50.8
Vote2Cap-DETR* [11]	3D	72.8	41.6	29.5	60.6	62.6	35.9	27.4	55.8
SpaCap3D [46]	3D	58.8	36.3	25.2	58.1	51.2	32.0	23.5	51.6
TOD^3*Cap* (Ours)	3D	85.3	43.0	29.9	60.5	74.4	39.4	27.2	55.4
Scan2Cap* [12]	2D+3D	60.6	41.5	28.4	58.6	62.5	39.2	26.4	56.5
X-Trans2Cap* [51]	2D+3D	99.8	45.9	35.5	66.8	92.2	43.3	34.7	65.7
Vote2Cap-DETR* [11]	2D+3D	110.1	48.0	44.4	67.8	98.4	46.1	41.3	65.1
TOD^3*Cap* (Ours)	2D+3D	**120.3**	**51.5**	**45.1**	**70.1**	**108.0**	**50.2**	**48.9**	**69.2**

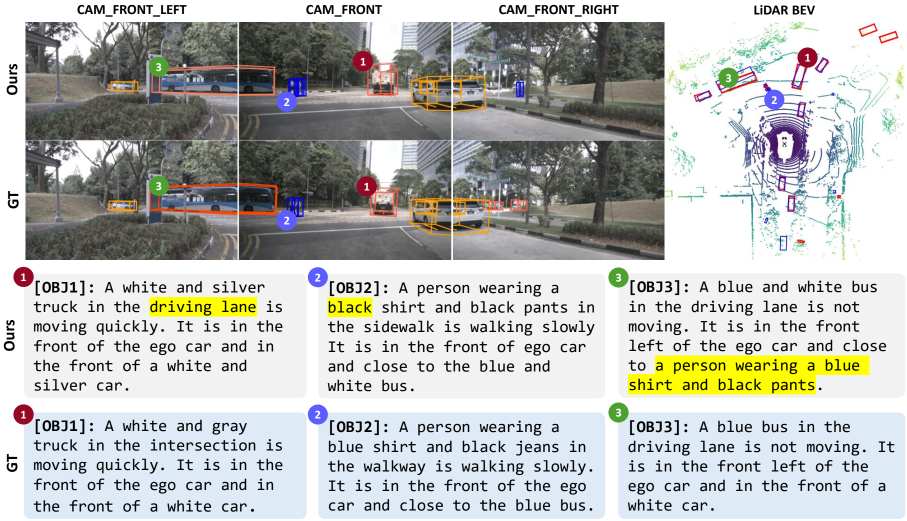
Baselines. From existing methods for 3D dense captioning, we take milestone methods and state-of-the-art methods [11, 12, 46, 51] for benchmarking: (1) Scan2Cap [12] utilizes the VoteNet [36] detector to localize objects in a scene and uses a graph-based relation module to explore object relations. (2) X-Trans2Cap [51] utilizes a teacher-student framework to transfer the rich appearance information from 2D images to 3D scenes. (3) Vote2Cap-DETR [11] adopts a one-stage architecture that applies two parallel prediction heads to decode the scene features into bounding boxes and the corresponding captions. (4) SpaCap3D [46] uses a spatiality-guided encoder and an object-centric decoder to generate spatially-enhanced object captions in 3D scenes.

Adaptation. These methods involve domain-specific design choices for 3D indoor scenes. However, directly applying them to outdoor scenes leads to suboptimal performance. A major challenge is that their detectors cannot effectively locate outdoor objects because of the varying sparsity of LiDAR point clouds and the limited number of camera viewpoints. For a fair comparison, we adapt these methods to the outdoor setting by (1) replacing their detector with the same one as ours and (2) loading our pre-trained detector weights. In this way, these methods obtain the same localization capabilities as ours. All these methods are then trained on the $\textit{TOD}^3\textit{Cap}$ dataset until convergence. In Tab.2, adapted baseline methods are marked with *. The comparison between baseline methods before and after adaptation is provided in the appendix, showing a substantial upgrade.

Protocol. For the proposed $\textit{TOD}^3\textit{Cap}$ network, we train the network in three stages to facilitate the optimization process. Firstly, the BEV-based detector is pre-trained on object detection task. We train the detector on the train split of nuScenes with 24 epochs and a learning rate of 2e-4. Then the weights of the BEV-based detector are frozen and the object box proposals are utilized to generate captions. We train this stage with 10 epochs and a learning rate of 2e-4. Finally, the entire model is finetuned with a smaller learning rate of 2e-5 for 10 epochs. We employ AdamW [31] with a weight decay of 1e-2 as the optimizer. The pre-trained LLaMA-7B [52] is taken as the LLM in our captioning decoder.

5.2 Comparing with State-of-the-art Methods

Quantitative Results. We show results separately for different input modalities, including (1) multi-view RGB images (denoted as 2D), (2) LiDAR point clouds (denoted as 3D), and (3) both images and point clouds (denoted as 2D+3D). The quantitative results are shown in Table. 2, demonstrating:
(1) $\boldsymbol{\textit{TOD}^3\textit{Cap}}$ network outperforms prior arts. Specifically, when taking 2D images and 3D point clouds (2D+3D) as input, the proposed $\textit{TOD}^3\textit{Cap}$ network outperforms Vote2Cap-DETR by 10.2 (9.26%) on C@0.25 and 9.6 (9.76%) on C@0.5. When taking only point clouds as input, our $\textit{TOD}^3\textit{Cap}$ network



Ours

1 [OBJ1]: A white and silver truck in the driving lane is moving quickly. It is in the front of the ego car and in the front of a white and silver car.

2 [OBJ2]: A person wearing a black shirt and black pants in the sidewalk is walking slowly It is in the front of ego car and close to the blue and white bus.

3 [OBJ3]: A blue and white bus in the driving lane is not moving. It is in the front left of the ego car and close to a person wearing a blue shirt and black pants.

GT

1 [OBJ1]: A white and gray truck in the intersection is moving quickly. It is in the front of the ego car and in the front of a white car.

2 [OBJ2]: A person wearing a blue shirt and black jeans in the walkway is walking slowly. It is in the front of the ego car and close to the blue bus.

3 [OBJ3]: A blue bus in the driving lane is not moving. It is in the front left of the ego car and in the front of a white car.

Fig. 4: Qualitative results for our proposed *TOD*^{3}*Cap* network. In the top left, we show our predicted bounding boxes and corresponding captions in the first row and ground truth in the second row. In the top right, we show our predicted bounding boxes in blue and the ground truth bounding boxes in red. In the bottom, we mark the wrong descriptions in yellow. The *TOD*^{3}*Cap* network produces impressive results except for a few mistakes.

achieves 12.5 (17.17%) and 11.8 (18.85%) improvement over Vote2Cap-DETR. These results indicates the effectiveness of the proposed *TOD*^{3}*Cap* network.

(2) The multi-modal input improves captioning performance. The performance of *TOD*^{3}*Cap* network with multi-modal input outperforms that with the camera-only or LiDAR-only input, indicating that the information from the camera and LiDAR are complementary to each other. For LiDAR-only results, the sparsity of LiDAR point clouds makes it challenging to capture the visual attributes and textures of objects. For camera-only results, it is difficult to capture distance information of objects solely based on images, which results in the poor captioning related to motion and environment. We provide qualitative results in the appendix to support the complementary nature of multi-modal inputs.

Qualitative Analysis. We show some qualitative results in Fig. 4, including the detection results and corresponding descriptions. We can see *TOD*^{3}*Cap* network accurately localizes most objects and provides sound descriptions, except for a few mistakes in small and remote objects, calling for future algorithmic development in this novel and important outdoor 3D dense captioning problem.

Table 3: Comparison of different relation modeling modules. The Relation Q-Former outperforms other relation modeling modules for its good context awareness.

Relation Module	C@0.25	B-4@0.25	C@0.5	B-4@0.5
Relational Graph	88.8	41.8	82.7	38.4
Transformer Decoder	94.9	44.3	90.0	41.7
Relation Q-Former (Ours)	**96.2**	**45.0**	**94.1**	**47.6**

5.3 Ablation Study

We conduct a comprehensive ablation study to investigate the effectiveness of the *TOD*^{3}*Cap* network design. Unless specified, we utilize the 2D images as input.

Effectiveness of Relation Q-Former. The relation modeling module is crucial for 3D dense captioning to model the intricate interactions between objects [50]. Prior arts focus on modeling the relation between different specific objects with "Graph" [7,12,25] or transformer decoder [4,11,46,54]. In this section, we conduct experiments to compare different relation modules. As shown in Tab. 3, the Relation Q-Former outperforms other relation modules, which is attributed to the good context awareness of the Relation Q-Former and the fact that *relational graph* and *transformer decoder* fail to incorporate information from BEV queries.

Comparisons with Different Language Decoders. The large foundation models have been proved effective for their generalization and commonsense understanding abilities. These abilities help *TOD*^{3}*Cap* network to well resolve long-tailed cases. To investigate the impact of the LLM decoder on *TOD*^{3}*Cap* network, we conduct experiments on different language decoders utilized in former dense captioning methods, including S&T [45] and GPT2 [38], apart from LLaMA in our original setting. The results in Tab. 4 shows that the model with LLaMA achieves higher performance than other language decoders. This demonstrates that our network design can fully unleash the the superior language generation capabilities of large language models. Note there exist domain gaps as the original LLaMA-adapter is meant to process visual features from RGB images and language prompts, while our design processes BEV queries from multi-modal inputs and object prompts.

Impact of Different Training Strategies. A critical issue in our network design is the alignment between object proposal prompts with language prompts. Thus it is difficult to directly optimize the entire network from the scratch. We utilize the training strategy that divides the optimization process into several stages. We take three steps to optimize the network, (1) we pre-train the BEV-based detector on object detection task; (2) we freeze the detector weights

Table 4: Comparison of different language decoders. The LLaMA achieves the best performance, demonstrating our network design can fully unleash the superior language generation capabilities of large language models, despite domain gaps.

Decoder	Adapter	C@0.25	B-4@0.25	C@0.5	B-4@0.5
S&T	Yes	81.2	32.0	78.6	29.8
GPT2	Yes	89.4	41.2	85.6	38.6
LLaMA (Ours)	Yes	**96.2**	**45.0**	**94.1**	**47.6**

Table 5: Comparison of different training strategies. We can see that the pretraining of detector and captioner could benefit the 3D dense captioning in outdoor scenes.

Detector	Captioner	Entire Model	C@0.25	B-4@0.25	C@0.5	B-4@0.5
	✓	✓	74.2	39.2	69.5	37.4
✓		✓	87.4	41.9	85.3	39.1
✓	✓	✓	**96.2**	**45.0**	**94.1**	**47.6**

and train the caption generation module; (3) the entire model is finetuned with a smaller learning rate. In this section, we investigate the effectiveness of the strategy we use, as shown in Tab. 5. We can see that the removal of each training phase leads to a significant performance decrease. For example, the results decrease by 8.8 on C@0.25 and by 8.8 on C@0.5 without the captioner pre-training stage. This indicates the necessities of all the pre-training.

6 Conclusions

In this study, we present the task of generating dense captions for outdoor 3D environments, utilizing both LiDAR-generated point clouds and RGB images from a panoramic camera rig. To support this task, we introduce the *TOD*^{3}*Cap* dataset, featuring 2.3 million detailed descriptions for over 64,300 outdoor objects across 850 scenes, derived from the nuScenes dataset. Our approach leverages the *TOD*^{3}*Cap* network, which employs a Relation Q-Former to understand the inter-object relationships and their contexts within a scene, and integrates with the LLaMA-Adapter for efficient caption generation without necessitating retraining of the underlying large language model. Through our contributions, we aim to facilitate advancements in outdoor 3D visual language research.

Acknowledgments

We would like to thank Dave Zhenyu Chen at Technical University of Munich for his valuable proofreading and insightful suggestions. We would also like to thank Lijun Zhou and the student volunteers at Li Auto for their efforts in building the *TOD*^{3}*Cap* dataset.

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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 422–440. Springer (2020)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16464–16473 (2022)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
7. 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. arXiv preprint arXiv:2112.01551, 2021.3 (2021)
8. Chen, D.Y., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3d model retrieval. In: Computer graphics forum. vol. 22, pp. 223–232. Wiley Online Library (2003)
9. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. arXiv preprint arXiv:2311.18651 (2023)
10. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11124–11133 (2023)
11. Chen, S., Zhu, H., Li, M., Chen, X., Guo, P., Lei, Y., Yu, G., Li, T., Chen, T.: Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. arXiv preprint arXiv:2309.02999 (2023)
12. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3193–3203 (2021)
13. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18109–18119 (2023)
14. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Can vision-language models think from a first-person perspective? arXiv preprint arXiv:2311.15596 (2023)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)

16. 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. Advances in Neural Information Processing Systems **36** (2024)
17. 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
18. Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
19. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36** (2024)
20. Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
21. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10608–10615. IEEE (2023)
22. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
23. Huang, X., Peng, Y., Yuan, M.: Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. IEEE transactions on cybernetics **50**(3), 1047–1059 (2018)
24. Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. arXiv preprint arXiv:2401.09340 (2024)
25. Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: More: Multi-order relation mining for dense captioning in 3d scenes. In: European Conference on Computer Vision. pp. 528–545. Springer (2022)
26. Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., Liu, J.: Adapt: Action-aware driving caption transformer. arXiv preprint arXiv:2302.00673 (2023)
27. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
28. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. Springer (2022)
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
30. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2774–2781. IEEE (2023)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

32. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
34. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer (2020)
35. Pidathala, P., Franz, D., Waller, J., Kushalnagar, R., Vogler, C.: Live captions in virtual reality (vr). arXiv preprint arXiv:2210.15072 (2022)
36. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
37. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4542–4550 (2024)
38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
39. Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Kochenderfer, M., Choi, C., Dariush, B.: Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7513–7522 (2024)
40. Saha, A., Mendez, O., Russell, C., Bowden, R.: Translating images into maps. In: 2022 International conference on robotics and automation (ICRA) (2022)
41. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
42. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. arXiv preprint arXiv:2312.14150 (2023)
43. Tian, X., Gu, J., Li, B., Liu, Y., Hu, C., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
44. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
45. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
46. Wang, H., Zhang, C., Yu, J., Cai, W.: Spatiality-guided transformer for 3d dense captioning on point clouds. arXiv preprint arXiv:2204.10688 (2022)
47. 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 the IEEE/CVF conference on computer vision and pattern recognition. pp. 6629–6638 (2019)

48. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
49. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. arXiv preprint arXiv:2312.14074 (2023)
50. Yu, T., Lin, X., Wang, S., Sheng, W., Huang, Q., Yu, J.: A comprehensive survey of 3d dense captioning: Localizing and describing objects in 3d scenes. IEEE Transactions on Circuits and Systems for Video Technology (2023)
51. 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8563–8573 (2022)
52. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
53. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3d objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15225–15236 (2023)
54. Zhong, Y., Xu, L., Luo, J., Ma, L.: Contextual modeling for 3d dense captioning on point clouds. arXiv preprint arXiv:2210.03925 (2022)
55. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10012–10022 (2020)
56. 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 the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)