TOD^3*Cap*: Towards 3D Dense Captioning in Outdoor Scenes

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

[1]Li Auto [2]AIR, Tsinghua University [3]Beihang University [4]HKUST [5]HKU

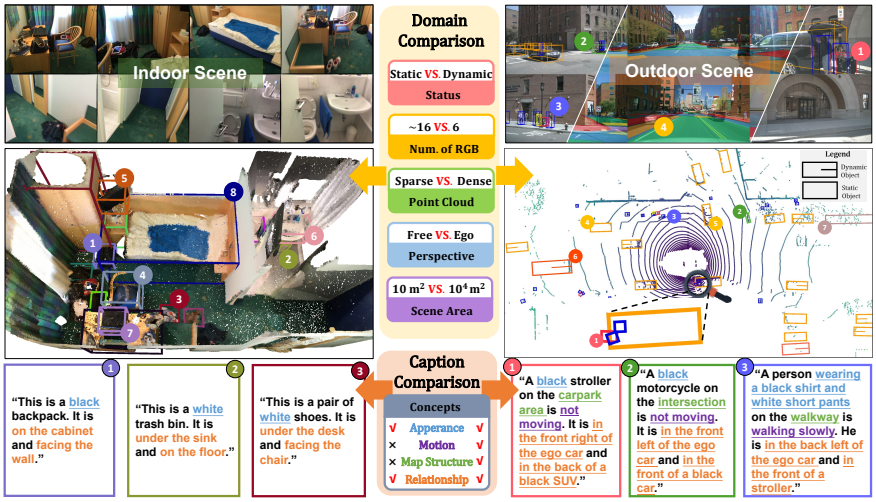

Fig. 1: We introduce the task of 3D dense captioning in outdoor scenes (right). Given point clouds (right middle) and multi-view inputs (right top), we predict box-caption pairs of all objects in the 3D outdoor scene. Due to the domain gap (middle) between indoor and out door scene, including Status, Num. of RGB, Point Cloud, Perspective, and Scene Area, our outdoor 3D dense captioning (right bottom) contains more concepts than indoor scenes (left bottom).

Abstract. 3D dense captioning stands as a cornerstone in comprehensively scene understanding by explicit natural language, it has seen remarkable achievements recently in indoor scenes. 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 sparse visual inputs and dynamics, making it difficult to directly transfer existing methods; 2) the **lack of data** with descriptive 3D-Language pair annotations specifically tailored for outdoor scenes. Hence, we introduce the new task of outdoor 3D dense captioning. As input, we assume a point cloud of a LiDAR swept 3D scene along with

a set of RGB images captured by ego-camera. To address this task, we propose *TOD*^{3}*Cap* network, leveraging the BEV representation to encode sparse outdoor scenes, and then combine Relation Q-Former with LLaMA-Adapter to capture spatial relationships and generate rich concept descriptions in the open-world outdoor environment. We also introduce *TOD*^{3}*Cap* dataset, the first million-scale effort to jointly perform 3D object detection and captioning in outdoor scenes, containing 2.3M descriptions of 64.3k outdoor objects from 850 scenes in nuScenes. Notably, ours *TOD*^{3}*Cap* network can effectively localize and describe 3D objects in outdoor, which outperforms indoor baseline methods by a significant margin (+9.76% CiDEr@0.5IoU).

Keywords: 3D dense captioning · Outdoor scenes · BEV · Dataset

1 Introduction

3D dense captioning [4, 7, 11, 12, 29, 51, 56] aims at detecting and describing each object in a 3D scene with detailed manner. By explicitly expressing its understanding of scenes, 3D dense captioning empowers machine to support a range of 3D vision-language learning tasks, including 3D grounding [21, 23, 44], 3D question answering [2, 19, 36], and 3D-assisted dialogues [22]. Consequently, it opens up avenues for diverse applications such as cross-modal retrieval [8, 27], robotic navigation [25, 47, 52, 60], interactive AR/VR [41], and autonomous driving [30, 48, 49]. So far, 3D dense captioning work has been exclusively conducted on indoor scenes, benefiting from the strength of powerful 3D detectors, captioners, and datasets specifically developed for indoor environments.

However, upon revisiting the different scene as shown in Fig. 1, the significant domain gap (listed below) between indoor and outdoor environments makes it impractical to directly apply indoor 3D dense captioning methods to outdoor domains.

- **Dynamic, not static.** Outdoor scenes are typically dynamic, necessitating the recognition and tracking of objects that undergo temporal changes.
- **Less RGB images available.** The limited of accessible RGB images for outdoor scenes results in a lack of visual cues, which can lead to the deficiency in appearance descriptions in the captioning.
- **Sparse LiDAR point clouds.** The utilization of sparse point clouds collected through LiDAR for outdoor scenes presents significant challenges in shape analysis and modeling, as the sparsity hinders accurate representation of the scene.
- **Ego-centric perspective.** The ego-centric view in outdoor scenes introduces unavoidable object occlusion, which adversely affects the completeness of the captured scene.
- **Larger areas.** The larger area of outdoor scenes poses difficulties in spatial perception and understanding.

Overall, these domain gap factors pose significant obstacles to achieving successful 3D dense captioning in outdoor scenes.

To this end, we introduce a new task of outdoor 3D dense captioning. It takes LiDAR point clouds and ego-centric images as inputs and the expected output is a set of object boxes with captions. We propose a transformer-based network, 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 detection head to generate the object proposals. We also employ a Relation Q-Former to model the object relationships and its contexts. The object features are finally utilized for the language model to generate dense captions. With an Adapter [57], TOD^{3}Cap network does not require a re-training of the language model and thus we can leverage the advanced large foundational models pre-trained on a large corpus of data.

Apart from the limitation of the indoor-outdoor domain gap, our proposed outdoor 3D dense captioning task also suffers from the data hungry [55] issue, i.e., the lack of aligned 3D-language pairs in the outdoor scenes. To facilitate our proposed tasks of outdoor 3D dense captioning, we collect the *TOD*^{3}*Cap* dataset, which provides natural language descriptions for LiDAR point cloud and ego-centric images from nuScenes [43]. In total, we acquire 2.3M descriptions of 63.4k outdoor instances and 1.1M objects. To the best of our knowledge, our *TOD*^{3}*Cap* dataset is the first million-scale effort that combines 3D detection and captioning tasks in outdoor scenes. To summarize, our contributions are as follows:

- We introduce the outdoor 3D dense captioning task to densely detect and describe 3D objects in LiDAR point clouds along with a set of RGB ego-centric images.
- We provide the *TOD*^{3}*Cap* dataset containing 2.3M descriptions of 63.4k instances and 1.1M objects in outdoor scenes and benchmark existing approaches on our proposed *TOD*^{3}*Cap* dataset.
- We show that our method effectively outperforms the baselines transferred from representative indoor scene methods by a significant margain (+**9.76% CiDEr@0.5IoU**).

2 Related Work

3D Dense Captioning Recently, the community has witnessed significant progress in 3D dense captioning [4, 7, 11, 12, 29, 51, 56]. There are mainly two paradigms in previous research: "detect-then-describe" [4, 7, 12, 29, 51, 56] and "set-to-set" [11]. The "detect-then-describe" first utilizes a detector to generate proposals and then employ a generator to generate captions. For example, Scan2Cap [12] utilizes a VoteNet [42] 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

Table 1: Overview of existing 3D captioning datasets. For App., Mot.,Env., Rel. denotes descriptive captioning objective Appearance, Motion, Environment and Relationship, respectively.

Dataset	Domain	Dense Capt.	App.	Mot.	Env.	Rel.	♯Scenes	♯ Frames	♯Annotations
Objaverse [17]	Object	✗	✓	✗	✗	✗	-	N/A	800k
SceneVerse [28]	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 [58]	Indoor	✓	✓	✗	✗	✓	800	N/A	61.9k
nuCaption [54]	**Outdoor**	✗	✗	✓	✓	✓	265	34.1k	420k
Rank2Tell [37]	**Outdoor**	✗	✓	✓	✓	✓	116	5.8k	-
***TOD*^{3}*Cap* (Ours)	**Outdoor**	✓	✓	✓	✓	✓	**850**	**34.1k**	**2.3M**

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, 22, 54, 61] 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 apply directly in outdoor scenes. In contrast, our proposed *TOD*^{3}*Cap* network is aimed at outdoor 3D dense captioning.

3D Captioning Datasets Obtaining aligned 3D language descriptions (object-centric and context-aware) is a significant but difficult task. 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 Objaverse [16, 17] have integrated large-scale object captioning for 3D-language alignment, they lacks scene context information. Recently proposed indoor scene datasets like SceneVerse [28], SceneFun3D [18], and Multi3DRefer [58] focus on large-scale scene-graph captioning, object part-level captioning, and multi-object relationship captioning, respectively. However, existing datasets mostly based on indoor scene, which carry dense 3D information, static layout, limited objects and fixed spatial relationships than outdoor scenes. nuCaption [54] and Rank2Tell [45] focus on outdoor scenes, they focus only on events-centric scene captioning, not dense captioning. Instead, our proposed *TOD*^{3}*Cap* dataset provides dense object-centered descriptive language annotations in outdoor scenes. We show the comparison of our dataset with existing 3D captioning datasets in Tab. 1.

BEV-based 3D Perception In recent years, there has been a rapid development and an increasing interest in BEV-based 3D perception techniques [26, 31, 32, 46], primarily because BEV representation has proven to be highly

beneficial for outdoor perception tasks such as 3D object detection and tracking. The Lift-Splat-Shoot [40] and its subsequent research [24,31] project image features into BEV pillar using predicted depth probabilities. BEVFormer [32] 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 modeling object motion. BEVFusion [34] combines point cloud features from LiDAR and image features to enhance the geometric information in the BEV space. In contrast, our method fuse features from LiDAR and multi-view images and utilize 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 [43] with dense captioning annotation. We introduce the data collection pipeline in Sec. 3.1, further 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 [43] encompassing 850 scenes for 3.4k frames. Each frame comprises 6 images taken from 6 cameras and point clouds from 1 LiDAR. The original dataset provides the annotation for 3D bounding box with 23 classes. We extend it to 3D dense captioning by annotating the appearance, motion, environment and relationship for all of the objects.

Collection Objective. 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 their appearance, activity, environment and relationship, respectively.

Appearance: The ability to describe what an object looks like is deeply embedded in human nature. 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: Considering the structural nature of outdoor scenes, we ask the annotators to position the object implicitly 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

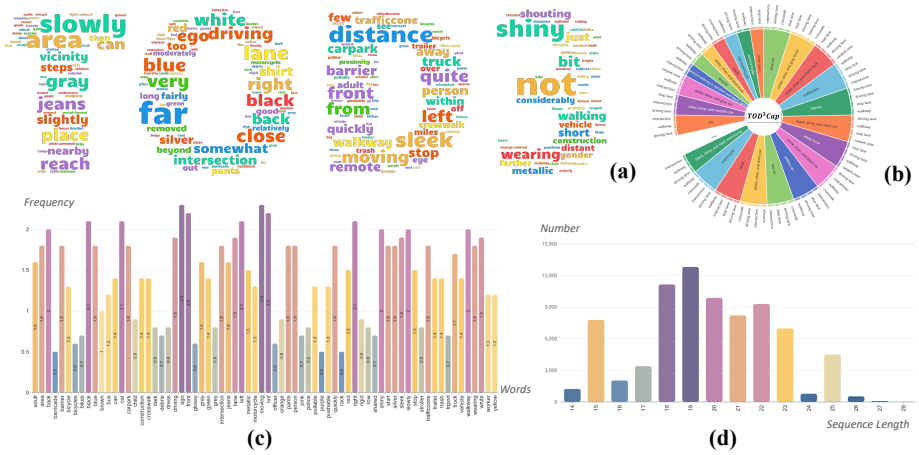

Fig. 2: The properties of *TOD*^{3}*Cap* dataset. (a) The word cloud distribution of *TOD*^{3}*Cap*. Each part represents the different components of sentence annotation and the last column represents the overall annotation. Note that in the third column we convert the Arabic numerals to text. (b) The distribution of Object's appearance and road map. (c) Statistics of the top-70 word frequency.

car". Following [1], we formulate the relation using the following compositional template:

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

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

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

To be mentioned, considering the significant progress of large foundation models in language annotation, 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 expert captioning model (e.g. LLaMA-Adapter [57]) to generate the captions for each object. Afterwards, we employ human annotators to perform a strict correction and refinement of the generated sentences. After labeling all of the four parts of the caption, we utilize AI tools [38] to concatenate them together. 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 Appendix.

3.2 Data Statistics

In general, we employ ten expert human annotators to work for about 2000 hours. The total number of language descriptions is about 2.3k, with an average of 67.4 descriptions per sample 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 vocabularies 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, activity, environment and relationship is 69.7%, 2.6%, 7.1% and 20.6%. Moreover, we find that humans use more words to describe the complex relations of different objects. The average words of different parts are 3.7, 2.0, 2.9 and 11.2. The diversity and complexity requires the model designed for this dataset to be capable of understanding the inter-object properties, object dynamics, object-object interactions and object-environment relationships. More details about the dataset are provided in the supplement.

4 *TOD*^{3}*Cap* Network

To deal with the challenging outdoor 3D dense captioning problem, we propose a new end-to-end baseline 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 detection head that generates a set of 3D object proposals from the BEV features (see Sec. 4.1). Secondly, to capture the relationships of different objects, 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 [57], the features are processed to be the soft object prompt 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 advanced large foundational 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 $F \in \mathbb{R}^{H_b \times W_b \times C}$ and generate object proposals.

For multi-view images I, following [32], 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 [32] to capture the spatial information, resulting in F_c:

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

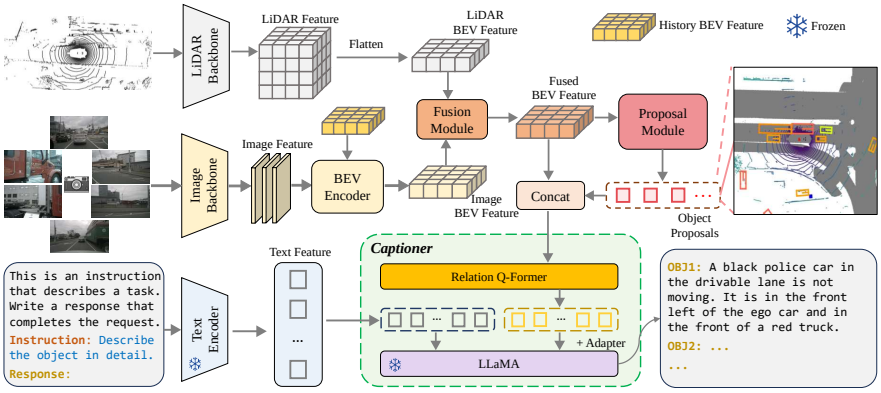

Fig. 3: Architecture of our proposed *TOD*^{3}*Cap* network. Firstly, BEV features are extracted from 3D LiDAR point cloud and 2D multi-view images, followed by a detection head that generates a set of 3D object proposals from the BEV features. Secondly, to capture the relationships of different objects, 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 the soft object prompt for the language model to generate dense captions. This formulation does not require a re-training process of the language model.

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.

For LiDAR input L, 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 proposal 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 object box description. 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.

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

The resulting object queries Q_B are taken as input to a captioning decoder for language sentences 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 [57] to align the object representation with language modeling, which bridges the modality gap. The adapted object features serve as the soft object prompt $\mathcal{V}$ for the LLM to generate corresponding natural language captions.

$$Q'_B = \text{MLP}(Q_B), \quad \mathcal{V} = \text{Adapter}(Q'_B),$$
$$\hat{\mathcal{C}} = \text{LLM}(\mathcal{T}, \mathcal{V}),$$

where $\mathcal{T}$ is the system text prompt 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" 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 [53] 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 proposals 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.

5 Experiments

We conduct a comprehensive evaluation of baseline methods on *TOD*^{3}*Cap*. In Sec. 5.1, we provide the evaluation metrics and implementation details of our model and other baselines. In Sec. 5.2, we compare different existing 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 take 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 corpus of the ground truth caption for the i-th object, respectively. 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 BLEU [39], METEOR [3], Rouge [33] and CIDEr [50], abbreviated as B, M, R, C, respectively.

Baselines. From the existing methods for 3D dense captioning, we take the most popular ones [11,12,56] for benchmarking. Scan2Cap [12] utilizes the Votenet [42] detector to localize objects in a scene and uses a pioneering graph-based relation module to explore object relations. X-Trans2Cap [56] utilizes a teacher-student framework to transfer the rich appearance information from 2D images to 3D scenes. Vote2Cap-DETR [11] adopts a one-stage architecture which applies two parallel prediction heads to decode the scene features into bounding boxes and the corresponding captions. These methods are well-designed framework for 3D indoor scenes. However, it is difficult to directly apply them to outdoor scenes. A major challenge is that their detectors cannot locate the outdoor objects precisely because of the sparsity of LiDAR point clouds and the limitation of

Table 2: Our *TOD*^{3}*Cap* benchmark. The "*" represents that we replace the scene encoder with BEV encoder. 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 remarkable margin.

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
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* [56]	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* network (Ours)	2D+3D	**120.3**	**51.5**	**45.1**	**70.1**	**108.0**	**50.2**	**48.9**	**69.2**

camera views, which is different from indoor scenes. 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 *TOD*^{3}*Cap* dataset until convergence.

For the proposed *TOD*^{3}*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 prediction results 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 [35] with a weight decay of 1e-2 as optimizer. The pre-trained LLaMA-7B [57] is taken as the LLM in our captioning decoder.

5.2 Comparing of Baselines

Quantitative Results. We show results separately for different input modalities, including (1) multi-view camera images, (2) LiDAR point clouds and (3) both images and point clouds.

The quantitative results are shown in Table. 2. It can be observed that:
(1) *TOD*^{3}*Cap* network outperforms prior arts. Specifically, when taking 2D images and 3D point chouds as input, the proposed *TOD*^{3}*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 *TOD*^{3}*Cap* network 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 the final results. 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. We attribute this to the shortcomings

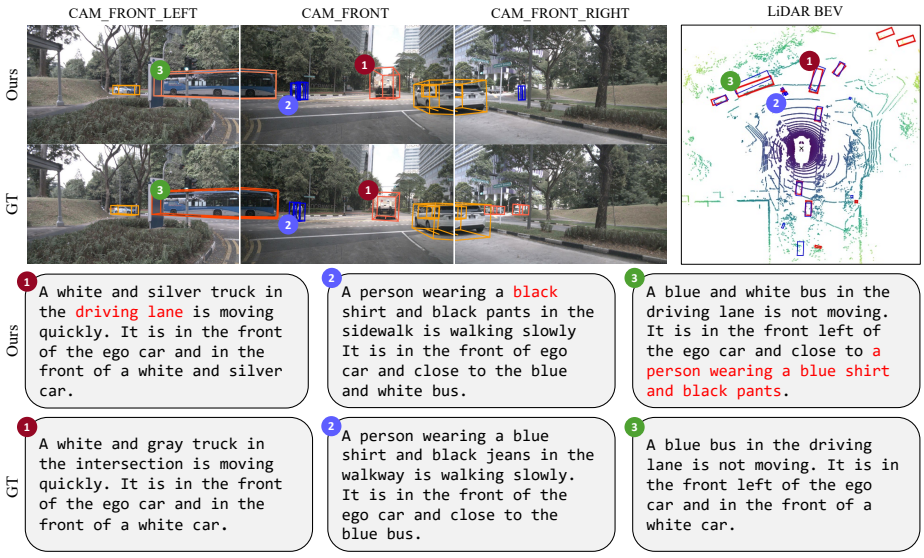

Fig. 4: Qualitative results for our proposed *TOD*^{3}*Cap* network. In the top left, we show our predicted bounding boxes and corresponding captions in the first row and ground truth in the second row. In the top right, we show our predicted bounding boxes in blue and the ground truth bounding boxes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

of single-modal input. For LiDAR-only results, the sparsity of LiDAR point clouds makes it neglect 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 results.

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.

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 module is crucial for 3D dense captioning to model the intricate connections and interactions between objects [55]. Prior arts focus on modeling the relation between different specific objects with "Graph" [7, 12, 29] or transformer decoder [4, 11, 51, 59]. In this section, we conduct experiments to compare different relation modules. As shown in

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

Table 4: Comparison of different language decoders. The LLaMA achieves the best performance, demonstrating the superior language generation capabilities of large language models.

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	Yes	96.2	45.0	94.1	47.6

Tab. 3, the Relation Q-Former outperforms other relation modules, which could be attributed to the good environment awareness of the Relation Q-Former.

Comparisons with Different Language Decoders. The large foundation models have been proved for their generalization and world understanding abilities, which helps *TOD*^{3}*Cap* network to solve the long-tailed problem. To investigate the impact of the LLM decoder on *TOD*^{3}*Cap* network, we conduct experiments on different traditional language decoders utilized in dense captioning literature, including S&T and GPT2, in contrast to 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 the superior language generation capabilities of large language models

Impact of Different Training Strategies. As the proposed network involves the process of aligning the object features with high-dimension language features, it is difficult to directly optimize the entire network from the scratch. Thus, 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 and train the caption generation module; (3) the entire model is finetuned with a small 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 that all the pre-training have positive impacts on the overall performance.

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

Table 6: Comparison of different model scales. The "Tuned Params" means the parameter size that is trainable. The smaller BEV resolution decreases the final performance while reducing the demand of memory.

BEV Resolution	Tuned Params	C@0.25	B-4@0.25	C@0.5	B-4@0.5
TOD^3*Cap*-Tiny	90.5M	90.0	42.2	87.3	41.0
TOD^3*Cap*-Small	115.4M	92.3	45.1	87.5	43.3
TOD^3*Cap*	124.5M	96.2	45.0	94.1	47.6

Efficiency Analysis of *TOD*^{3}*Cap* network. In this study, we investigate the impact of the model scale by varying the BEV resolution, as shown in Tab. 6. Following [32], our default setting is that we use Resnet101 [20] as our backbone, 200*200 BEV resolution, 1600*900 input resolution and multi-scale image feautres. The *TOD*^{3}*Cap*-small means we utilize small BEV resolution (150*150), smaller input resolution (1280*720) and single scale image features. The *TOD*^{3}*Cap*-tiny means we utilize smaller backbone (Resnet 50), smaller BEV resolution (50*50), smaller input resolution (800*450) and single scale features. We can see that the smaller BEV resolution decreases the final performance while reducing the demand of memory. Specifically, the *TOD*^{3}*Cap*-Small decreases 3.9 C@0.25 and 6.6 C@0.5 and reduces 9.1M tuned parameters. The *TOD*^{3}*Cap*-Tiny decreases 6.2 C@0.25 and 6.8 C@0.5 and reduces 34.0M tuned parameters.

6 Conclusions

In this paper, we introduce the task of outdoor 3D dense captioning with point cloud of a LiDAR swept 3D scene and a set of RGB images captured by ego-camera. We collect the *TOD*^{3}*Cap* dataset which contains 2.3M rich descriptions for 64.3K outdoor objects from 850 scenes in nuScenes [43]. We propose *TOD*^{3}*Cap* network for outdoor 3D dense captioning with Relation Q-Former to model the object relationships and its contexts, then combine with LLaMA-Adapter to generate captioning without re-training the large language model. Overall, we hope that our new method and dataset will enable future research in the 3D visual language field in outdoor domain.

Acknowledgement

We express our acknowledge to Dave Zhenyu Chen at Technical University of Munich for his valuable proofreading and insightful suggestions.

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. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19107–19117 (2022)
3. 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)
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., Sumner, R., Tombari, F., Pollefeys, M., Engelmann, F.: Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2024)
19. Etesam, Y., Kochiev, L., Chang, A.X.: 3dvqa: Visual question answering for 3d environments. In: 2022 19th Conference on Robots and Vision (CRV). pp. 233–240. IEEE (2022)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hong, Y., Du, Y., Lin, C., Tenenbaum, J., Gan, C.: 3d concept grounding on neural fields. Advances in Neural Information Processing Systems **35**, 7769–7782 (2022)
22. 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)
23. Hsu, J., Mao, J., Wu, J.: Ns3d: Neuro-symbolic grounding of 3d objects and relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2614–2623 (2023)
24. 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)
25. 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)
26. 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)
27. 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)
28. 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)
29. 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)

30. 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)
31. 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)
32. 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)
33. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
34. 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)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
37. Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J.: Drama: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1043–1052 (2023)
38. 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)
39. 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)
40. 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)
41. Pidathala, P., Franz, D., Waller, J., Kushalnagar, R., Vogler, C.: Live captions in virtual reality (vr). arXiv preprint arXiv:2210.15072 (2022)
42. 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)
43. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. arXiv preprint arXiv:2305.14836 (2023)
44. Roh, J., Desingh, K., Farhadi, A., Fox, D.: Languagerefer: Spatial-language model for 3d visual grounding. In: Conference on Robot Learning. pp. 1046–1056. PMLR (2022)
45. 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)
46. Saha, A., Mendez, O., Russell, C., Bowden, R.: Translating images into maps. In: 2022 International conference on robotics and automation (ICRA) (2022)

47. 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)
48. 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)
49. 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)
50. 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)
51. Wang, H., Zhang, C., Yu, J., Cai, W.: Spatiality-guided transformer for 3d dense captioning on point clouds. arXiv preprint arXiv:2204.10688 (2022)
52. 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)
53. 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)
54. 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)
55. 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)
56. 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)
57. 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)
58. 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)
59. Zhong, Y., Xu, L., Luo, J., Ma, L.: Contextual modeling for 3d dense captioning on point clouds. arXiv preprint arXiv:2210.03925 (2022)
60. 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)
61. 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)