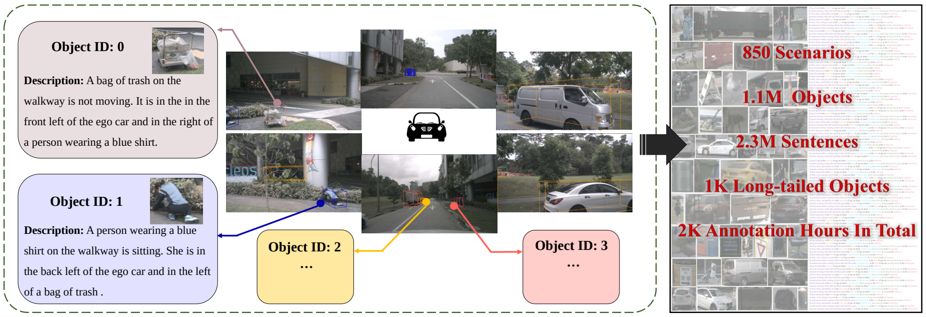
Supplementary Materials of TOD3Cap

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



Description: A bag of trash on the walkway is not moving. It is in the in the front left of the ego car and in the right of a person wearing a blue shirt.

Description: A person wearing a blue shirt on the walkway is sitting. She is in the back left of the ego car and in the left of a bag of trash .

Fig. 1: *TOD*^{3}*Cap*: Towards 3D Dense Captioning in outdoor scenes.

A Annotation Details

In this section, we provide the annotation details of *TOD*^{3}*Cap* dataset. We deploy a web-based annotation system to collect the object descriptions in outdoor scenes, which provides an interactive interface for human workers to annotate with high efficiency. We show the overall annotation pipeline in Fig. 2.

A.1 Pre-labeling

Considering the significant progress of large foundation models in language annotation, we deploy a semi-automatic pipeline for the pre-labeling, where the human annotators would be able to choose and refine the outputs of vision-language models [1, 2]. To make the model focus on a specific object, we crop the whole camera image to an image patch that focuses on the target object and its surroundings. The image patches are then passed as input to the vision-language models to generate descriptions. For the different parts of the caption (Appearance, Motion, Environment, and Relationship), we prompt the model

[⋆] indicates the corresponding author.

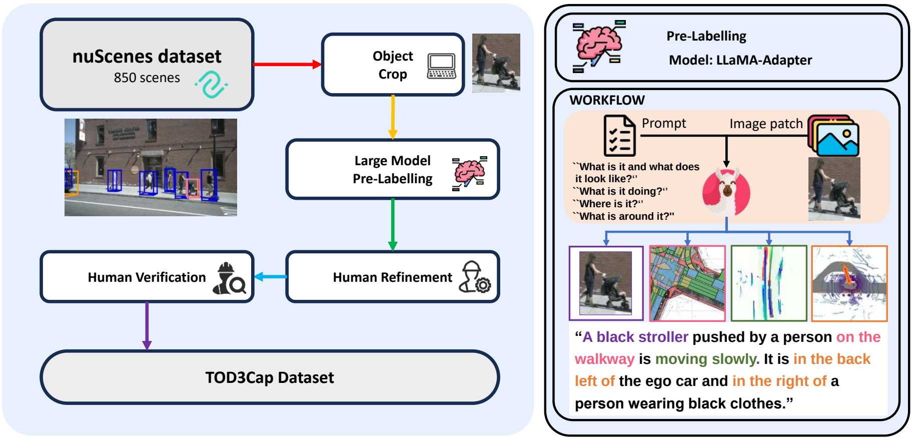

nuScenes dataset
850 scenes
Object Crop
Large Model Pre-Labelling
Human Refinement
Human Verification
TOD3Cap Dataset
Pre-Labelling
Model: LLaMA-Adapter
WORKFLOW
Prompt
Image patch
``What is it and what does it look like?‘’
``What is it doing?‘’
``Where is it?‘’
``What is around it?"
“A black stroller pushed by a person on the walkway is moving slowly. It is in the back left of the ego car and in the right of a person wearing black clothes.”

Fig. 2: The Annotation Process of *TOD*^{3}*Cap*. Generally, there are three steps: pre-labeling, human annotation, and human verification, each of which is important to the effectiveness and efficiency of the whole process.

with different questions: “What is it and what does it look like?”, “What is it doing?”, “Where is it?” and “What is around it?”. The text outputs would be displayed when the human workers annotate the target object.

A.2 Additional Information

We also provide other information of the object to help the annotators analyze the properties of objects. To be mentioned, these additional information is based on the ego-car coordinate system.

Viewing Direction. We define the viewing direction of an object O as the angle between $\overrightarrow{P_{ego}P_O}$ and the orientation of the vehicle R_{ego}, formulated as:

$$\theta = \cos^{-1} \frac{(P_O - P_{ego}) \cdot R_{\text{ego}}}{\| (P_O - P_{ego}) \|_2 \, \| R_{\text{ego}} \|_2}, \tag{1}$$

where P_O and P_{ego} represent the position of target object and ego car, respectively. For example, when the object is in the front of the car, the θ is 0. When when the object is in the back of the car, the θ is 180°.

Distance. The distance of an object is calculated by the euclidean distance between the bounding box centers of the target and ego car, formulated as:

$$d = \|P_O - P'_O\|_2, \tag{2}$$

where P_O and P_{ego} represent the position of target object and ego car.

Speed. We also provide the speed of each object. We first differentiate the trajectory with respect to time to obtain the velocity of the object:

$$V_O = \frac{\delta \text{trajectory}}{\delta t}, \tag{3}$$

where δtrajectory represents the position change of the ego car in δt duration. The velocity V_O is a vector on the global coordinate system, which is challenging for a human annotator to directly comprehend. Thus we only provide the speed of an object, defined as $||V_O||_2$.

A.3 Human Annotation

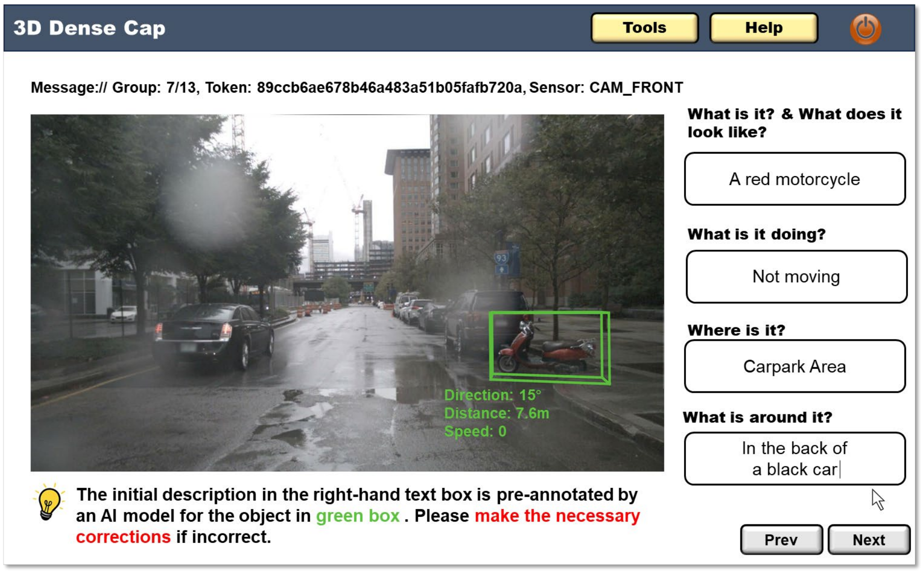
We request the annotators to follow the following instructions:
(1) Describe the appearance of the object such as shape, color, material and so on.
(2) Describe the motion of the object, e.g., "the car is stopped" or "the man is walking".
(3) Describe the environment of the object especially its map structure, e.g., "the truck is in the carpark area" or "the child is in the crosswalk".
(4) Describe its relative position to ego car and other objects in the scene. For instance, "the car is in left of the ego car and in the back of a bus".

During annotating, annotators can refer to the information we provided previously, including the pre-labeling sentences and detailed information, as shown in Fig. 3.

A.4 Verification

After labeling all of the four parts of the caption, we utilize AI tools [1] to concatenate them, referred to as “raw captions”. Subsequently, another three human workers are employed to check the fluency and readability of the raw captions, producing “refined captions”. The annotation will not be reserved until three annotators reach an agreement.

We provide both the raw captions and the refined captions in our dataset to ensure the descriptions are diverse and linguistically rich.



Message:// Group: 7/13, Token: 89ccb6ae678b46a483a51b05fafb720a, Sensor: CAM_FRONT

What is it? & What does it look like?

A red motorcycle

What is it doing?

Not moving

Where is it?

Carpark Area

What is around it?

In the back of a black car

The initial description in the right-hand text box is pre-annotated by an AI model for the object in green box . Please make the necessary corrections if incorrect.

Fig. 3: The web-based UI of Annotation.

B Statistics

In this section, we provide detailed statistics of the proposed *TOD*^{3}*Cap* dataset.

B.1 Number of Instances and Objects

Our *TOD*^{3}*Cap* dataset consists of 1.1M objects and 64K instances. Our annotations and experiments are performed on the object level. For instance, should an object be present in a scene spanning 40 frames, it will have 40 separate object descriptions to account for possible movements to different locations.

B.2 Number of Sentences

The total collected sentences of *TOD*^{3}*Cap* is about 2.3M, with 1.97 sentences per object, 67.4 sentences per frame and 2705.9 sentences per scene. The total vocabulary consists of about 2k words. In Fig. 4, we show the top-200 word frequency, whose frequency represents the logarithmic percentage of each word. We observe that the proportion of the first 100 words accounts for more than 90% of the total, while some words only appear in specific scenes, like ambulance.

B.3 Sentence Length Frequency

We provide the statistics of the length of all caption sentences in Fig. 5. The sentences length ranges from 15 to 28 and the frequency represents the logarithmic percentage of each length. It can be seen that the number of words in most caption sentences is 15 to 25.

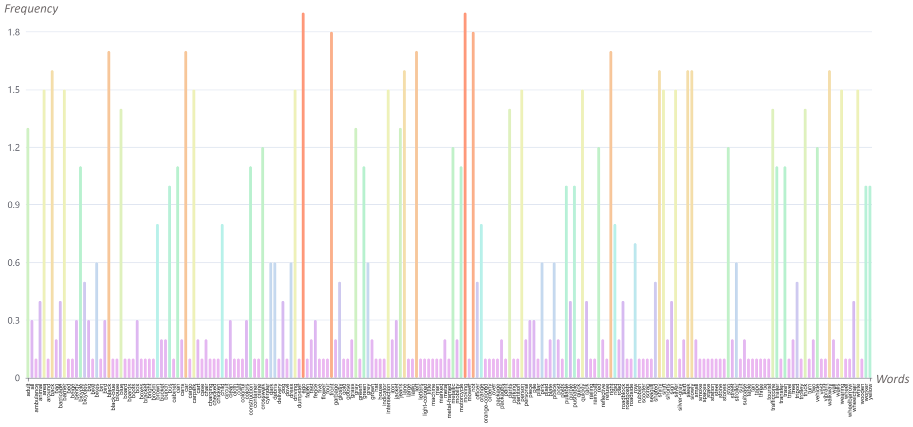


Fig. 4: Statistics of word frequency.

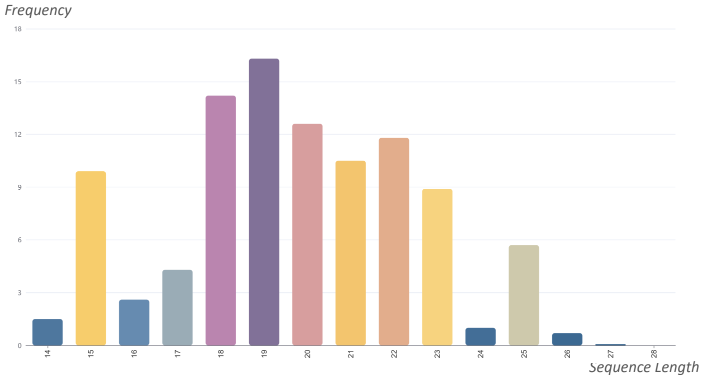
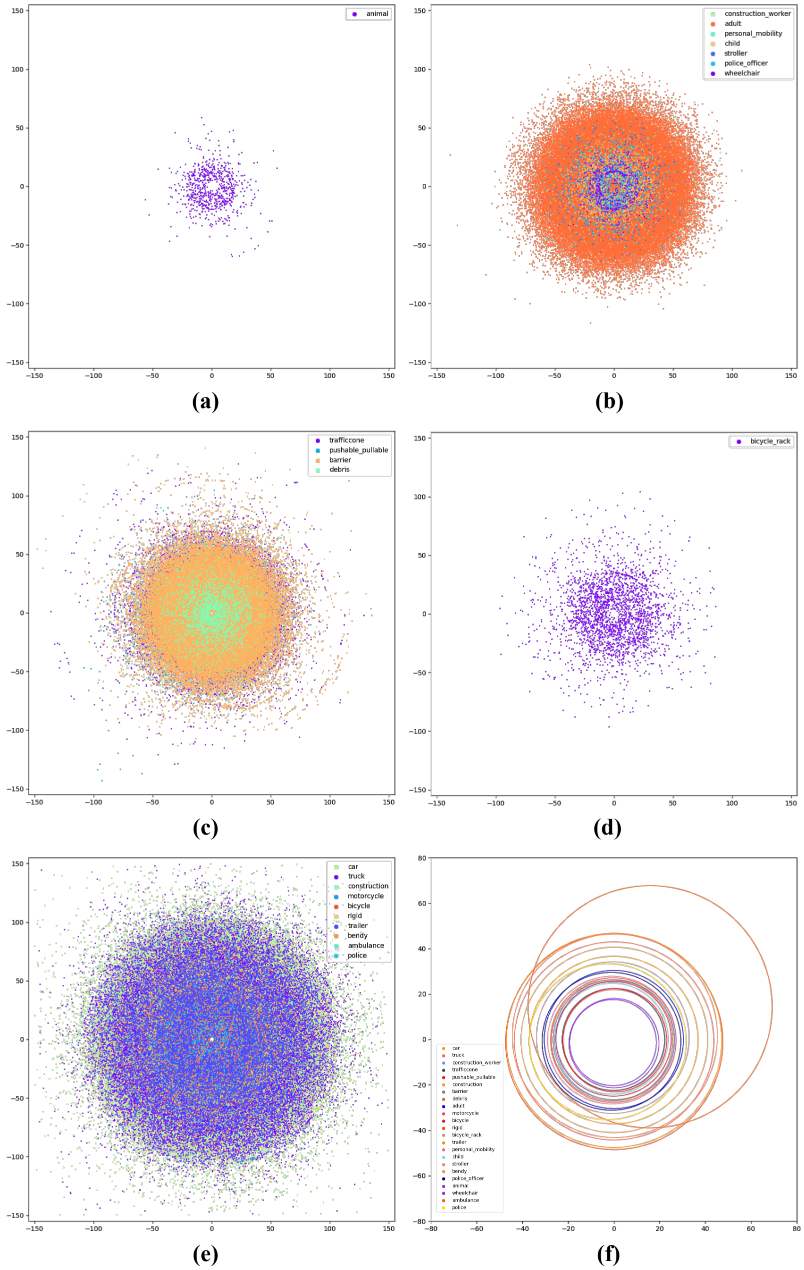


Fig. 5: Statistics of sentence length frequency.

B.4 Distribution of objects' spatial position

Given the object's viewing angle and distance from the ego car, the projection of the distance in the ego car direction and perpendicular to the ego car direction can be calculated. Following this, we visualize the distribution of objects' spatial position separately according to their category, as shown in Fig. 6 where (a) - (e) represents the position distribution of animals, humans, movable objects, static objects, and vehicles relative to the ego car respectively. Besides, the visualization of the mean distribution of each kind of object is shown in Fig. 6 (f), which demonstrates the even distribution of all kinds except the ambulance. (This is because ambulances rarely appear in the dataset.)



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6: Visualization of objects' spatial position. (a) - (e) represents the distribution of objects' position. (f) shows the mean distribution of objects' position.

References

1. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2**, 13 (2023)
2. 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)