Talk to Parallel LiDARs: A Human-LiDAR Interaction Method Based on 3D Visual Grounding

Yuhang Liu[1,2], Boyi Sun[1,2], Yishuo Wang[3], Jing Yang[1], Xingxia Wang[1], and Fei-Yue Wang[1]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing 100049, China
{liuyuhang2021, sunboyi2024, yangjing2020, wangxingxia2022, feiyue.wang}@ia.ac.cn
[2] Zhongke Jingyu (Beijing) Sensing Technology Co., Ltd, Beijing 101407, China
[3] School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
1320211115@bit.edu.cn

Abstract. LiDAR sensors play a crucial role in various applications, especially in autonomous driving. Current research primarily focuses on optimizing perceptual models with point cloud data as input, while the exploration of deeper cognitive intelligence remains relatively limited. To address this challenge, parallel LiDARs have emerged as a novel theoretical framework for the next-generation intelligent LiDAR systems, which tightly integrate physical, digital, and social systems. To endow LiDAR systems with cognitive capabilities, we introduce the 3D visual grounding task into parallel LiDARs and present a novel human-LiDAR interaction paradigm for 3D scene understanding. We propose Talk2LiDAR, a large-scale benchmark dataset for 3D visual grounding in autonomous driving. Additionally, we present a two-stage baseline approach and an efficient one-stage method named BEVGrounding, which significantly improves grounding accuracy by fusing coarse-grained sentences and fine-grained word embeddings with visual features. Our experiments on Talk2Car-3D and Talk2LiDAR datasets demonstrate the superior performance of BEVGrounding, laying a foundation for further research in this domain. We will release all datasets, code, and checkpoints at https://github.com/liuyuhang2021/Talk2LiDAR.

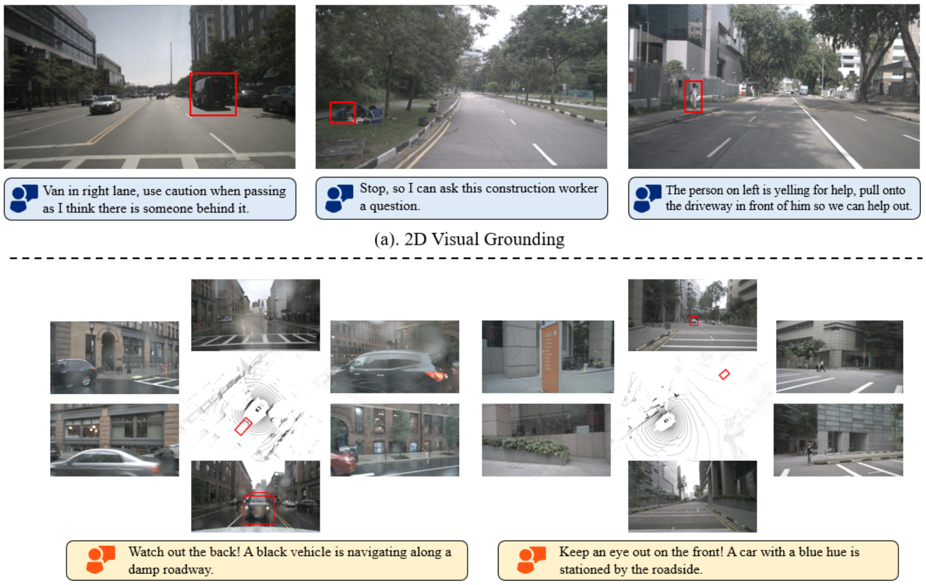
Keywords: Autonomous Driving · Parallel LiDARs · 3D Scene Understanding · 3D Visual Grounding.

1 Introduction

Autonomous driving is experiencing rapid development, with high-performance sensing systems being a critical step towards achieving L4 or L5 autonomous driving [15]. LiDAR sensors play a crucial role in vehicular sensing systems, which can collect point clouds with precise spatial information [29]. However, current LiDAR systems suffer from the separation of software and hardware development, severely limiting the system's intelligence. LiDAR manufacturers prioritize hardware optimization, while autonomous driving companies focus solely on software development. To redefine and build a new generation of intelligent LiDAR systems, we propose the framework of parallel LiDARs as theoretical guidance and have constructed a prototype based on the DAWN experimental platform [17, 21]. It tightly couples physical and virtual spaces, enabling joint optimization of sensing and perception links through virtual-real interaction. Previous work has explored leveraging perceptual information to enhance data utilization efficiency and found that it can significantly improve perceptual accuracy [21]. However, a serious issue remains that these operating modes cannot cognize and reason about the scene, for example, anticipating potentially dangerous areas to avoid accidents. Therefore, this work introduces user instructions to grant cognitive abilities of the LiDAR system and focuses on the 3D visual grounding task in autonomous driving.

The 3D visual grounding task aims to identify the referred objects according to textual descriptions. It takes point clouds and text instructions as input and spits out 3D bounding boxes, which can be regarded as an image-based 2D visual grounding extension. Indoor 3D visual grounding has gained significant attention in recent years due to its promising applications in embodied intelligence [24]. Several high-quality indoor datasets, such as ScanRefer [6] and Sr3d [1], have been released to facilitate systematic research in this field. Following that, various methods have been developed to enhance grounding accuracy and they have already achieved notable progress [14, 25, 38]. However, existing works primarily focus on indoor environments with dense point clouds captured by RGBD sensors. They overlook outdoor scenarios with sparse LiDAR point clouds, which are crucial for applications like autonomous driving. Thus, there is an urgent need to investigate 3D visual grounding in the context of autonomous driving, as illustrated in Fig. 1. This will pave the way for the development of an interactive guidance paradigm for parallel LiDARs.

This work advances the research of 3D visual grounding in driving scenarios, focusing on both dataset creation and method development. Concerning the dataset aspect, Talk2Car [11] emerges as an outstanding dataset for 2D visual grounding in autonomous driving, enabling the association of matching point cloud data. However, its limited size, comprising 11,950 prompts across 7,818 frames, leads to poor generalizability. To address this issue, we propose a novel large-scale 3D visual grounding dataset based on nuScenes [3], named ***Talk2LiDAR***. It consists of 59,207 prompts across 6,419 scenes considering the diversity of viewpoint. To build our dataset cost-effectively, we introduce advanced MLLMs (Multimodal Large Language Models) [16] and LLMs (Large Language Models) [32] for automatic text prompt generation, followed by manual verification. In the method aspect, due to the scarcity of research in driving scenarios, we first propose a baseline approach. It adopts a two-stage processing pipeline based on the prevalent "detect-then-match" strategy. Then we introduce ***BEVGrounding***, a novel single-stage method that significantly boosts 3D vi-


Van in right lane, use caution when passing as I think there is someone behind it.
Stop, so I can ask this construction worker a question.
The person on left is yelling for help, pull onto the driveway in front of him so we can help out.
Watch out the back! A black vehicle is navigating along a damp roadway.
Keep an eye out on the front! A car with a blue hue is stationed by the roadside.

(a). 2D Visual Grounding

(b). 3D Visual Grounding

Fig. 1: Introduction for the visual grounding task in autonomous driving. 2D visual grounding utilizes an image and language prompt as input (Fig.1a), while 3D visual grounding utilizes multi-view images, point clouds, and prompts as input (Fig.1b).

sual grounding accuracy. Specifically, it utilizes a progressive fusion mechanism to achieve fine-grained alignment among text, image, and point cloud data. Extensive experiments on Talk2Car-3D and Talk2LiDAR datasets effectively validate the superior performance of our proposed BEVGrounding method, laying a foundation for future research in this field. The main contribution of this work can be summarized as follows:

- We innovatively introduce the 3D visual grounding task into parallel LiDARs, endowing LiDAR systems with cognitive capabilities through human-LiDAR interaction.
- We propose ***Talk2LiDAR***, a new large-scale benchmark for 3D visual grounding in autonomous driving.
- We develop a two-stage baseline approach and an efficient one-stage method, called ***BEVGrounding***. It utilizes coarse-grained sentence and fine-grained word embeddings to fuse visual and textual features.

2 Related Work

2.1 Parallel LiDARs

Parallel LiDAR emerges as a novel class of intelligent 3D sensors built upon parallel intelligence which is capable of capturing both physical and semantic information of 3D scenes [17, 19]. Parallel intelligence is an innovative theoretical framework proposed by Prof. Wang in 2004 [33]. It integrates cyber, physical, and social spaces for intelligent systems catering to biological humans, robots, and digital humans [35]. Currently, it has garnered widespread attention and found applications in various fields, such as autonomous driving [7, 34], sensing [18, 31], and manufacturing [39]. To facilitate research on parallel sensing, we have established a comprehensive experimental platform, **DAWN**, short for **D**igital **A**rtificial **W**orld for **N**atural. It supports exploration in various subprojects such as parallel LiDARs and parallel light fields. Parallel LiDAR was originally proposed in [17] which consists of three main parts: descriptive, predictive, and prescriptive LiDARs. Descriptive LiDARs focus on constructing digital LiDAR representations; predictive LiDARs emphasize the importance of computational experiments in cyberspace; while prescriptive LiDARs facilitate real-time interaction between the physical and digital LiDAR systems. [21] proposed a software-defined parallel LiDAR model and constructed a hardware prototype in the DAWN platform. It allows for dynamic adjustment of laser beam distribution to optimize the utilization of sensing resources. To provide more comprehensive information, [20] presented a novel HPL-ViT method to enhance the perception accuracy of heterogeneous parallel LiDARs in V2V. Furthermore, [22] discusses accurate modeling of parallel LiDARs in adverse weather. This paper delves into the 3D visual grounding task, aiming to further refine sensing resource allocation through human-LiDAR interaction.

2.2 Visual Grounding in Autonomous Driving

Visual grounding is crucial for autonomous driving, as it enhances the efficiency of human-computer interaction between drivers and vehicles [10]. Current research primarily focuses on 2D object detection and tracking based on language references using images or videos. The Talk2Car dataset, built upon nuScenes, serves as a pioneering benchmark that introduces the visual grounding task within the context of autonomous driving [11]. Significant advancements have been made in enhancing visual grounding accuracy, with notable progress achieved [9,26,30]. [12] extended the Talk2Car dataset to Talk2Car-Destination, enabling language-guided destination prediction, while [41] introduced a novel 3D visual grounding task utilizing a single image as input and established the Mono3DRefer dataset for evaluation. However, these approaches are limited to grounding individual objects, which falls short of capturing the complexities of real-world environments. To address this limitation, [36] proposed the Refer-KITTI dataset, enabling the grounding of multiple objects with a single prompt. Building upon this work, [37] constructed the Nuprompt dataset and introduced PromptTrack, a method that leverages multi-view images for 3D tracking of referred objects. Notably, MSSG [8] stands as the sole approach that incorporates LiDAR point clouds for 3D visual grounding. Nevertheless, it suffers from limited details on the experimental setup and evaluation metrics, hindering replication by subsequent researchers. This work delves into both data and methodological aspects of the 3D visual grounding task in autonomous driving, establishing a solid foundation for future research.

2.3 3D Visual Grounding

3D visual grounding aims to pinpoint objects according to the user's textual descriptions. Compared to complex outdoor environments, indoor settings have received more research attention due to their simpler scene structures. Multiple datasets like ScanRefer [6], Sr3d, and Nr3d [1] have been released to provide robust evaluation benchmarks for indoor 3D visual grounding. Two-stage methods have dominated this landscape, achieving promising results in indoor scenes. These methods typically utilize a pre-trained object detector to generate candidate regions and extract prompt embeddings through a frozen text encoder. The second stage focuses on matching the proposals with textual features to identify the final referred object. [13, 43] employ self-attention and cross-attention mechanisms for improved feature fusion, and [4,40] incorporate additional image information. However, a crucial limitation lies in their inability to recover missed objects during the initial stage. To address this issue, single-stage methods have emerged and demonstrated competitive results on public datasets. 3D-SPS [25] stands as the pioneer single-stage method, utilizing text features to progressively guide key point selection. Similarly, BUTD-DETR [14] leverages a Transformer decoder for referred object prediction. Recent advancements like EDA [38] introduce a text decoupling module, enabling finer-grained alignment by decomposing sentences into semantic components. While these methods hold significant promise for indoor environments, they often overlook the vast potential of outdoor applications, particularly in autonomous driving. This work bridges this gap by proposing BEVGrounding, a novel method specifically designed for 3D visual grounding in autonomous driving.

3 Talk2LiDAR Dataset

In this section, we will first introduce the statistics of our proposed Talk2LiDAR dataset. Then we present its construction process, highlighting the role of foundation models in its development.

3.1 Dataset Statistics

The Talk2LiDAR dataset is the first large-scale dataset specifically designed for LiDAR-based 3D visual grounding in autonomous driving. It's built on the nuScenes dataset, a classic autonomous driving dataset collected in Boston and Singapore. Tab.1 presents a detailed comparison of Talk2LiDAR with other leading visual grounding datasets. Talk2LiDAR establishes itself as the largest benchmark for 3D visual grounding in autonomous driving. It features 59,207 language prompts across 6,419 scenes, with an average of 9.2 prompts per scene. Notably, Talk2LiDAR surpasses prior datasets confined to front-view images. Referred

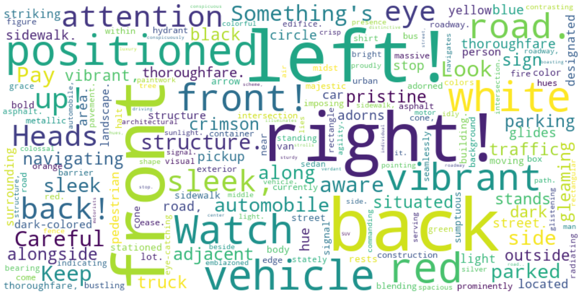


Fig. 2: Word cloud of language prompts in the Talk2LiDAR dataset.

objects are distributed around the ego vehicle, viewable from six distinct image perspectives. It offers a more comprehensive and realistic representation of real-world driving scenarios. Assisted by advanced foundation models [16, 32], Talk2LiDAR boasts a diverse vocabulary within its language prompts, revealing its rich incorporation of information regarding location, category, and color. Fig. 2 depicts the word cloud of language prompts in the Talk2LiDAR dataset.

Table 1: A comprehensive comparison of visual grounding datasets.

Dataset	Publication	Scene	Scene Num.	Prompt Num.	Basic Tasks	Input Modality	Views
ScanRefer [6]	ECCV2020	Indoor	800	51583	3D Det	PC	-
Sr3d [1]	ECCV2020	Indoor	1273	83572	3D Det	PC	-
Nr3d [1]	ECCV2020	Indoor	707	41503	3D Det	PC	-
Multi3DRefer [42]	ICCV2023	Indoor	800	61926	3D Det	PC	-
Talk2Car [11]	EMNLP2019	Outdoor	7818	11959	2D Det	Img	1
Refer-KITTI [36]	CVPR2023	Outdoor	6650	818	2D MOT	Img	1
Mono3DRefer [41]	AAAI2024	Outdoor	2025	41140	3D Det	Img	1
NuPrompt [37]	arXiv2023	Outdoor	34149	35367	3D MOT	Img	6
Talk2Car-3D	**-**	**Outdoor**	**5534**	**8352**	**3D Det**	**PC, Img**	**1**
Talk2LiDAR	**-**	**Outdoor**	**6419**	**59207**	**3D Det**	**PC, Img**	**6**

3.2 Dataset Construction

Prior visual grounding datasets relied heavily on the manual generation of language prompts, leading to significant time consumption and labor costs. Additionally, manual annotation often led to an abundance of repetitive words, hindering vocabulary diversity. To address these limitations, we introduce a novel three-step data annotation pipeline assisted by powerful foundation models. By leveraging the cognitive capabilities of these models, we significantly reduce the workload associated with creating triplet data pairs (text, image, and point cloud). Fig. 3 illustrates the overall construction workflow, followed by a detailed discussion of each step. In total, we hired five interns to construct the Talk2LiDAR dataset in a month.

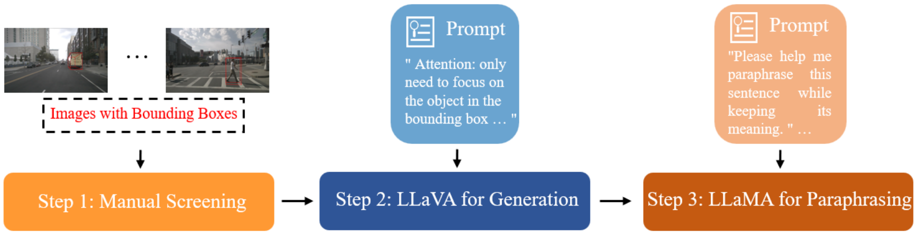

...
Images with Bounding Boxes
Prompt
" Attention: only need to focus on the object in the bounding box ... "
Prompt
"Please help me paraphrase this sentence while keeping its meaning. " ...
Step 1: Manual Screening
Step 2: LLaVA for Generation
Step 3: LLaMA for Paraphrasing

Fig. 3: The Talk2LiDAR dataset construction process.

Step 1: Although Talk2LiDAR focuses on LiDAR-based 3D visual grounding, we describe referred objects using multi-view images, mirroring how drivers perceive their environment. Since the nuScenes dataset contains multiple continuous segments, we randomly sample 20% of all frames to eliminate redundancy. Subsequently, we visualize the annotations on the images and manually filter out ambiguous samples, such as one of a series of consecutively placed traffic cones.

Step 2: Following manual filtering, we utilize the advanced multimodal foundation model LLaVA [16] to automatically generate initial textual descriptions. We design the following prompt for LLaVA:

- *Attention: only need to focus on the object in the bounding box. Please use one or two sentences to describe the object in the red bounding box with greater detail, including its precise location, type, and color characteristics.*

We fed it along with images into LLaVA to generate object descriptions. However, we observed significant hallucinations in LLAVA's outputs, which tend to prioritize prominent objects in the image or macroscopic factors like weather conditions. To address this issue, we manually verified the alignment between text descriptions and referred objects, removing instances with clear hallucinations.

Step 3: We have obtained preliminary image-point cloud-text triplet data pairs by step 2. However, foundation models often employ probabilistic token prediction, leading to repetitive phrasings and a limited vocabulary in the generated prompts. To address this limitation, we utilize LLaMA2 [32] for refinement, aiming to enhance the diversity of descriptions. We provide LLaMA2 with the following prompts:

- *Please help me paraphrase this sentence while keeping its meaning.*
- *Please help me reword a sentence with richer vocabulary but keep its meaning.*
- *Help me reword a sentence, you should describe it in a different way.*

We can obtain descriptions with a richer vocabulary after prompt paraphrasing. Additionally, we integrate viewpoint information into the descriptions to incorporate more comprehensive spatial information. Below are some examples of the original (O) and paraphrased (P) prompts:

- *O: A car is driving down the street at night.*

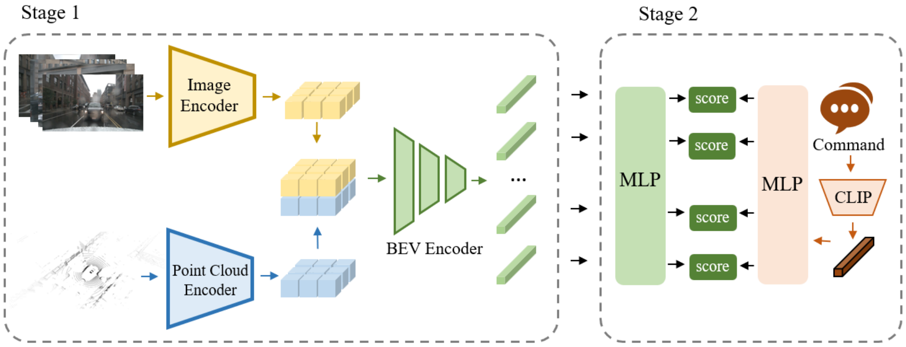


Fig. 4: The architecture of the two-stage baseline method.

- *R: Be aware of the back right! A luxuriant automobile is navigating the boulevard under the cover of darkness.*
- *O: A woman in a black dress is standing in the middle of a street.*
- *R: Look out for the front! A statuesque woman in a sleek black gown stands majestically in the bustling street, her poise and elegance commanding attention.*

4 Methods

4.1 Problem Definition

LiDAR-based 3D visual grounding presents a novel and promising task in autonomous driving. Given a textual instruction T, point cloud data P, and images I, the objective is to precisely localize the referred object B within the 3D scene. T is a textual description consisting of L words, formally represented as $T = \{w_1, ..., w_L\}$. $P = \{p_1, \ldots, p_n\}$ is a single frame of point cloud data. Each point $p_i(i = 1, ..., n)$ is a 4-tuple (x_i, y_i, z_i, i_i), where x_i, y_i, z_i specifies its spatial coordinates and i_i represents its intensity information. Image set $I = \{I_1, \ldots, I_6\}$ comprises a collection of six images captured from different viewpoints surrounding the vehicle. The 3D bounding box B of the referred object is represented as $(x, y, z, l, w, h, \alpha)$. Here, x, y, z denotes the center location of the bounding box, l, w, h represents its dimensions, and α corresponds to its rotation angle.

4.2 Baseline

Given the nascent state of 3D visual grounding research in autonomous driving, we establish a baseline method adhering to the dominant two-stage paradigm. As illustrated in Fig. 4, the baseline method utilizes BEVFusion [23] for 3D object detection in the first stage, generating candidate proposals along with their extracted features. The second stage leverages a pre-trained language encoder to extract sentence-level embeddings from the textual description. We then introduce a lightweight matching network to identify the referred object. Specifically, the object and language features are fed into separate MLPs for feature alignment in the matching network. We then compute the final score using matrix multiplication and select the candidate object with the highest score as the text-referred one:

$$s = E_l(f_{lang}) * E_o(f_{obj}) \tag{1}$$

E_l and E_o represent the MLPs for language and object features, each consisting of two fully connected layers. $*$ denotes the matrix multiplication operation, and s represents the final score for each candidate object. During training, the weights of the language encoder are frozen, while only the weights of the lightweight matching network are updated. The cross-entropy loss function is employed for optimization.

4.3 BEVGrounding

The baseline method adopts a common two-stage approach, while it poses challenges for practical applications due to its training and deployment complexities. To address this issue, we propose BEVGrounding, a novel one-stage method designed for 3D visual grounding in autonomous driving. The overall architecture of BEVGrounding is shown in Fig. 5.

Unimodal Encoder BEVGrounding, being a multimodal method, leverages images, point clouds, and text data as input. The point cloud branch employs a grid-based encoder E_P to address the high computational cost associated with point-based encoders. We utilize sparse convolution to extract voxel features, which are subsequently flattened to obtain f_{bev}^P. The image branch utilizes a Swin-Transformer architecture as the encoder E_I to extract features from multi-view images. These features are then projected onto the BEV (Bird's-Eye View) space, resulting in f_{bev}^I. For the text instruction, we experiment with CLIP [28] as the text encoder E_T to learn both word-level and sentence-level embeddings, facilitating a more fine-grained feature interaction between different modalities:

$$f_{bev}^P = E_P(P); f_{bev}^I = E_I(I); f_{sen}, f_{word} = E_T(T) \tag{2}$$

Trimodal BEV Encoder Following the extraction of individual modality features, we design a trimodal BEV encoder for global feature fusion. Considering the real-time requirements, our BEV encoder leverages a purely CNN architecture, eschewing the use of a Transformer-based approach. We first perform a preliminary fusion of f_{bev}^P and f_{bev}^I to obtain the initial fused representation, denoted as $f_{bev}^{'}$. Subsequently, f_{sen} obtained from the text encoder is broadcasted to match the spatial dimensions of the BEV feature maps and then concatenated with $f_{bev}^{'}$. To alleviate the computational burden, a 1x1 convolution layer is employed to reduce the number of feature channels. Finally, we adopt a classic feature pyramid network to achieve a global coarse-grained fusion, resulting in the representation $f_{bev}^{''}$.

Grounding Decoder Guided by the heatmap scores, we select a subset of proposals from $f_{bev}^{''}$ that exhibit a high likelihood of corresponding to the text-referred objects. These selected proposal features are then fed into the grounding decoder for further fine-grained feature fusion. Our proposed grounding decoder consists of four key blocks: two self-attention (SA) blocks, a spatial cross-attention (SPCA) block, and a semantic cross-attention (SECA) block. Each block consists of a multi-head attention layer and a FFN layer. The SA block effectively captures the global dependencies within the proposal features. The SPCA block fuses proposal features f_{pro} and $f_{bev}^{''}$, while SECA facilitates interaction between f_{pro} and the word-level embeddings f_{word}, providing more fine-grained textual features for candidate objects. The formulation of the attention layer in each module is as follows:

$$SA(f_{pro}) = \sigma(Q(f_{pro})K(f_{pro}^{T}))V(f_{pro}) \tag{3}$$

$$SPCA(f_{pro}, f_{bev}^{''}) = \sigma(Q(f_{pro})K(f_{bev}^{''}))V(f_{bev}^{''}) \tag{4}$$

$$SECA(f_{pro}, f_{word}) = \sigma(Q(f_{pro})K(f_{word}))V(f_{word}) \tag{5}$$

Here, σ is the softmax function. Q, K, and V correspond to the query, key, and value transformation layer, respectively. Finally, we utilize a standard detection head to predict the 3D bounding boxes of the referring objects [2]. During the training phase, the loss function incorporates three key parts:

$$L_{all} = L_{heatmap} + L_{cls} + L_{reg} \tag{6}$$

Consistent with prior work, $L_{heatmap}$ leverages Gaussian focal loss for proposal filtering, L_{cls} uses the focal classification loss, and L_{ref} adopts a L1 loss for bounding box regression. Furthermore, inspired by DETR [5], we incorporate the Hungarian algorithm for bipartite matching during the training process

5 Experiments

5.1 Implementation Details

We conduct experiments on both the Talk2Car-3D and our proposed Talk2LiDAR datasets. Talk2Car-3D is derived from the original Talk2Car through a three-step preprocessing procedure. First, we categorize the referred objects into 10 standard categories according to common 3D object detection conventions. Second, we constrain their positions, retaining only objects in the range requirement of [-54, -54, -5] <[x, y, z] <[54, 54, 3], where [x, y, z] represent the center location. Third, we filter objects based on the number of point clouds inside each

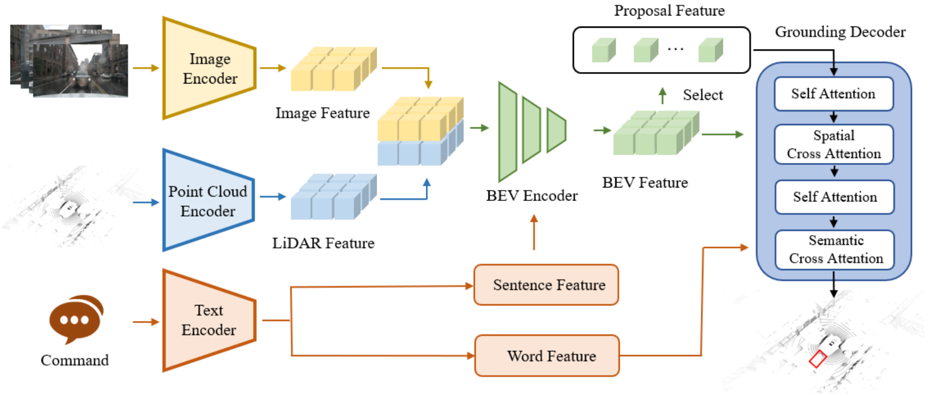


Fig. 5: The architecture of our proposed one-stage BEVGrounding.

object, keeping only those with at least one point. The processed Talk2Car-3D comprises 8,352 data frames. Following the original Talk2Car split, we utilize 7,332 frames for training and the remaining 1,020 frames for testing. To provide additional context for referred objects, we define two attribute labels for each sample: *"unique"* and *"multiple"*. The *"unique"* label indicates that the referred object is the only category-matched target in the frame, while *"multiple"* signifies the presence of several objects in the same category. The training and testing sets of Talk2Car-3D contain 836 and 106 frames with the *"unique"* attribute, respectively. We apply similar processing steps to the Talk2LiDAR dataset. The processed Talk2LiDAR training set comprises 48,813 frames, while the testing set comprises 12,394 frames. Within its training and testing sets, 2,344 and 694 frames are labeled as *"unique"*, respectively.

5.2 Quantitative Analysis

We design multiple two-stage methods to enable more fair evaluations and compare them with our proposed BEVGrounding. **GT-Rand** randomly selects a ground truth box as the prediction result, while **Pred-Rand** randomly selects a predicted proposal as the referred object. **Pred-Best** chooses the candidate with the highest confidence score among all detection boxes. **Baseline** is the method we proposed in Sec. 4.2 and **-L** denotes that only point cloud data is used.

Performance on Talk2Car-3D Tab.2 presents the accuracy of all methods from BEV and 3D perspectives. The overall performance of GT-Rand and Pred-Rand is poor, with accuracies below 10%, effectively demonstrating the difficulty of the 3D visual grounding task. Although Pred-Best shows slight improvement, it still exhibits significant randomness due to the lack of textual features. The baseline significantly enhances the accuracy metrics compared to

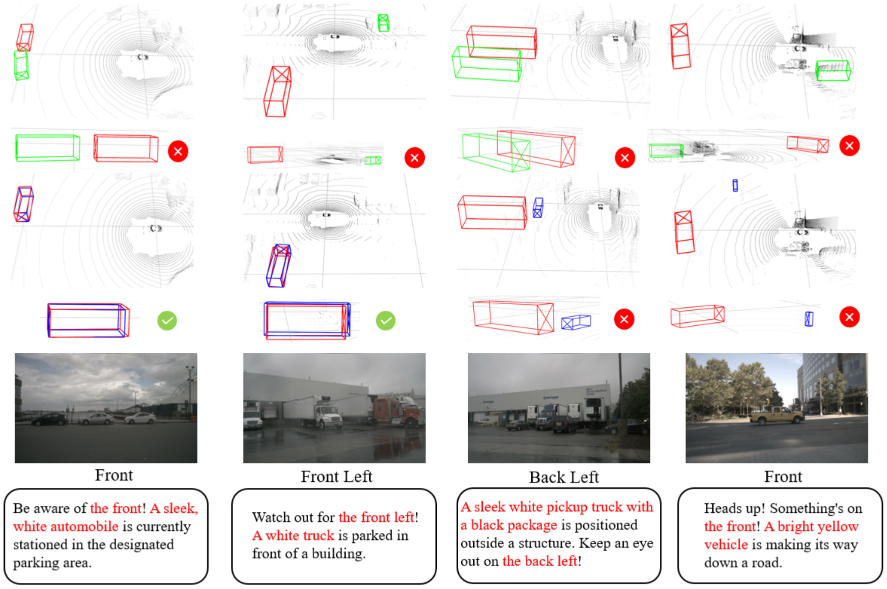
Table 2: Comparasion with other methods on Talk2Car-3D.

	Method	Type	Unique		Multiple		Overall	
			Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)
BEV	GT-Rand	Two-Stage	9.43	9.43	6.78	6.78	7.06	7.06
	Pred-Rand	Two-Stage	6.60	4.72	7.44	5.25	7.35	5.20
	Pred-Best	Two-Stage	0.94	0.94	14.33	12.91	12.94	11.67
	Baseline-L	Two-Stage	2.83	1.89	27.46	26.59	24.90	24.02
	Baseline	Two-Stage	13.21	9.43	30.53	27.35	28.73	25.49
	BEVGrounding-L	**One-Stage**	**32.08 (+18.87)**	**16.98 (+7.55)**	**43.76 (+13.23)**	**28.12 (+0.77)**	**42.55 (+13.82)**	**26.96 (+1.47)**
	BEVGrounding	**One-Stage**	**33.02 (+19.81)**	**17.92 (+8.49)**	**45.30 (+14.77)**	**29.76 (+2.41)**	**44.02 (+15.29)**	**28.53 (+3.04)**
3D	GT-Rand	Two-Stage	9.43	9.43	6.78	6.78	7.06	7.06
	Pred-Rand	Two-Stage	6.60	1.89	7.11	3.94	7.06	3.73
	Pred-Best	Two-Stage	0.94	0.94	14.11	12.69	12.75	11.47
	Baseline-L	Two-Stage	2.83	0.94	25.38	21.77	23.04	19.61
	Baseline	Two-Stage	13.26	7.55	28.01	26.48	26.37	24.51
	BEVGrounding-L	**One-Stage**	**24.53 (+11.27)**	**7.55 (-)**	**36.32 (+8.31)**	**18.38 (-)**	**35.10 (+8.73)**	**17.25 (-)**
	BEVGrounding	**One-Stage**	**25.57 (+12.31)**	**8.49 (-)**	**37.64 (+9.63)**	**20.90 (-)**	**36.37 (+10.00)**	**19.61 (-)**

the aforementioned methods. BEVGrounding outperforms all other methods on almost all metrics, except for the 3D Acc@0.5. Notably, it achieves 44.02% on BEV Acc@0.25, exceeding the second-best method by 15.29%. BEVGrounding-L, which utilizes only point cloud data as input, significantly outperforms the multimodal fusion baseline, demonstrating the superiority of our single-stage architecture. However, it experiences a slight decrease in 3D Acc@0.5. We speculate that it could be attributed to BEVGrounding's emphasis on semantic alignment without fully considering fine-grained geometric features, leading to negative effects at higher IoU thresholds. Furthermore, we observe an anomaly where samples labeled as *"unique"*, intuitively simpler, exhibit lower accuracy. We conducted a statistical analysis of the category distribution and found a significant bias between these two attribute labels. Cars are the most common objects in the dataset, yet they don't appear in *"unique"* samples. Additionally, we find that the average distance for *"unique"* samples is 24.8 meters, higher than the 21.6 meters for *"multiple"* samples. These factors pose greater challenges for the former, resulting in decreased model accuracy.

Table 3: Comparasion with other methods on Talk2LiDAR.

	Method	Type	Unique		Multiple		Overall	
			Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)
BEV	GT-Rand	Two-Stage	6.20	6.20	3.79	3.79	3.88	3.88
	Pred-Rand	Two-Stage	5.48	4.47	4.44	3.85	4.49	3.89
	Pred-Best	Two-Stage	2.88	2.59	6.95	6.79	6.72	6.55
	Baseline-L	Two-Stage	4.76	4.32	8.42	8.17	8.21	7.96
	Baseline	Two-Stage	9.22	8.07	10.68	10.32	10.61	10.20
	BEVGrounding-L	**One-Stage**	**14.12 (+4.90)**	**12.97 (+4.90)**	**16.39 (+5.71)**	**11.97 (+1.65)**	**16.27 (+5.66)**	**12.02 (+1.82)**
	BEVGrounding	**One-Stage**	**15.99 (+6.77)**	**14.99 (+6.92)**	**18.19 (+7.51)**	**13.39 (+3.07)**	**18.07 (+7.46)**	**13.48 (+3.28)**
3D	GT-Rand	Two-Stage	6.20	6.20	3.79	3.79	3.88	3.88
	Pred-Rand	Two-Stage	5.48	3.31	4.28	3.35	4.35	3.35
	Pred-Best	Two-Stage	2.88	2.31	6.91	6.48	6.68	6.24
	Baseline-L	Two-Stage	4.76	3.46	8.34	7.94	8.62	7.69
	Baseline	Two-Stage	8.65	7.35	10.05	9.67	9.97	9.54
	BEVGrounding-L	**One-Stage**	**13.83 (+5.18)**	**6.77 (-)**	**16.13 (+6.08)**	**7.68 (-)**	**16.00 (+6.03)**	**7.63 (-)**
	BEVGrounding	**One-Stage**	**15.27 (+6.62)**	**8.65 (+1.30)**	**17.49 (+7.44)**	**8.72 (-)**	**17.36 (+7.39)**	**8.71 (-)**



Be aware of the front! A sleek, white automobile is currently stationed in the designated parking area.

Watch out for the front left! A white truck is parked in front of a building.

A sleek white pickup truck with a black package is positioned outside a structure. Keep an eye out on the back left!

Heads up! Something's on the front! A bright yellow vehicle is making its way down a road.

Fig. 6: Visualization results of the two-stage baseline and one-stage BEVGrounding method. Red, green, and blue boxes denote the ground truth, predicted boxes by the baseline, and predicted boxes by BEVGrounding, respectively.

Performance on Talk2LiDAR Tab.3 illustrates the results of all methods on the Talk2LiDAR dataset. Compared to their performance on Talk2Car-3D, all methods show a significant drop in accuracy. This decline can be attributed to the increased complexity of language prompts and object locations in Talk2LiDAR, indicating that there is substantial room for improvement in future research. Among all the methods, BEVGrounding stands out by achieving SOTA performance on most evaluation metrics, with an average improvement of 5%-7%. However, it also demonstrates a slight decrease in 3D Acc@0.5, consistent with the trend observed on Talk2Car-3D.

Ablation Studies

- **Text Encoder:** We use CLIP as the language encoder in our experiments. However, prior research often utilizes a GRU-based language encoder with GloVE embeddings [27]. To investigate the impact of different text encoders, we conduct ablation experiments on Talk2Car-3D and present the results in Tab.4. Our findings indicate that CLIP achieves superior performance, surpassing the GloVE-based approach by an average of 5%.
- **Module Components:** To assess the contribution of each module in BEVGrounding, we conduct ablation studies, and the results are detailed in Tab.5. It reveals that the encoder module exerts the most significant in-

Table 4: Ablation studies on text encoder.

	Text Encoder	Unique		Multiple		Overall	
		Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)	Acc@0.25 (%)	Acc@0.5 (%)
BEV	GloVe-50b	16.98	8.49	28.99	22.65	27.75	21.18
	GloVe-100b	27.36	12.26	32.82	23.96	32.25	22.75
	GloVe-200b	25.47	8.49	37.20	25.93	35.98	24.12
	CLIP	**33.02**	**17.92**	**45.30**	**29.76**	**44.02**	**28.53**
3D	GloVe-50b	16.04	2.83	27.24	15.75	26.08	14.41
	GloVe-100b	16.98	3.77	28.56	16.63	27.32	15.29
	GloVe-200b	16.04	3.77	33.48	18.71	31.67	17.16
	CLIP	**25.41**	**8.49**	**37.64**	**20.90**	**36.37**	**19.61**

Table 5: Ablation studies on BEVGrounding's module.

EN	SPCA	SECA	Overall@BEV		Overall@3D	
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
			31.86	20.10	27.65	15.00
✓			38.43	25.49	33.43	17.06
✓	✓		37.84	24.90	31.37	16.96
✓	✓	✓	**44.02**	**28.53**	**36.37**	**19.61**

fluence on the model's performance, leading to a 12.16% accuracy improvement. Interestingly, SPCA and SECA modules exhibit comparable impacts, suggesting that both spatial and semantic feature interactions are crucial for crucial for accurate object grounding.

5.3 Qualitative Analysis

Fig. 6 illustrates the visualization results for both the baseline method and our proposed BEVGrounding approach. We can observe that one-stage BEVGrounding can generate more accurate 3D bounding boxes for the referred objects compared to the two-stage baseline. It stems from BEVGrounding's capability to effectively extract and integrate richer semantic information about the scene. However, both methods still encounter numerous failures during testing, indicating significant room for improvement in 3D visual grounding for autonomous driving.

6 Conclusions

This work introduces the 3D visual grounding task into parallel LiDARs, aiming to equip sensors with a degree of cognitive capability. It provides a novel human-machine interaction approach for LiDAR-based 3D scene understanding. We establish the Talk2LiDAR dataset, a large-scale benchmark for 3D visual grounding, and propose BEVGrounding, a novel one-stage method that demonstrates promising results. Our future work will further explore integrating MLLMs to elevate the cognitive intelligence of parallel LiDAR systems.

Acknowledgements

This work is supported by the Beijing Natural Science Foundation (L245025) and the Joint Development of Multi-modal Parallel LiDARs with Waytous Inc.

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. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022)
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, L., Zhang, Y., Tian, B., Ai, Y., Cao, D., Wang, F.Y.: Parallel driving os: A ubiquitous operating system for autonomous driving in cpss. IEEE Transactions on Intelligent Vehicles **7**(4), 886–895 (2022). `https://doi.org/10.1109/TIV.2022.3223728`
8. Cheng, W., Yin, J., Li, W., Yang, R., Shen, J.: Language-guided 3d object detection in point cloud for autonomous driving. arXiv preprint arXiv:2305.15765 (2023)
9. Dai, H., Luo, S., Ding, Y., Shao, L.: Commands for autonomous vehicles by progressively stacking visual-linguistic representations. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 27–32. Springer (2020)
10. Deruyttere, T., Vandenhende, S., Grujicic, D., Liu, Y., Van Gool, L., Blaschko, M., Tuytelaars, T., Moens, M.F.: Commands 4 autonomous vehicles (c4av) workshop summary. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 3–26. Springer (2020)
11. Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F.: Talk2car: Taking control of your self-driving car. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2088–2098 (2019)

12. Grujicic, D., Deruyttere, T., Moens, M.F., Blaschko, M.B.: Predicting physical world destinations for commands given to self-driving cars. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 715–725 (2022)
13. He, D., Zhao, Y., Luo, J., Hui, T., Huang, S., Zhang, A., Liu, S.: Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2344–2352 (2021)
14. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: European Conference on Computer Vision. pp. 417–433. Springer (2022)
15. Li, Y., Moreau, J., Ibanez-Guzman, J.: Emergent visual sensors for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems **24**(5), 4716–4737 (2023)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916 (2023)
17. Liu, Y., Shen, Y., Fan, L., Tian, Y., Ai, Y., Tian, B., Liu, Z., Wang, F.Y.: Parallel radars: from digital twins to digital intelligence for smart radar systems. Sensors **22**(24), 9930 (2022)
18. Liu, Y., Shen, Y., Guo, C., Tian, Y., Wang, X., Zhu, Y., Wang, F.Y.: Metasensing in metaverses: See there, be there, and know there. IEEE Intelligent Systems **37**(6), 7–12 (2022). `https://doi.org/10.1109/MIS.2022.3221844`
19. Liu, Y., Shen, Y., Tian, Y., Ai, Y., Tian, B., Wu, E., Chen, L.: Radarverses in metaverses: A cpsi-based architecture for 6s radar systems in cpss. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(4), 2128–2137 (2022)
20. Liu, Y., Sun, B., Li, Y., Hu, Y., Wang, F.Y.: Hpl-vit: A unified perception framework for heterogeneous parallel lidars in v2v. In: 2024 International Conference on Robotics and Automation (ICRA) (2024)
21. Liu, Y., Sun, B., Tian, Y., Wang, X., Zhu, Y., Huai, R., Shen, Y.: Software-defined active lidars for autonomous driving: A parallel intelligence-based adaptive model. IEEE Transactions on Intelligent Vehicles **8**(8), 4047–4056 (2023). `https://doi.org/10.1109/TIV.2023.3289540`
22. Liu, Y., Tian, Y., Sun, B., Wang, Y., Wang, F.Y.: Parallel lidars meet the foggy weather. IEEE Journal of Radio Frequency Identification **6**, 867–870 (2022). `https://doi.org/10.1109/JRFID.2022.3203733`
23. 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 (2023). `https://doi.org/10.1109/ICRA48891.2023.10160968`
24. Lu, Z., Pei, Y., Wang, G., Li, P., Yang, Y., Lei, Y., Shen, H.T.: Scaneru: Interactive 3d visual grounding based on embodied reference understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3936–3944 (2024)
25. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16454–16463 (2022)
26. Luo, S., Dai, H., Shao, L., Ding, Y.: C4av: learning cross-modal representations from transformers. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 33–38. Springer (2020)

27. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Roriz, R., Cabral, J., Gomes, T.: Automotive lidar technology: A survey. IEEE Transactions on Intelligent Transportation Systems **23**(7), 6282–6297 (2022). `https://doi.org/10.1109/TITS.2021.3086804`
30. Rufus, N., Nair, U.K.R., Krishna, K.M., Gandhi, V.: Cosine meets softmax: A tough-to-beat baseline for visual grounding. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 39–50. Springer (2020)
31. Shen, Y., Liu, Y., Tian, Y., Na, X.: Parallel sensing in metaverses: Virtual-real interactive smart systems for "6s" sensing. IEEE/CAA Journal of Automatica Sinica **9**(12), 2047–2054 (2022). `https://doi.org/10.1109/JAS.2022.106115`
32. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
33. Wang, F.Y.: Parallel system methods for management and control of complex systems. Control and Decision. **19**, 485–489 (2004)
34. Wang, F.Y., Zheng, N.N., Cao, D., Martinez, C.M., Li, L., Liu, T.: Parallel driving in cpss: a unified approach for transport automation and vehicle intelligence. IEEE/CAA Journal of Automatica Sinica **4**(4), 577–587 (2017). `https://doi.org/10.1109/JAS.2017.7510598`
35. Wang, F.: The emergence of intelligent enterprises: From cps to cpss. IEEE Intelligent Systems **25**(4), 85–88 (2010). `https://doi.org/10.1109/MIS.2010.104`
36. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14633–14642 (2023)
37. Wu, D., Han, W., Wang, T., Liu, Y., Zhang, X., Shen, J.: Language prompt for autonomous driving. arXiv preprint arXiv:2309.04379 (2023)
38. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19231–19242 (2023)
39. Yang, J., Wang, X., Zhao, Y.: Parallel manufacturing for industrial metaverses: A new paradigm in smart manufacturing. IEEE/CAA Journal of Automatica Sinica **9**(12), 2063–2070 (2022). `https://doi.org/10.1109/JAS.2022.106097`
40. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1856–1866 (2021)
41. Zhan, Y., Yuan, Y., Xiong, Z.: Mono3dvg: 3d visual grounding in monocular images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6988–6996 (2024)
42. 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)
43. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)