LLM-RG: Referential Grounding in Outdoor Scenarios using Large Language Models

Pranav Saxena[1] Avigyan Bhattacharya[2] Ji Zhang[2] Wenshan Wang[2]

***Abstract*— Referential grounding in outdoor driving scenes is challenging due to large scene variability, many visually similar objects, and dynamic elements that complicate resolving natural-language references (e.g., "the black car on the right"). We propose LLM-RG, a hybrid pipeline that combines off-the-shelf vision–language models for fine-grained attribute extraction with large language models for symbolic reasoning. LLM-RG processes an image and a free-form referring expression by using an LLM to extract relevant object types and attributes, detecting candidate regions, generating rich visual descriptors with a VLM, and then combining these descriptors with spatial metadata into natural-language prompts that are input to an LLM for chain-of-thought reasoning to identify the referent's bounding box. Evaluated on the Talk2Car benchmark, LLM-RG yields substantial gains over both LLM and VLM-based baselines. Additionally, our ablations show that adding 3D spatial cues further improves grounding. Our results demonstrate the complementary strengths of VLMs and LLMs, applied in a zero-shot manner, for robust outdoor referential grounding.**

I. INTRODUCTION

Enabling autonomous systems to ground referring expressions to real-world entities in complex settings is a crucial step toward safe and natural interactions with humans. In contrast to indoor settings, which have been the focus of most prior works, outdoor scenes pose distinct challenges due to larger scales, greater object diversity, and more complex, dynamic environments like roads and intersections. Referential expressions in this setting frequently rely on high-level attributes (such as *color*, *orientation*, or *type*) and relative spatial relations (such as "on the right" or "behind the van"), which are harder to resolve than the structured references typically found indoors.

In recent years, significant progress has been made in grounding referential language within indoor environments. Large-scale 3D datasets such as Matterport3D [1], ScanNet [2], and HM3D [3] have enabled tasks including visual grounding, embodied instruction following, and object-goal navigation. Methods developed on these datasets often combine object detection with geometric or spatial reasoning modules [4]–[6], or leverage pretrained language models to link natural language queries to structured scene graphs [7], [8]. More recently, large multimodal models have been explored for zero-shot grounding in 3D indoor spaces [9], [10]. These approaches have shown strong performance in resolving references to small household objects and reasoning about relations such as "next to the chair" or "on top of the table." However, they remain heavily tuned to the relatively constrained and repetitive structure of indoor scenes, where object categories are limited, contexts are predictable, and the variability of natural language references is narrower.

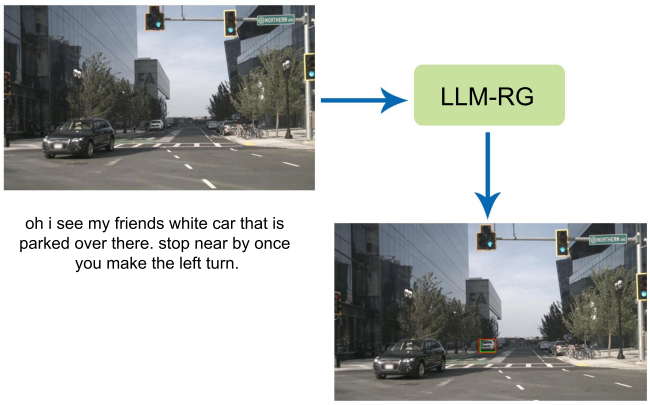

Fig. 1: An example output of LLM-RG on a scene from Talk2Car. Red box denotes Ground Truth bounding box from Talk2Car, green box denotes the predicted bounding box using LLM-RG.

In contrast, outdoor referential grounding has received considerably less attention, even though it is essential for applications in autonomous driving, mobile robotics, and delivery systems. Outdoor environments are inherently more complex: they contain a larger and more open-ended vocabulary of objects (e.g., cars, trucks, pedestrians, bicycles, traffic lights), involve greater scene variability (e.g., urban streets, intersections, crosswalks, parking lots), and are subject to dynamic changes such as moving vehicles or occlusions. Furthermore, outdoor language queries tend to be more diverse and ambiguous, often requiring fine-grained disambiguation across multiple similar objects (e.g., "the black car on the right" when several black cars are present) or reasoning over higher-level semantics (e.g., "the car waiting at the stop sign"). Datasets such as Talk2Car [11] address this gap by providing natural language commands linked to visual driving scenes, but methods specifically designed for outdoor referential grounding remain scarce.

To address these challenges, we propose to leverage recent advances in vision-language models (VLMs) for extracting fine-grained object attributes and large language models (LLMs) for reasoning over natural language queries. We evaluate this approach on the Talk2Car dataset, which provides a realistic and challenging benchmark for outdoor

[1]Pranav Saxena is with Birla Institute of Technology and Science Pilani, K.K Birla Goa Campus, Goa, India f20220257@goa.bits-pilani.ac.in

[2]The authors are with Carnegie Mellon University, Robotics Institute, Pittsburgh, PA. {avigyanb, zhangji, wenshanw}@andrew.cmu.edu

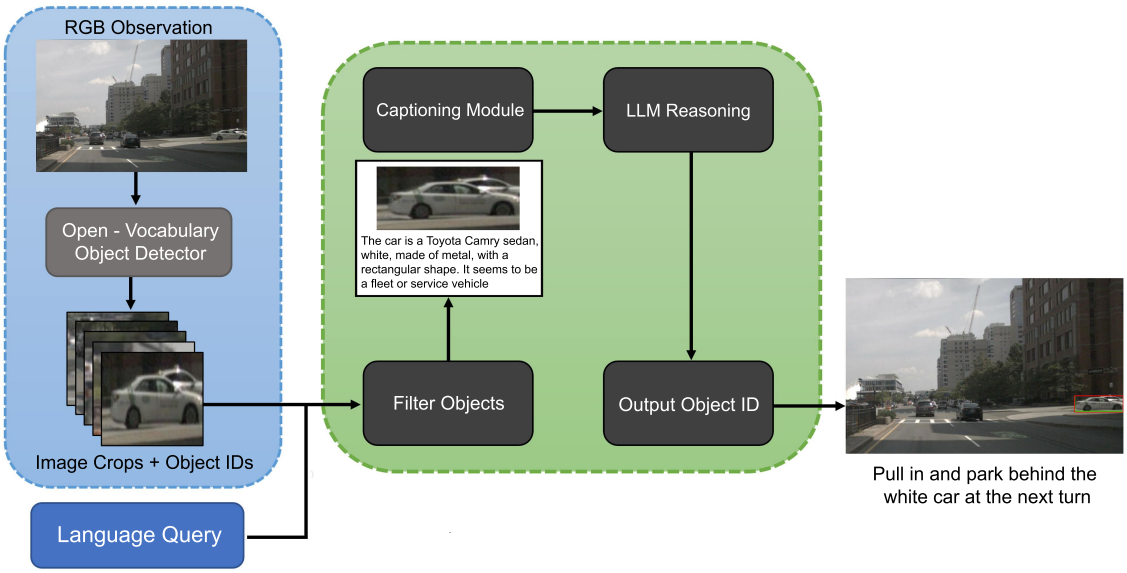

Fig. 2: **Architecture of LLM-RG: (A)** A large language model (LLM) processes the referring expression to identify relevant object types and attributes, generating a shortlist of candidate objects. **(B)** MMDetection is used to detect objects and obtain 2D bounding boxes. **(C)** Object crops for each detection are extracted and passed to a vision-language model (VLM) which provides fine-grained descriptions of each candidate object, capturing properties such as color, type, orientation, and contextual details. **(D)** The LLM combines object IDs, spatial locations, and object descriptions to reason over the referring expression and identify the bounding box of the target object.

referential grounding in driving scenarios.

Our work, based on SORT3D [10], introduces these three key contributions -

1) We present a novel pipeline that combines VLM-based object attribute extraction with LLM-based reasoning for outdoor referential grounding.
2) We show that our approach works without any task-specific fine-tuning, making it deployable across unseen datasets and robotic setups.
3) We provide an extensive evaluation demonstrating the effectiveness of this hybrid approach and highlight its potential for natural human-vehicle interaction in real-world contexts.

II. Related Work

A. Referential Grounding in Indoor Scenes

Referential grounding in indoor environments has become a major focus, aiming to connect natural language with complex 3D scenes. **Scene-LLM** [12] integrates large language models with 3D visual data for tasks like scene captioning, object identification, and interactive planning. **IRef-VLA** [13] introduces a benchmark with over 11.5K scanned rooms and 4.7M referential statements for precise language grounding to objects and actions. **3D-GRAND** [14] provides a large-scale 3D-text dataset with 40K+ scenes and 6.2M instructions, supporting object reference, spatial reasoning, and training of 3D-LLMs.

While indoor benchmarks have advanced referential grounding, our focus is on outdoor driving scenes, which present greater diversity, dynamics, and ambiguity relevant to our application.

B. Referential Grounding in Outdoor Scenes

Although relatively underexplored, referential grounding in outdoor environments introduces distinct challenges stemming from the large scale and variability of scenes. **LidaRefer** [15] addresses these challenges with a transformer-based 3D visual grounding framework tailored for large-scale outdoor scenes, tackling issues such as high computational demands and ambiguous object identification in sparse LiDAR point clouds. Meanwhile, **Grounded 3D-LLM** [16] introduces the use of scene referent tokens to reference 3D scenes, enabling the processing of interleaved 3D and textual data sequences and unifying multiple 3D vision tasks within a generative framework. While these methods rely on extensive training to adapt to outdoor domains, our pipeline remains training-free, enabling flexible zero-shot deployment across new datasets and scenarios.

C. LLMs and VLMs for Referential Grounding

The integration of large language models (LLMs) and vision-language models (VLMs) has led to significant progress in referential grounding across both indoor and outdoor environments. Recent approaches leverage the complementary strengths of language understanding and visual perception to address increasingly complex grounding tasks. **SpatialVLM** [17] introduces a vision-language model capable of interpreting spatial relationships directly from language, without relying on explicit 3D scene representations, enabling the resolution of spatial queries through purely linguistic and visual cues. **GLaMM** [18] demonstrates robust multimodal capabilities by supporting scene-level understanding, region-level interpretation, and pixel-level ground-

ing, effectively addressing a broad range of visual grounding challenges across different granularities. Extending these ideas into the 3D domain, **ScanReason** [19] proposes a 3D visual grounding framework that combines multimodal LLMs with a visual-centric reasoning module and a 3D grounding module, enhancing geometric understanding and fine-grained object localization within complex 3D scenes.

Unlike prior work, our framework adopts a modular chaining strategy, where a VLM generates fine-grained visual descriptions and an LLM performs symbolic reasoning, unifying perception and reasoning in a zero-shot manner.

III. Methodology

In this work, we present **LLM-RG**, a hybrid LLM-VLM pipeline for outdoor referential grounding, specifically designed to handle the scale, diversity, and dynamic nature of real-world driving scenes. The system takes a driving scene and a free-form referring expression as input and outputs the bounding box corresponding to the referenced object. By leveraging large language models for reasoning and vision-language models for fine-grained attribute extraction, our approach enables robust zero-shot grounding in challenging urban environments.

Figure 2 provides an overview of the proposed pipeline. Each component of the system is described in detail in the following subsections.

A. Object Detection and Localization

Given an RGB image paired with a referential statement, we first use a Large Language Model (LLM) to extract the relevant object categories mentioned in the query. This acts as a textual filter, narrowing down candidate objects in complex scenes. For instance, considering the referential statement "Park near the car under the tree," the extracted relevant objects are ["car", "tree"].

The filtered categories are then passed to an open-vocabulary object detector [20], which predicts bounding boxes and class labels for each candidate object. This combination of textual filtering and visual detection allows the system to focus on relevant objects efficiently.

B. Object Feature Extraction

For each detected object, we crop the corresponding image region and associate it with its class label. We then use a large vision-language model [21] to generate fine-grained attributes such as color, material, shape, affordances, and other meaningful attributes. These object descriptors provide richer semantic information than bounding boxes alone, enabling the system to capture details critical for disambiguating similar objects. This step mirrors human perception, where attributes help distinguish between multiple candidate objects in a scene. An example caption is shown in Figure 3.

C. Prompt Construction and LLM Reasoning

The extracted object attributes, bounding box coordinates, and class labels are formatted into a natural language prompt for the LLM [22] to reason about the scene, represented as *[id, 'name', 'caption', [x,y]]*. To improve reasoning, we include example input-output pairs, which we find helps increase the reasoning significantly, and instruct the LLM to use chain-of-thought reasoning. This structured prompt enables the LLM to interpret visual and spatial information in textual form, thereby providing the necessary context to accurately identify the correct referent from the candidate objects.

The Truck is a white FedEx vehicle. It is white, made of metal, rectangular in shape. It has a boxy design with a flatbed at the back for transporting goods. Affordances include the ability to transport various items and deliver packages.

Fig. 3: Example of an object caption from a VLM that includes fine-grained attributes to be used for further reasoning.

The LLM processes the prompt to identify the object that best matches the referential statement. Once the LLM outputs the object ID, it is mapped back to the corresponding bounding box.

IV. Experiments and Results

A. Evaluation

Dataset: We evaluate on the Talk2Car dataset, which provides real-world driving scenes from nuScenes [23] paired with free-form referring expressions and corresponding ground-truth bounding boxes.

Metrics: We report Accuracy (Acc@0.5), the percentage of predictions with Intersection over Union (IoU) ≥ 0.5 with the ground truth.

B. Baselines

We compare our results against a recent work and against different baselines that we design:

(i) **LLM-Wrapper [24]:** It adapts off-the-shelf VLMs for referring expression comprehension in a black-box setting. The method converts candidate bounding boxes from a VLM into textual prompts and uses an LLM to reason over them, selecting the most relevant box. We compare against its zero-shot version.

(ii) **Naive-VLM:** We utilize a VLM (Gemini 2.5-flash) by passing the RGB image and the referential statement directly. The model outputs a bounding box without any additional reasoning or context. This serves as a simple baseline for zero-shot comprehension.

(iii) **Image Crops + VLM:** We crop the image into object regions and assign each region an object ID. These crops with IDs are passed to the VLM, which is asked to output the best matching object ID. We then compare the IoU of the predicted bounding box corresponding to the best object ID with the ground-truth bounding box.

Fig. 4: Qualitative results of LLM-RG on Talk2Car (first row) and mecanum robot (second row). Red box denotes Ground Truth bounding box from Talk2Car, green box denotes the predicted bounding box using LLM-RG.

(iv) **Image with bounding box + captions + VLM:** We pass the entire image with detected bounding boxes along with object IDs and captions to a VLM (Gemini 2.5-flash) and ask it output the best object ID. We then compare the IoU of the predicted bounding box corresponding to the best object ID with the ground-truth bounding box.

Method	val	test
Naive-VLM	37.13	39.23
Image Crops + VLM	47.32	48.11
Image with bounding box + Captions + VLM	63.12	63.78
LLM-Wrapper (Zero-Shot)	55.37	58.44
LLM-RG	**64.72**	**67.91**

TABLE I: Results of LLM-RG on Talk2Car Dataset.

C. Ablation Study

We perform an ablation study II to evaluate the impact of incorporating 3D spatial information on reasoning performance. Specifically, we consider the following variants:

(i) We extend the Talk2Car framework to the NuScenes dataset by projecting 2D bounding boxes onto the corresponding LiDAR point cloud to obtain partial 3D bounding boxes. The centroid of each resulting 3D box is then used to estimate the approximate 3D location of the object.

(ii) We directly utilize the ground-truth 3D bounding boxes provided by NuScenes, bypassing the detection and projection steps, and perform reasoning based on these boxes.

We observe that having 3D Spatial Information helps improve reasoning significantly

D. Qualitative Results

We show the qualitative results of LLM-RG on Talk2Car dataset, as well as on a real-life mecanum robot, highlighting its zero-shot adaptability on a new robotic setup (Fig. 4).

Method	Accuracy
LLM-RG	64.72
LLM-RG + LIDAR	70.23
LLM-RG + GT Bounding boxes from NuScenes	77.93

TABLE II: Comparison of reasoning across different methods

V. Conclusion And Future Work

We present LLM-RG, a hybrid referential grounding pipeline for outdoor driving scenes that leverages off-the-shelf vision-language models for detailed visual description and large language models for flexible, symbolic reasoning. Operating zero-shot without task-specific fine-tuning, LLM-RG parses free-form referring expressions into object-level attributes, generates candidate detections, and converts visual descriptors and spatial metadata into natural language prompts, framing grounding as a language-guided selection problem. Evaluations on the Talk2Car benchmark show this modular approach substantially improves over simple VLM baselines and related methods.

Our results highlight two key strengths: the complementary capabilities of VLMs, for extracting fine-grained image attributes, and LLMs, for compositional reasoning over them; and the practicality of a modular pipeline that can exploit black-box models without end-to-end retraining. These features make LLM-RG well-suited for outdoor driving scenarios requiring fast domain adaptation and interpretable outputs. Our ablations further reveal that incorporating 3D spatial information significantly improves grounding accuracy.

In future work, we plan to incorporate richer multimodal signals, such as depth maps and radar, to improve robustness in real-world driving. We also aim to extend LLM-RG to dynamic environments by integrating object tracking and temporal reasoning, enabling reliable operation in scenes with moving objects and complex interactions.

REFERENCES

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[3] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[4] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.

[6] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019.

[7] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10737–10746, 2020.

[8] Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding, 2025.

[9] Jiading Fang, Xiangshan Tan, Shengjie Lin, Igor Vasiljevic, Vitor Guizilini, Hongyuan Mei, Rares Ambrus, Gregory Shakhnarovich, and Matthew R. Walter. Transcrib3d: 3d referring expression resolution through large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9737–9744, 2024.

[10] Nader Zantout, Haochen Zhang, Pujith Kachana, Jinkai Qiu, Guofei Chen, Ji Zhang, and Wenshan Wang. Sort3d: Spatial object-centric reasoning toolbox for zero-shot 3d grounding using large language models, 2025.

[11] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. 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)*, pages 2088–2098, 2019.

[12] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206, 2025.

[13] Haochen Zhang, Nader Zantout, Pujith Kachana, Ji Zhang, and Wenshan Wang. Iref-vla: A benchmark for interactive referential grounding with imperfect language in 3d scenes, 2025.

[14] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F. Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29501–29512, 2025.

[15] Yeong-Seung Baek and Heung-Seon Oh. Lidarefer: Context-aware outdoor 3d visual grounding for autonomous driving, 2025.

[16] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens, 2024.

[17] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, June 2024.

[18] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018, 2024.

[19] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities, 2024.

[20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[21] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[22] Mistral AI Team. Mistral large 2: The new generation of flagship model. https://mistral.ai/news/mistral-large-2407/, 2024.

[23] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020.

[24] Amaia Cardiel, Eloi Zablocki, Elias Ramzi, Oriane Siméoni, and Matthieu Cord. Llm-wrapper: Black-box semantic-aware adaptation of vision-language models for referring expression comprehension, 2025.

[25] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024.

[27] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15503–15512, 2022.

[28] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2899–2909, 2023.

[29] Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, 2023.

[30] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701, 2024.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[32] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[33] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20623–20633, 2024.