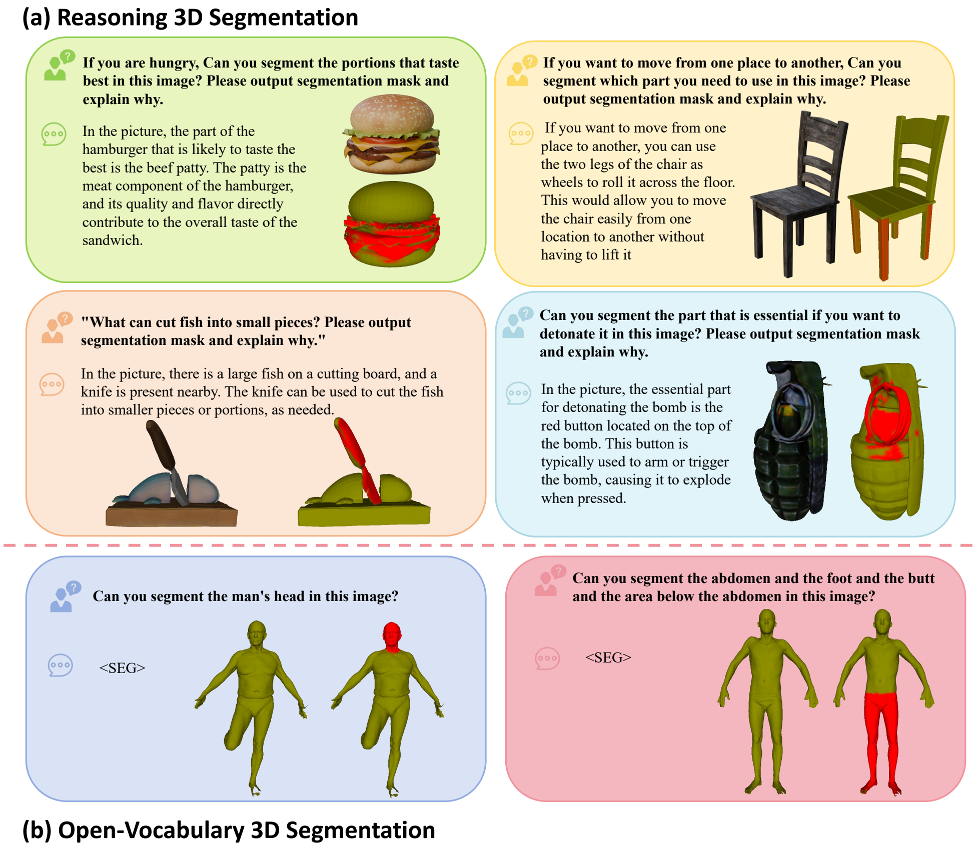
# Reasoning3D - Grounding and Reasoning in 3D: Fine-Grained Zero-Shot Open-Vocabulary 3D Reasoning Part Segmentation via Large Vision-Language Models

**Tianrun Chen**[1,6]* **Chunan Yu**[2]* **Jing Li**[2] **Jianqi Zhang**[2] **Lanyun Zhu**[3]
Deyi Ji[4] Yong Zhang[2] Ying Zang[2]† Zejian Li[5] Lingyun Sun[1]

[1]College of Computer Science and Technology, Zhejiang University
[2]School of Information Engineering, Huzhou University
[3]Singapore University of Technology and Design
[4]University of Science and Technology of China
[5]School of Software Technology, Zhejiang University
[6]KOKONI3D, Moxin (Huzhou) Technology Co., LTD.
tianrun.chen@kokoni3d.com 02750@zjhu.edu.cn



(a) Reasoning 3D Segmentation

**If you are hungry, Can you segment the portions that taste best in this image? Please output segmentation mask and explain why.**

In the picture, the part of the hamburger that is likely to taste the best is the beef patty. The patty is the meat component of the hamburger, and its quality and flavor directly contribute to the overall taste of the sandwich.

**If you want to move from one place to another, Can you segment which part you need to use in this image? Please output segmentation mask and explain why.**

If you want to move from one place to another, you can use the two legs of the chair as wheels to roll it across the floor. This would allow you to move the chair easily from one location to another without having to lift it

**"What can cut fish into small pieces? Please output segmentation mask and explain why."**

In the picture, there is a large fish on a cutting board, and a knife is present nearby. The knife can be used to cut the fish into smaller pieces or portions, as needed.

**Can you segment the part that is essential if you want to detonate it in this image? Please output segmentation mask and explain why.**

In the picture, the essential part for detonating the bomb is the red button located on the top of the bomb. This button is typically used to arm or trigger the bomb, causing it to explode when pressed.

**Can you segment the man's head in this image?**

<SEG>

**Can you segment the abdomen and the foot and the butt and the area below the abdomen in this image?**

<SEG>

(b) Open-Vocabulary 3D Segmentation

Figure 1: In this work, we propose a new task: reasoning 3D segmentation. We also propose a method that can segment 3D object parts with explanations based on various criteria such as reasoning, shape, location, function, and conceptual instructions.

## Abstract

In this paper, we introduce a new task: Zero-Shot 3D Reasoning Segmentation, a new paradigm in 3D segmentation that goes beyond traditional category-specific methods. We propose a baseline method, Reasoning3D, that leverages pre-trained 2D segmentation networks powered by Large Language Models (LLMs) to interpret user queries and segment 3D meshes with contextual awareness. This approach enables fine-grained part segmentation and generates natural language explanations without requiring extensive 3D datasets. Experiments demonstrate that Reasoning3D can effectively localize and highlight parts of 3D objects. Our training-free method allows rapid deployment and serves as a universal baseline for future research in various fields such as robotics, object manipulation, autonomous driving, AR/VR, and medical applications. The code and the user interface have been released publicly.

## 1 Introduction

Traditional 3D segmentation approaches are typically confined to fixed object categories. With recent breakthroughs in Multi-modal Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) Liu et al. (2023b); Shen et al. (2023); Lin et al. (2021); Wang et al. (2024); Zheng et al. (2024); Zhu et al. (2024b), were extending these capabilities Lai et al. (2024); Yang et al. (2024b); Zhu et al. (2024a) to 3D.

However, the scarcity of 3D data with question-and-answer pairs stopped us from performing large-scale training. Inspired by research that has tackled similar challenges in 3D generation, we propose to leverage off-the-shelf 2D networks to perform the task in a zero-shot manner. This approach, which we named Reasoning3D, renders a 3D model from multiple viewpoints and applies a pre-trained reasoning segmentation network to each 2D view. By fusing these individual masks and explanations, we create a comprehensive 3D segmentation mask.

While Reasoning3D is a straightforward baseline method, we believe it serves as a good starting point for researchers to explore and expand the future of 3D part segmentation, paving the way for future research in 3D part segmentation. To spark further innovation, were releasing the implementation and benchmark code.

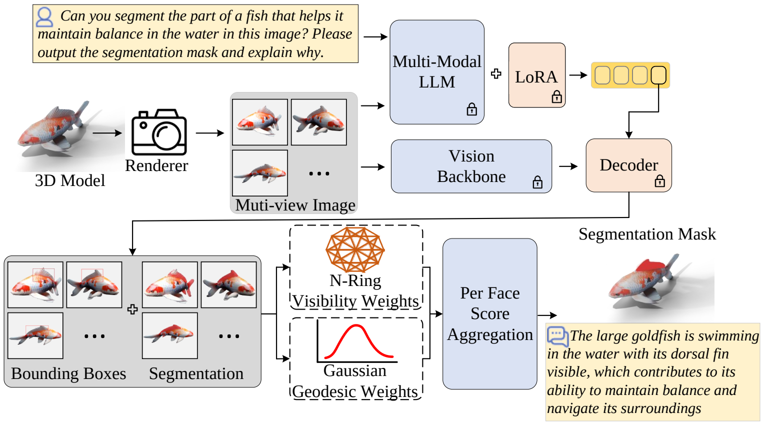
Below in Tab. 1, we show how our approach differs from some existing LLM-based 3D segmentation method (including SQA3D Ma et al. (2022), 3D-VisTA Zhu et al. (2023), ViewRefer Guo et al. (2023b), Point-Bind Guo et al. (2023a), 3D-OVS Liu et al. (2023a), OpenMask3D Takmaz et al. (2023), PLA Ding et al. (2023), OpenScene Peng et al. (2023), Chat-3D Wang et al. (2023), M3DBench Li et al. (2023), LLM-Grounder Yang et al. (2024a), 3D-LLM Hong et al. (2023), LL3DA Chen et al. (2024), PointLLM Xu et al. (2025), PARIS3D Kareem et al. (2025)).

Table 1: Recently, there has been a significant increase in comparative studies of 3D segmentation models and large multimodal models (LMMs), highlighting their potential for 3D reasoning and conversations. In reasoning queries, these models need to autonomously analyze tasks and generate text or perform corresponding actions. In terms of segmentation, some models respond using 3D segmentation masks, while others focus on providing conversation-style answers.

| | Method | SQA 3D | 3D-VisTA | View Refer | Point-Bind | 3D-OVS | Open Mask3D | PLA | Open Scene | Chat-3D | M3D Bench | LLM-Grounder | 3D-LLM | LL 3DA | Point LLM | PARIS 3D | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Input | Input Type | Scene | **Object** | Scene | **Object** | Scene | Scene | Scene | Scene | Scene | Scene | Scene | Scene | Scene | **Object** | Scene | **Object** |
| | Reasoning Query | Yes | Yes | No | No | No | No | No | No | Yes | No | No | Yes | No | No | Yes | Yes |
| Task | Conversation | No | No | No | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| | Segmentation | No | No | No | No | Yes | Yes | Yes | Yes | No | No | No | No | No | No | Yes | Yes |
| | Explanation | No | No | No | No | No | No | No | No | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes |

## 2 Method

Reasoning3D begins with a mesh input fed into the renderer for viewpoint rendering, generating the face id for each corresponding viewpoint. Next, the rendered viewpoints and the user-input prompt are processed by the pre-trained 2D reasoning segmentation network Lai et al. (2024), which segments the image to extract the desired parts and output explanations. Finally, using the mapping relationship between each viewpoint and its corresponding mesh face id, the segmented parts are reconstructed back onto the mesh with a specially designed multi-view fusion mechanism. Following Abdelreheem et al. (2023), we smooth and refine the segmentation boundaries, reducing noise and errors with Gaussian Geodesic Reweighting. Subsequently, we apply the Visibility Smoothing technique to eliminate discontinuities caused by changes in viewpoints, ensuring that the segmented mesh appears natural and coherent from all angles. Finally, we use a Global Filtering Strategy that filters out the masked regions with low confidence scores. The threshold is the mean confidence score calculated for every face.


*Can you segment the part of a fish that helps it maintain balance in the water in this image? Please output the segmentation mask and explain why.*

*The large goldfish is swimming in the water with its dorsal fin visible, which contributes to its ability to maintain balance and navigate its surroundings*

Figure 2: The overview of Reasoning3D. First, a 3D model represented by 3D meshes is fed into a renderer to obtain multi-view images. Then, each image goes through a vision backbone and a multi-modal LLM along with user input queries. The decoder decodes the final layer embedding which contains the extra token, thus producing K segmentation masks. We also extract the bounding boxes in this stage. Finally, a specially designed mask-to-3D segmentation algorithm elevates the projections back into the 3D space.

## 3 Experiments

We first evaluated the zero-shot open-vocabulary segmentation performance on the FAUST Bogo et al. (2014) benchmark (an open-vocabulary 3D segmentation benchmark) proposed in SATR Abdelreheem et al. (2023). We also validated the effectiveness of our method on reasoning 3D segmentation by our collected in-the-wild data from SketchFab. During the rendering process, we centered the input mesh at the origin and normalized it within a unit sphere. We evenly sample 8 images horizontally around all 360 degrees with a resolution of 1024×1024 and set a uniform black background color. Multiple reasons (or explanations) will be generated in each view to give a comprehensive understanding of the object, and users can choose one as the desired answer.

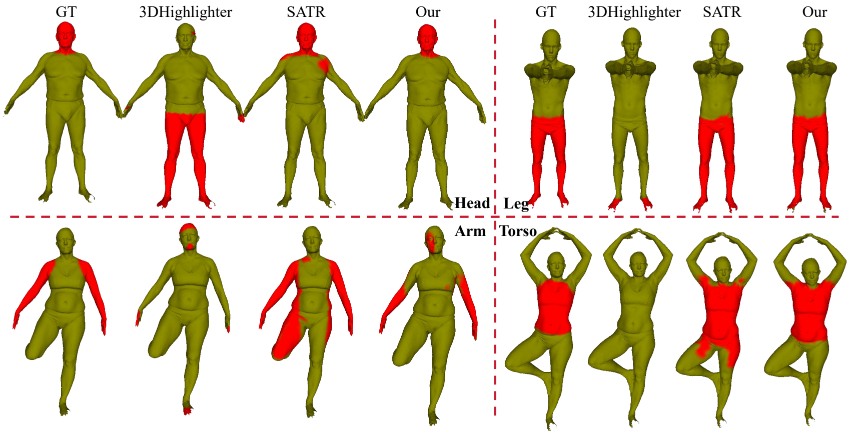
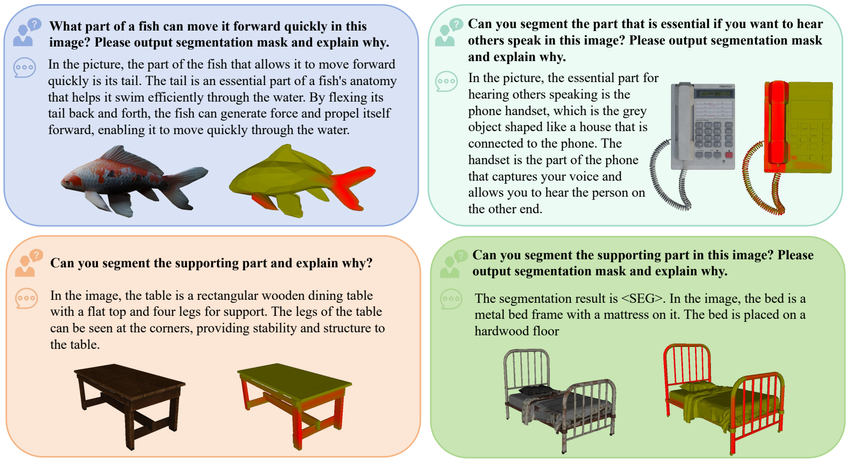


Figure 3: Qualitative results and comparison between our method and baseline method in FAUST benchmark. The segmented regions are shown in red.

The examples in Fig. 1 ane Fig. 4 show that Reasoning3D has the capabilities to offer in-depth reasoning, 3D understanding, part segmentation, and conversational abilities. The model can output



**What part of a fish can move it forward quickly in this image? Please output segmentation mask and explain why.**

In the picture, the part of the fish that allows it to move forward quickly is its tail. The tail is an essential part of a fish's anatomy that helps it swim efficiently through the water. By flexing its tail back and forth, the fish can generate force and propel itself forward, enabling it to move quickly through the water.

**Can you segment the part that is essential if you want to hear others speak in this image? Please output segmentation mask and explain why.**

In the picture, the essential part for hearing others speaking is the phone handset, which is the grey object shaped like a house that is connected to the phone. The handset is the part of the phone that captures your voice and allows you to hear the person on the other end.

**Can you segment the supporting part and explain why?**

In the image, the table is a rectangular wooden dining table with a flat top and four legs for support. The legs of the table can be seen at the corners, providing stability and structure to the table.

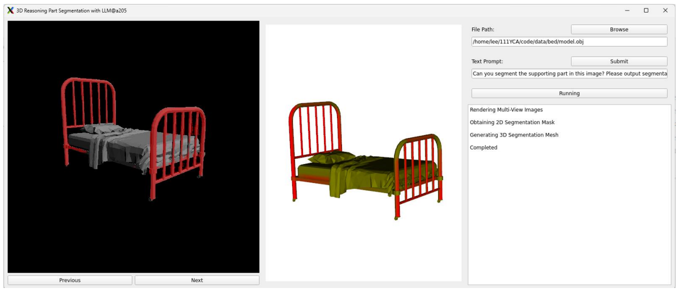
**Can you segment the supporting part in this image? Please output segmentation mask and explain why.**

The segmentation result is <SEG>. In the image, the bed is a metal bed frame with a mattress on it. The bed is placed on a hardwood floor

Figure 4: This figure shows Reasoning3D's ability to segment 3D object parts (in a fine-grained manner) from in-the-wild samples, including real-world scanned data (samples are randomly collected from SketchFab). These examples highlight Reasoning3D's advanced capabilities in in-depth reasoning, comprehensive 3D understanding, precise part segmentation, and robust conversational abilities. The original mesh and the segmentation result are visualized, and the segmented region is highlighted in Red.

the segmentation masks and the explanation as we desire. We also show that though not designed for open-vocabulary segmentation tasks and without fine-tuning or specially designed structure, our method achieves satisfactory performance in the open-vocabulary segmentation benchmark compared to existing open-vocabulary 3D segmentation models such as SATR Abdelreheem et al. (2023) and 3DHighlighter Decatur et al. (2022) (Fig. 3).

To better allow users to interact with our system, we designed a User Interface (UI) so that users can input arbitrary 3D models and their desired prompt to segment the desired region. (Fig. 5)



3D Reasoning Part Segmentation with LLM@a205

File Path: /home/lee/111YCA/code/data/bed/model.obj

Text Prompt: Can you segment the supporting part in this image? Please output segmenta

Rendering Multi-View Images

Obtaining 2D Segmentation Mask

Generating 3D Segmentation Mesh

Completed

Figure 5: We offer a user-friendly and open-sourced interface designed for users to interactively segment 3D objects.

## 4 Conclusion

This paper introduces a new task: Zero-Shot 3D Reasoning Segmentation for part searching and localization within 3D objects. The proposed method, Reasoning3D, leverages pre-trained 2D segmentation networks and large language models to enable zero-shot 3D segmentation. This allows for effective part-level 3D understanding with limited 3D datasets. Experiments demonstrate that Reasoning3D can accurately localize and identify parts of 3D objects based on textual queries, including articulated objects and real-world scans. The method can also produce natural language explanations for the segmented 3D models and their components. The training-free approach facilitates rapid deployment and provides a robust baseline for future research in part-level 3D object understanding. This has potential applications across various domains such as robotics, object manipulation, and AR/VR. We have released the code, UI, and more experimental results publicly.

## References

Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. Satr: Zero-shot semantic segmentation of 3d shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15166–15179, 2023.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2014. doi: 10.1109/cvpr.2014.491.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024.

Dale Decatur, Itai Lang, and Rana Hanocka. 3d highlighter: Localizing regions on 3d shapes via text descriptions, 2022.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7010–7019, 2023.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023a.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15372–15383, 2023b.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Amrin Kareem, Jean Lahoud, and Hisham Cholakkal. Paris3d: Reasoning-based 3d part segmentation using large multimodal model. In *European Conference on Computer Vision*, pp. 466–482. Springer, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts, 2023.

Junyang Lin, An Yang, Jinze Bai, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Yong Li, Wei Lin, Jingren Zhou, and Hongxia Yang. M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining, 2021.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023a.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language, 2023b.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. Oct 2022.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023.

Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2025.

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)*, pp. 7694–7701. IEEE, 2024a.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model, 2024b.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2024.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation, 2024a.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024b.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2911–2921, 2023.