

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

Tianrun Chen^{1,7*} Chunan Yu^{2,3*} Jing Li² Jianqi Zhang² Lanyun Zhu⁴
Deyi Ji⁵ Yong Zhang² Ying Zang^{2†} Zejian Li⁶ Lingyun Sun¹

¹College of Computer Science and Technology, Zhejiang University

²School of Information Engineering, Huzhou University

³School of Computer Science and Engineering, Nanjing University of Science and Technology

⁴Singapore University of Technology and Design

⁵University of Science and Technology of China

⁶School of Software Technology, Zhejiang University

⁷KOKONI3D, Moxin (Huzhou) Technology Co., LTD.

tianrun.chen@kokoni3d.com

02750@zjhu.edu.cn

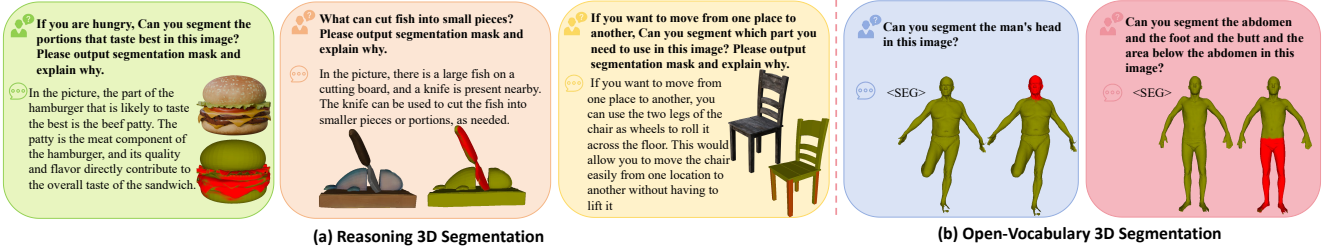


Figure 1. In this work, we propose a new task: reasoning 3D segmentation. We also propose a method that can segment 3D object parts with explanations based on various criteria such as reasoning, shape, location, function, and conceptual instructions.

Abstract

In this paper, we introduce a new task: Zero-Shot 3D Reasoning Segmentation for parts searching and localization for objects, which is a new paradigm to 3D segmentation that transcends limitations for previous category-specific 3D semantic segmentation, 3D instance segmentation, and open-vocabulary 3D segmentation. We design a simple baseline method, Reasoning3D, with the capability to understand and execute complex commands for (fine-grained) segmenting specific parts for 3D meshes with contextual awareness and reasoned answers for interactive segmentation. Specifically, Reasoning3D leverages an off-the-shelf pre-trained 2D segmentation network, powered by Large Language Models (LLMs), to interpret user input queries in a zero-shot manner. Previous research has shown that extensive pre-training endows foundation models with prior world knowledge, enabling them to comprehend complex commands, a capability we can harness to "segment anything" in 3D with limited 3D

datasets (source efficient). Experimentation reveals that our approach is generalizable and can effectively localize and highlight parts of 3D objects (in 3D mesh) based on implicit textual queries, including these articulated 3D objects and real-world scanned data. Our method can also generate natural language explanations corresponding to these 3D models and the decomposition. Moreover, our training-free approach allows rapid deployment and serves as a viable universal baseline for future research of part-level 3D (semantic) object understanding in various fields including robotics, object manipulation, part assembly, autonomous driving applications, augmented reality and virtual reality (AR/VR), and medical applications. The code and the user interface have been released publicly.

* These authors contributed equally to this work.

† Corresponding author.

Code: <http://tianrun-chen.github.io/Reason3D/>

1. Introduction

The importance of 3D segmentation cannot be overstated - it is foundational in fields like robotics, autonomous driving, and augmented reality[12, 35, 39]. Traditional approaches have often required extensive manual labeling or complex rule-based algorithms that struggle to generalize to diverse real-world scenarios[21, 24, 38, 40]. The sheer complexity of 3D data, combined with the inherent ambiguities and varying viewpoints, has posed significant challenges in developing robust and generalizable 3D segmentation techniques.

In this work, we introduce fine-grained **Zero-Shot 3D Reasoning Segmentation** for parts in 3D objects, which aims to bring 3D segmentation to a new level. Imagine instructing a system with words like "segment the part of the chair where you would sit" or "highlight the nutritious parts of this vegetable" and watching it magically understand and perform the task in the 3D world. You can have natural conversations with the system and see it output the segmentation mask along with explanations (See Fig. 1 for examples). It is a future where 3D systems can intuitively understand and respond to intricate queries – The possibilities are endless.

However, achieving this vision is no small feat. Traditional 3D segmentation approaches are typically confined to fixed object categories, severely limiting their flexibility. Recent endeavors in open-vocabulary segmentation can handle a broader range of labels but are still limited to dealing with straightforward tasks like labeling "the apple" and cannot handle complex, reasoning-based queries. Asking a system to perform more nuanced operations like "segment edible parts of a fruit" requires a level of contextual understanding and reasoning that current methods do not possess.

Thanks to the recent advancements in Multi-modal Large Language Models (LLMs)[19, 22, 27, 31, 42, 44], we can now bring our aforementioned vision to life. Recently, Large Vision-Language Models (LVLM) have shown remarkable capabilities in comprehending 2D images, excelling in tasks that require complex reasoning, multi-turn conversations, and explanatory answers[15, 37, 43]. We aim to extend their capabilities into the 3D realm, and we believe that this transition is promising with much practical value – Never forget that we live in a 3D world!

However, extending the success of reasoning segmentation from 2D to 3D domains also presents substantial challenges. The scarcity of available 3D data and ground-truth Question-and-answer pairs stopped us from performing large-scale training. The added dimension also increases the computational demands of 3D architectural components. Here, inspired by research that has tackled similar challenges in 3D generation[7, 8, 28, 30, 33, 41] – using network models in 2D and then lifting some information to 3D, we introduce our approach to leverage off-the-shelf 2D models to perform the task in a zero-manner. This approach, which we named Reasoning3D, allows us to circumvent the limitations im-

posed by the scarcity of extensive 3D datasets and the high computational costs with its training-free property and 2D pre-training.

Specifically, our Reasoning3D approach involves rendering a 3D model from multiple viewpoints and applying a pre-trained reasoning segmentation network to each 2D view based on the given query input. By doing so, we generate segmentation masks and accompanying text explanations for each perspective. These individual masks and explanations are then fused to produce a comprehensive 3D segmentation mask (labels are assigned to the vertices of the 3D model). We have evaluated our approach in various models in the wild, both with and without textures. We have also tested our approach in existing open-vocabulary segmentation benchmarks, which validates the effectiveness of our approach.

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. We will release the implementation code and the benchmark code publicly to facilitate future research, with the hope that our initial step sets the stage for further innovation and refinement, and eventually bring us closer to a future where 3D computer vision systems are as versatile and perceptive as human cognition, capable of revolutionizing a myriad of applications across various fields.

Below in Tab. 1, we show how our approach differs from some existing LLM-based 3D segmentation method (including SQA3D [25], 3D-VisTA [45], ViewRefer [9], Point-Bind [10], 3D-OVS [20], OpenMask3D [29], PLA [6], OpenScene [26], Chat-3D [32], M3DBench [18], LLM-Grounder [36], 3D-LLM [11], LL3DA [3], PointLLM [34], PARIS3D [13]).

2. Method

As illustrated in Fig. 2, 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, 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.

2.1. Multi-View Image Rendering and Face ID Generation.

Human interaction with the 3D environment often involves dynamic exploration, amalgamating viewpoints from various angles to construct a cohesive 3D comprehension, rather than assimilating a 3D setting instantaneously. Our methodology advocates for 3D reasoning cultivated from multi-perspective imagery. This strategy also leverages the extensive 2D pre-

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	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
Conversation	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Task	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

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.

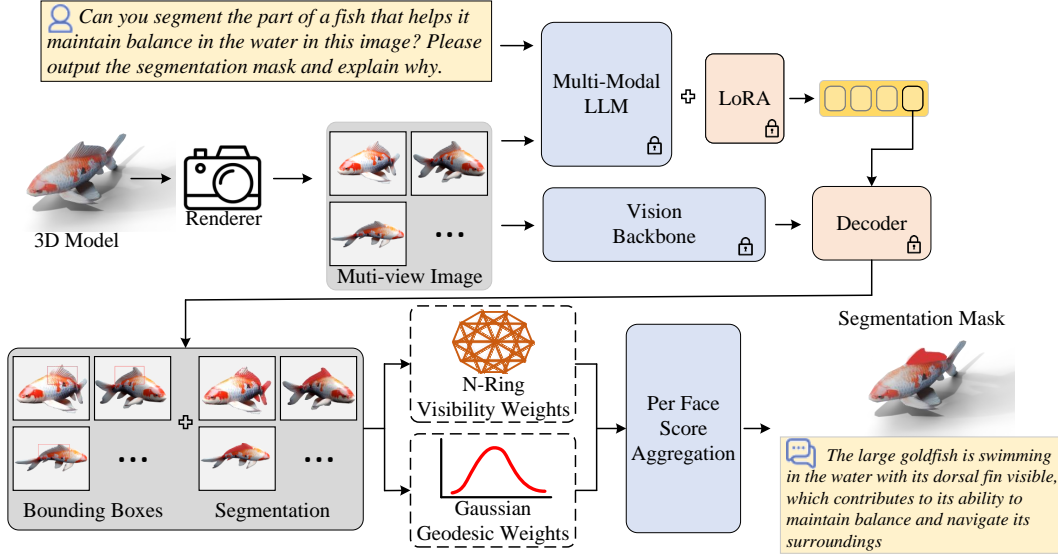


Figure 2. The overview of Reasoning3D. First, a 3D model represented by 3D meshes is fed into a renderer to obtain multi-view images. Then, each image goes through a vision backbone and a multi-modal LLM along with user input queries. The decoder decodes the final layer embedding which contains the extra 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.

training accessible in vision-language models, akin to prior methodologies capitalizing on pre-trained vision-language models for 3D visual tasks. The input for this process is mesh $F = \{f_n\}_{n=1}^N$, which is composed of N sets of faces f_n . During this rendering process, the 3D model is converted into multiple 2D images $X_{img} = \{x_i\}_{i=1}^8$ from various perspectives. In addition to generating these 2D images, the rendering process also produces corresponding face ids for each image. These face ids serve as a crucial link between the 2D images and the original 3D mesh. Specifically, they form a mapping matrix W_{pf} that connects each pixel $P = \{p_i\}_{i=k}^M$ in the 2D images to a specific face f_n on the 3D mesh, ensuring ensures that the 2D and 3D data remain accurately aligned. The process is formulated as follows:

$$P = W_{pf} \sum_{n=1}^k f_n \quad (1)$$

where P represents the pixels in the rendered images, and

f_n , represents the faces of the 3D mesh from the viewpoint. Denoted by k , the face id W_{pf} connects the pixels P in the rendered image and the faces of the 3D mesh f_n visible in the current view.

2.2. Reasoning and Segmenting with User Input Prompt

Unlike previous methods (CLIPSeg[23], LSeg[16], and GLIP[17]) which can handle open-vocabulary explicit prompt, our method aims to handle the implicit prompt such as "Can you segment the appropriate parts of the image containing a 'caged bird'?" Here, we leverage the recent advances of large foundation models to perform this multimodal reasoning task.

Following Lai[15], we extend the original LLM vocabulary with a new token, SEG , which denotes a request for segmentation output. Given the user-input prompt question $X_{question}$ and the input image X_{img} , these are input into the multimodal LLM F_{MM} , which subsequently outputs the

textual response Y_{answer} . The process is formulated as:

$$Y_{answer} = F_{MM}(X_{img}, X_{question}) \quad (2)$$

Next, the generation of segmentation masks corresponding to the input images involves a series of steps. Firstly, we extract the embedding \hat{E}_{answer} corresponding to the SEG_i token from the output answer textual Y_{answer} . This step enables us to capture information relevant to the segmentation task from the language prompts. Subsequently, we process \hat{E}_{answer} through the MLP γ projection layer to obtain the feature vector E_{answer} . Concurrently, utilizing a visual backbone network F_{vb} , we extract visual embeddings E_{img} from the visual inputs X_{img} . Finally, we feed both the feature vector E_{answer} and the visual embeddings E_{img} into the decoder F_{dec} . The decoder F_{dec} utilizes these features to generate the final segmentation mask M and confidence scores S_M for each mask. This yields segmentation results based on both language prompts and visual information, where each segmentation mask is accompanied by its respective confidence score and corresponding answer textual. The detailed structure of the decoder follows Segment Anything[14]. The process is formulated as follows:

$$\begin{aligned} E_{answer} &= \gamma \hat{E}_{answer} \\ E_{img} &= F_{vb}(X_{img}) \\ M, S_M &= F_{dec}(E_{answer}, E_{img}) \end{aligned} \quad (3)$$

2.3. Mask Fusion and Refinement in 3D

The obtained 2D segmentation mask eventually needed to be fused in 3D space to obtain the desired 3D segmentation result. We find that the result from directly merging the multi-view segmentation may not be coherent and high-quality due to the accumulated error and lack of comprehensive multi-view 3D information. Therefore, we designed a multi-stage fusion and refinement mechanism to fully exploit the semantic information and viewpoint information to obtain better 3D segmentation results.

First, we use the top-k method to filter the generated masks to reduce errors in 2D segmentation. Specifically, if the area difference between two masks is greater than a certain threshold T , we select $k=1$, indicating that is the mask (the most salient part) we want, we generate a bounding box that fits with the mask, otherwise, we select multiple masks and generate multiple bounding boxes. The filtered top-k masks S_M , the corresponding confidence scores S_M , and the face id is then used as the input to the fusion algorithm. We use the mapping relationship W_{pf} that maps the 2D image mask regions onto the faces of the 3D mesh, resulting in an initial segmented mesh. Note that only the masks within the generated bounding box are involved in the fusion process.

Following [1], 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.

Specifically, for each 2D mask M , we estimate its central face G_i^j , where i denotes the view and j denotes the mask within the view. For the 3D mesh under the current view, we retrieve the vertices of all faces corresponding to the current mask and compute their area-weighted average. This average point is then projected onto point F , and the face F_i^j containing this projection is identified as the central face for the current view and mask. Subsequently, we calculate the geodesic distance vector $d_i^j \in R^N$ from the central face G_i^j to F_i^j for all faces in $f \in F_i^j$. Here, N represents the number of faces in the mesh for the current mask.

$$f(x) = \begin{cases} gdist(G_i^j, f), & \text{if } f \in F_i^j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $gdist(\cdot, \cdot)$ represents the geodesic length between two faces computed using a heat method [4] on mesh F . The geodesic distance between mesh faces measures the path length along the surface from one face to another.

Next, we fit a Gaussian distribution on the distances and calculate the corresponding probability density values given the geodesic distances between each face and the uppercase face.

$$r_i^j = \xi[(d; \mu_i^j, (\sigma_i^j)^2), d \in d_i^j] \quad (5)$$

where μ_i^j and σ_i^j represent the mean and standard deviation of the distances to d_i^j , respectively. Subsequently, we tally the number of times n each face in the mesh is segmented in each view. Finally, we multiply the frequency of each face by the corresponding probability density, and then by the corresponding confidence score S_M , to obtain the final confidence for each mesh face.

However, using only the above method may result in insufficient weighting around the central face G_i^j , especially in regions where the average distances between faces are large. To address this issue, we compute its local neighborhood, where neighbors are determined by mesh connectivity: if two faces share at least one vertex, then face m is considered a neighbor of face n . To achieve this, we construct a q -rank neighborhood $N_q(n)$ ($q = 5$) as follows. For face $m \in F$, if there exists a path on the graph connecting m and n with at most q other vertices along the path, then we include face $n \in F$ in the neighborhood.

Finally, we adopt a global filtering using the calculated threshold. We filter out masked regions with low confidence scores. The threshold is the mean confidence score calculated for every face.

3. Experiment

3.1. Experimental Setup

3.1.1. Dataset and Evaluation Metric:

Since there are no existing zero-shot reasoning 3D segmentation benchmarks, we first evaluated the zero-shot open-vocabulary segmentation performance on the FAUST[2] benchmark (an open-vocabulary 3D segmentation benchmark) proposed in SATR[1]. We also validated the effectiveness of our method on reasoning 3D segmentation by our collected in-the-wild data from SketchFab. The FAUST dataset consists of manually annotated registered meshes of human body scans, re-meshed independently for each scan to contain approximately 20K triangular faces. We randomly collected samples from the 3D modeling website SketchFab and asked human volunteers to give implicit segmentation commands. For the evaluation metric, we employ the mean Intersection over Union (mIoU) for semantic segmentation as described in [39] for qualitative evaluation for each semantic category across all test shapes in open-vocabulary 3D segmentation. For the reasoning 3D segmentation, the result is visualized and rated by the user.

3.1.2. Implementation Details:

We utilized a single NVIDIA A100 GPU for each set of experiments. 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, maintaining consistency in viewpoints across all experiments. During the rendering process, we used 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.

3.1.3. Comparison Experiments for Open-Vocabulary Segmentation:

Since there is no existing reasoning 3D segmentation approach that can be compared, we first compared our method with existing open-vocabulary 3D segmentation models such as SATR[1] and 3DHighlighter[5] following the protocol in [1] but use the same rendering protocol in our method. As illustrated in Tab. 2 and Tab. 3, We show that though not designed for open-vocabulary segmentation tasks and without fine-tuning or specially designed structure, our method achieves competitive performance in the open-vocabulary segmentation benchmark. As shown in Fig. 3.

Model	Backbone	Arm	Head	Leg	Torso
3DHighlighter	CLIP	28.60	14.20	14.90	8.20
SATR	GLIP	61.54	76.89	87.41	52.32
Ours	LISA	52.73	70.61	80.10	46.69

Table 2. Performance on the coarse-grained semantic segmentation on FAUST dataset

Model	Arm	Belly button	Chin	Ear	Elbow	Eye	Foot	Forehead	—
3DHighlighter	18.39	1.99	0.46	0.72	0.08	0	20.81	0.70	—
SATR	24.23	22.00	26.53	34.55	33.67	22.55	75.20	30.35	—
Ours	21.65	1.87	3.36	10.61	18.18	2.77	71.85	6.56	—

Model	Hand	Head	Knee	Leg	Mouth	Neck	Nose	Shoulder	Torso
3DHighlighter	0.02	3.49	6.17	3.91	0.05	1.94	0.07	0.04	7.28
SATR	75.11	40.31	46.96	56.5	20.46	22.01	37.41	24.41	50.52
Ours	43.15	38.56	13.95	59.58	4.12	11.88	5.5	9.6	45.07

Table 3. Performance on the fine-grained semantic segmentation on FAUST dataset

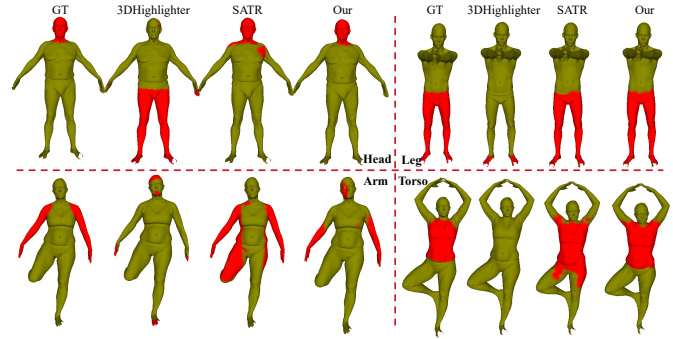


Figure 3. Qualitative results and comparison between our method and baseline method in FAUST benchmark. The segmented regions are shown in red.

3.2. Performance in Reasoning 3D Segmentation

A better property that our method has compared to existing open-vocabulary segmentation is that our method can use natural language as the input information. The LLM parses the natural language and gives the segmentation result directly, which enables a more natural and convenient computer-human interaction experience. An example is shown in Fig. 4. The models are from the FAUST dataset.

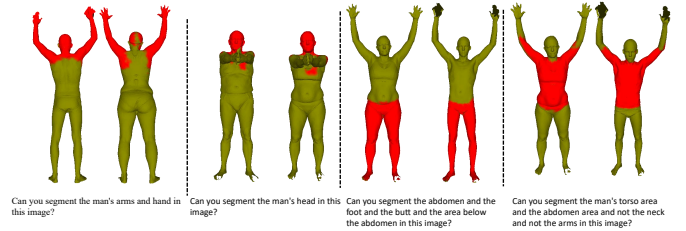


Figure 4. A natural language command can make the model segment the desired regions. The segmented regions are shown in red.

In the open-vocabulary segmentation, only explicit segmentation command is given, in which Reasoning3D's potential has not been fully exploited. We randomly collect 3D models from the 3D modeling website SketchFab perform the assessment with these in-the-wild 3D models and let hu-

man volunteers give "implicit" segmentation commands. Fig. 1 and Fig. 6 show some examples. The examples show that Reasoning3D has the capabilities to offer in-depth reasoning, 3D understanding, part segmentation, and conversational abilities. The model can output the segmentation masks and the explanation as we desire.

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

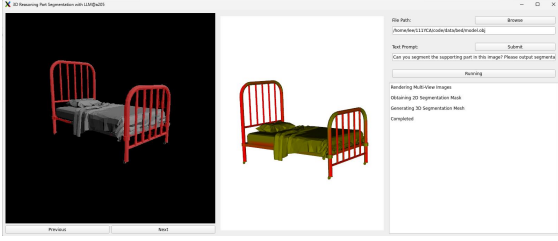


Figure 5. We offer a user-friendly interface designed for performance assessment, facilitating the easy upload of 3D models and prompts by users. It enables the swift acquisition of 3D segmentation outcomes. This tailored software is available as open-source.

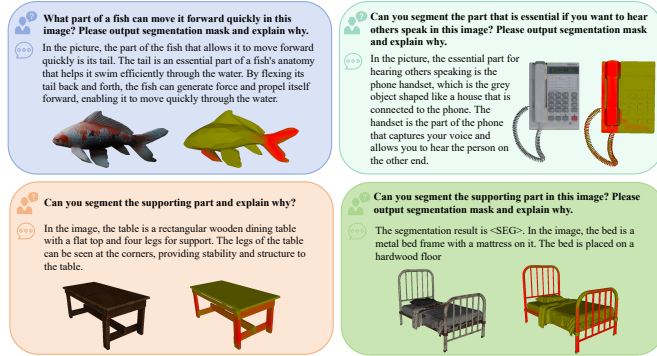


Figure 6. This figure shows Reasoning3D’s ability to segment 3D object parts (in a fine-grained manner) from in-the-wild samples, including real-world scanned data (samples are randomly collected from SketchFab). These examples highlight Reasoning3D’s advanced capabilities in in-depth reasoning, comprehensive 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.

4. Discussion and Limitations

This research represents preliminary findings in the task of reasoning 3D segmentation, and several areas require further exploration and validation. One major aspect is the need for comprehensive benchmarking to rigorously evaluate our method’s performance. Additionally, conducting user studies will provide valuable insights into the practical applicability and usability of our approach.

Our findings indicate that view information plays a critical role in the performance of 3D segmentation tasks. Optimizing view selection to align with the pre-trained vision encoder could significantly enhance outcomes. This suggests that a strategic approach to view selection is essential for leveraging the full potential of the pre-trained models.

The flexibility of our method is noteworthy, as the LVLM can perform zero-shot inference without the need for additional training. While fine-tuning with data could potentially improve performance, we observed that fine-tuning with a very small dataset might negatively impact the network’s generalization ability, sometimes resulting in poorer performance compared to fine-tuning. It is also worth noting that our multi-view 2D segmentation and 3D projection method can be applied to scenes, which will be beneficial for more real-world applications.

To foster further advancements and collaborative innovation in 3D reasoning and segmentation, we are releasing our code. We encourage the community to build upon our work and develop improved methods.

5. Conclusion

This paper introduces a new task: Zero-Shot 3D Reasoning Segmentation for part searching and localization within objects. This new approach moves beyond the constraints of traditional category-specific 3D semantic segmentation, 3D instance segmentation, and open-vocabulary 3D segmentation. We have developed Reasoning3D, a simple yet effective baseline method that can understand and perform complex commands to segment specific parts of 3D meshes with contextual understanding and reasoned outputs for interactive segmentation.

Reasoning3D leverages pre-trained 2D segmentation networks in conjunction with Large Language Models (LLMs) to interpret user queries in a zero-shot manner. Previous studies have shown that extensive pre-training equips foundational models with a broad understanding of the world, enabling them to process complex commands. Our method harnesses this capability, allowing for effective 3D segmentation with limited 3D datasets, making it a resource-efficient solution.

Our experiments demonstrate that Reasoning3D is generalizable and capable of accurately localizing and identifying parts of 3D objects based on implicit textual queries. This includes both articulated 3D objects and real-world scanned data. Additionally, our method can produce natural language explanations for the segmented 3D models and their components. The training-free nature of our 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, part assembly, autonomous driving, augmented and virtual reality (AR/VR), and medical fields.

We have released the code, model weights, deployment guide, and evaluation protocol to encourage further innovation and collaboration.

References

- [1] Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. SATR: zero-shot semantic segmentation of 3d shapes. In *ICCV*, pages 15120–15133. IEEE, 2023. 4, 5
- [2] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: dataset and evaluation for 3d mesh registration. In *CVPR*, pages 3794–3801. IEEE Computer Society, 2014. 5
- [3] 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 *CVPR*, pages 26418–26428. IEEE, 2024. 2
- [4] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. The heat method for distance computation. *Commun. ACM*, 60(11):90–99, 2017. 4
- [5] Dale Decatur, Itai Lang, and Rana Hanocka. 3d highlighter: Localizing regions on 3d shapes via text descriptions. In *CVPR*, pages 20930–20939. IEEE, 2023. 5
- [6] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: language-driven open-vocabulary 3d scene understanding. In *CVPR*, pages 7010–7019. IEEE, 2023. 2
- [7] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *3DV*, pages 1–11. IEEE, 2022. 2
- [8] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panopticnerf-360: Panoramic 3d-to-2d label transfer in urban scenes. *CoRR*, abs/2309.10815, 2023. 2
- [9] 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 *ICCV*, pages 15326–15337. IEEE, 2023. 2
- [10] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *CoRR*, abs/2309.00615, 2023. 2
- [11] 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. In *NeurIPS*, 2023. 2
- [12] R. Kenny Jones, Aalia Habib, and Daniel Ritchie. SHRED: 3d shape region decomposition with learned local operations. *ACM Trans. Graph.*, 41(6):186:1–186:11, 2022. 2
- [13] Amrin Kareem, Jean Lahoud, and Hisham Cholakkal. PARIS3D: reasoning-based 3d part segmentation using large multimodal model. In *ECCV*, pages 466–482. Springer, 2024. 2
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *CoRR*, abs/2304.02643, 2023. 4
- [15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. In *CVPR*, pages 9579–9589. IEEE, 2024. 2, 3
- [16] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*. OpenReview.net, 2022. 3
- [17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pages 10955–10965. IEEE, 2022. 3
- [18] 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. *CoRR*, abs/2312.10763, 2023. 2
- [19] 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. *CoRR*, abs/2110.03888, 2021. 2
- [20] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El-Saddik, Christian Theobalt, Eric P. Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. In *NeurIPS*, 2023. 2
- [21] Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *CVPR*, pages 21736–21746. IEEE, 2023. 2
- [22] 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. *CoRR*, 2023. 2
- [23] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7076–7086. IEEE, 2022. 3
- [24] Yecheng Lyu, Xinming Huang, and Ziming Zhang. Learning to segment 3d point clouds in 2d image space. In *CVPR*, pages 12252–12261, 2020. 2
- [25] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *ICLR*. OpenReview.net, 2023. 2
- [26] Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas A. Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824. IEEE, 2023. 2
- [27] 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. In *NeurIPS*, 2023. 2

- [28] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*. OpenReview.net, 2024. [2](#)
- [29] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *NeurIPS*, 2023. [2](#)
- [30] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from A single image with diffusion prior. In *ICCV*, pages 22762–22772. IEEE, 2023. [2](#)
- [31] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024. [2](#)
- [32] 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. *CoRR*, abs/2308.08769, 2023. [2](#)
- [33] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. In *ICCV*, pages 2383–2393. IEEE, 2023. [2](#)
- [34] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, pages 131–147. Springer, 2024. [2](#)
- [35] Yuheng Xue, Nenglu Chen, Jun Liu, and Wenyun Sun. Zeros: High-quality cross-modal knowledge transfer for zero-shot 3d part segmentation. *CoRR*, abs/2311.14262, 2023. [2](#)
- [36] Jianing Yang, Xuwei 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 *ICRA*, pages 7694–7701. IEEE, 2024. [2](#)
- [37] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *CoRR*, abs/2312.17240, 2023. [2](#)
- [38] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *CVPR*, pages 20612–20622. IEEE, 2024. [2](#)
- [39] Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, and Kai Xu. Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *CVPR*, pages 9491–9500, 2019. [2](#), [5](#)
- [40] Fenggen Yu, Yiming Qian, Francisca Gil-Ureta, Brian Jackson, Eric P. Bennett, and Hao Zhang. HAL3D: hierarchical active learning for fine-grained 3d part labeling. In *ICCV*, pages 865–875. IEEE, 2023. [2](#)
- [41] Shangzhan Zhang, Sida Peng, Tianrun Chen, Linzhan Mou, Haotong Lin, Kaicheng Yu, Yiyi Liao, and Xiaowei Zhou. Painting 3d nature in 2d: View synthesis of natural scenes from a single semantic mask. In *CVPR*, pages 8518–8528. IEEE, 2023. [2](#)
- [42] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative tokens. *CoRR*, abs/2310.02239, 2023. [2](#)
- [43] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llaifs: When large language models meet few-shot segmentation. In *CVPR*, pages 3065–3075. IEEE, 2024. [2](#)
- [44] 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. In *CVPR Workshops*, pages 1624–1633. Computer Vision Foundation / IEEE, 2025. [2](#)
- [45] 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 *ICCV*, pages 2899–2909. IEEE, 2023. [2](#)