Geometric Shape Reasoning for Zero-Shot Task-Oriented Grasping

Samuel Li, Sarthak Bhagat, Joseph Campbell, Yaqi Xie, Woojun Kim, Katia Sycara, Simon Stepputtis

***Abstract*— We present a novel zero-shot task-oriented grasping method leveraging a geometric decomposition of a target object into shape primitives that we represent in a graph structure. Our approach employs only the object name and task along with this graph to engage the commonsense reasoning capabilities of large language models to dynamically assign semantic meaning and subsequently task suitability to each decomposed part. Through extensive real-world robotic experiments, we demonstrate that our approach is capable of identifying the correct part in 92% and successfully lifting the object in 82% of the tasks we evaluate. Additional videos, experiments, code, and data are available on our project website: `https://shapegrasp.github.io/`.**

I. Introduction

Interacting with novel objects in unstructured environments, such as households, is an essential skill for robots operating in the real world. In particular, grasping objects in a way that facilitates a certain task, *task-oriented grasping*, requires a system to not only detect an object but also to reason over the utility of its parts. For example, when picking up a hammer with the goal to "hand it over" (see Fig. 2, left), the robot should grasp the hammer by the head to promote ease and safety for the human receiving it. Large Language Models (LLMs) provide the capability of such commonsense reasoning and can be utilized for task-oriented grasping with only a minimal set of contextual information, namely the object's name and desired task [1], [2].

However, zero-shot task-oriented grasping remains challenging, particularly since current approaches are computationally expensive and may require additional information for high performance, such as object part names [1], [2], limiting zero-shot performance. Other techniques based on semantic knowledge graphs [3], [4], [5], shape segmentation techniques [6], and physics simulators [7] have been considered. These approaches are limited in their ability to generalize to unseen objects due to requiring computationally expensive training. Additionally, LLMs and VLMs have successfully been applied to the robotics domain [8], [9], [10], [11], [12], including for task-oriented grasping [2], [13]. In such context, LLMs have been integrated into planning procedures in various ways, such as providing semantic knowledge [8] and performing complex reasoning in the form of an inner monologue [10]. Despite the benefits of utilizing LLMs and VLMs, which include no need for additional training and providing commonsense reasoning capability, naive utilization of LLMs and VLMs still have limitations stemming from their inherent shortcomings, such as indecisiveness, lack of domain knowledge, hallucination, and the black-box problem [14]. We address these limitations by infusing the LLM with a symbolic graph representing a target object's geometric composition. The infusion of such structured knowledge has been shown to be effective in preventing LLMs from deviating into the realm of fictitious information, thereby ensuring a connection to factual data [15], [16], [14], [17] and enhancing reasoning capabilities for task-oriented grasping.

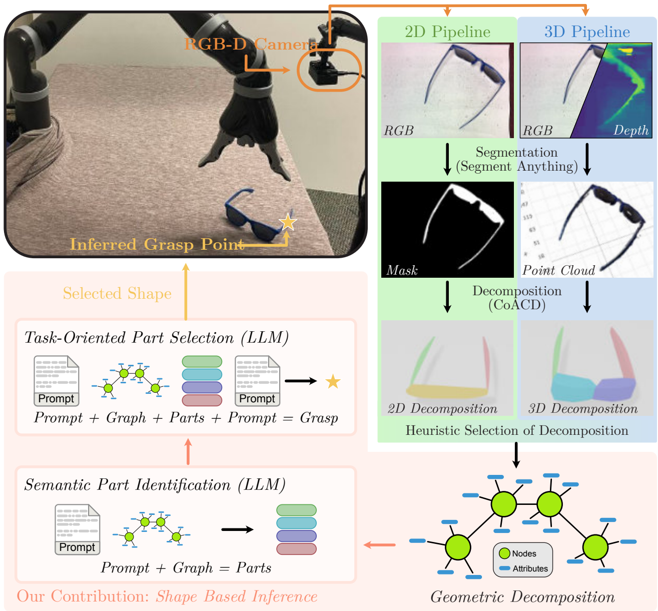

Fig. 1: **The `ShapeGrasp` Pipeline**: Given a target object, our RGB+D-based approach decomposes the object into basic convex parts. We propose a heuristic approach to select an appropriate decomposition that we then convert into a shape graph, allowing an LLM to utilize its commonsense reasoning to determine part semantics and task suitability.

In this extended abstract, we propose `ShapeGrasp`, **a robust and efficient task-oriented grasping framework based on representing a target object's decomposed convex parts in a symbolic graph that facilitates shape-based semantic part reasoning using LLMs**, inspired by the human ability to analyze a novel object's geometry, relating it to prior knowledge, and inferring part utilities [18] to identify a suitable part to grasp for an intended task.

II. Shape-Based Grasping

`ShapeGrasp`, our approach $\boldsymbol{g}, \theta = f_{SG}(\boldsymbol{I})$ to zero-shot task-oriented grasping, utilizes a passive monocular

All authors are with the Robotics Institute, Carnegie Mellon University. `{swli, sarthakb, jacampbe, yaqix, woojunk, sycara, sstepput}@andrew.cmu.edu`

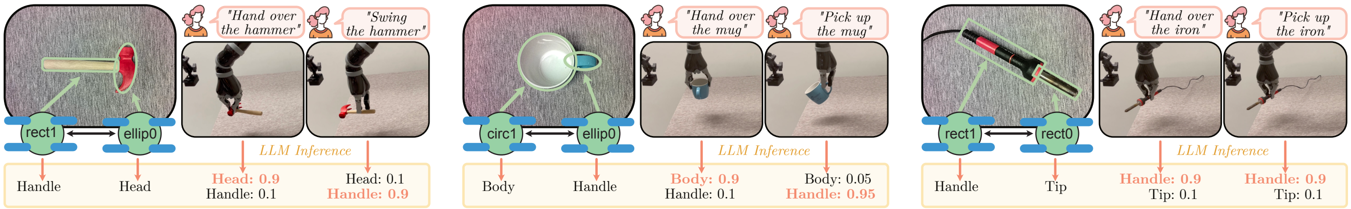

Fig. 2: Different resulting grasps given our shape-based inference pipeline. Parts in orange-boldface are ultimately grasped. Green circles and blue lines represent the part-graph decomposition with each entity's associated attributes.

RGB+D image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times C}$, to predict a task-oriented grasp location $\boldsymbol{g} \in \mathbb{R}^3$ and rotation $\theta \in \mathbb{R}^1$. The function $f_{SG}(\dots)$ represents the ShapeGrasp pipeline (visualized in Figure 1), composed of the following modules: Starting with a single static RGB+D image, we use Segment Anything Model (SAM) [19] to obtain the object's segmentation mask and point cloud. We then employ CoACD [20] in 2D (from the masked image) and 3D (from the masked point cloud) mode to decompose the target object into convex parts. At this stage, we develop an automatic heuristic to set decomposition thresholds and select between the 2D and 3D modes to identify a decomposition that results in parts suitable for semantic reasoning (see project website for more details). Leveraging the resulting decomposition, we present a novel approach to construct a graph-based representation of the object using each part's shape approximation, geometric attributes, and spatial relationships. We then use an LLM for multi-step reasoning over this structure, inspired by Chain of Thought [21], to 1) assign semantic meaning to each part of the object and 2) reasons over utility of each part to select the most suitable part to grasp for the task (see Sec. II-B).

A. Structured Visual Object-Graph Representations

After selecting an appropriate decomposition $\mathcal{C}^*$, we project it back onto the original input image $\boldsymbol{I}$ and create an object-graph $\mathcal{G}$ describing the composition of the target object. Each decomposed part is represented as a node in the graph accompanied by attributes derived from the segmented image. The primary attribute is an approximate shape primitive, chosen from an isosceles triangle, rectangle, circle, or ellipse. To select the appropriate shape, we approximate each convex hull with a simplified polygon given a pre-defined approximation threshold ε. The resulting points from this simplification dictate the fitted shape as follows: **Isosceles triangles** are formed by modifying any three points to equalize the leg lengths and adjust the base. **Rectangles** are formed by finding the rotated rectangle of the minimum area enclosing part. **Circles** are classified checking if the ratio of a part's area to that of its bounding circle is greater than 0.9. **Ellipses** are formed by fitting the part inside a rectangle (see above) if the fit reduces errors further. Each part is represented as a node in the graph $\mathcal{G}$ with attributes shape, aspect ratio, angle, centroid, and area calculated from the masked image. Part color is also included, obtained by bucketing the RGB color spectrum based on the standard 16 web colors and selecting the most prevalent color. Edges within graph $\mathcal{G}$ are drawn to connect nodes whose convex parts share boundaries or intersect in the decomposition.

B. Grasp Inference through Shape Reasoning

To determine a task-oriented grasp, we propose to leverage the commonsense knowledge encoded in LLMs across two interaction stages to reason about each part's semantic meaning and task utility. We leverage a prompt template that depends only on the graph $\mathcal{G}$, target object and task. We utilize TypeChat [22] to ensure the intended output structure.

1) Semantic Part Identification: In the first stage, the LLM is instructed to reason about the nodes in the graph and what semantic part (e.g., "handle" and "blade", for a knife) each may represent in the target object. To conduct this reasoning, we first ask for an unstructured, free-form answer in which the LLM explicitly explains its thoughts. As a follow-up, the LLM is then tasked to assign a single semantic label to each graph node in a structured manner.

2) Task-Oriented Part Selection: After the semantic reasoning is complete, the LLM is instructed to reason about the task utility of each part, using the semantics assigned in the first stage in addition to the graph representation. Similar to the first stage, this is accomplished in two steps: a free-form reasoning and explanation stage that the LLM then uses to assign a final task-oriented suitability score to each node.

C. Selecting a Grasp Pose

Finally, we select the graph node with the highest task-score. To grasp the selected node, we calculate the 3D coordinates of the part segment's centroid using the depth information in input image $\boldsymbol{I}$. To consider rotations, we calculate the principle components of the segment in the point cloud and orient the gripper along the largest component.

III. Experiments

We evaluate our approach in real-world experiments and demonstrate its effectiveness in grasping 38 household objects covering 12 general categories and 49 tasks, inspired by the LERF-TOGO [2] dataset (see project page). Section III-A demonstrates our approach on a real-world robotic platform and compares it against state-of-the-art baselines followed by additional qualitative experiments. All experiments are conducted with a Kinova Jaco robotic arm equipped with a three-finger gripper and coupled with a fixed Oak-D SR passive stereo-depth camera for RGB and depth perception.

We employ three metrics to evaluate our approach ShapeGrasp: *Part Identification (Part ID)* measures the

Model	Part ID	Part Sel.	LS	Time
1 GraspGPT [23]	N/A	0.37	0.31	150
2 GPT4-Vision [24]	N/A	0.82	0.73	20
3 ShapeGrasp (Starling)	0.54 (0.63)	0.65	0.57	25
4 ShapeGrasp (GPT-4)	**0.84 (0.90)**	**0.92**	**0.82**	30

TABLE I: Results on ShapeGrasp compared to GraspGPT and GPT4-V baselines. "Part ID" is measured across all parts (and across only target parts) and "Time" is the typical inference time in seconds for each method.

accuracy of the semantic label assigned to the object parts. *Part Selection (Part Sel.)* quantifies the proficiency of our model to select the correct part to grasp for the task. *Lift Success (LS)* indicates the percentage of objects successfully lifted at the correct part for the given task.

A. Zero-Shot Task-Oriented Grasping

Results for ShapeGrasp on the evaluated metrics are shown in Table I. We employ two baselines for comparison against ShapeGrasp: **GraspGPT [23]**, which is a current state-of-the-art approach for zero-shot task-oriented grasping and **GPT4-V [24]**, a foundation model trained with internet-scale data with visual input modality, prompted with language instructions to select the correct task-oriented part.

Our empirical findings indicate a significant performance advantage of our method over GraspGPT [23] (55% and 51% for the "Part Selection" and "Lift Success" metrics, respectively; see rows 1 and 4 in Table I), underscoring the efficacy of our structured graph in conjunction with multi-step LLM reasoning. We note that GraspGPT depends on GraspNet [13] for grasp sampling, which may be inaccurate or fail on certain objects when the depth quality is noisy or poor, which may occur due to our static monocular depth camera. While GraspGPT is limited to tasks and objects related to previously known concepts, ShapeGrasp demonstrates robustness to the same noisy depth inputs, while featuring zero-shot and being more lightweight than GraspGPT (see "Time" in Table I).

An important hypothesis that motivates our image-to-graph construction is that directly processing object part features and spatial relationships and providing this information in a structured way for LLM reasoning is more robust and performant than relying on VLMs for end-to-end reasoning. Though VLMs are considerably larger and more expensive models, performance on low-level features and relationships within parts of an object image may be unreliable and subject to hallucinations [25]. To directly compare our graph-construction and reasoning pipeline to a VLM, we establish a privileged GPT4-Vision baseline that directly uses the same heuristic-selected object segmentations to select a part to grasp. This baseline is grounded by coloring each part and assigning them integer index labels for clarity. We confirm GPT4-Vision's capability to interpret segmented and grounded input object images through a series of questions and human-verified responses. We use the same method to determine the grasp pose for the GPT4-Vision selected part to ensure comparability. Our method shows significant success rate gains over GPT4-Vision, by 10% and 9% on the evaluation metrics (see rows 2 and 4 in Table I).

It is interesting to note that while ShapeGrasp using the heuristic decomposition achieves 0.92% "Part Selection" accuracy, the performance drops to 0.86% and 0.73% when 2D- and 3D-only decompositions are respectively used. This result demonstrates the efficacy of the heuristic and the flexibility of ShapeGrasp to dynamically adapt to settings where depth information may be low quality or unsuitable, such as with concave, reflective, or transparent surfaces, where the 2D pipeline excels due to the sole reliance on RGB data. The complexity and prevalent noise in real-world settings often necessitates a reliance on the 2D mode for accuracy and robustness, while leveraging depth in the 3D mode when deemed reliable and beneficial by the heuristic.

We further test the modularity of ShapeGrasp by evaluating the full pipeline using Starling [26], a much smaller and more efficient open-source LLM, as the inference backend instead of GPT-4. Performance across all metrics, while lower than the larger and more powerful GPT-4, remains meaningful and higher than the GraspGPT [23] baseline.

Qualitative Results. The flexibility and generalizability of ShapeGrasp allow us to explore more complex interactions by incorporating additional object and robot attributes. For example, Fig. 2 shows how LLM semantic reasoning over our shape graph enables effective execution of the "hand over" task. As the LLM is able to identify the part corresponding to an object's "handle", it can prioritize that part for the human in the "hand over" interaction. Additional attributes can also be included in an object-specific way; while both the "mug" and the "soldering iron" are given the attribute of being "hot", the task-oriented reasoning stage can make the commonsense inference that the level of heat and risk exhibited by these two objects differ dramatically. For the "hand over" task, while the mug is grasped by the hot body, which "minimizes the risk of spilling hot liquid and ensures a comfortable handover", the soldering iron is still grasped by the handle, which "positions the hot tip away from both the robot and the human" (see Fig. 2).

IV. Conclusion

In this work, we present ShapeGrasp, a novel approach that represents an object's convex decomposition as a graph of basic shapes, allowing an LLM to effectively perform fine-grained semantic and task suitability reasoning over each part to identify a task-oriented grasp. Through extensive experiments on real-world hardware, we demonstrated that our approach can efficiently utilize a single, static RGB+D camera image for zero-shot task-oriented grasping and outperform current state-of-the-art approaches.

Acknowledgements

We want to acknowledge the support from DARPA under grant FA8750-23-2-1015, AFOSR under grants FA9550-18-1-0251 and FA9550-18-1-0097, and ARL under grant W911NF-19-2-0146 and W911NF-2320007.

References

[1] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same object, different grasps: Data and semantic knowledge for task-oriented grasping," in *Conference on Robot Learning*, 2020.

[2] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *Conference on Robot Learning*, 2023.

[3] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same object, different grasps: Data and semantic knowledge for task-oriented grasping," in *Conference on robot learning*. PMLR, 2021.

[4] S. Bhagat, S. Stepputtis, J. Campbell, and K. Sycara, "Sample-efficient learning of novel visual concepts," in *Proceedings of The 2nd Conference on Lifelong Learning Agents*, 2023.

[5] S. Bhagat, S. Stepputtis, J. Campbell, and K. P. Sycara, "Knowledge-guided short-context action anticipation in human-centric videos," *ArXiv*, vol. abs/2309.05943, 2023.

[6] Y. Lin, C. Tang, F.-J. Chu, and P. A. Vela, "Using synthetic data and deep networks to recognize primitive shapes for object grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 494–10 501.

[7] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, pp. 202 – 216, 2018.

[8] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[9] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[10] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Conference on Robot Learning*. PMLR, 2023, pp. 1769–1782.

[11] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-conditioned imitation learning for robot manipulation tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 139–13 150, 2020.

[12] Y. Zhou, S. Sonawani, M. Phielipp, H. Ben Amor, and S. Stepputtis, "Learning modular language-conditioned robot policies through attention," *Autonomous Robots*, vol. 47, no. 8, pp. 1013–1033, 2023.

[13] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[14] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *ArXiv*, vol. abs/2306.08302, 2023.

[15] W. Ding, S. Feng, Y. Liu, Z. Tan, V. Balachandran, T. He, and Y. Tsvetkov, "Knowledge crosswords: Geometric reasoning over structured knowledge with large language models," *ArXiv*, vol. abs/2310.01290, 2023.

[16] F. Moiseev, Z. Dong, E. Alfonseca, and M. Jaggi, "SKILL: Structured knowledge infusion for large language models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[17] L. F. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling," *ArXiv*, vol. abs/2306.11489, 2023.

[18] H. P. O. de Beeck, K. Torfs, and J. Wagemans, "Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway," *Journal of Neuroscience*, vol. 28, no. 40, 2008.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[20] X. Wei, M. Liu, Z. Ling, and H. Su, "Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search," *ACM Transactions on Graphics*, vol. 41, no. 4, p. 1–18, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1145/3528223.3530103

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] Microsoft. [Online]. Available: https://microsoft.github.io/TypeChat/

[23] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automation Letters*, 2023.

[24] OpenAI, "Gpt-4 technical report," 2023.

[25] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *ICLR*, 2023.

[26] B. Zhu, E. Frick, T. Wu, H. Zhu, and J. Jiao, "Starling-7b: Improving llm helpfulness & harmlessness with rlaif," November 2023.