Combining VLM and LLM for Enhanced Semantic Object Perception in Robotic Handover Tasks

Jiayang Huang[1,4], Christian Limberg[2], Syed Muhammad Nashit Arshad[3], Qifeng Zhang[4], Qiang Li[1*]

***Abstract*— We are utilizing a combination of Large Language Model (LLM) and Vision Language Model (VLM) to perform a robot-to-human handover task with semantic object knowledge. Current object perception systems for this task often work with a fixed set of objects and primarily consider geometric properties, neglecting semantic knowledge about where or where not to grasp an object. By applying LLM and VLM in a zero-shot fashion, we demonstrate that our approach can identify optimal and semantically correct handover parts for both the robot and the human in this handover task. We validate our approach quantitatively across several object categories.**

I. INTRODUCTION

In recent years, we witnessed the remarkable advancements in the field of AI, particularly in the development of Large Language Models (LLMs) and Vision-Language Models (VLMs). LLMs, such as GPT-4 and BERT, have made significant strides in Natural Language Processing (NLP) tasks, enabling more sophisticated understanding and generation of human-like text. VLMs, such as CLIP, LlaVa and BLIP, combine visual and textual data to perform tasks like image captioning and visual question answering, effectively bridging the gap between language and vision.

Handover tasks are fundamental interactions in robotics, involving the transfer of objects between entities (robots or humans). These tasks necessitate precise coordination, comprehension of object properties, and situational awareness for both safety and efficiency.

A critical challenge for service robots assisting humans is delivering everyday objects. This requires the robot to grasp objects in a user-friendly manner for handover, allowing for seamless and appropriate human use. To achieve this, robots need a sophisticated understanding of object semantics and their intended human interaction. In this work, we explore how LLMs and VLMs can be leveraged to infer optimal grasping locations and execute successful handovers.

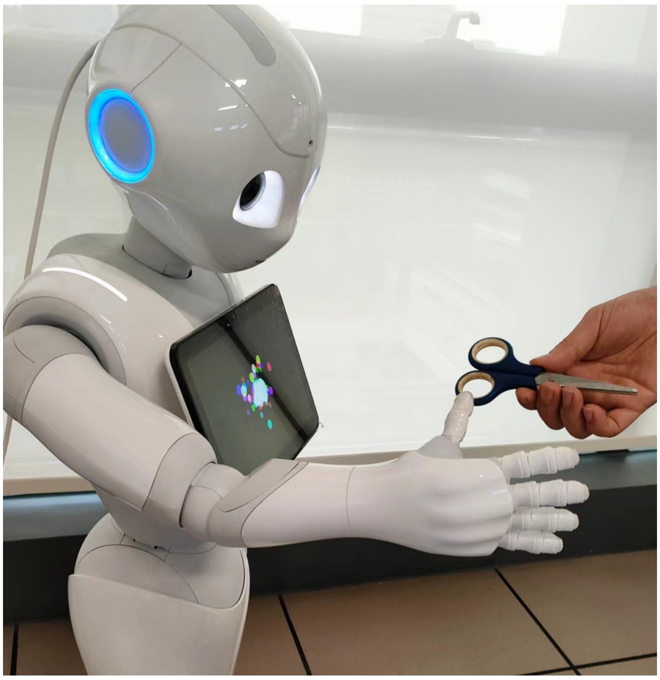


Fig. 1: Robotic handover tasks. The semantic object's information is provided by VLM and LLM.

Consider a typical use case: a human commands the robot to hand over a pair of scissors. The robot navigates the environment using a VLM to locate the queried object category in typical places. Upon finding the scissors, the robot queries an LLM to identify the best handover parts for both the robot and the human. This allows the robot to perform a semantic segmentation of the object and pass the information to a grasp planner, enabling an optimal handover. For instance, the robot might grasp the blades, allowing the human to comfortably and immediately grasp the handle, ready for use.

Our work focuses on leveraging these advancements to improve robotic perception and interaction, particularly in the context of handling everyday objects and tools. We introduce an effective approach that combines VLMs and LLMs for the perception part of the handover. By utilizing prompts, OWL-ViT and Llama3 models are capable of comprehending objects in the scene, specifically identifying appropriate grasping parts for handover tasks. These identified grasping parts are then subjected to instance segmentation using Grounded-SAM[1].

Our approach demonstrates the capability of integrating VLMs and LLMs to enhance robotic perception and interaction in dynamic environments. This integration allows for the efficient and accurate execution of handover tasks involving a variety of household objects and tools, without the need

*This work was supported by Chinese State Key Laboratory of Robotics

[1] College of Big Data and Internet, Shenzhen Technology University (SZTU), Guangdong, China

[2] National Institute of Informatics (NII), Japan

[3] Department of Robotics and Artificial Intelligence, National University of Sciences and Technology (NUST), Islamabad, Pakistan

[4] State key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences

*Corresponding authors, College of Big Data and Internet, Shenzhen Technology University (SZTU), Guangdong, China liqiang1@sztu.edu.cn

for additional training. The proposed method showcases a significant step forward in enabling robots to understand and interact with their surroundings more effectively, paving the way for more seamless human-robot collaboration. In summary, we make the following contributions:

- We propose a novel approach that utilizes foundation models for zero-shot identification of semantic object grasping parts.
- We validate our method by combining VLM with LLM, demonstrating its enhanced perception capabilities.
- We demonstrate that the presented approach can effectively handle a variety of tools and everyday objects for robotic handover perception without additional training.

II. RELATED WORK

A. Zero-shot Object Detection

Recent advancements in zero-shot object detection (ZSD) have shown significant progress, especially in handling object classes that were not labeled during training. One notable approach, GroundVLP[2], leverages visual grounding abilities from pre-trained models on image-text pairs and open-vocabulary object detection data. This method, which was presented at AAAI 2024, integrates visual-linguistic pre-training with open-vocabulary detection to effectively identify and localize objects without explicit training on those categories.

Another important development is DINO[3], which achieves state-of-the-art performance on the COCO dataset by incorporating improved anchor boxes and denoising techniques. This model emphasizes end-to-end training, robustness, and fast convergence, significantly enhancing zero-shot object detection capabilities.

Furthermore, some works like LAN-Grasp[4] utilizes OWL-ViT[5] as a zero-shot detector to detect best grasping parts. This model serves as a zero-shot, text-conditioned object detection system. It facilitate the detection of unseen object classes, proving superb effectiveness in robotic applications where a dynamic and versatile object handling is required.

B. LLMs and VLMs in Robotics

LLMs and VLMs play a critical and innovative role in the field of robotics. The paradigm shift [6] from training task-specific models to training a foundation model and just querying it represents a significant evolution in the field of machine learning. Traditionally, the development of machine learning models involved collecting domain-specific data, selecting and training a model for each particular task, and then deploying this model in a production environment. This approach, while effective, is resource-intensive and requires substantial effort for each new task.

In contrast, the foundation model paradigm leverages extensive pretraining on diverse, large-scale datasets to create models with broad generalization capabilities. These models, such as GPT-4[7] and BERT[8], can then be adapted to specific tasks either through minimal fine-tuning or by employing natural language prompts, thereby significantly reducing the need for task-specific training data and computational resources. This paradigm not only enhances efficiency and scalability but also enables rapid adaptation to new tasks, facilitating more dynamic and versatile applications across various domains. By leveraging these advanced models researchers have developed robots that can comprehend and respond to human instructions with greater accuracy. These models enable robots to interpret nuanced prompt inputs and execute corresponding actions, showcasing significant potential for application in specific tasks, including handover and grasping.

Our approach introduces the innovative use of LLMs for handover tasks, setting a new standard in robotic interaction. By integrating VLM (OWL-ViT) with LLM (Llama3), our method allows robots to deeply understand both the context of the task and the semantic objects. This combination ensures that robots can identify the most appropriate grasping parts for safe and efficient handover, even without prior training on specific objects.

C. Handover Tasks

Handover tasks have traditionally relied on extensive data collection, simulation and pre-training methods to achieve effective and safe transfers between robots and humans[9]. These methods often involved the creation of large, task-specific datasets, followed by the simulation of handover scenarios to train models that could generalize to real-world applications. Despite their effectiveness, these approaches required significant time and resources to ensure the models could handle the variability and unpredictability of real-world environments.

Recent advancements in robotic manipulation have led to a paradigm shift in how handover tasks are approached. Particularly, the integration of Transformer models and Diffusion models has revolutionized the way robots infer goal states and actions. Transformer models, known for their powerful sequence modeling capabilities, can process and understand complex task instructions and contextual information. Meanwhile, Diffusion models, which excel at generating high-quality data samples, enable robots to visualize and achieve desired goal states through iterative refinement.

In the domain of reasoning grasping, recent work by Jin et al. has introduced a novel task where robots generate grasp poses based on indirect verbal instructions or intentions[10]. This study presented an end-to-end reasoning grasping model that integrates a multi-modal LLM with a vision-based robotic grasping framework, setting a new benchmark for reasoning grasping tasks.

Additionally, the concept of integrating language models with robotic task planning has been further explored in the work on Vision-Language Interpreters for robot task planning [11]. This research proposes a Vision-Language Interpreter (ViLaIn) framework, which generates problem descriptions (PDs) from language instructions and scene observations, driving symbolic planners in a language-guided framework. The ViLaIn framework refines PDs using feedback from symbolic planners and has shown impressive accuracy in generating syntactically correct problems and valid plans.

These advancements highlight the significant potential of combining multi-modal LLMs with vision-based models and symbolic planners to enhance the adaptability and robustness of robotic systems in human-centric environments.

III. METHOD DESCRIPTION

In this section, we outline our novel approach for augmenting robotic perception and interaction, with a specific focus on handover tasks. Our method leverages the synergistic strengths of LLMs and VLMs to achieve precise and context-aware object manipulation. The following subsections outline the key components and steps involved in our methodology. Fig. 2 shows our work's pipeline.

A. VLM Module

The initial phase of our approach involves the utilization of the OWL-ViT model, a VLM, for object detection. This process begins with inputting a prompt and an image of the object into OWL-ViT. The prompt consists of the object's name, such as "scissors". The VLM processes the image and the prompt to identify and localize the object within the scene. OWL-ViT's capability to perform zero-shot object detection ensures that it can recognize and locate objects even without prior training on specific categories, thereby enhancing the versatility of our approach in dynamic environments.

B. LLM Module

Following object detection, we leverage the Llama3 model, a sophisticated LLM, to identify the optimal grasping points for both the robot and the human. The prompt scheme is designed to be compatible with the Llama3 model, ensuring seamless integration within our pipeline. The prompt is structured as follows:

- **Prompt:**
  "Imagine you are a highly advanced robot companion, designed to assist humans in various tasks."
  "Your goal is to hand over an object to a human in a way that ensures a comfortable and secure grasp, allowing them to use it efficiently."
  "Your task is to determine the optimal part for you to grasp the object and the corresponding part on the object where the human should receive it for a smooth handover."
  "Please respond in the following format: ['Object Part Robot Grasping', 'Object Part Human Receiving']. For example, if the object is a scissors, you might respond with ['blade', 'handle']."

Upon processing this prompt, the Llama3 model generates a response indicating the most appropriate handover parts for the object, tailored to facilitate a smooth and effective handover. This step ensures that the robot can comprehend the semantic properties of the object and make informed decisions about grasping and receiving parts.

C. Segmentation of Handover parts

The detected object and its identified grasping parts, as output by OWL-ViT and Llama3, are subsequently subjected to instance segmentation using Grounded-SAM. Grounded-SAM employs instance segmentation techniques to delineate the specific regions of the object identified by OWL-ViT and Llama3. This precise segmentation is essential for accurately executing the grasping and handover actions, ensuring that the robot can interact with the object in a manner that aligns with the human's expectations and ergonomic requirements.

IV. EXPERIMENTAL EVALUATION

In this section, we present the details of our experiments and results to demonstrate the effectiveness of our proposed method for robotic handover tasks.

A. Dataset

Affordance[12] refers to how the characteristics and shape of an object suggest how it will be used. This is particularly important for handover tasks, as a good grasping part helps the object to be passed more smoothly and used more fluently.

We collected 9 different objects that require such a semantic object understanding into a small dataset. These objects were selected to represent a range of common tools and utilities, each with distinct affordances that make them suitable for various handover scenarios.

The dataset includes objects such as scissors, claw hammer, knife, brush, flat-nose pliers, needle-nose pliers, goblet, screwdriver and wrench.

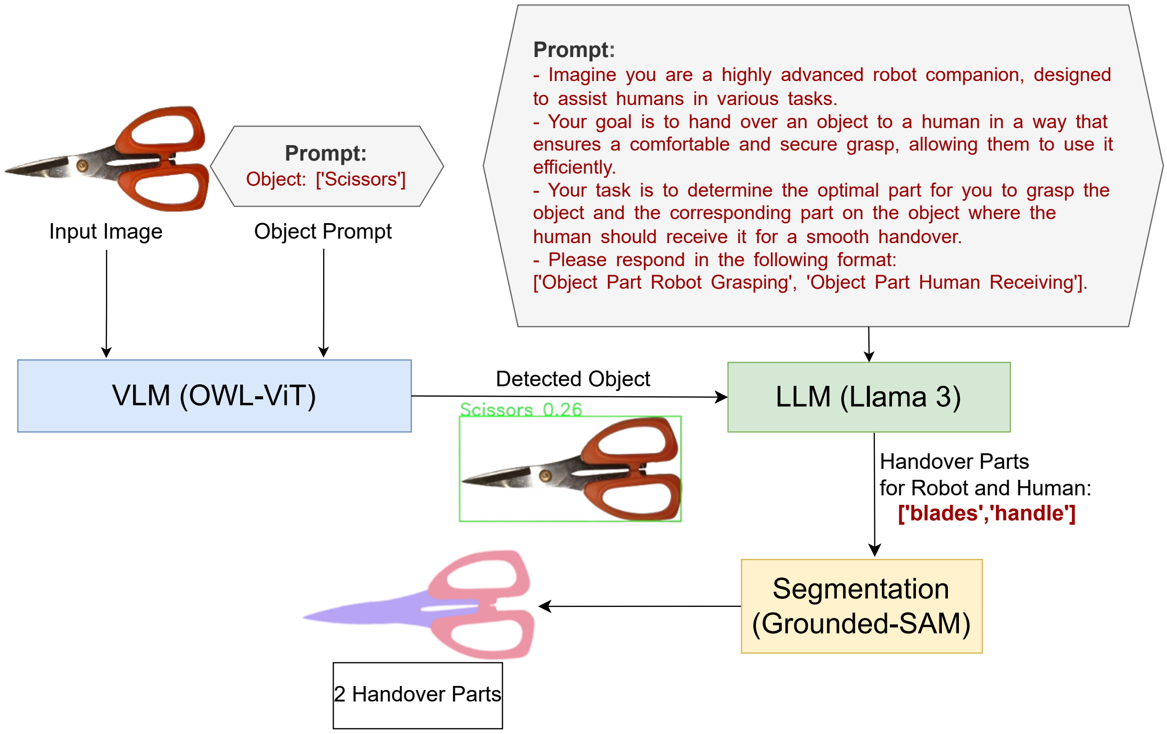
B. Experimental Setup

Our experimental setup includes a camera to capture images of objects. The whole system integrates OWL-ViT (for object detection), Llama3 (for identifying grasping parts), and Grounded-SAM (for segmenting handover parts). Experiments are conducted in a mixed environment of simulation and reality to verify the zero-shot ability and practicality of our method and lay the foundation for future practical applications.

C. Qualitative Results

Our qualitative results demonstrate the ability of our system to accurately detect objects and identify optimal grasping parts for handover. We present several examples to illustrate the performance of our approach.

- Scissors: The OWL-ViT model detected the scissors with a confidence level of 0.26. The Llama3 model identified the blade as the part for the robot to grasp and the handle for the human to receive.
- Claw Hammer: Detected with a confidence level of 0.08, the claw hammer's claws were identified for robot grasping, and the handle for human receiving.
- Knife: Detected with a confidence level of 0.12, the knife's hilt was identified for robot grasping, and the handle for human receiving.



Prompt:
- Imagine you are a highly advanced robot companion, designed to assist humans in various tasks.
- Your goal is to hand over an object to a human in a way that ensures a comfortable and secure grasp, allowing them to use it efficiently.
- Your task is to determine the optimal part for you to grasp the object and the corresponding part on the object where the human should receive it for a smooth handover.
- Please respond in the following format:
['Object Part Robot Grasping', 'Object Part Human Receiving'].

Fig. 2: The results perform a robot-to-human handover task by combining VLM (OWL-ViT) and LLM (Llama3) for object detection, handover parts identification, and instance segmentation, enabling precise and efficient handover.

For each object, the detected bounding box and segmented grasping parts are visualized in the results. As you can see in Fig. 3

D. *Quantative Results*

We present the quantitative results of our object detection and handover parts identification system. Table I shows the confidence levels for different objects along with their respective handover parts. The results highlight the varying confidence levels achieved for each object, indicating the performance of our detection model across different items.

TABLE I: Confidence levels for different objects and their inferred handover parts for grasping.

Object	**Confidence**	**Handover parts**
Scissors	0.26	blade, handle
Claw Hammer	0.08	claws, handle
Knife	0.12	hilt, handle
Brush	0.02	bristles, handle
Flat-nose Pliers	0.04	jaws, handle
Needle-nose Pliers	0.06	jaws, handle
Goblet	0.15	rim, handle
Screwdriver	0.10	shaft, handle
Wrench	0.23	barrel, handle

As shown in Table I, the model achieves varying confidence levels for different objects, reflecting its ability to accurately identify handover parts across a diverse set of items. These results demonstrate the robustness of our detection model in handling complex tasks involving semantic understanding of objects and their functional components.

E. *Discussion*

Our experimental results demonstrate that the proposed method can effectively handle a variety of objects. The integration of VLMs and LLMs allows for zero-shot detection and grasping part identification, significantly improving the robot's ability to interact with a wide range of objects without additional training. This capability is crucial for dynamic and unpredictable environments where robots need to adapt to new objects and scenarios efficiently.

V. CONCLUSIONS

In this paper, we have presented a novel approach that leverages the combined strengths of VLM and LLM to enhance robotic perception in the context of handover tasks. Our method employs the OWL-ViT model for object detection and the Llama3 model for identifying optimal grasping parts, ensuring smooth and safe handovers. By integrating Grounded-SAM for precise instance segmentation of grasp parts, our approach effectively handles a variety of objects without requiring additional training.

Our method's ability to understand the semantic context of objects allows it to perform zero-shot evaluations, enabling robots to recognize and interact with previously unseen objects based on their semantic attributes. This capability is crucial for handover tasks, where the robot must adapt to a wide range of objects and scenarios without extensive retraining.

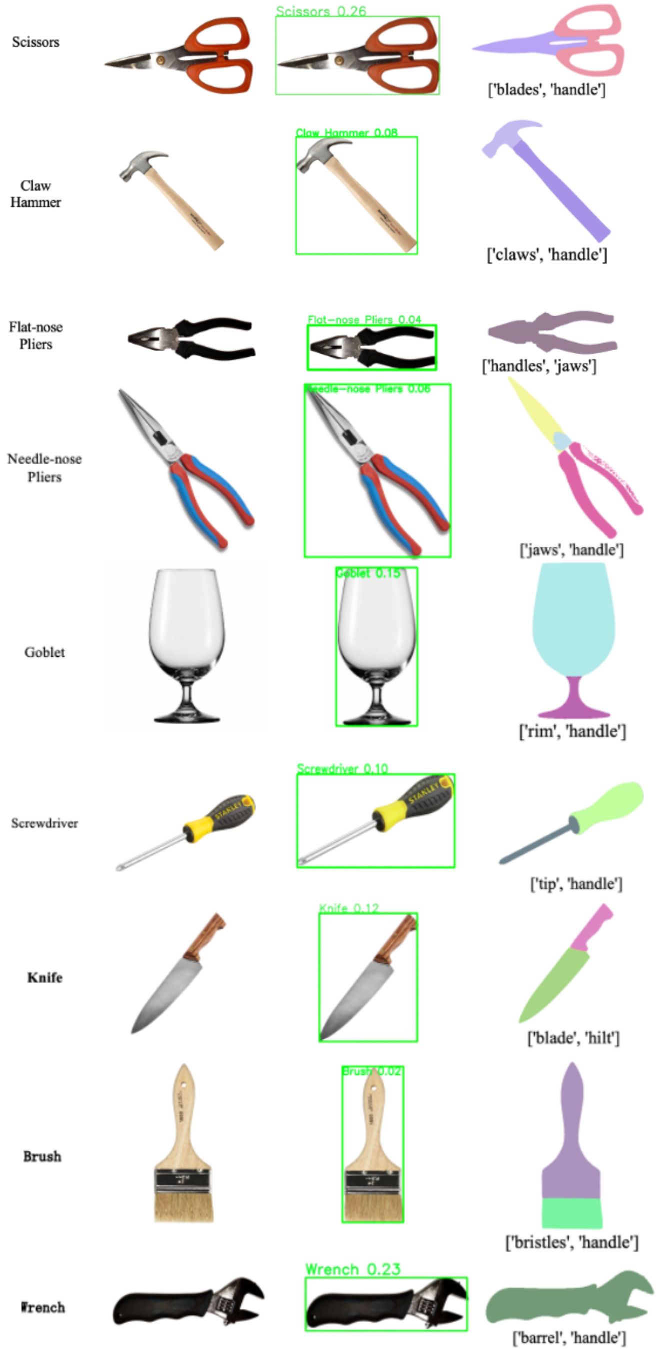


Fig. 3: These are the results of our work. From left to right, we have the original object image, the image with the detected bounding box by OWL-ViT, and the image segmented into two handover parts by Llama3 and Grounded-SAM.

Our experiments demonstrate the effectiveness of our method in both simulated and real-world environments, highlighting its ability to handle a wide range of objects in handover tasks efficiently and accurately. The semantic understanding capabilities of our approach allow the robot to adapt to new and unseen objects, making it versatile in dynamic settings. This work paves the way for more seamless human-robot collaboration by enabling robots to better interpret and respond to their surroundings. Future research can build upon our findings to further refine and expand the capabilities of robotic systems, emphasizing the significance of semantic understanding and zero-shot learning for advanced and flexible robotic applications.

VI. Future Work

In future work, we aim to address several challenges to enhance the performance and robustness of our system. We plan to explore more efficient methods for information transfer and optimization between multiple models to improve their collaborative performance. To stay current with emerging VLMs and LLMs, we will develop adaptive mechanisms that simplify the integration of new models, thus maintaining cutting-edge performance. Additionally, we will investigate strategies to reduce our reliance on pre-trained models, thereby enhancing the system's robustness across various scenarios and objects.

References

[1] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[2] Haozhan Shen, Tiancheng Zhao, Mingwei Zhu, and Jianwei Yin. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4766–4775, 2024.

[3] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[4] Reihaneh Mirjalili, Michael Krawez, Simone Silenzi, Yannik Blei, and Wolfram Burgard. Lan-grasp: Using large language models for semantic object grasping. *arXiv preprint arXiv:2310.05239*, 2023.

[5] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[6] Christian Limberg, Artur Goncalves, Bastien Rigault, and Helmut Prendinger. Leveraging yolo-world and gpt-4v lmms for zero-shot person detection and action recognition in drone imagery. In *ICRA 2024 First Workshop on Vision-Language Models for Navigation and Manipulation*, 2024.

[7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Valerio Ortenzi, Akansel Cosgun, Tommaso Pardi, Wesley P Chan, Elizabeth Croft, and Dana Kulić. Object handovers: a review for robotics. *IEEE Transactions on Robotics*, 37(6):1855–1873, 2021.

[10] Shiyu Jin, Jinxuan Xu, Yutian Lei, and Liangjun Zhang. Reasoning grasping via multimodal large language model. *arXiv preprint arXiv:2402.06798*, 2024.

[11] Keisuke Shirai, Cristian C Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. Vision-language interpreter for robot task planning. *arXiv preprint arXiv:2311.00967*, 2023.

[12] Paola Ardón, Maria E Cabrera, Eric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, Katrin S Lohan, and Maya Cakmak. Affordance-aware handovers with human arm mobility constraints. *IEEE Robotics and Automation Letters*, 6(2):3136–3143, 2021.