

Towards decoding Human Intent from Gaze: LLM-Driven Few-Shot Action Generation for Assistive Robotics in Complex Tasks

Abstract—Assistive robotic systems, such as exoskeletons and prosthetics, offer significant potential for human motor augmentation, restoration, and rehabilitation. However, their effectiveness depends on accurately interpreting human intentions through human-robot interfaces. Gaze-based interfaces offer a noninvasive way to read out human intentions but the Midas Touch problem and the complexity of tasks involving reaching-and-grasping make reliable decoding challenging. We introduce a framework using Large Language Models (LLMs) for gaze-driven assistive robotics, leveraging their generative and reasoning capabilities to infer the user’s intent from gaze, action history, and context, thereby suggesting appropriate assistive actions. Unlike hand-coded action-grammars, our approach generalizes across tasks through few-shot examples, obviating the need for predefined rules. Unlike the state-of-the-art generalist robot policy models that use explicit text inputs, our method derives intent directly from gaze signals, facilitating natural and intuitive human-robot interactions without a user interface. Evaluations on structured and human-in-the-wild tasks reveal GPT-4o achieves near-perfect accuracy, while Llama-3B handles simpler tasks, struggling with increasing complexity.

I. INTRODUCTION

Assistive robotic systems, such as exoskeletons [2] and prosthetics [3], hold significant promise for rehabilitating individuals with upper extremity motor impairment due to spinal cord injuries, amputations, degenerative diseases, and strokes [4], [5]. The effectiveness of these systems depends on human-robot interfaces that accurately interpret user intentions. Traditional neural interfaces [6], [7], despite their high data rates, require invasive procedures and extensive training, making wide adoption impractical [8]. Alternatively, gaze-based intention decoding retains goal-oriented functionality despite motor impairments [9], enabling "Zero UI" control where natural gaze guides the robot without explicit gestures.

Gaze-based intention decoding has been used in both low-level control, like navigating a wheelchair [10] or identifying robotic manipulator endpoints [11], [12], and high-level control for understanding abstract human intentions, facilitating actions such as "pouring a bottle" [1] or "navigating to the next room" [13]. High-level control is intuitive, decoding overall intentions ("drink sip of tea") instead of detailed physical actions, and is adaptable to diverse human interactions.

Challenges in gaze-based control include the Midas Touch Problem, where gaze does not always indicate action intent. Classifiers can identify when a user intends to interact with an object [1], with advanced methods distinguishing between inspection and interaction eye movements using deep learning [14]. Another challenge is predicting higher-level intents when observing an object. Studies suggest human actions follow stereotypical sequences describable by formal grammars

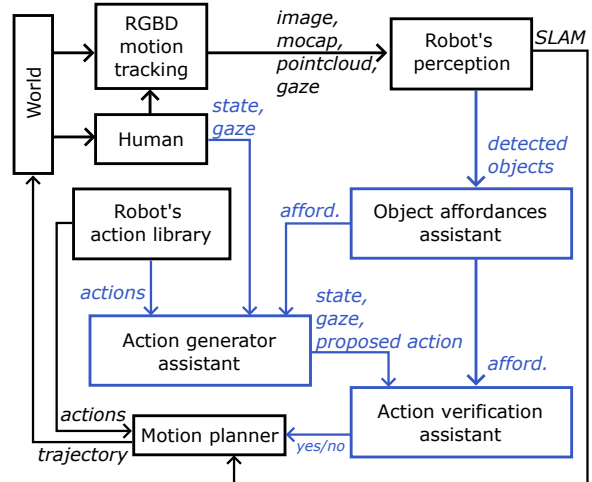
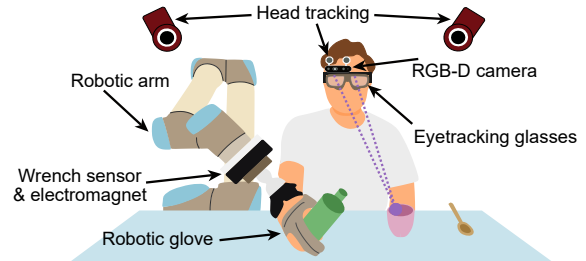


Fig. 1: Top: Robotic assistance setup [1], Bottom: System overview - blue boxes are LLM modules, arrows in blue indicate JSON inputs/outputs of LLM modules.

[15]–[19], which have been applied in robotics and reinforcement learning [20]–[23]. Action-grammars, analogous to language components, represent objects, affordances, and actions to describe behaviors in assistive robotics [1], [24]. However, generating these grammars is resource-intensive. We propose using Large Language Models (LLMs), which leverage extensive internet-scale knowledge, as substitutes for predefined action-grammars.

LLMs have shown promise in robot control, affordance classification, and human behavior modeling. RAIL classified object affordances [25], while other studies explored grounding affordances using Vision Transformers [26]–[28]. LLMs facilitate action planning [29], [30], and SayCan [31] combined LLMs with vision to select symbolic actions, extended by NLMaP [32] for unexplored environments. Further, VLA models predict end-effector trajectories [33]–[35]. LLMs also shown impressive performance in Theory of Mind tasks [36] and as human models for human-robot interactions [37].

```

<start> := {
  "Pick"
  (
    <pickable target> |
    <pourable source> {"Pour"} <pourable target> } |
    <mixable source> {"Mix"} <mixable target> }
  )
  ["Place"] <droppable target> ]
}
<pickable target> := <pourable source> | <mixable source> | "orange"
<pourable source> := "cup" | "bottle"
<pourable target> := "cup" | "bowl"
<mixable source> := "teaspoon" | "spoon"
<mixable target> := "cup" | "bowl"
<droppable target> := "table" | "bowl"

```

(...) - non-terminal, "..." - terminal, (...) - grouping, | - or
 [...] - optional (none or once), {...} - repetition (none or more)
 "Action", <affordance>, "object"

Fig. 2: Action-grammars of "tea-making" in Extended Backus-Naur Form

In our research, we replace hand-coded action-grammar controllers with LLMs (GPT-4o and Llama3.2-3B-Instruct) in a gaze-driven assistive robotic framework [1]. We test LLMs' ability to infer human intent from human's gaze, state and action history, generating object affordances and proposing assistive actions. These LLM-based controllers are validated across multiple assistive robotic tasks and natural behavior sequences, confirming their utility in enhancing robotic assistance.

II. METHODS

A. Background

The focus is on robotic decision-making; thus, only a brief outline of the experimental setup is presented (detailed in [1] and illustrated in Fig. 1-top). The setup includes a Bioservo Carbonhand soft robotic glove attached to a UR10 robot via an electromagnet and wrench sensor that automatically detaches the user if wrench exceeds safety limits. Pupil Labs' Core eye trackers are employed, along with an Intel RealSense D435i RGBD camera and passive optical trackers. OptiTrack Flex 13 cameras are utilized for head pose tracking. User gaze is mapped to the RGBD camera video stream, and objects are detected and labeled using Mask R-CNN. A machine-learned classifier determines gaze intention or inspection, which, combined with an action-grammar controller, informs the robot's assistive actions. Robot's trajectory planning considers user's ergonomic comfort (using Rapid Upper Limb Assessment [38]) and collision avoidance.

Action grammars, akin to context-free language grammars [39], validate human action sequences [16], [19]. For instance, the "making tea" sequence can be modeled using Extended Backus-Naur Form (EBNF) to define action-grammar production rules - see Fig. 2. These grammars monitor the user's state (e.g., "Pick bottle") and object affordances (e.g., "cup" is "pourable") to suggest and select the appropriate next action and state (e.g., "Pick bottle Pour cup").

B. LLMs as Action-Grammar Substitutes

Action-grammars require significant labor and data to develop, necessitating large annotated datasets. We hypothesize that Large Language Models (LLMs) can generate affordances and actions given their training on extensive text data involving scenes and human intents and actions. We propose three testable hypotheses to check whether LLMs can be used as action-grammar substitutes:

- LLM can generate exhaustive objects' affordances.
- LLM can infer human intent from human's action history, state and gaze and propose an appropriate action.,
- LLM can sanity-check its own decisions to ensure that the proposed action is safe for the user attached to the assistive robot.

Our novel assistive robotic control framework, depicted in Fig. 1, incorporates key components: "Object Affordance Assistant," "Action Generation Assistant," and "Action Verification Assistant," targeting the hypotheses above. We deploy two LLMs within this framework: GPT-4o and Llama3.2-3B-Instruct. GPT-4o is a comprehensive, multimodal model requiring internet access. Llama-3B runs locally on an RTX 4070 GPU but is limited to text processing. Unlike NLMMap [32], which relies on explicit text prompts for action selection, our method infers user intent directly from gaze data for natural and implicit control. We chose not to utilize vision+text input trajectory generation as in OpenVLA [34] for three reasons: avoidance of explicit text input, inclusion of a user physically attached to the robot - outside OpenVLA's training distribution - and the necessity for trajectory safety constraints not addressed by OpenVLA.

LLM assistants are assigned distinct roles using specific text instructions and few-shot prompt examples to guide their responses, ensuring fair comparison by providing identical instructions for both models. GPT-4o showed greater tolerance to instruction variations, leading us to focus on engineering instructions for Llama-3B. Each assistant received two examples: one analogous to test cases but with different items ("making a sandwich") and one real-world based ("making an omelette"). Due to space constraints, only a sample input/output is provided (full details in accompanying video). Inputs and outputs were in JSON to include task execution data and reasoning, fostering a chain-of-thought in LLMs. While data was shared among assistants, reasoning was not, ensuring independent, modular processing.

Object Affordance Assistant determines possible one-handed interactions with objects detected within the user's fovea or currently grasped. For example, a cup can be grasped, poured into, or stirred. This assistant identifies all feasible one-handed actions for objects in the gaze of the user.

- Input:

```

{"objects": ["bread", "butter", "table"
"cheese", "ham", "knife"]}

```

- Output:

```

[{"grasp": ["bread"]}, {"grasp": ["knife"]},
... (omitted for brevity) ...
{"scoop": ["knife", "butter"]},
{"reason": "(omitted for brevity)"}]

```

Action Assistant infers the human’s intended action using gaze, state, history, and available affordances. It chooses the most suitable action, restricted to the action library of the robotic arm: "pick," "place," "pour," and "mix." The assistant ensures compatibility by filtering out unfeasible actions and converting synonyms of these actions into the supported set.

- Input:

```
{
  "history": "grasp knife",
  "grasping": "knife",
  "gaze": "butter",
  "affordances": (see above)
}
```

- Output:

```
{
  "place": ["knife", "butter"],
  "reason": ["(omitted for brevity)"]
}
```

Action Verification Assistant assesses the feasibility of a proposed action given gaze, state, history, and available affordances. Although similar to the Action Assistant, it prioritizes safety to ensure that the proposed action is appropriate and secure.

- Input:

```
{
  "history": "grasp knife",
  "grasping": "knife",
  "gaze": "butter",
  "affordances": (see above)
  "action": {
    "place": ["knife", "butter"]
  }
}
```

- Output:

```
{
  "accept": True,
  "reason": ["(omitted for brevity)"]
}
```

III. EXPERIMENTS

This study evaluated the effectiveness of our LLM-based system via three experiments: affordance generation, assistive robotic tasks, and human-in-the-wild tasks.

Affordance Generation. We assessed whether the affordance generation assistant could comprehensively list object affordances (e.g., table, cup, teaspoon, bottle) for tasks like Pouring+Mixing describable by EBNF in Fig. 2. Through up to 10 re-prompts per object, and repeated 5 times each, we aimed to optimize initial prompt responses to produce exhaustive affordances without iteration.

Assistive Robotic Tasks. Using tasks from [1] described by "tea-making" action-grammars EBNF in Fig. 2, we evaluated task performance:

- Pouring task: pick up bottle, pour bottle into a cup, and place the bottle on the table.
- Mixing task: pick up a spoon, mix bowl with a spoon, and place the spoon on the table.
- Pouring+Mixing task: pick up bottle, pour bottle into a cup, place the bottle on the table, pick up a teaspoon, mix cup with a teaspoon, and place the teaspoon on the table.

Human-In-The-Wild Tasks. Tasks derived from the Human Ethome database [40], collected in living lab conditions, were used to test LLM reasoning over complex human activity sequences. Unlike the robotic tasks these tasks were not limited to the robot’s action library:

- Make and Eat cereal task: pick a box of cereal, pour cereal into the bowl, place box of cereal on table, pick

jug of milk, pour milk into bowl, place jug of milk on table, pick spoon, scoop cereal with spoon, eat, place spoon on table.

- Make and eat jam toast task: Open a jam jar, pick knife, scoop jam with knife, spread on toast with knife, place knife on table, pick toast, eat toast.

IV. RESULTS

Affordance Generation. Fig. 3 illustrates the average number of affordances generated across 5 trials with up to 10 re-prompts. The assistant consistently generated more affordances than the minimum needed, including additional logical actions like pushing or tipping, not originally in the action-grammars (Fig. 2). The number of affordances plateaued after initial prompts, with later responses often producing synonyms or affordances requiring unavailable objects. Crucially, both GPT-4o and Llama-3B generated essential affordances necessary for tasks within the first prompt.

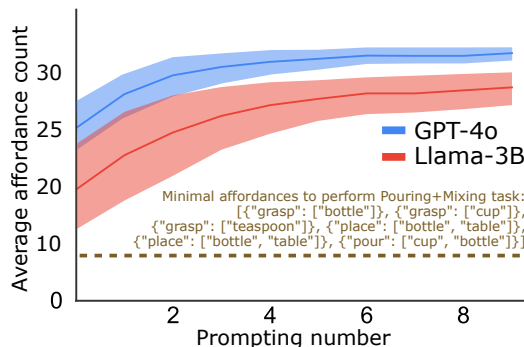


Fig. 3: Mean and std. cumulative number of affordances of objects for Pouring+Mixing task. The total number of generated affordances is always higher than minimal affordances necessary to perform the task. The total number of affordances also plateau at different levels for the two LLMs evaluated.

Assistive Robotic Tasks. Table I presents the performance results. GPT-4o achieved 100% accuracy in all tasks and successfully completed validation trials using the full robotic setup, as shown in Fig. 4 and the accompanying video. Conversely, Llama-3B showed lower accuracy, particularly in the Pouring+Mixing task, frequently rejecting correct actions or suggesting incorrect ones. However, the Llama-3B verification assistant often rejected incorrect proposals.

Human-in-the-Wild Tasks. Table II shows results for GPT-4o, which accurately identified and verified human actions in long sequences. Both in "eating-cereal" and "eating-jam-toast" tasks, the assistant proposed only correct actions, though 10% of these were wrongly rejected by the verification assistant, reflecting cautious behavior due to safety protocols tailored for the assistive environment.

V. DISCUSSION

Our LLM-driven method for human-robot assistive interaction shows promise, with a significant performance

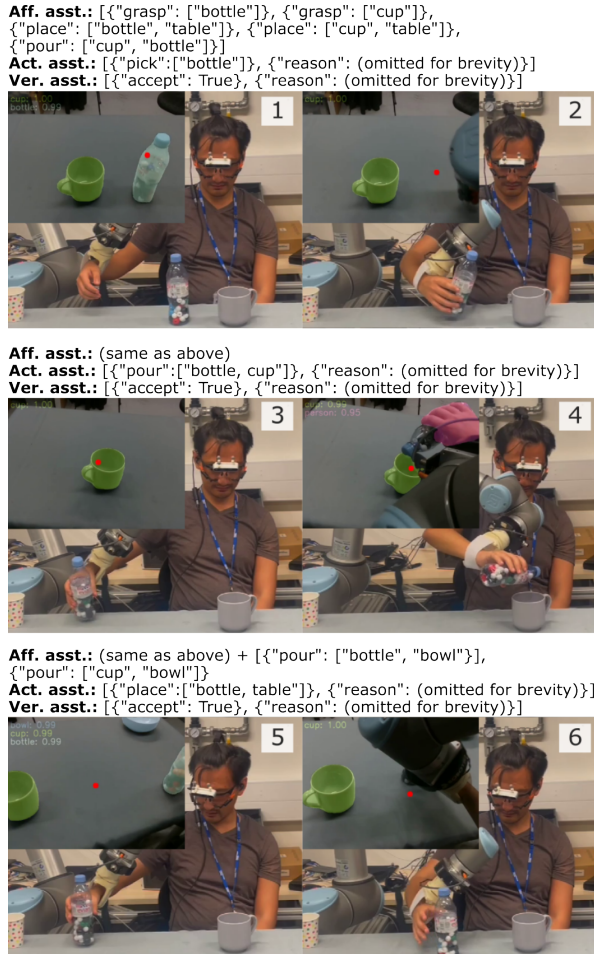


Fig. 4: Pouring task performed with GPT-4o assistants with a real world assistive robotic setup. Gaze intention is detected at 1, 3, 5 and corresponding affordance, action and verification assistant outputs are shown above subfigures. Assistive robotic actions are performed in 2, 4, 6.

disparity between GPT-4o and Llama-3B assistants. GPT-4o model effectively replaces action-grammar controllers by identifying object affordances and interpreting human intentions from actions, state, and gaze. Its consistent accuracy and contextually appropriate responses make it potentially suitable for real-world applications. Despite generating correct minimal affordances, Llama-3B often fails to select suitable actions, demonstrating flawed reasoning, such as misinterpreting object interactions. Its reasoning deteriorates with more objects, indicating scalability challenges.

Our approach differs from OpenVLA [34], which achieves a 70% success in similar tasks but is pre-trained on large robot demonstration datasets, unlike our few-shot prompting approach. GPT-4o handled lengthy human-in-the-wild tasks with 90% success, comparable to NLMap’s [32] 90% planning accuracy; while NLMap and OpenVLA uses text+vision input our approach uses human action history and gaze for, showcasing a subtle Theory of Mind [41].

In 10% of cases, GPT-4o’s verification assistant rejected a correct action in the human-in-the-wild tasks. While this

TABLE I: Assistive robotic tasks

Task	GPT-4o		Llama-3B	
	Accept	Reject	Accept	Reject
Pouring				
Correct Action	1	0	0.5	0.2
Incorrect Action	0	0	0.1	0.2
Mixing				
Correct Action	1	0	0.4	0.2
Incorrect Action	0	0	0.2	0.2
Pouring+Mixing				
Correct Action	1	0	0.2	0.15
Incorrect Action	0	0	0.25	0.4

TABLE II: Human-in-the-wild tasks

Task	GPT-4o	
	Accept	Reject
Make & eat cereal		
Correct Action	0.9	0.1
Incorrect Action	0	0
Make & eat jam toast		
Correct Action	0.9	0.1
Incorrect Action	0	0

may lead to user frustration due to the need for re-fixating, it does not cause a complete failure in executing an action sequence. Llama-3B proposed and accepted several incorrect actions; in these scenarios: actions outside the robot’s library would be ignored (for example "scoop" is not part of robot’s action library), while others nonsensical actions would fail during motion planning ("pick table" - would not pass motion planning) or would halted by wrench sensor’s safety limit.

Our results show that interpretation and reasoning about human action for robotic purposes is a non-trivial task for LLMs, that requires among the most powerful LLMs currently available to perform correctly with few-shot learning – despite the deceptively "simpler" structure action grammars when compared to the complexity of grammars of human written language.

VI. CONCLUSION

We have proven that LLMs can function as controllers for gaze-driven assistive robotics by effectively generating affordances, interpreting human intent from actions, state, and gaze, and proposing suitable assistive actions, hinting at an implicit Theory of Mind of user intentions. Compared to the action-grammar methods [1], our LLM-based approach generalizes across tasks with minimal adaptation through few-shot examples, offering more flexibility. In contrast, action-grammar-based methods require task-specific engineering.

Unlike models like NLMap [32] and OpenVLA [34] that rely on structured text input, our method uses neurobehavioral signals from eye tracking to derive user inputs naturally and dynamically. This approach adapts intuitively to real-time scenarios without needing explicit commands, making it suitable for applications in assistive robotics and human robotic augmentation [42].

REFERENCES

- [1] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," *ICRA*, 2019.
- [2] Z. Li, Z. Huang, W. He, *et al.*, "Adaptive impedance control for an upper limb robotic exoskeleton using biological signals," *IEEE Trans. on Industrial Electr.*, 2017.
- [3] D. Farina, N. Jiang, H. Rehbaum, *et al.*, "The extraction of neural information from the surface emg for the control of upper-limb prostheses: Emerging avenues and challenges," *IEEE Trans. on Neur. Sys. and Rehab. Eng.*, 2014.
- [4] J. Zariffa, A. Curt, M. C. Verrier, *et al.*, "Predicting task performance from upper extremity impairment measures after cervical spinal cord injury," *Spinal Cord*, 2016.
- [5] D. Cattaneo, I. Lamers, R. Berti, *et al.*, "Participation restriction in people with multiple sclerosis: Prevalence and correlations with cognitive, walking, balance, and upper limb impairments," *Arch. of Phys. Med. and Rehab.*, 2017.
- [6] A. B. Ajiboye, F. R. Willett, D. R. Young, *et al.*, "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: A proof-of-concept demonstration," *The Lancet*, 2017.
- [7] L. R. Hochberg, D. Bacher, B. Jarosiewicz, *et al.*, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, 2012.
- [8] T. R. Makin, F. De Vignemont, and A. A. Faisal, "Neurocognitive barriers to the embodiment of technology," *Nature Biomed. Eng.*, 2017.
- [9] W. W. Abbott and A. A. Faisal, "Ultra-low-cost 3D gaze estimation: An intuitive high information throughput compliment to direct brain-machine interfaces," *J. of Neur. Eng.*, 2012.
- [10] S. I. Ktena, W. Abbott, and A. A. Faisal, "A virtual reality platform for safe evaluation and training of natural gaze-based wheelchair driving," in *IEEE Neural Engineering (NER)*, IEEE, 2015.
- [11] P. M. Tostado, W. W. Abbott, and A. A. Faisal, "3d gaze cursor: Continuous calibration and end-point grasp control of robotic actuators," in *2016 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2016.
- [12] R. O. Maimon-Mor, J. Fernandez-Quesada, G. A. Zito, *et al.*, "Towards free 3d end-point control for robotic-assisted human reaching using binocular eye tracking," *ICORR*, 2017.
- [13] M. Subramanian and A. A. Faisal, "Natural gaze informatics: Toward intelligence assisted wheelchair mobility," in *Eye Tracking: Background, Methods, and Applications*, Springer, 2022.
- [14] P. Festor, A. Shafti, A. Harston, *et al.*, "Midas: Deep learning human action intention prediction from natural eye movement patterns," *arXiv:2201.09135*, 2022.
- [15] D. Haber, A. A. Thomik, and A. A. Faisal, "Unsupervised time series segmentation for high-dimensional body sensor network data streams," in *IEEE Body Sensor Networks (BSN)*.
- [16] D. Stout, T. Chaminade, A. Thomik, *et al.*, "Grammars of action in human behavior and evolution," *BioRxiv*, 2018.
- [17] R. T. Lange and A. Faisal, "Action grammars: A cognitive model for learning temporal abstractions," *CoRR*, 2019.
- [18] F. Wörgötter, F. Ziaetabar, S. Pfeiffer, *et al.*, "Humans Predict Action using Grammar-like Structures," *Scientific Reports*, 2020.
- [19] D. Stout, T. Chaminade, J. Apel, *et al.*, "The measurement, evolution, and neural representation of action grammars of human behavior," *Scientific Reports*, 2021.
- [20] K. Lee, T.-K. Kim, and Y. Demiris, "Learning action symbols for hierarchical grammar induction," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, IEEE, 2012.
- [21] K. Lee, Y. Su, T.-K. Kim, *et al.*, "A syntactic approach to robot imitation learning using probabilistic activity grammars," *Robotics and Autonomous Systems*, 2013.
- [22] R. T. Lange and A. Faisal, "Semantic rl with action grammars: Data-efficient learning of hierarchical task abstractions," *arXiv preprint arXiv:1907.12477*, 2019.
- [23] P. Christodoulou, R. T. Lange, A. Shafti, *et al.*, "Reinforcement learning with structured hierarchical grammar representations of actions," *arXiv preprint arXiv:1910.02876*, 2019.
- [24] R. Lioutikov, G. Maeda, F. Veiga, *et al.*, "Learning attribute grammars for movement primitive sequencing," *Int. J. of Rob. Research*, 2020.
- [25] C. Zhang, X. Meng, D. Qi, *et al.*, *Rail: Robot affordance imagination with large language models*, 2024. arXiv: 2403.19369 [cs.RO].
- [26] C. Chen, Y. Cong, and Z. Kan, *Worldafford: Affordance grounding based on natural language instructions*, 2024. arXiv: 2405.12461 [cs.CV].
- [27] G. Li, V. Jampani, D. Sun, *et al.*, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," *CVPR*, 2023.
- [28] G. Li, D. Sun, L. Sevilla-Lara, *et al.*, "One-shot open affordance learning with foundation models," *CVPR*, 2024.
- [29] I. Singh, V. Blukis, A. Mousavian, *et al.*, "Progprompt: Generating situated robot task plans using large language models," *ICRA*, 2023.
- [30] K. Lin, C. Agia, T. Migimatsu, *et al.*, "Text2motion: From natural language instructions to feasible plans," *Aut. Rob.*, 8 2023.
- [31] M. Ahn, A. Brohan, N. Brown, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *PMLR*, 2023.
- [32] B. Chen, F. Xia, B. Ichter, *et al.*, "Open-vocabulary queryable scene representations for real world planning," *ICRA*, 2023.
- [33] B. Zitkovich, T. Yu, S. Xu, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *CoRL*, 2023.
- [34] M. Kim, K. Pertsch, S. Karamcheti, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv:2406.09246*, 2024.
- [35] K. Black, N. Brown, D. Driess, *et al.*, " π_0 : A vision-language-action flow model for general robot control," *arXiv:2410.24164*, 2024.
- [36] J. W. Strachan, D. Albergo, G. Borghini, *et al.*, "Testing theory of mind in large language models and humans," *Nature Human Behaviour*, 2024.
- [37] B. Zhang and H. Soh, "Large language models as zero-shot human models for human-robot interaction," *IROS*, 2023.
- [38] L. McAtamney and N. Corlett, "Rapid upper limb assessment (rula)," in *Handbook of human factors and ergonomics methods*, 2004.
- [39] N. Chomsky, "Three models for the description of language," *IRE Trans. on Inf. Theory*, 1956.
- [40] I. Galea, "Towards the human ethome: Human kinematics study in daily life environments," Ph.D. dissertation, Sep. 2012.
- [41] A. I. Goldman *et al.*, *Theory of mind*. Oxford handbook of philosophy and cognitive science, 2012.
- [42] A. Shafti, S. Haar, R. Mio, *et al.*, "Playing the piano with a robotic third thumb: Assessing constraints of human augmentation," *Scientific reports*, 2021.