From Gaze to Action: Leveraging Affordance Grounding for Human Intention Understanding

Yiding Wang[1], Bukeikhan Omarali[1], and A. Aldo Faisal[1,2,3]

***Abstract*— We present an early-stage investigation into gaze-driven intention recognition for assistive robotics, with the goal of overcoming persistent challenges in the Midas Touch Problem—distinguishing intention gaze from inspection gaze—and Intention Decoding—inferring the user's goal from gaze. Our prior work addressed these challenges separately using supervised classifiers and action-grammars or LLMs, but limitations remain in generalizability and ambiguity resolution. Here, we propose a unified approach inspired by affordance grounding—the visual identification of object regions responsible for specific actions. We study whether humans fixate on affordance regions prior to interaction and whether this signal can be used to infer intention in real-world scenarios. Using large egocentric datasets, we analyze gaze-object-action relationships across time. We benchmark automated annotations against human-labeled data, assess the applicability of existing affordance models (e.g., LOCATE, OOAL) in egocentric settings, and explore models' capacity to resolve ambiguous intentions. Our work offers insights into integrating gaze, affordance, and language models for more robust human-in-the-wild intention decoding.**

I. INTRODUCTION

Assistive robotic technologies, including exoskeletons [1] and prosthetic devices [2], show considerable potential as rehabilitation tools for individuals with motor impairments. Motor dysfunction in the upper limbs is a prevalent condition resulting from spinal cord injuries, amputations, neurodegenerative diseases, and stroke [3], [4]. Impairments in upper limb function impose significant barriers to daily living and profoundly reduce quality of life.

The utility and adoption of assistive robotic systems depends on the development of safe and reliable human-robot interfaces capable of accurately inferring user intent given limited input modalities of disabled users. The human gaze provides a high-bandwidth, information-rich signal [5] that maintains goal-directed properties even in the presence of motor impairments and is inherently aligned with user intention. This supports the development of "Zero-UI" paradigms, wherein natural gaze behavior is directly interpreted to control robotic systems [6], [7], eliminating the need for deliberate gaze gestures or screen-based interactions.

Zero-UI gaze-based robotic control faces two principal challenges. The first is the Midas Touch problem [8], wherein not all objects or locations fixated by the user are intended to prompt robotic actions. The second challenge concerns Intention Decoding—the need to infer the user's higher-level goals associated with a physical object [6]. For instance, when a user fixates on a cup, it remains ambiguous whether the intended action is to grasp the cup, pour a liquid into it, or perform some other related task.

Current state-of-the-art approaches treat the Midas Touch problem and Intention Decoding as two separate tasks. Typically, models first classify fixations as either "intention" or "inspection" and subsequently contextualize the user's intent based on the gazed object, user state, and action history [6]. For example, a "intention"/"inspection" classification can be performed learning patterns in gaze sequences and object identities through supervised deep learning [8]. While effective in controlled laboratory environments, its generalization to real-world, "in-the-wild" human behavior remains limited. Intention Decoding problem can be partially resolved using action-grammars [6]. Analogous to natural language sentence structures: nouns, adjectives, and verbs [9]; action-grammars model interactions by linking objects, their affordances, and possible actions, thereby providing a structured representation of human behavior [10]. In assistive robotic applications, action-grammars describe the user's current state and valid potential future actions, enabling intention inference based on gaze information [6], [11]. However, these frameworks struggle when multiple plausible intentions coexist. For instance, if a user holds a spoon and gazes at a cup, it is equally plausible that they intend to place the spoon into the cup or to stir its contents—an ambiguity that existing models find difficult to resolve.

In this work, we present our methodology and early findings on whether an affordance grounding based approach could address both the Midas Touch and Intention Mapping challenges simultaneously.

Affordance grounding [12]–[15] is the process of identifying and localizing regions within an image that correspond to specific affordances and actions; for instance, the handle of a cup is the region associated with a "grasp" action. In robotic control, affordance grounding enables robots to localize areas appropriate for certain actions — such as where to grasp a cup given a high-level command "grasp cup." An object can have multiple separate regions of affordances - a cup can be grasped, poured into, drank from, etc.

Prior studies in human vision and action [16]–[18] demonstrate that humans use gaze to build and update internal maps and models of spaces that decay over time. Humans often fixate on a target object or interaction region immediately before precision actions to reduce uncertainty in their internal representations and complete the hand-eye coordination loop. This raises an interesting question: do humans specifically

[1]Brain & Behaviour Lab: Dept. of Computing, Imperial College London, UK; [2]Brain & Behaviour Lab: Dept. of Bioengineering, Imperial College London, UK; [3]Chair in Digital Health, Universität Bayreuth, Germany
Correspondence: a.faisal@imperial.ac.uk

Acknowledgements: UKRI Turing AI Fellowship (EP/V025449/1) to AAF.

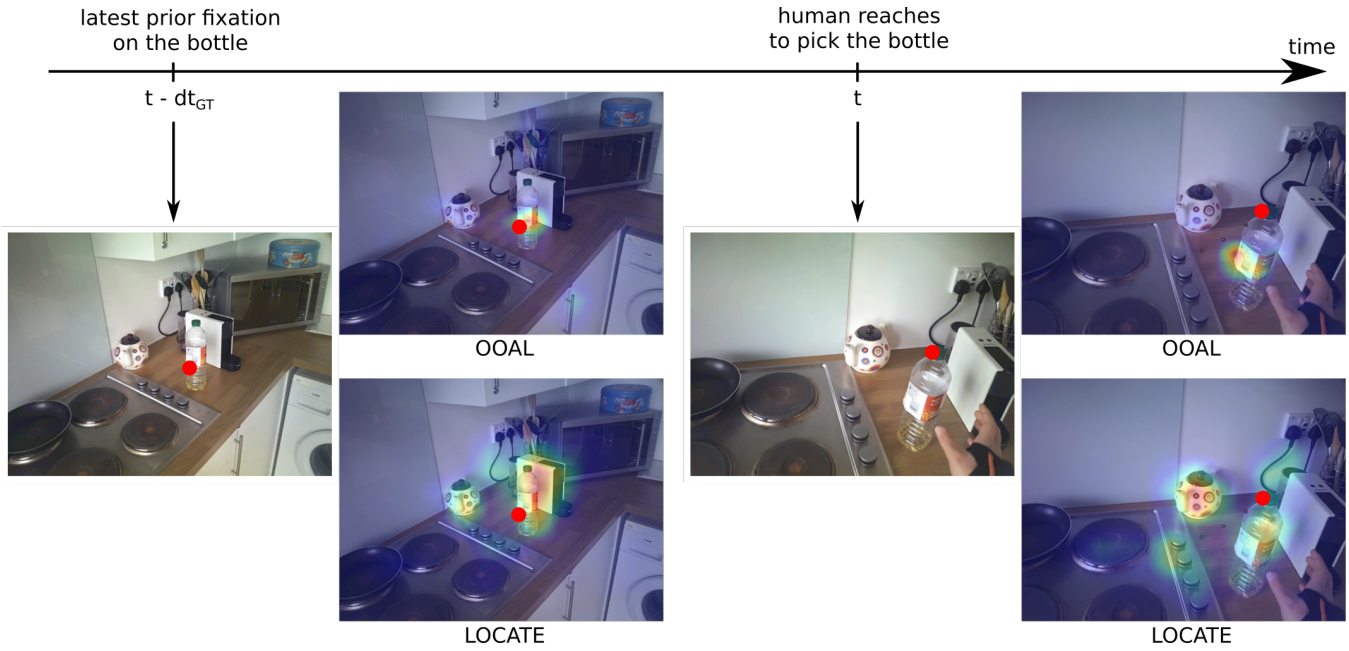

Fig. 1: An example of gaze affordance grounding. The red dot indicates the human's gaze fixation. OOAL and LOCATE "pick" affordance heatmaps are overlaid over the original image. At time t the human reaches to "pick" the bottle but the gaze is fixated on the cap (responsible for the following "open" manipulation). However, the latest prior fixation on the bottle at time $t - dt_{GT}$ was on the area responsible for the "pick" affordance.

fixate on the affordance regions of objects that correspond to their intended actions (e.g., gazing at a cup's handle before grasping it)? If the answer is yes, we hypothesize that detecting fixations on specific affordance regions can potentially help reduce the ambiguity in both Midas Touch and Intention Decoding problems.

However, humans are also capable of interacting with objects without directly fixating on them [16]; for example, opening a door without looking at the handle. Such interactions rely on internal models of the environment, allowing individuals to reach toward and interact with objects based on memory built from prior fixations, proprioceptive feedback and tactile cues. Consequently, learning from natural human-in-the-wild gaze and behavior may require analyzing not only the frames corresponding to the moment of interaction but also preceding frames to model the human's spatial representation and memory of the environment.

We investigate the relationship between object affordances and human gaze using natural human behavior data from large-scale datasets. Specifically, we utilize samples from our internal Deep Omelette dataset, which comprises egocentric recordings with gaze tracking of participants engaged in an omelette preparation. Additionally, publicly available resources such as Meta's Aria Everyday Activities Dataset [19] offer complementary data for analysis.

Analyzing natural human-in-the-wild gaze presents several technical challenges. Accurate annotation of object-action pairs from egocentric video recordings is both time consuming and difficult, as interactions may occur outside the camera's field of view, may be occluded or gaze tracking can be inaccurate. Segmenting unstructured environments for automated gaze detection is also problematic; conventional closed-vocabulary models like Mask R-CNN [20] are limited by their label sets and struggle to generalize over complex, natural scenes, while open-vocabulary vision-language model-based approaches (e.g., GSAM2 [21], GPT-4o) are susceptible to hallucinations. Moreover, existing affordance grounding methods (Locate [13], WorldAfford [14], and OOAL [15]) are typically trained on datasets like AGD20K [12] featuring single isolated objects, and may perform poorly on egocentric video data that can contain multiple instances of objects, motion blur, and dynamic environments. Addressing these technical challenges is crucial for advancing our research objectives.

Accordingly, this paper presents our preliminary approach, addressing the relationship between gaze and affordances at the moment of interaction as well as prior to interaction. Additionally we present complementary results of a comparison between human and automated annotation of action-object pairs and gaze fixations in large-scale egocentric datasets of natural human behavior.

II. METHODS

A. Gaze affordance grounding

The first objective of our study is to determine whether humans fixate on the area of an object corresponding to their intended action affordance at the moment of interaction. If not, we investigate whether fixation on the relevant affordance area occurred during preceding gazes—suggesting that the information was memorized into the individual's internal spatial representation for later use. To this end, we employ

Fig. 2: Automated annotation with GPT-4o in "action-gazed object" format as well as segmentation and gaze detection using Grounded-SAM2 and MRCNN.

two established affordance grounding methods: LOCATE [13] and OOAL [15], both trained on the AGD20K dataset [12].

To conduct this study, we utilize our internal Deep Omelette dataset, which contains egocentric video, gaze tracking, and full-body motion capture recordings of 20 participants preparing an omelette in a kitchen environment. The dataset includes 24 fps egocentric video recordings with gaze data captured at 120 fps, allowing for accurate identification of fixations and saccades. As an alternative, we considered Meta's Aria Everyday Activities dataset [19], which provides egocentric recordings at 30 fps with gaze tracking at 20 fps. However, the lower gaze sampling rate in Aria may be insufficient to reliably distinguish saccades from fixations.

We manually annotate the user's actions following a methodology similar to [6], breaking down each human interaction with the environment into "grasp object"-"manipulate object"-"release object" sequences. For example: "grasp spatula", "stir omelette", and "place spatula". Each action sequence begins with the human hand making contact with the object, followed by the interaction, and ending with the release. In addition to annotating action-object interactions, we also manually annotate the latest prior fixations on the object up to 10 seconds - i.e. if and when did participant fixate on the spatula prior to picking it.

Participants naturally perform a variety of manipulations during omelette preparation, some more discrete (e.g., "crack an egg") and others more continuous (e.g., "stir omelette"). These different actions are associated with distinct gaze patterns, such as fixation on the area of interaction or a guiding/tracking gaze [16]. However, not every annotated action from Deep Omelette has corresponding examples in AGD20K [12]. Therefore, for simplicity, we initially focus the gaze-affordance grounding only on the "pick" action — a discrete, common action that has different affordance regions across various objects, with corresponding examples available in the AGD20K dataset. Hence, given the action-object annotations prior and during the interaction, we checked whether human gaze indeed fixated on the "pick"-affordance areas considered appropriate by the affordance grounding.

B. Automated annotation and gaze detection

Manual annotation of action-object pairs and gaze fixations is a labor-intensive process. As egocentric data collection becomes more widespread with the adoption of smart eyewear and mixed reality headsets, large-scale manual annotation becomes increasingly infeasible. To address this, we present our initial approaches for automating both action-object annotation and gaze detection.

We leverage the vision capabilities of multimodal models such as CLIP and GPT-4o, which can perform image-based interpretation and description. Specifically, we use GPT-4o—the most capable and widely available vision-language model—to generate automated annotations of human action-object interactions. GPT-4o received egocentric video frames (processed as individual images) with the user's gaze indicated by a red dot, and was prompted to return a structured JSON object containing the gazed object, the inferred action, and the interacted object. To control complexity and reduce token usage, we use GPT-4o in "low" detail mode with no-shot instructions i.e. we provided no examples of input image

and output annotation) and without any prior history of gaze, actions, objects or frames seen prior.

In addition to automated annotations we employed automated image segmentation and gaze detection on interacted objects. We used MRCNN — a conventional closed-vocabulary label-based segmenter (similar to one used in gaze-driven assistive robotic setup in [6]) and Grounded-SAM2, a open-vocabulary Vision Transformer-based segmenter that combines Grounded Dino [22] and SegmentAnything2 [23]. The MRCNN used the COCO90 dataset labels, and the Grounded-SAM2 was provided with the list of interacted objects from manual annotations to guide the segmentation. Similarly to the manual annotation case the automated annotation checked fixations on the interacted object up to 10 seconds prior to interaction.

III. Preliminary Results

An example of gaze and affordance maps for grasp is shown in Fig. 1. In this case, we present two egocentric frames. At time t, the human reached for and grasped the bottle. At time $t - dt_{GT}$ (GT - human-experimenter annotated ground truth), the most recent prior fixation on the bottle occurred. Notably, the human fixated on the area marked as the "pick" affordance by OOAL at $t - dt_{GT}$ (we focus on OOAL for reporting, as it outperforms LOCATE). However, during the actual action at time t, the human's gaze was fixated on the bottle's cap, corresponding to a future "open" manipulation. This may reflect the human's internal affordance map of the environment — human did not fixate on the affordance region during the interaction but instead relied on prior observations. While we are still processing the entire dataset, this observation is noteworthy. It implies that in human-robot interaction, robots may need to consider the extended history of human gaze to estimate the human's internal model of the space, enabling more accurate deduction of the human's intention.

The results of automated annotation, segmentation, and gaze detection are demonstrated in Fig. 2. GPT-4o demonstrates promising ability to annotate human activity from egocentric video, although instruction refinement and fine-tuning are costly. As a result, local vision transformers such as Llama-11B are considered for future work. Meanwhile, automated gaze detection with MRCNN proves impractical, as many objects in the human natural environment are not present in the COCO90 dataset. Grounded-SAM2 performs better, as it can leverage the language transformer component to generate previously unseen object labels, albeit not always accurately.

The fixation detection prior to and during interaction is shown in Fig. 3, where we present the dt between times of fixations on an object and interaction time (similar to what was demonstrated in Fig. 1). The critical aspect is whether the automated annotation detected the same fixations as the ground truth. GPT-4o often suggested that the gaze was on an object, even when the gaze did not directly fall on it. Meanwhile, Grounded-SAM2 frequently mislabeled objects and/or struggled to differentiate between different instances of the same type of object (e.g., detecting fixations on different eggs in a carton).

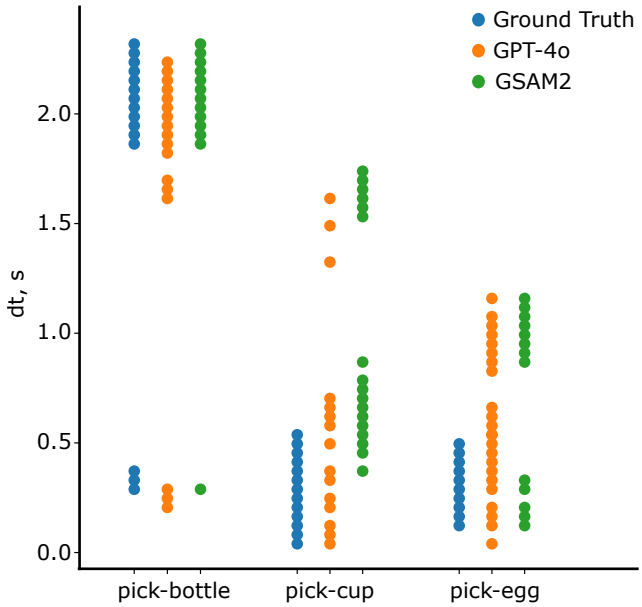

Fig. 3: Samples of the time difference dt between the human interacting with an object and prior gaze fixations on the object - ground truth annotated manually by human; GPT-4o used both for annotation and fixation detection; GSAM2 used for prior fixations detection given interacted object at t.

Overall, while modern Vision Transformers show promise in automated annotation and fixation detection, there remain significant technical hurdles to be overcome before they can be reliably used.

IV. Conclusion

Our work demonstrates the potential of gaze-affordance grounding for enhancing our understanding of natural human intent for human-robot interaction. Unlike traditional black-box machine learning methods, such as [8] that lack interpretability, gaze-affordance grounding is more explainable and generalizable. By linking gaze patterns to object affordances, it provides a more intuitive approach to understanding human behavior, making it a powerful tool for robotic systems.

However, several methodological and technical challenges remain. The complexity of natural human behaviour datasets and recording limitations (actions happening out-of-view, occlusions, gaze-tracking inaccuracies) make manual annotation arduous. Meanwhile automated annotation and segmentation methods require refinement as their accuracy is currently too low to be used reliably.

A potential insight emerging from this work is that perhaps the affordance grounding should be learned from human natural interactions - both gaze and manipulation with objects rather than manual annotations such as AGD20K [12]. This shift in perspective could significantly accelerate the affordance grounding research and improve the alignment between robotic systems and human behaviors in real-world environments.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] D. Cattaneo, I. Lamers, R. Bertoni, *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.

[5] 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.

[6] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," *ICRA*, 2019.

[7] C. Auepanwiriyakul, A. Harston, P. Orlov, *et al.*, "Semantic Fovea: Real-time annotation of ego-centric videos with gaze context," *ETRA*, 2018.

[8] P. Festor, A. Shafti, A. Harston, *et al.*, "Midas: Deep learning human action intention prediction from natural eye movement patterns," *arXiv:2201.09135*, 2022.

[9] N. Chomsky, "Three models for the description of language," *IRE Trans. on Inf. Theory*, 1956.

[10] D. Stout, T. Chaminade, J. Apel, *et al.*, "The measurement, evolution, and neural representation of action grammars of human behavior," *Scientific Reports*, 2021.

[11] R. Lioutikov, G. Maeda, F. Veiga, *et al.*, "Learning attribute grammars for movement primitive sequencing," *Int. J. of Rob. Research*, 2020.

[12] H. Luo, W. Zhai, J. Zhang, *et al.*, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] G. Li, V. Jampani, D. Sun, *et al.*, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," *CVPR*, 2023.

[14] C. Chen, Y. Cong, and Z. Kan, *Worldafford: Affordance grounding based on natural language instructions*, 2024. arXiv: 2405.12461 [cs.CV].

[15] G. Li, D. Sun, L. Sevilla-Lara, *et al.*, "One-shot open affordance learning with foundation models," *CVPR*, 2024.

[16] M. M. Hayhoe, "Vision and action," *Annual review of vision science*, 2017.

[17] M. Hayhoe and D. Ballard, "Modeling task control of eye movements," *Current Biology*, 2014.

[18] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in cognitive sciences*, 2005.

[19] Z. Lv, N. Charron, P. Moulon, *et al.*, "Aria everyday activities dataset," *arXiv preprint arXiv:2402.13349*, 2024.

[20] K. He, G. Gkioxari, P. Dollár, *et al.*, "Mask R-CNN," in *ICCV*, 2017.

[21] T. Ren, S. Liu, A. Zeng, *et al.*, *Grounded sam: Assembling open-world models for diverse visual tasks*, 2024. arXiv: 2401.14159 [cs.CV].

[22] S. Liu, Z. Zeng, T. Ren, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, *et al.*, *Sam 2: Segment anything in images and videos*, 2024. arXiv: 2408.00714 [cs.CV].