Ego-Tutor: Multimodal Reasoning for Dexterous Mobile Robots

Mohamed Malek Abid[1] Julie Terrassier[1] Konstantin Lucny[1] Dragos Chileban[1]
[1]ETH Zürich
{moabid, jterrassier, klucny, dchileban}@ethz.ch
Supervisors: Isar Meijer, Jeffrey Delmerico, Oier Mees (Microsoft Research)

Abstract

Understanding how humans perceive and describe tasks can help improve embodied reasoning models for robotic manipulation. Existing policies such as Embodied Chain-of-Thought (E-CoT) are mostly trained on robot-centric datasets and lack exposure to real egocentric human demonstrations. We present Ego-Tutor, a system that enhances Vision-Language-Action models by leveraging rich multimodal signals from Meta Aria glasses to provide human-aligned reasoning for robot learning. Crucially, we develop and iterate upon a Mixed Reality application on the Microsoft HoloLens for interactive inspection and modification of generated model reasoning, enabling on-the-fly policy and data corrections. The iteration process involved user studies and analysis, allowing us to evaluate the usability and effectiveness of our mixed reality interface for data annotation and policy improvement. Training code and our MR application are available at https://github.com/julie-trrsr/ego-tutor.

1. Introduction

Vision-Language-Action (VLA) models have been an increasingly popular area of research over the past couple years; as the most promising approach for open-world robotic manipulation, they enable robots to understand natural language instructions and execute corresponding actions by leveraging large-scale Vision-Language Model (VLM) backbones to improve Out-of-Domain (OOD) generalization to tasks. Prior approaches have struggled with this OOD generalization since the early days of robotics and symbolic AI, where symbolic approaches struggled with brittleness and representing complex environments. Recent work on Embodied Chain-of-Thought (E-CoT) [9] has shown that explicitly generating intermediate reasoning steps improves both interpretability and task performance compared to base VLAs. However, these models and their associated intermediate reasonings are still predominantly trained on robot-centric datasets captured from fixed exocentric video of robot arms with coupled trajectories, which fundamentally differ from how humans perceive and reason about manipulation tasks.

We present Ego-Tutor, a system that bridges this gap by leveraging multimodal egocentric data from Meta Aria glasses [7] to enhance VLA reasoning. Our pipeline makes full use of all relevant data that the Meta Aria glasses provide: namely, synchronized RGB video, eye gaze tracking, hand pose estimation, and spoken task narrations that we collect during natural human demonstrations.

To enable human-in-the-loop refinement, we developed, iterated on, and presented a demo of our Mixed Reality application on the Microsoft HoloLens, which allows users to visualize model predictions directly in Augmented Reality (AR) and provide corrections to detected objects, subtasks, and all other aspects of generated reasonings. This creates a feedback loop where human expertise can be injected to improve the policy's performance via augmented intermediate model reasonings in complex settings and tasks. Our contributions include:

- A pipeline for collecting and processing egocentric multimodal demonstrations using Meta Aria glasses, extracting synchronized gaze, hand tracking, and speech signals.
- Improvements to the existing reasoning generation (considering past/present frames as well as incorporating eye gaze to augment reasonings).
- A HoloLens Mixed Reality application for interactive inspection and correction of model reasoning, enabling on-the-fly data augmentation.

We evaluate our approach through training comparisons showing improved convergence with our collected data and improved reasoning generation for fine-tuning, as well as a user study demonstrating the usability of our mixed reality interface.

2. Related Work

Vision-Language-Action Models. OpenVLA [5] introduced an open-source VLA architecture combining a vision encoder with a language model backbone for end-to-end robotic control. Embodied Chain-of-Thought (E-CoT) [9] extended this 7B parameter model by generating explicit reasoning chains before action prediction, thus improving interpretability and performance. Gemini Robotics 1.5 [**?**] introduces a similar approach which they call a "Thinking VLA", which also interleaves natural language reasoning with actions, and uses *2D pointing* as an intermediate representation for embodied reasoning. We were partially inspired by this pointing, to utilize the human gaze data as a similar signal to ground reasoning in what humans actually look at during manipulation.

Egocentric Perception and Datasets. Meta's Project Aria [7] glasses are equipped with many sensors including stereo RGB cameras, eye tracking, and IMUs. Large-scale egocentric datasets like Ego4D [2] and Ego-Exo4D [3] have advanced action recognition and video understanding, but lack native 3D hand pose annotations, a limitation for learning dexterous manipulation. EgoDex [4] addresses this by collecting 829 hours of egocentric video with paired hand tracking using Apple Vision Pro, but no comparable dataset exists for Aria glasses. Off-the-shelf hand detectors suffer from poor accuracy on egocentric views due to heavy occlusion and limited viewpoints, we were thus motivated to collect our own Aria dataset with *native hand tracking from Aria MPS* which could then be mapped to approximate gripper/end-effector positions and states more accurately than an off-the-shelf solution.

Open-Vocabulary Object Detection. Grounding DINO [6] enables zero-shot object detection from natural language descriptions by combining DINO's self-supervised features with grounded pre-training. We use Grounding DINO to detect objects mentioned in spoken task descriptions, then apply gaze-based classification to determine object relevance.

3. Data Collection

We collected a custom dataset using Aria smart glasses, comprising approximately 15 stationary scenes in which a user interacted with everyday objects placed on a table. The recorded activities involved simple manipulation actions such as picking up objects, placing them on top of one another, and repositioning them in the scene. These controlled yet natural interactions were designed to capture multimodal signals relevant to embodied perception and action understanding.

For data processing, we leveraged the Aria Machine Perception Services (MPS) outputs, in particular the provided hand pose estimates. The stereo RGB streams were first undistorted and spatially aligned to ensure geometric consistency across views. Multiple modalities were then fused into a unified representation by integrating 3D gaze rays and hand skeletons into each video frame. Gaze information was projected into the image plane to enable intuitive 2D visualization, while hand skeletons were projected after undistortion of the RGB images.

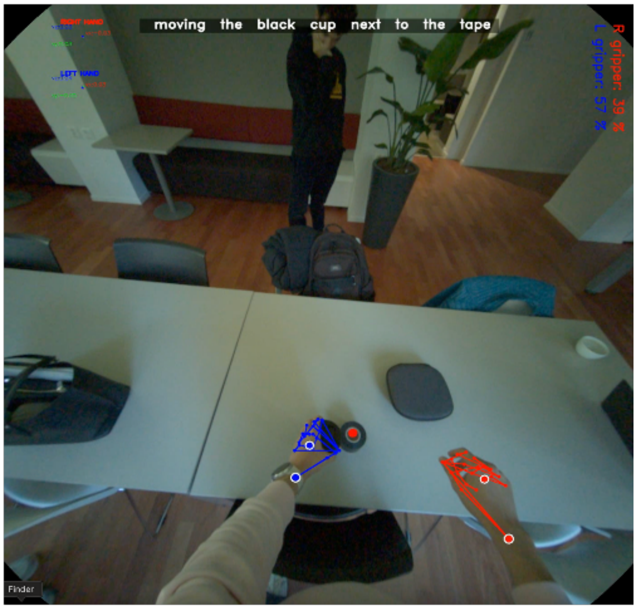


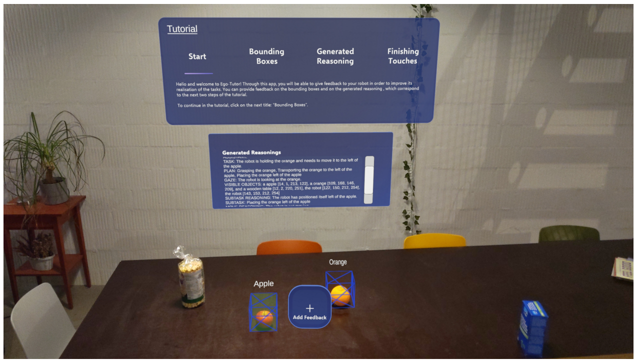
Figure 1. Visualization of Aria recording multi-modal data overlay.

To further characterize hand–object interaction, we implemented an algorithm to estimate hand rotation and compute a grasping coefficient, defined as a continuous value between 0 and 100 representing the degree of hand closure. In parallel, we applied the Whisper [8] speech-to-text model to automatically transcribe the user's verbal narration of task execution, producing natural-language descriptions aligned with the visual data.

All modalities were temporally synchronized and stored in RLDS format. In addition, we developed visual inspection tools 2 that overlay gaze points, hand skeletons, and textual task instructions onto the video frames, enabling efficient analysis and validation of the collected dataset.

4. Methodology

Improved E-CoT Reasoning Generation Pipeline. We use gaze information to classify detected objects as either primary objects (relevant to the immediate subtask), or contextual, providing explicit attention reasonings that are absent in standard E-CoT annotations. This approach was inspired by Gemini robotics ER[1], which adds "pointing" as a prediction in intermediate model reasonings. Humans naturally attend to task-relevant objects through eye gaze, de-



Tutorial

Hello and welcome to Ego-Tutor! Through this app, you will be able to give feedback to your robot in order to improve its realisation of the tasks. You can provide feedback on the bounding boxes and on the generated reasoning , which correspond to the next two steps of the tutorial.

To continue in the tutorial, click on the next title: "Bounding Boxes".

Generated Reasonings

TASK: The robot is holding the orange and needs to move it to the left of the apple.
PLAN: Grasping the orange, Transporting the orange to the left of the apple, Placing the orange left of the apple
GAZE: The robot is looking at the orange.
VISIBLE OBJECTS: a apple [14, 1, 213, 122], a orange [109, 168, 146, 209], and a wooden table [12, 2, 220, 251], the robot [122, 150, 212, 254], the robot [143, 153, 212, 254]
SUBTASK REASONING: The robot has positioned itself left of the apple.
SUBTASK: Placing the orange left of the apple
MOVE REASONING: The robot is not moving

Figure 2. Screenshot of our HoloLens Application.

scribe their intentions through speech, and coordinate hand movements with visual attention. We argue that incorporating these egocentric human demonstrations into reasonings for training could provide stronger supervision for learning which objects are relevant at each stage of a task, as well as more human-aligned reasonings. Furthermore, apart from updating the LLM used to the latest feasible/fast model (Gemini 2.5 Flash), we also modify the reasoning generation strategy to provide the model with the past $k = 3$ reasonings as well as the first generated reasoning; this leverages the longer context windows of the updated LLM model and allows the reasonings to be more relevant with respect to the overall task as well as maintain some brief causal context when things go wrong due to an incorrect previous reasoning step.

5. Mixed Reality Application

5.1. HoloLens

To support interactive inspection and correction of model reasoning, we developed an immersive mixed-reality application on Microsoft HoloLens. This enables users to see E-CoT predictions directly in their environment and correct them in real-time, creating a human-in-the-loop feedback pipeline for on-the-fly data augmentation and policy improvement. This tool provides an intuitive way to understand and refine embodied reasoning outputs, making it easier to diagnose failures and inject human knowledge into the learning process.

Reasoning visualization. The interface spatially overlays the model's predicted reasoning output onto the real environment. Detected objects are visualized as bounding boxes anchored to their physical counterparts, each annotated with the predicted object label. The generated reasoning chain, including the current task and subtasks, is displayed in a dedicated panel within the user's field of view. This combined visualization allows users to simultaneously inspect what the model perceives and how it decomposes the task, facilitating identification of perceptual and reasoning errors.

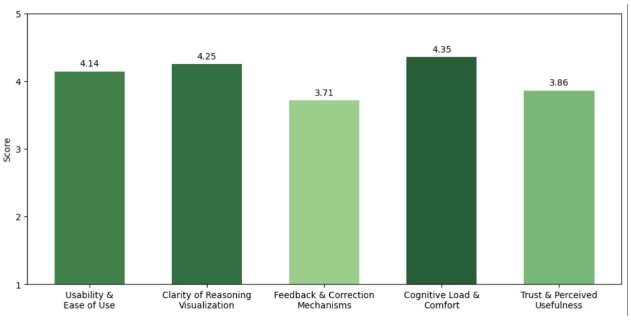


Figure 3. Qualitative user study results.

Voice corrections. Users can interact with the system through natural voice command to correct reasoning outputs. By selecting either a bounding box or the generated reasoning chain panel, users can dictate corrections such as object relabeling or subtask description refining. Corrections are recorded and associated with the selected element, enabling targeted feedback without requiring manual text input or external annotation tools. This voice-based interaction supports fluid hands-free supervision in situ.

Live Feedback Loop. All recorded corrections are aggregated and submitted explicitly by the user, after which they are transmitted to the backend. These corrections are fed back into the training pipeline to improve reasoning annotations and refine perception-reasoning alignment, enabling iterative policy improvement through human-in-the-loop data augmentation.

The HoloLens application provides a mechanism for spatially grounded visualization and correction of embodied reasoning output. It supports the efficient identification of perception and reasoning errors and facilitates their incorporation into the training pipeline. This design improves the interpretability and controllability of the embodied policies.

5.2. User Study

To evaluate the usability, clarity, and overall effectiveness of the proposed system, we conducted a qualitative user study with approximately 20 participants. The study was carried out primarily during the course demo presentation, where a diverse group of visitors interacted with the system at our stand in the faculty hallway. Most participants had a technical background, including students and researchers, while others had limited prior exposure to mixed reality systems. During the evaluation, participants interacted with physical objects placed on a table and explored the system's predictions and feedback mechanisms in a realistic, hands-on setting.

Study design and questionnaire. Participants interacted freely with the system and subsequently completed a structured questionnaire consisting of Likert-scale (1–5) and open-ended questions. The quantitative evaluation covered the following aspects:

1. Usability and ease of use, including intuitiveness of the MR interface, learnability, and confidence during interaction.
2. Clarity of reasoning visualization, assessing the comprehensibility of the displayed reasoning, spatial placement of overlays, and comparison to traditional 2D interfaces.
3. Feedback and correction mechanisms, focusing on the ease and effectiveness of providing voice-based corrections.
4. Cognitive load and comfort, evaluating information density, distraction, mental effort, and physical comfort.
5. Trust and perceived usefulness, measuring user trust in the model, perceived involvement, and applicability to real-world robot supervision.

Additional questions addressed overall satisfaction, preference for MR-based feedback over screen-based interfaces, and potential application scenarios.

Quantitative results. The aggregated results of the Likert-scale questions are shown in Fig 3. Overall, participants reported consistently high scores across all categories, with mean values ranging from 3.71 to 4.35 out of 5. Cognitive Load and Comfort received the highest score (4.35), indicating that the amount of information presented in MR was generally perceived as manageable and non-overwhelming. Clarity of Reasoning Visualization (4.25) and Usability and Ease of Use (4.14) were also rated positively, suggesting that spatialized reasoning and bounding box visualization effectively supported model understanding. Slightly lower scores for Feedback and Correction Mechanisms (3.71) and Trust and Perceived Usefulness (3.86) highlighted opportunities for improving interaction feedback and transparency.

Qualitative feedback and system improvements. Open-ended responses provided valuable insights into user expectations and system limitations. Several participants indicated the need for an integrated tutorial to explain interaction steps without requiring external guidance. Based on this feedback, we implemented an in-app guided tutorial that introduces system functionality in a step-by-step manner. Additional feedback concerned the readability of reasoning text, particularly regarding color contrast and font size, which was addressed by refining the visual design of text elements. Further, we have changed the positioning of buttons connected to the feedback loop for more natural interaction. In the course of this we have added a menu, appearing next to the user's wrist, which contains general functionality of the application. Finally, users requested clearer feedback regarding when voice input was active. Consequently, we added a green microphone indicator to explicitly signal when the system is listening for verbal corrections. Overall, the feedback confirms the potential of MR-based reasoning visualization while guiding concrete improvements to the interface.

In conclusion, the user study demonstrates that the proposed mixed reality interface effectively supports intuitive interaction, clear reasoning visualization, and user-driven model correction. The consistently high quantitative scores and constructive qualitative feedback indicate that participants found the system both usable and informative, while also highlighting concrete areas for improvement. The implemented refinements based on user feedback further suggest that such MR-based interfaces have strong potential to enhance transparency, trust, and human-in-the-loop learning in real-world robotic and embodied AI applications.

6. Experiments

We compare models fine-tuned on standard E-CoT data against our Ego-Tutor generated reasonings from our collected Aria-E-CoT data and resulting reasonings. Both use the ecot-openvla-7b-bridge base model with LoRA adapters (rank 32) trained for 100 steps on an 80GB A100 GPU.

Setup. We generate two versions of reasoning annotations for our Aria dataset: (1) *Base E-CoT*, using Gemini 2.5 Flash but following the original E-CoT reasoning format and prompts for generating reasonings, and (2) *Our improved Ego-Tutor CoT*, using our full pipeline with object importance classification and gaze-grounded reasoning generation. We also ablate adding the past two frames and reasonings to the model when generating reasonings for fine-tuning, and find that this qualitatively improved the accuracy and relevance of generated subtasks, so we included this modification in our prompt as well. Both runs were fine-tuned with the same hyperparameters: learning rate 0.001, batch size 8.

Results Figure 5 shows training curves for action token accuracy; note that this is evaluated on a constrained task where our goal was to overfit via fine-tuning as efficiently as possible. The gaze-aware model exhibits faster initial convergence and achieves, potentially due to our egocentrically collected data aligning with the optimal intermediate reasoning steps that a human would perform for the evaluation tasks.

Qualitative Analysis We observe notable differences in the generated reasoning chains. Baseline E-CoT produces generic spatial descriptions, while gaze-aware reasoning explicitly references attention/the relevant object that is to be interacted with in a given instant. The gaze-aware model also produces more structured MOVE explanations that describe action phases (approaching, grasping, transporting) rather than simple directional commands, improving interpretability and flexibility for more complex tasks. More examples can be found here.

These results suggest that egocentrically collected hu-

Baseline E-CoT

TASK: Take the yellow sponge and put it on the green napkin.

PLAN: Grasp the yellow sponge. Move the yellow sponge to the green napkin.

SUBTASK REASONING: The yellow sponge is the first object that needs to be grasped in order to complete the task.

SUBTASK: Grasp the yellow sponge.

MOVE REASONING: The yellow sponge is below the robotic arm.

MOVE: Move down.

GRIPPER POSITION: [24, 173]

VISIBLE OBJECTS:
yellow sponge [168, 153, 187, 171]
green napkin [171, 173, 222, 227]
metal sink [85, 142, 177, 225]

// generic reasoning, gripper pos far from relevant object
// no object prioritization

Ego-Tutor (Ours)

TASK: The yellow sponge needs to be grasped and then placed onto the green napkin.

PLAN: 1. Grasp the yellow sponge. 2. Move sponge to the green napkin. 3. Release the yellow sponge.

GAZE POINT: [137, 150]

SUBTASK REASONING: The spoken instruction "putting it on the green napkin" sets the goal. Gaze at [137, 150] is on the yellow sponge, confirming it as the immediate target for grasping.

SUBTASK: Grasping the yellow sponge.

MOVE REASONING: Stopping to maintain stable pre-grasp position directly above the yellow sponge. Gaze on PRIMARY confirms alignment for precise grasp initiation.

MOVE: Stop. Preparing to grasp.

GRIPPER POSITION: [127, 152]

VISIBLE OBJECTS:
yellow sponge [132, 142, 142, 157] — **PRIMARY**
green napkin [171, 173, 222, 227] — AUXILIARY
table [77, 124, 203, 220] — AUXILIARY

Figure 4. Reasoning chain comparison for task "taking the yellow sponge and putting it on the green napkin". Our multimodal reasoning provides explicit gaze reasonings, spoken instruction grounding, PRIMARY/AUXILIARY object classification, and detailed move explanations.

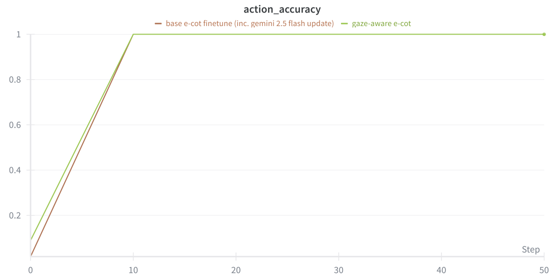


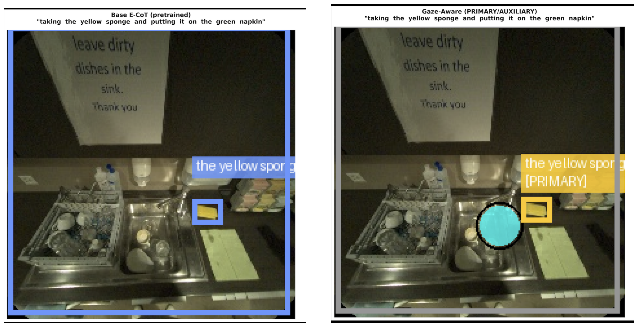
Figure 5. Training curve comparing Base E-CoT (brown) vs Gaze-Aware E-CoT (green) action token accuracy.

man data with our pipeline from Aria glasses provide meaningful supervision for learning how to interact with and reason given object manipulation tasks. Due to limited compute resources and time restrictions, we were unable to run additional ablations and experiments to verify the scaling even further beyond our current dataset; However, our pipeline is compatible with this logical next step, as the aforementioned EgoDex dataset from the Apple Vision Pro could be used in lieu of our manually collected Aria dataset of 15 scenarios.

7. Discussion

Our work demonstrates that egocentric human data with multimodal signals can provide meaningful supervision for embodied reasoning in robotic manipulation. We discuss key strengths, limitations, and directions for future work.

Strengths. (1) Incorporating eye gaze as an explicit attention signal for reasonings provides a natural mechanism for the phenomena of "pointing" or object prioritization that baseline E-CoT lacks[1]. Extending this to improving our object classification scheme grounds reasoning in human attention patterns, making the model's decision process more interpretable and aligned with how humans approach manipulation tasks. Additionally, providing the model with


Base E-CoT (pretrained)
"taking the yellow sponge and putting it on the green napkin"
leave dirty dishes in the sink. Thank you
the yellow spor
Gaze-Aware (PRIMARY/AUXILIARY)
"taking the yellow sponge and putting it on the green napkin"
leave dirty dishes in the sink. Thank you
the yellow spor
[PRIMARY]

(a) Base E-CoT

(b) Ego-Tutor (with gaze)

Figure 6. Visual comparison of gaze-classified bounding box outputs. Base E-CoT (left) produces generic spatial descriptions while Ego-Tutor (right) incorporates gaze-guided reasonings and object prioritization.

context on the first reasoning of the main task as well as the past two subtasks improves the quality and groundedness of intermediate reasoning steps by mimicing short-term causal memory of how the robot got to the state that it is at. When it comes to our Mixed Reality aplications, interface enables intuitive human-in-the-loop correction of model outputs. The user study results (mean scores 3.71–4.35/5 across categories) indicate that even users without prior MR experience can effectively inspect and correct reasoning chains, creating a practical pathway for iterative policy improvement. Finally, leveraging native hand tracking from Aria MPS yields more reliable gripper state estimates compared to off-the-shelf detectors that struggle with egocentric viewpoint occlusions [4].

Limitations and Future Work. Due to the limited time and scope of our project, there are some avenues that could offer major improvements. First, our dataset comprises only 15 stationary tabletop scenes, limiting the diversity of tasks and environments. Scaling data collection using larger egocentric datasets such as EgoDex [4] or Ego-Exo4D [3] could improve generalization, especially when coupled with more Aria collected data. Third, while the HoloLens application enables real-time visualization, the current feedback loop requires explicit user submission of corrections. An end-to-end system that continuously learns from implicit user attention during MR interaction [1] would reduce annotation burden. Additionally, the current VLM backbone lacks specialized tokens for bounding box coordinates, potentially limiting spatial reasoning precision; building upon a more recent architecture e.g Pi 0.5 could improve the action accuracy performance and intermediate reasoning steps further. Furthermore, while we collect data with the Aria, it would be interesting to explore a data collection method with the HoloLens that would allow users to edit their collected data in real time; this should be possible due to the HoloLens having gaze, hand, and voice/speech recognition support much like the Aria.

Conclusion. Ego-Tutor bridges egocentric human perception and robot policy learning through both context and gaze-aware reasoning generation; our primary contribution is coupling this pipeline with an interactive Mixed Reality app for the Hololens that augments the data collection and policy correction for Embodied Chain-of-Thought. Our experiments show faster convergence to perfect accuracy on limited scenarios using our Aria-collected data, and our user studies confirm the intuitive usability of spatial and verbal reasoning visualization in AR. As embodied AI systems scale toward real-world deployment, grounding model reasoning in human attention and enabling intuitive correction interfaces for on-the-fly policy correction and context injection may be increasingly important for operating reliable and trustworthy robotics in complex or dangerous environments.

References

[1] Gemini Robotics Team. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer, 2025. 2, 5, 6

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[3] Kristen Grauman et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *International Journal of Computer Vision*, 2024. 2, 6

[4] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 2, 6

[5] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Arjun Punnakkal, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024. 2

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2024. 2

[7] Meta. Project aria research kit. `https://www.projectaria.com/`, 2024. 1, 2

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2

[9] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Embodied chain-of-thought reasoning for vision-language-action models. In *Conference on Robot Learning*, 2024. 1, 2