2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

EgoMI: Learning Active Vision and Whole-Body Manipulation from Egocentric Human Demonstrations

Justin Yu[1,2*†], Yide Shentu[1,2*], Di Wu[2], Pieter Abbeel[1], Ken Goldberg[1], Philipp Wu[2]
[1]UC Berkeley, [2]xdof.ai

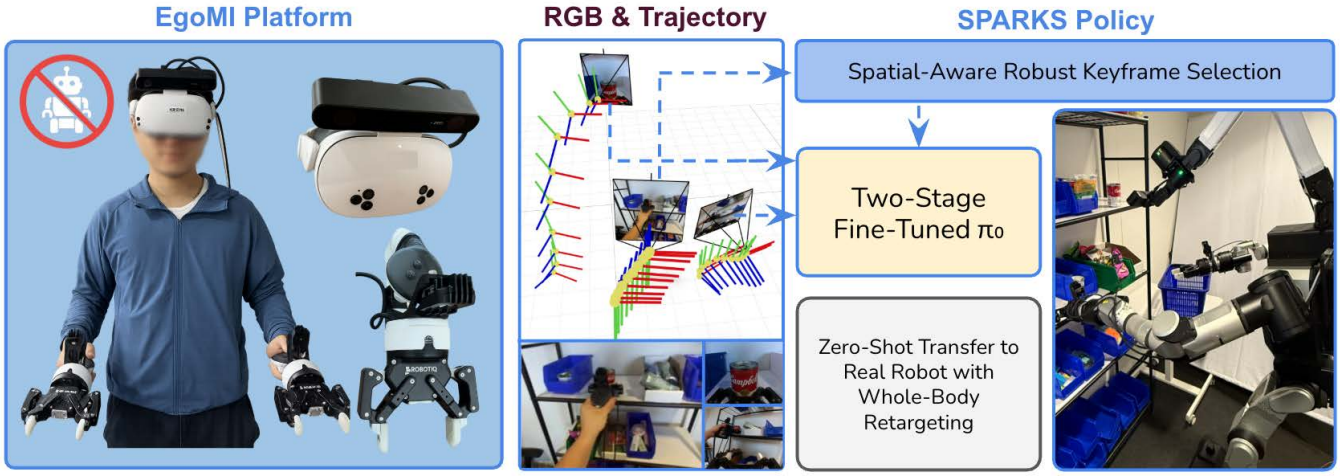

Fig. 1: Overview of the EgoMI framework. EgoMI captures egocentric human demonstrations with synchronized head and hand tracking. To handle rapid viewpoint changes from head motion, demonstrations are processed via *Spatial-Aware Robust Keyframe Selection* (SPARKS). A two-stage fine-tuning procedure then adapts a pre-trained absolute joint-space foundation model (π_0) into *Relative-Operation Space*. The learned policy remarkably transfers zero-shot to real robots through whole-body retargeting, without requiring any visual augmentation, explicit visual alignment, or on-embodiment data collection.

***Abstract*— Imitation learning from human demonstrations offers a promising approach for robot skill acquisition, but egocentric human data introduces fundamental challenges due to the embodiment gap. During manipulation, humans actively coordinate head and hand movements, continuously reposition their viewpoint and use pre-action visual search strategies to locate task-relevant objects. These behaviors create dynamic, task-driven head motions that static robot sensing systems cannot replicate, leading to a significant distribution shift that degrades policy performance. We present EgoMI (Egocentric Manipulation Interface), a framework that captures synchronized end-effector and active head trajectories during manipulation tasks, resulting in data that can be retargeted to compatible semi-humanoid robot embodiments. To handle rapid and wide-spanning head viewpoint changes, we introduce a memory-augmented policy that selectively incorporates context from historical observations. We evaluate our approach on a bimanual robot equipped with an actuated camera head and find that policies with explicit head-motion modeling consistently outperform baseline methods. Results suggest that coordinated hand–eye learning with EgoMI effectively bridges the human-robot embodiment gap for robust imitation learning on semi-humanoid embodiments. Project page: `https://egocentric-manipulation-interface.github.io`**

I. INTRODUCTION

Learning from large-scale human demonstrations represents a powerful pathway toward scalable robot skill acquisition. Recent advances in imitation learning have shown promising results in training robots directly from human collected data. However, a significant barrier persists in the form of the *embodiment gap*, the fundamental mismatch between human demonstrators and robotic platforms. During manipulation tasks, humans actively integrate head and eye movements with hand actions, dynamically adjusting their viewpoint to maintain visual contact with task-relevant objects and resolve occlusions. This active perception strategy is fundamental to how we understand and interact with our environment.

In contrast, many contemporary robotic systems rely on static-external camera configurations that cannot replicate this coordinated visual behavior. This creates a severe distribution shift when learning from egocentric data, as the task-driven viewpoint changes inherent in human demonstrations cannot be reproduced by fixed sensing systems. Various methods attempt to minimize the embodiment gap through strategies such as restricting to wrist mounted cameras [1], [2], [3], [4] or projecting top camera views into a coordinate-invariant representation [5], [6]. However, for more complex tasks that require searching or looking with the head, the learned policy is unable to reproduce the demonstrated behavior. Furthermore, the inability of standard policies to maintain a persistent spatial memory of past observations exacerbates this gap, leading to context loss during rapid head movements.

To address this challenge, we introduce the Egocentric

* Equal contribution
† Work done during internship at xdof.ai.

Manipulation Interface (EgoMI), as illustrated in Figure 1, a framework that captures the full degrees of freedom of human perceptual movements with minimal embodiment gap. EgoMI simultaneously records both end-effector and head movements during human demonstrations. Proprioception and camera data streams for both the wrist and head result in more complete conditioning information for the downstream policy. We can then retarget a policy's prediction of the whole-body motion data onto a robot. We use a modified Rainbow-RBY1 (wheeled, semi humanoid robot), equipped with a 6-DoF arm that functions as a neck to enable faithful retargeting of human head and end-effector motions to the robot.

A primary challenge of utilizing an actuated head is the potential for context loss due to large viewpoint shifts. To address this, we present Spatial-Aware Robust Keyframe Selection (SPARKS), a lightweight algorithm that selects a compact set of past head keyframe images to mitigate the loss of context caused by rapid head viewpoint shifts, while avoiding out-of-distribution failures. SPARKS is a simple yet effective mechanism that emphasizes keyframes rich in spatial information, and scores past frames using viewpoint novelty, temporal recency, and motion smoothness as a training-free proxy for visual informativeness. By embedding spatial memory into robot policies, SPARKS enables more stable long-horizon reasoning and resilience to viewpoint shifts.

Real world robot experiments show that incorporating head observations and movement from the egocentric demonstration is crucial during both the data collection stage and policy deployment stage. Policies trained without the head trajectory data or memory consistently fail due to a lack of critical context. Remarkably, our framework achieves zero-robot-data transfer *without relying on any* augmentations, in-painting, or viewpoint re-rendering, demonstrating that capturing egocentric head and hand motions and observations is sufficient to bridge the embodiment gap. In summary, we:

- Demonstrate the importance of an actuated head for imitation learning in everyday robotic manipulation tasks.
- Introduce a simple yet effective approach for training robot policies with spatial memory, addressing the challenges posed by rapid perspective changes from the egocentric head camera.
- Develop a data collection device that records key egocentric data.
- Evaluate the approach in real-world experiments, demonstrating capabilities enabled by head retargeting and memory-aware policies.
- Release code, hardware designs, and experiments to facilitate reproducibility and further research.

II. Related Work

A. Data collection devices for imitation learning

Progress in imitation learning is closely intertwined with advancements in the devices used to capture demonstrations, as the choice of interface heavily influences both the quality of collected data and scalability of the process. Devices used for teleoperation vary widely, ranging from joysticks and 3D spacemice [7], [8], VR controllers [9], [10], [11], and camera based sensors [1], [12], [13], [14]. While effective for many tasks, these devices typically require control abstractions to be simplified into end-effector space due to mismatches between human morphology and robot kinematics. Leader–follower systems such as ALOHA [15], GELLO [16] and AirExo-2 [17] mitigate this issue by matching the teleoperation device more directly to the robot morphology. UMI [2] represents a middle ground, using only the robot gripper as the input device. Large-scale datasets such as EgoDex [18] exemplify this approach. However, these methods still exhibit a large embodiment gap, since human motions, perspectives, and physical morphology differ substantially from a robot's own embodiment.

B. Imitation learning approaches from human data

Recent advances in large-scale imitation learning have resulted in robot foundation model approaches that integrate pre-trained vision and language representations with behavior cloning to scale imitation learning across numerous tasks [19], [20], [21], [22], [23]. Follow-on systems such as RoboFlamingo [24], Octo [25], OpenVLA [26] π_0 [27] [28], $\pi_{0.5}$[29], and Seed GR-3 [30] confirm that combining demonstration data with foundation models yields strong zero-shot generalization and versatile open-world behavior.

While the above largely use teleoperated data, others focus on data from human egocentric data. One notable example is EgoMimic [31], which captures egocentric video and detailed 3D hand tracking using augmented reality (AR) glasses and trains policies on both human and robot data to improve generalization. DexCap [6] introduces scalable finger-level demonstrations collected via motion-capture gloves alongside a specialized imitation algorithm to precisely capture fine manipulation details. Building upon their works, ARCap [5] harnesses AR feedback to guide novice operators in recording robot-executable trajectories, improving demonstration quality and usability. These efforts collectively push forward scalable and accessible high-fidelity data collection for imitation learning. Our approach additionally considers head motion and memory.

C. Policy learning with Active Vision

Policy learning with active vision seeks to endow agents with the ability to intelligently control their viewpoints dynamically, coordinating sensory actions (e.g., camera movements or gaze shifts) with motor actions to reveal task-relevant information and mitigate occlusions. Early approaches relied on fixed [32], [15], [33] or eye-in-hand cameras [34], while more recent work combines reinforcement and imitation learning to jointly optimize perception and action.

ViA [35] uses a 6-DoF head-mounted camera to imitate human head motion, reducing occlusions and selecting informative viewpoints, but leverages a robot embodiment specific teleoperation data collection system. EyeRobot [36] equips a robotic system with a 2DoF "eyeball" camera and learns gaze behavior through a combined BC-RL (behavior cloning and reinforcement learning) objective, while

TABLE I: Comparison of teleoperation systems and their features. For on-embodiment data collection, teaching device accuracy is less critical since the robot records data from sensor feedback rather than depending on the fidelity of the teaching device itself. Our proposed method, EgoMI, is the only system that simultaneously captures head and hand trajectories, supports true gripper actions, and enables whole-body retargeting, bridging the embodiment gap between human demonstrations and robotic execution.

System	Error (mm) ($\text{avg}_{\pm\text{std}}$)	Robot-free Collection	Head Pose Tracking	True Gripper Action
Gello	N/A	✗	✗	✓
ALOHA	N/A	✗	✗	✓
UMI	$8.855_{\pm 3.228}$	✓	✗	✗
AirExo-2	$1.737_{\pm 1.713}$	✓	✗	✗
VIA	N/A	✗	✓	✓
EgoMI	$2.126_{\pm 1.216}$	✓	✓	✓

manipulation is learned through behavior cloning. Look, Focus, Act [37] incorporates human gaze data into a foveated vision encoder, showing how explicit fixation signals improve policy robustness. In contrast, EgoMI directly captures human head motion for natural active vision and combines it with SPARKS, a lightweight memory mechanism to learn policies in a robot agnostic manner while still maintaining a minimal embodiment gap.

III. EgoMI System

In this section, we detail the EgoMI system, complete with the hardware device, data processing, and robot. The goal of the EgoMI design is to enable easy data collection for humans while minimizing the embodiment gap. Policies trained on EgoMI demonstrations transfer to semi-humanoid robotic embodiments *without* requiring extensive data domain adaptation or the inclusion of teleoperated data on the embodiment of the deployed robot.

A. Data Collection Hardware

The EgoMI collection system integrates commercially available hardware with custom components to capture synchronized head, hand, and visual data in a format directly compatible with downstream robot execution, shown in Figure 1. The core of the system is a Meta Quest 3S VR headset, which provides 6-DoF tracking of the operator's head and hand controllers. A camera (ZED 2i) is rigidly mounted above the headset to record first-person video that is aligned with head movements. Each VR hand controller is augmented with two custom hardware features to further align with the robot embodiment. First, a mounting point for wrist cameras provides a view closely aligned with the robot's wrist-mounted cameras for fine-grained manipulation tasks. Second, a mechanical flange interface with a standard mounting pattern allows the system to interface with off the shelf gripper systems (in our case, a Robotiq 2F-85).

During data collection, the triggers on each VR hand controller are mapped to real-time drive-by-wire control of the robot gripper actuation. For a comparison of EgoMI with respect to other data collection systems, refer to Table I. Notably, EgoMI enables robot free data collection, synchronized streams of head pose, hand trajectories, gripper action, proprioception, and egocentric and wrist videos, all while physically matching the geometry and visual appearance of the target robot platform.

B. Operator Gaze and Active Vision Data

While the EgoMI device captures operator head motion, one practical challenge is the absence of explicit eye-gaze tracking in our current hardware stack. Humans naturally fixate their gaze on task-relevant objects before acting [38]. To approximate this behavior, we overlay a fixed visual reticle at the center of the passthrough view and instruct operators to align it with manipulation targets and placement locations. This addition imposes little cognitive load, formalizes natural gaze behavior, and enables head orientation to serve as a reliable proxy for the center of visual attention, following the human look-then-reach behavior observed in motor studies.

Centering provides significant benefits for downstream policies by driving task-relevant visual features into the middle of the observation space and creating an object-centric representation that couples perception and action. In contrast, fixed external-cameras or unactuated head-cameras scatter task-relevant visual features across the image plane, forcing models to rely on weaker positional encodings. Qualitatively, we observe that policies trained without the reticle often fail completely, likely due to free gaze variability and weak head–gaze correlation, which caused trajectories and observations to drift out of the demonstration distribution at deployment.

C. Data Reformatting and Cleaning

We develop a high-throughput conversion pipeline that discovers and validates trajectory episodes, filtering out low-quality or corrupted data via video timing checks, and trajectory smoothness thresholds (SE(3) translation/rotation deltas). A key step is applying transforms that re-orient the raw data to minimize the proprioceptive gap between the capture system and the target robot. Because demonstrations are collected in a VR system with its own arbitrary world frame, we align all poses to the robot's canonical coordinate system based on the first time step data sample per-episode. This is done by applying a homogeneous transform that aligns the first timestep head position in the horizontal plane with the world-frame origin and re-orienting the forward-facing direction determined by the combined orientation of the gripper end-effectors. Specifically, let the raw VR-frame poses at time t be

$$T_V^L(t),\ T_V^R(t),\ T_V^H(t) \in SE(3)$$

where $R \in SO(3)$ and $p \in \mathbb{R}^3$. Here, L and R refer to the left and right end-effectors, and H to the head.

1) Forward direction estimation: We first extract the forward-facing axis (the z-axes in our case) of the two end-effector frames at the first timestep:

$$z_L(t) = R_V^L(t)\, e_z, \quad z_R(t) = R_V^R(t)\, e_z, \quad e_z = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top.$$

We project each vector onto the xy-plane and normalize:

$$\bar{z}_L = \frac{\Pi_{xy} z_L(0)}{\|\Pi_{xy} z_L(0)\|}, \qquad \bar{z}_R = \frac{\Pi_{xy} z_R(0)}{\|\Pi_{xy} z_R(0)\|},$$

where the projection operator is

$$\Pi_{xy} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}.$$

We then compute the yaw angles of each vector and take the circular mean to avoid discontinuities:

$$\theta_L = \text{atan2}(\bar{z}_{L,y}, \bar{z}_{L,x}), \quad \theta_R = \text{atan2}(\bar{z}_{R,y}, \bar{z}_{R,x}),$$

$$\theta = \text{atan2}(\sin\theta_L + \sin\theta_R,\ \cos\theta_L + \cos\theta_R).$$

We define this re-orientation component as a yaw rotation about the z-axis $R_z(\theta)$.

2) Base origin from head position: We place the base-frame origin at the x-y position of the head from the first timestep: $t_B = \begin{bmatrix} p_{H,x} & p_{H,y} & 0 \end{bmatrix}^\top$.

3) VR to base transform: We define the base-to-VR and its inverse VR-to-base transform as:

$$T_V^B = \begin{bmatrix} R_z(\theta) & t_B \\ 0 & 1 \end{bmatrix}, \; T_B^V = (T_V^B)^{-1} = \begin{bmatrix} R_z(\theta)^\top & -R_z(\theta)^\top t_B \\ 0 & 1 \end{bmatrix}.$$

4) Applying calibration and offsets: Let T^L_{flange} and T^R_{flange} be the VR controller-to-robot flange calibration transforms, and $T_{\text{flange}\to\text{TCP}}$ be the fixed tool-center-point (TCP) offset for the end-effectors. The calibrated, base-frame end-effector poses are then

$$T_B^{L/R}(t) = T_B^V \, T_V^{L/R}(t) \, T^{L/R}_{\text{flange}} \, T_{\text{flange}\to\text{TCP}}$$

For the head, let $T_{\text{head}\to\text{cam}}$ be the fixed transform from the VR headset to the camera:

$$T_B^H(t) = T_B^V \, T_V^H(t) \, T_{\text{head}\to\text{cam}}$$

This process re-expresses all trajectories in a consistent robot-centric frame, minimizing the proprioceptive gap between VR-collected demonstrations and the target robot's embodiment.

D. Robot Setup: Wheeled-Humanoid with Active Head

To successfully transfer whole-body human movement to a robotic setup, it is necessary to have hardware that can simultaneously capture and reproduce the agility of both head and hand motion. Since humans naturally engage their waist, shoulders, and other joints during demonstrations, a robot system with more than a fixed torso and arms is necessary to realize such movements.

Moreover, a fully actuated neck is required, as humans move their heads freely in 6D space during search and manipulation phases. For our experiments, we use a modified Rainbow RBY1 robot equipped with a 6-DoF torso and 2× 7-DoF arms. On top of its torso, we mount an I2RT YAM [39] robot with a ZED2i camera [40] as the active vision head, enabling us to simultaneously track and replicate the full 6DoF head and hand movements collected from EgoMI demonstrations. See Fig. 3 for the detailed system setup.

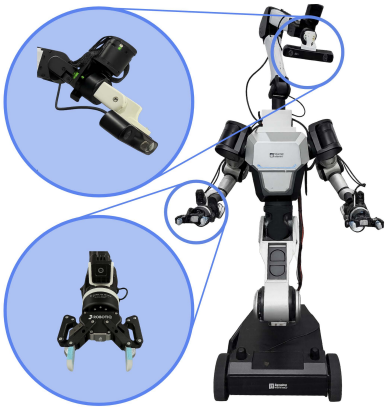

Fig. 3: EgoMI policy deployment setup. We use a modified Rainbow RBY1 robot with a 6-DoF YAM [39] + ZED2i [40] camera mounted on top as the fully actuated head. The gripper configuration is identical to the human demonstration setup, minimizing the embodiment gap.

IV. Method

This section details the model training and inference approach to enable learning policies on EgoMI data. Furthermore, we detail Spatial-Aware Robust Keyframe Selection (SPARKS), as a simple but effective method to train policies with spatial memory.

A. Policy Interface: 29D Action and State Representation

a) Dataset action / state format (absolute, world-frame): In addition to synchronized image observations, each timestep of our dataset encodes a 29D *action* vector and *state* (proprioception) vector.

$$\underbrace{\Big[\overbrace{r_6^L}^{\text{rot6}}, \overbrace{p_3^L}^{\text{pos}}, \overbrace{g_1^L}^{\text{grip}}, \overbrace{r_6^R}^{\text{rot6}}, \overbrace{p_3^R}^{\text{pos}}, \overbrace{g_1^R}^{\text{grip}}, \overbrace{r_6^H}^{\text{rot6}}, \overbrace{p_3^H}^{\text{pos}}\Big]}_{\text{29D dataset vector}}$$

where $r^{(\cdot)}_{\cdot} \in \mathbb{R}^6$ is a 6D rotation vector (first two columns or rows of the $\mathbb{R}^{3\times3}$ rotation matrix concatenated), $p^{(\cdot)}_{\cdot} \in \mathbb{R}^3$ is position, and $g^{(\cdot)} \in \mathbb{R}$ is the continuous gripper signal. For actions, g^L, g^R come from operator intent (VR triggers); for state, they come from measured hardware closure/aperture.

b) Model input space (relative, inter-gripper): For model training, we keep the proprioception right hand in the world frame and compose the left hand and head as poses *relative to the right*:

$${}^R T_L = \left({}^W T_R\right)^{-1} {}^W T_L, \qquad {}^R T_H = \left({}^W T_R\right)^{-1} {}^W T_H,$$

We then parameterize $({}^R T_L, {}^R T_H)$ by $(r_6^{L:R}, p_3^{L:R})$ and $(r_6^{H:R}, p_3^{H:R})$ using the same 6D rotation + 3D position ordering convention:

$$\underbrace{\left[\, r_6^{L:R}, p_3^{L:R}, g_1^L, r_6^R, p_3^R, g_1^R, r_6^{H:R}, p_3^{H:R} \,\right]}_{\text{29D proprio model input}}$$

For action chunk samples, we transform from absolute to *relative* parameterization over all SE(3) components excluding gripper actions, as introduced in [2].

c) Deployment space (absolute, world-frame): The inverse transforms reproject policy outputs back to absolute world-frame commands. Predicted 6D rotations are expanded into valid 3×3 rotation matrices via Gram-Schmidt orthonormalization.

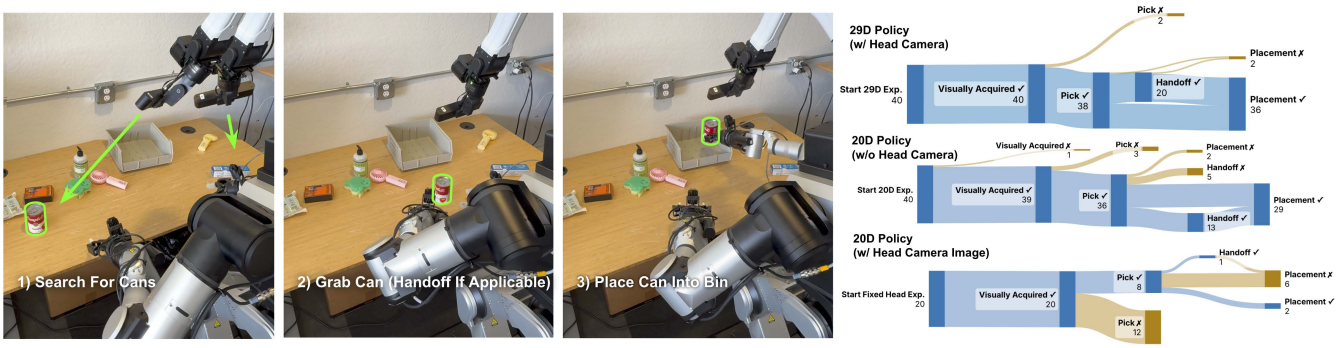

Fig. 4: Tabletop Task Rollout Sequence: *(Left).* The images show a real 29D policy evaluation rollout where the robot (1) scans for target cans across a cluttered workspace, (2) grasps the correct item with potential handoff between grippers, and (3) places it into the designated bin. *(Right).* The Sankey diagrams illustrate failure modes between policies with full 29D action space and active head-camera versus reduced 20D wrist camera-only and 20D + head-camera images baselines.

d) $SO(3)$ Representation: We choose this 6D vector representation for $SO(3)$ over other discontinuous representations such as Euler angles or double cover representations like quaternions, critical for the stability of neural network gradient-based optimization [41].

B. Spatial-Aware Robust Keyframe Selection (SPARKS)

Natural egocentric head motion produces rapid, task-driven viewpoint changes as the operator scans, fixates, and repositions their head during manipulation. Critical task information is often first revealed under viewpoints that differ dramatically from the current frame, such as when a human leans or turns their head to disambiguate occlusions or search for an object. In such settings, policies trained to only take a single timestep—or fixed windows—suffer from severe *context loss*, as temporally distant but visually important observations are dropped from the model's conditioning. SPARKS leverages the head trajectory to select a compact set of past frames for memory, avoiding costly learned or recurrent modules. In other words, SPARKS is a direct benefit of our active head sensing design, and would be neither necessary nor effective if the robot had only a static viewpoint.

At time t, given a causal head pose history in a short lookback window $\{{}^{W}T_H(\tau) \,|\, \tau \le t\}$, SPARKS assigns past frame τ a score combining three factors:

$$J(\tau) = \underbrace{\phi\big(\angle(\hat{z}_H(\tau), \hat{z}_H(t))\big)}_{\text{viewpoint novelty}} + \underbrace{\psi(t-\tau)}_{\text{recency}} + \underbrace{\rho\big(\angle(\hat{z}_H(\tau-1), \hat{z}_H(\tau))\big)}_{\text{motion smoothness}},$$

where $\hat{z}_H$ is the forward axis of the head camera transform. The three terms encourage novel viewpoints, recency, and low angular velocity (avoiding blurred frames, along with the insight that frames are likely more informative when the operator's gaze is fixated). Only frames exceeding diversity thresholds (angular displacement $> \alpha \cdot \text{FOV}$ or translation displacement $> \delta$) are added to a FIFO buffer.

SPARKS is run offline to precompute keyframe indices to preserve IID minibatch sampling during training, and run online at deployment in $O(L)$ per step, where L is the lookback length.

C. Policy Training

Starting with a capable pre-trained foundation model is crucial for achieving strong final performance. We initialize our approach from pre-trained π_0 model weights [27]. However, π_0 was originally trained only on *absolute robot joint positions*, which introduces a mismatch with our target representation. To bridge this gap and map π_0's output space into the *relative Cartesian space*, we adopt a **two-stage finetuning process**:

1) **General Multi-Task Finetuning to 29D action space.** We first finetune π_0 end-to-end on a diverse multi-task dataset from our in-house data bank, adapting π_0 from absolute joint outputs into the **29-dimensional relative Cartesian action space**. This dataset consists of approximately **200 hours** of EgoMI demonstrations.
2) **Task-Specific Finetuning.** After this adaptation, we further finetune the model end-to-end with *task-specific datasets*. This stage ensures that the model not only operates correctly in the relative Cartesian space but also achieves **maximal performance** on the target tasks.

During both training and inference, the selected SPARKS head keyframe images are integrated directly into the Pali-Gemma vision-language model as additional context image tokens without requiring changes to the core network or introducing learned memory modules.

Fig. 5: Randomization distribution and example initial configuration of the tabletop environment highlighting the wide distribution of object positions and clutter scenarios used during evaluation rollouts. *(Right).* Initial configurations for target object may be outside of the immediate field of view of the initialized robot during experimentation. Target object and placement location may also reside on opposite ends of the workspace requiring a bi-manual handoff maneuver.

D. Policy Deployment

The action control frequency is set to 25Hz while the policy model asynchronously inferences at approximately 40 ms latency on

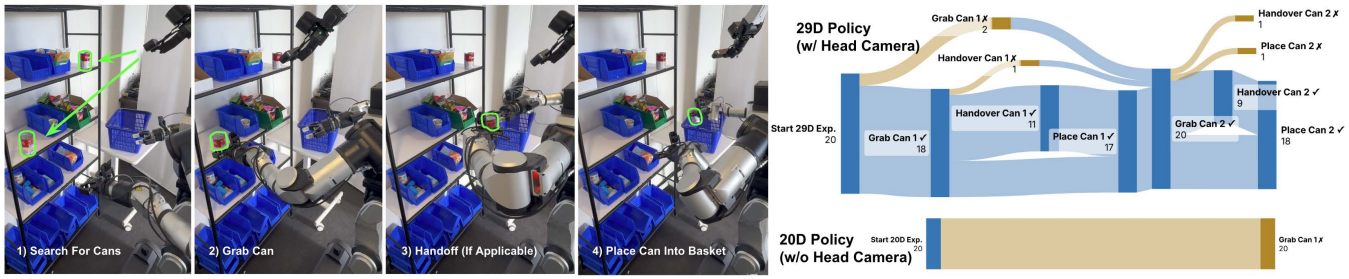

Fig. 6: Shelf Task Rollout Sequence. *(Left).* The images show a real 29D policy evaluation rollout where the robot (1) scans across multiple shelf tiers to locate target cans, (2) reaches and grasps the selected item, (3) performs a mid-air inter-gripper handoff, and (4) places the item into a shopping basket, then repeats on the remaining can. *(Right).* The Sankey diagrams show failure modes for the 29D active-head, whole-body retargeted policy compared to the 20D wrist camera-only policy.

an RTX 5090 GPU. The policy outputs a horizon of 40 future end-effector and head poses and gripper commands. We adopt temporal chunk ensembling [42] as our default control strategy.

To map these targets (left hand, right hand, and head) into the robot's specific joint configurations, we employ a differentiable inverse kinematics (IK) solver (Pyroki) [43]. Unlike analytical IK, which fails or returns nulls on unreachable poses, a differentiable approach treats IK as a weighted cost-minimization problem. By optimizing for end-effector pose error alongside posture regularization, the system achieves "graceful degradation" reaching as close as physically possible to the demonstrated pose rather than experiencing execution errors. This ensures robust transfer of unconstrained human demonstrations to robot embodiments with varying kinematic limits without requiring manual trajectory filtering or retraining.

We maintain hardware modularity by using the Viser [44] toolkit for high-level scene-graph management. Rather than authoring a monolithic URDF, we programmatically align the mounting interfaces of the Rainbow RBY1 and I2RT YAM actuated neck within a global coordinate frame. During deployment, two independent Pyroki processes resolve the torso/arm and head trajectories in parallel. This decoupled architecture ensures the 29D action representation remains platform-agnostic, facilitating zero-shot transfer to heterogeneous hardware configurations without retraining.

V. Experiments

We evaluate EgoMI on a suite of real-world manipulation tasks designed to test the impact of active head retargeting and memory-augmented policies. Our evaluation focuses on two key aspects: (1) the role of explicit head pose retargeting and head-camera observations in enabling robust wide-spanning bimanual manipulation, and (2) the necessity of SPARKS for handling tasks requiring visual memory under partial observability. All experiments are conducted on the robot platform described in Sec §III-D, with policies trained from only demonstrations collected from our VR-based data collection device, with zero teleoperated on-embodiment data.

A. Data Collection and Policy Training

For each task in the sections below, 1-1.5 hours of in-domain task-specific data was collected on the EgoMI device described in Sec. §III-A. Evaluation policies were trained for 40k steps, taking approximately 50 hours each.

B. Searching Tasks

1) Task Setup: The searching experiments evaluate the EgoMI framework in robot capability to localize and manipulate a target object in a wide workspace where the target may initially lie outside the field of view of both head and wrist cameras. Two searching tasks are considered:

(1) Tabletop Search: A soup can is placed within a large randomization distribution on a 30" × 60" table along with up to eight distractor objects. Importantly, the soup can can may start outside of the immediate field of view of the initalized robot. The robot must locate the soup can and place it into a designated organizing bin. When the target and bin reside on opposite sides of the table, operators are instructed to execute a transfer maneuver: pick with the closer hand, place the object temporarily at the center, then regrasp with the opposite hand to complete the placement. This produces natural and diverse bi-manual behaviors that emphasize the need for wide-spanning search and hand-to-hand coordination.

(2) Shelf Search: A tall shelving unit (height $\approx 2.4\,\text{m}$) contains distractor items and 2 soup cans, which may be placed at any shelf location at two tiers near shoulder height. The robot must identify the target cans and place it in a shopping basket located on a nearby table to the right. To succeed, the robot must search vertically and horizontally, and additionally execute a precise inter-gripper handoff when target objects are picked from the far side of the shelf.

2) Results: We compare and report two policy configurations:

- **29D Policy:** Includes head SE(3) action outputs for active head actuation and head-camera images.
- **20D Policy:** Excludes head SE(3) action outputs and excludes head-camera images, using only wrist-camera observations and gripper SE(3) commands.

On the tabletop task, the 29D policy achieves a success rate of **36/40**, outperforming the 20D wrist camera-only policy, which achieves **29/40**. Failure analysis shows that the 20D policy struggles primarily in wide-spanning scenarios requiring hand transfers, often failing to coordinate across the workspace due to incomplete scene context within only the wrist-views. By contrast, the 29D policy leverages the operator's natural pre-attentive head motion: as operators look toward the placement location before moving their hands, head-centering ensures that the placement location is already within the observation context for the policy model.

To further isolate the role of active head motion, we introduce an additional experiment in which the model is still provided with a head-camera image stream, but the ability to dynamically reposition the viewpoint is removed by fixing the head target. Performance drops achieving only **2/20** successful trials. Without active head control, the policy struggles heavily with object grasping and placement accuracy.

On the shelf task, the benefits of active head modeling are even more pronounced. The 29D policy achieves **35/40** success points (each can placed correctly counts as one point), while the 20D policy achieves **0/40**. Without head pose and gaze retargeting, the 20D model fails immediately, unable to localize off-screen targets or coordinate vertical and lateral search motions. This demonstrates that natural operator head motion provides essential cues for planning long-range reaching and handoffs without requiring explicit operator instruction.

3) Discussion: These results confirm that head pose retargeting and active head image observations are critical for bridging the embodiment gap in wide-spanning manipulation. Without them, the robot cannot reason about objects outside the initial wrist camera

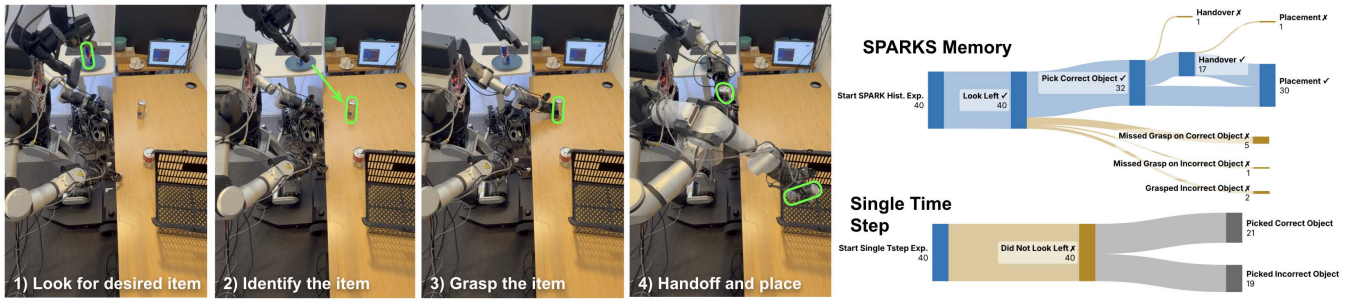

Fig. 7: Memory Task Rollout Sequence. *(Left).* The images show a real 29D policy evaluation rollout where the robot starts facing the table, then must (1) look left to identify the desired picking item, (2) look back to the table, (3) grasp the correct item (4) place the item into the bin with a mid-air handover if necessary. *(Right).* The Sankey diagrams show failure modes for policies leveraging SPARKS compared to single time-step conditioning.

view, leading to catastrophic failure on tasks involving search or cross-workspace coordination.

C. Memory Tasks

1) Setup: The memory task evaluates whether SPARKS enables the robot to maintain spatial memory across spatial and temporal gaps. The setup consists of a forward table with two objects (a soup can and a drink can) and a placement basket on the right. A second side table, positioned at a 90° angle to the left and *initially out of view*, contains either the soup can or the drink can (but never both).

The task proceeds as follows: the robot must first *look left* to inspect the side table, remember which object is there, then return to the forward table and pick and place the correct item into the basket. In the case that the correct object is on the left side, the operator is instructed to perform an inter-gripper handover before completing the placement into the shopping basket to the right

This design explicitly requires maintaining a temporally persistent memory of the visual information acquired during the initial head turn.

2) Results: We compare a baseline single-timestep policy (no memory) to our SPARKS memory-augmented policy:

- **Single-Timestep Policy:** Conditions only on the current head and wrist camera images.
- **SPARKS-Augmented Policy:** Selects and retains past frames using the SPARKS keyframe selector, providing a compact history buffer as additional input.

The single-timestep policy achieves a success rate of **21/40**, which is near random chance. Qualitative analysis shows that this policy fails to look left, instead directly picking from the forward table based on an ambiguous current view. By contrast, the SPARKS-augmented policy achieves **31/40** success, consistently looking left, adding the keyframe to memory, and then using it to disambiguate which object to select after the side table leaves the field of view.

3) Discussion: This experiment demonstrates the necessity of memory for tasks involving occluded or off-screen observations. Without SPARKS, the policy is forced into a Markovian assumption, treating the problem as if all relevant information is visible at once, which leads to high failure rates.

Together, these findings validate EgoMI as a scalable approach for closing the embodiment gap and enabling complex whole-body manipulation policies trained entirely from egocentric human demonstrations. Notably, our method required neither visual augmentation nor on-embodiment data. We hypothesize that the perspective changes introduced by the actuated head—as opposed to a fixed overhead camera—encourage the policy to learn robust visual features that ignore irrelevant context and focus on task-relevant information. This underscores the importance of active perception in imitation learning and suggests that explicit head modeling can reduce dependence on data augmentation.

VI. Conclusion

While EgoMI significantly narrows the embodiment gap, several limitations remain. The system is heavy and difficult to use for long durations for certain users. Physical embodiment mismatches also persist: for example, the robot's actuated head can move beyond natural human ranges, so retargeting may constrain performance. SPARKS, though effective, uses a fixed scoring heuristic and does not adaptively update memory; more intelligent conditioning mechanisms could further improve performance.

EgoMI enables learning active vision and whole-body manipulation from egocentric human demonstrations. By synchronizing head and hand motion, leveraging spatial memory with SPARKS, EgoMI achieves embodiment transfer to real robots without additional robot data. These results highlight egocentric demonstration as a scalable path for bridging the human–robot embodiment gap and enabling more general robot behavior.

References

[1] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4978–4985, 2020.

[2] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.

[3] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[4] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," 2020.

[5] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu, "Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," 2024.

[6] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[7] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," *RSS*, 2023.

[8] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," *CoRL*, 2022.

[9] D. Rakita, B. Mutlu, and M. Gleicher, "A motion retargeting method for effective mimicry-based teleoperation of robot arms," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 361–370.

[10] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[11] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto, "Holo-dex: Teaching dexterity with immersive mixed reality," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5962–5969.

[12] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[13] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," *arXiv preprint arXiv:2202.10448*, 2022.

[14] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *RSS*, 2023.

[15] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.

[16] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2023.

[17] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, L. Yang, W. Wang, C. Lu, and H.-S. Fang, "Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," *arXiv preprint arXiv:03081*, 2025.

[18] R. Hoque*, P. Huang*, D. J. Yoon*, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," 2025.

[19] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," Oct. 2022.

[20] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," *ArXiv*, vol. abs/2202.02005, 2022.

[21] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," 2022.

[22] A. B. et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 2023.

[23] C. Lynch and P. Sermanet, "Language conditioned imitation learning over unstructured data," *Robotics: Science and Systems*, 2021.

[24] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li, and T. Kong, "Vision-language foundation models as effective robot imitators," *arXiv preprint arXiv:2311.01378*, 2023.

[25] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[26] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[27] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024.

[28] D. Driess, J. T. Springenberg, B. Ichter, L. Yu, A. Li-Bell, K. Pertsch, A. Z. Ren, H. Walke, Q. Vuong, L. X. Shi, and S. Levine, "Knowledge insulating vision-language-action models: Train fast, run fast, generalize better," 2025.

[29] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025.

[30] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma, H. Niu, W. Ou, W. Peng, Z. Ren, H. Shi, J. Tian, H. Wu, X. Xiao, Y. Xiao, J. Xu, and Y. Yang, "Gr-3 technical report," 2025.

[31] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," 2024.

[32] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *ArXiv*, vol. abs/2304.13705, 2023.

[33] K. Shaw, Y. Li, J. Yang, M. K. Srirama, R. Liu, H. Xiong, R. Mendonca, and D. Pathak, "Bimanual dexterity for complex tasks," in *8th Annual Conference on Robot Learning*, 2024.

[34] Y. Liu, X. Xu, W. Chen, H. Yuan, H. Wang, J. Xu, R. Chen, and L. Yi, "Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1106–1113, 2024.

[35] H. Xiong, X. Xu, J. Wu, Y. Hou, J. Bohg, and S. Song, "Vision in action: Learning active perception from human demonstrations," 2025.

[36] J. Kerr, K. Hari, E. Weber, C. M. Kim, B. Yi, T. Bonnen, K. Goldberg, and A. Kanazawa, "Eye, robot: Learning to look to act with a bc-rl perception-action loop," 2025.

[37] I. Chuang, A. Lee, D. Gao, J. Zou, and I. Soltani, "Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers," 2025.

[38] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[39] I2RT-Robotics, "Yam – 6-dof robotic arm," https://i2rt.com/products/yam-manipulator, 2025.

[40] S. Inc., "Zed 2i stereo camera," https://www.stereolabs.com/store/products/zed-2i, 2025.

[41] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5745–5753.

[42] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[43] C. M. Kim, B. Yi, H. Choi, Y. Ma, K. Goldberg, and A. Kanazawa, "Pyroki: A modular toolkit for robot kinematic optimization," *arXiv preprint arXiv:2505.03728*, 2025.

[44] B. Yi, C. M. Kim, J. Kerr, G. Wu, R. Feng, A. Zhang, J. Kulhanek, H. Choi, Y. Ma, M. Tancik, and A. Kanazawa, "Viser: Imperative, web-based 3d visualization in python," *arXiv preprint arXiv:2507.22885*, 2025.