Where Do We Look When We Teach? Analyzing Human Gaze Behavior Across Demonstration Devices in Robot Imitation Learning

Anonymous Author(s)
Affiliation
Address
email

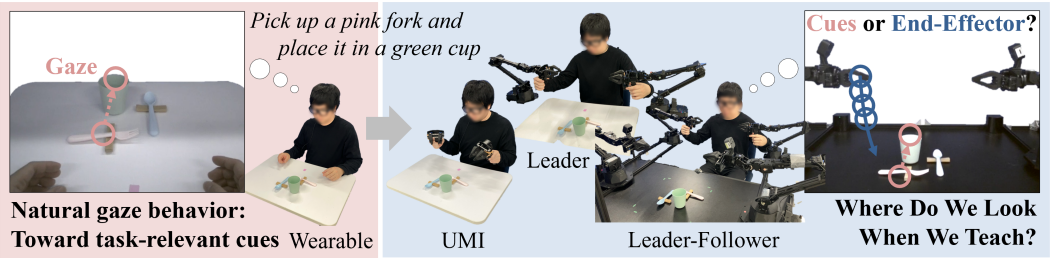


Figure 1: Illustration of the research question: Insights from cognitive science (left) — fixations focus on task-relevant cues; With demonstration devices (right) — how is gaze behavior influenced?

Abstract:

Imitation learning for acquiring generalizable policies often requires a large volume of demonstration data, making the process significantly costly. One promising strategy to address this challenge is to leverage the cognitive and decision-making skills of human demonstrators with strong generalization capability, particularly by extracting task-relevant cues from their gaze behavior. However, imitation learning typically involves humans collecting data using demonstration devices that emulate a robot's embodiment and visual condition. This raises the question of how such devices influence gaze behavior. We propose an experimental framework that systematically analyzes demonstrators' gaze behavior across a spectrum of demonstration devices. Our experimental results indicate that devices emulating (1) a robot's embodiment or (2) visual condition impair demonstrators' capability to extract task-relevant cues via gaze behavior, with the extent of impairment depending on the degree of emulation. Additionally, our proof-of-concept experiments reveal that gaze data collected using devices that capture natural human behavior improves the task success rate of imitation learning policies from 18.8% to 68.8% under environmental shifts.

Keywords: Gaze Behavior, Demonstration Devices, Imitation Learning

1 Introduction

End-to-end visuomotor imitation learning has gained significant attention for enabling robots to perform complex and dexterous manipulation tasks autonomously. Current mainstream imitation learning policies are trained in a supervised manner, where visual observations serve as inputs and actions as outputs. Typically, policies trained on large-scale datasets with higher data collection cost [1] generalize better than those trained on small-scale datasets with lower data collection cost [2]. This presently observed trade-off between generalization and cost is expected to be resolved.

The difficulty of realizing generalizations from small-scale datasets stems from their tendency to lack diversity compared to large-scale datasets. In this situation, the model is prone to learn infor-

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

mation containing biases unrelated to the task [3, 4]. In imitation learning, if an object consistently appears in the same shelf location, the policy might learn the visual spatial relationship between the end-effector and shelf rather than the relationship between the end-effector and object [5].

To extract task-relevant cues even from small-scale datasets, one promising strategy is to leverage the cognitive and decision-making skills of human demonstrators with strong generalization capabilities [6, 7]. In particular, cognitive science studies have shown that eye movements are tightly coupled with motor tasks [8, 9, 10, 11]. Gaze tends to be directed toward task-relevant cues, as humans prioritize the object being manipulated and naturally filter out task-irrelevant information [10]. In imitation learning, there are various potential applications for acquiring generalizable policies from small-scale datasets, including extracting task-critical observations [10], mitigating the impact of irrelevant environmental variations (thus reducing computational resources) [7], and supporting hierarchical policy modeling by capturing subgoals [12, 13].

However, in the context of imitation learning, where demonstrations are collected using specific demonstration devices, *does simply measuring demonstrators' gaze behavior actually improve policy performance?* Previous cognitive science studies [6, 7, 8, 9, 10, 11, 12, 13] primarily examined scenarios in natural, unconstrained settings for humans. In contrast, imitation learning commonly involves humans collecting data using demonstration devices that emulate the robot's embodiment [14, 15, 16, 17, 18] and visual condition [19, 20] to minimize domain gaps for the policy. Therefore, prior studies do not necessarily provide insights into the demonstrators' gaze behavior for all device types. In fact, relevant studies [21, 22] have reported that gaze behavior differs when individuals control a robot compared to when they move their own embodiment naturally.

To address this question, we propose an experimental framework that systematically analyzes demonstrators' gaze behavior across a spectrum of demonstration devices from those capturing natural human behavior to those emulating the robot's embodiment and visual condition. Our experimental results suggest that devices that capture natural human behavior enable demonstrators to extract task-relevant cues via gaze behavior more effectively than devices that emulate the robot's embodiment and visual condition. Additionally, our proof-of-concept experiments reveal that imitation learning policies trained with gaze behavior collected using devices that capture natural human behavior improve task success rate from 18.8% to 68.8% under environmental shifts.

Our key contributions are as follows:

- Providing a review that connects insights on eye movement from cognitive science with the use of demonstration devices in imitation learning.
- Introducing a novel experimental framework for analyzing demonstrators' gaze behavior across a range of demonstration devices.
- Identifying suitable demonstration devices that enable gaze behavior to highlight task-relevant cues effectively.
- Demonstrating that gaze data collected from such suitable devices improves the robustness of imitation learning policies.

2 Related Work

2.1 Analysis of Eye Movements in Cognitive Science

A study of eye movements has been explored extensively over the past five decades [23]. Early studies historically identified two primary components of eye movements: *saccades*, which rapidly redirect the gaze toward visual information, and *fixations*, which stabilize the gaze to extract that information. Later studies identified that task instructions play a critical role in determining when and where fixations occur [24]. Fixations tend to be directed toward task-relevant cues that optimize task performance regarding spatial and temporal demands rather than the most visually salient features [7, 10]. Approximately one-third of object fixations support four key monitoring functions: *locating* upcoming objects, *guiding* hand movement, *aligning* objects, and *checking* states [10]. Interestingly, while specific cognitive events can elicit certain fixations, the fixations themselves do

not uniquely determine the underlying cognitive events [7]. This suggests that fixations provide spatiotemporal coordinates of task-relevant cues but do not directly provide particular information being extracted.

Previous eye movement studies in cognitive science have primarily focused on natural human embodiment and visual condition scenarios. In contrast, our study investigates gaze behavior using a range of demonstration devices that emulate a robot's embodiment and visual condition, providing novel insights that are directly applicable to imitation learning.

2.2 Data Collection in Imitation Learning

In manipulation-focused imitation learning, data is commonly collected using demonstration devices that emulate the robot's embodiment and visual condition. Examples of embodiment emulation devices include the universal manipulation interface (UMI) [14], leader-only [15], and leader-follower [16, 17, 18]. These devices employ either a robot-mimetic mobile gripper or an actual robot to capture the coupling between visual observations and actions. An example of a visual condition emulation device is a head-mounted display (HMD). This device immerses the operator in the robot's visual observations, while enabling gesture-based control of the robot's actions [19, 20]. An alternative paradigm that does not emulate embodiment or visual condition collects egocentric video using wearable cameras [25, 26].

These commonly used demonstration devices in imitation learning have not been used to collect gaze data. In contrast, our study not only provides the first systematic analysis of demonstrator gaze behavior across different device types, but also establishes a concrete methodology for its collection.

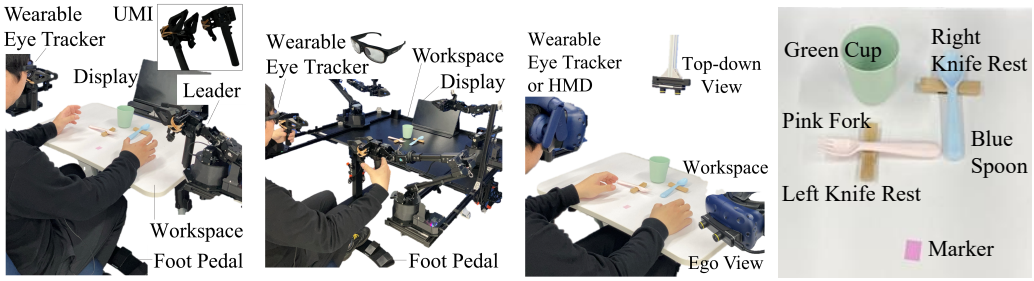
2.3 Leveraging Gaze Behavior in Machine Learning

Several studies have leveraged drivers' gaze behavior to improve machine learning models for autonomous driving [27, 28]. These studies demonstrate that gaze behavior can enhance the performance of behavior cloning and enable accurate modeling of driver attention. Similarly, in the context of Atari video games, prior studies have constructed gaze behavior datasets from human players and shown that incorporating gaze behavior can improve the performance of behavioral cloning [29, 30]. In the domain of robot learning with manipulators, one study has analyzed demonstrators ' gaze behavior and utilized it for subtask prediction and reward learning [22]. While insightful, this study lies outside the core context of imitation learning and is limited in terms of the range of demonstration device types. While another study successfully incorporated gaze information into imitation learning, it has not analyzed the properties of gaze behavior [15].

These prior studies demonstrate the effectiveness of incorporating gaze behavior into machine learning pipelines. However, to the best of our knowledge, no previous study has investigated how different types of demonstration devices influence the gaze behavior of human demonstrators in imitation learning. Our study is the first to systematically analyze this effect and to examine how understanding such characteristics can inform the design of gaze-informed imitation learning frameworks.

3 Proposed Method

We propose an experimental framework that comprehensively analyzes the demonstrators' gaze behavior across a spectrum of demonstration devices. Following the demonstration device paradigms in recent imitation learning, we selected three representative conditions: (A) wearable cameras for collecting egocentric video, (B) devices for embodiment emulation, such as UMI, leader-only or leader-follower, and (C) devices for visual condition emulation, such as HMDs. We regard (A) as a baseline that best captures natural human behavior. We hypothesize that both embodiment differences (A vs. B) and visual condition differences (A vs. C) independently influence the demonstrators' gaze-based task-relevant cue extraction during demonstration.



(a) (b) (c) (d)

Figure 2: Proposed experimental framework. (a) Wearable, UMI, and Leader setup, and (b) Leader-Follower setup (both in the embodiment experiment). (c) Wearable, HMD-Ego, and HMD-Top-down setup in the visual condition experiment. (d) Objects used in the experiments.

To examine this hypothesis, we pose the following research questions (RQs) and corresponding experimental designs:

- **RQ1:** How do embodiment differences in demonstration devices influence gaze-based task-relevant cue extraction?
- **RQ2:** How do visual condition differences in demonstration devices influence gaze-based task-relevant cue extraction?

3.1 An experimental design to investigate how embodiment differences influence gaze behavior

Focus: We design an experiment that provides a finer-grained analysis of embodiment differences. We list the demonstration devices by their degrees of embodiment emulation and examine how gaze behavior varies across these degrees. Ordered from the lowest degree of emulation, the devices are: **Wearable** (wearable cameras for collecting egocentric video [25, 31, 26], Fig. 2a), **UMI** (Fig. 2a) [14], **Leader** (leader-only [15], Fig. 2a), and **Leader-Follower** (Fig. 2b) [16, 17, 18]. As summarized in Tab. 1, these devices impose progressively stronger embodiment constraints on the demonstrator, ranging from Wearable to Leader-Follower.

Task: Following a previous study involving skewering task [21], we use a pick-and-place task for evaluation for several reasons: (1) Pick-and-place is a substantial portion of real-world robotic tasks [32, 33]. (2) Pick-and-place requires more precise perception of hand-object and object-object spatial relationships compared to the skewering task. (3) Pick-and-place can better reveal embodiment-induced gaze behavioral differences by designing the environment to require dynamic changes in the end-effector's pose between pick and place phases. We use commonly available

Table 1: Constraint characteristics of demonstration devices with different embodiments.

Constraints	Demonstration Devices			
	Wearable	UMI	Leader	Leader-Follower
Offset gripper[1]		✓	✓	✓
Low-DoF gripper[2]		✓	✓	✓
Low-DoF arm[3]			✓	✓
Control latency[4]				✓
Distant view[5]				✓
Without haptics[6]				✓

[1]Offset gripper: The offset distance between the demonstrator's finger and robot's gripper.
[2]Low-DOF gripper: A reduction in DoF from the human's five fingers to the robot's parallel gripper.
[3]Low-DoF arm: A reduction in DoF from the human's seven DoF arm to the robot's six DoF arm.
[4]Control latency: The delay in motion transmission from the leader to the follower.
[5]Distant view: Observing the robot from a distance (Fig. 2b).
[6]Without haptics: The lack of haptic feedback between the leader and follower.
[7]Egocentric view: The demonstrator performs the task from an egocentric view (Fig. 2c).
[8]Using an HMD: The demonstrator wears a heavy HMD with rendering latency.
[9]Top-down view: The demonstrator performs the task from a top-down view (Fig. 2c).

Table 2: Characteristics of demonstration devices with different visual conditions.

Conditions	Demonstration Devices		
	Wearable	HMD-Ego	HMD-Top-down
Egocentric view[7]	✓	✓	
Using an HMD[8]		✓	✓
Top-down view[9]			✓

household objects as task objects (Fig. 2d). The picked up *targets* are a pink fork and a blue spoon (IKEA KALAS), and their placement *destinations* are a green cup (IKEA KALAS) and left/right knife rests (IKEA TEJSTEFISK). The pick phase continues until the *target* is grasped, and the place phase ends when the *target* is successfully placed in/on the *destination*.

Metrics: Based on insights from cognitive science (described in Sec. 2.1), where gaze-based monitoring functions such as *locating* objects and *guiding* hand movements were identified, we define task-relevant cue extraction as fixating on the *target* during the pick phase and *destination* during the place phase. In contrast, a previous study [21] reported that the gaze of an operator tends to be located at the robot's end-effector during teleoperation. Therefore, we compute two Euclidean distances in the 2D image plane: (1) the distance between the gaze location and *target* or *destination* object and (2) the distance between the gaze location and end-effector. (The end-effector refers to the demonstrator's hand in the Wearable setting, or the robot's end-effector in all other settings.) These distances are averaged over each pick-and-place trial. By comparing these averaged values, we determine whether participants fixate more on the task-relevant cue or the end-effector. We quantified their capability to extract task-relevant cues by counting the number of trials in which they could fixate on the cue.

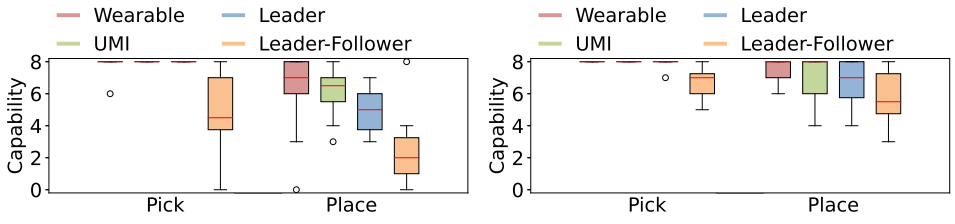
Instruction: Participants first perform the task with only high-level task instructions. In a second condition, they receive additional gaze-relevant instructions directing their attention toward task-relevant objects (for example, look at the *target* during the pick phase and *destination* during the place phase). This evaluates whether simple verbal instruction can help align gaze with task-relevant cues.

3.2 An experimental design to investigate how visual condition differences influence gaze behavior

Focus: We examine how gaze behavior varies across demonstration devices with different visual conditions: **Wearable** (wearable cameras for collecting egocentric video [25, 31, 26], Fig. 2c), **HMD-Ego** (Egocentric view displayed on an HMD [19, 20], Fig. 2c), and **HMD-Top-down** (Robot's top-down view displayed on an HMD, Fig. 2c). As summarized in Tab. 2, the Wearable and HMD-Ego differ in the device type, while the HMD-Ego and HMD-Top-down differ in the viewpoint. All other aspects of the experimental setup, including the task, metrics, and instruction, are the same as described in Sec. 3.1.

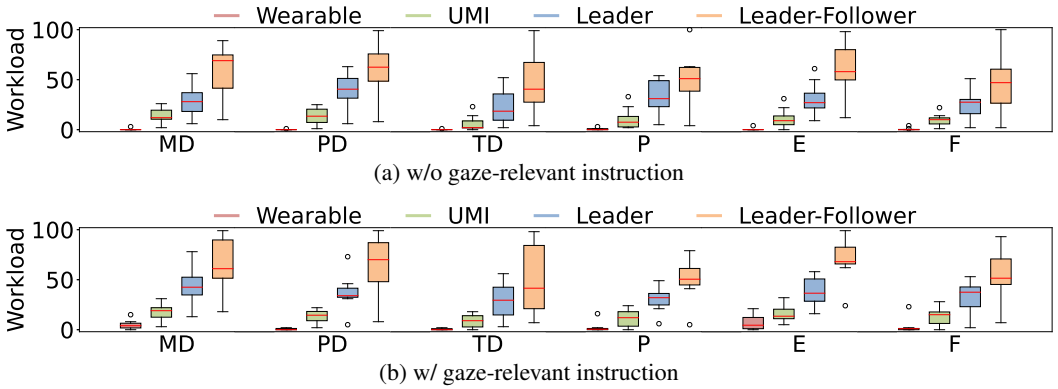
4 Experiments

The following experiments were approved by the institution's research ethics review board.



(a) w/o gaze-relevant instruction (b) w/ gaze-relevant instruction

Figure 3: Effect of embodiment emulation devices on task-relevant cue extraction capability. Capability refers to the number of trials (out of eight) in which the demonstrator fixated on the *target* or *destination* (defined in Sec. 3.1).



(a) w/o gaze-relevant instruction

(b) w/ gaze-relevant instruction

Figure 4: Effect of embodiment emulation devices on NASA-TLX sub-indicator

4.1 An experiment on RQ1: Investigating how embodiment differences influence the capability to extract task-relevant cues

The detailed experiment setup (participants, equipment, procedure, data collection, and annotation) is described in Appendix B.

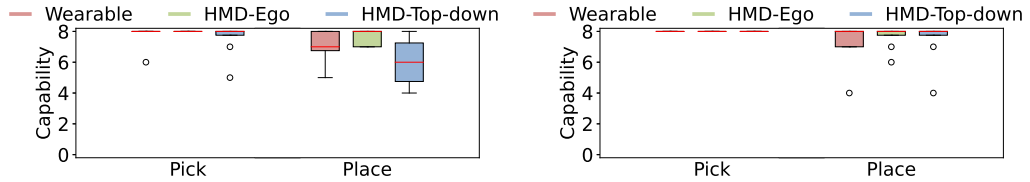
Results: *(A) Analysis on RQ1:* As previously defined, we measure the capability to extract task-relevant cues as the participant fixating on the *target* during the pick phase or the *destination* during the place phase (Sec. 3.1). Figure 3 shows the effect of embodiment emulation devices on the task-relevant cue extraction capability. The Wearable consistently exhibited the highest capability, followed by UMI, Leader, and Leader-Follower, under both gaze-relevant instruction conditions (Fig. 3a and Fig. 3b). Without gaze-relevant instructions, the Leader-Follower in the pick phase and the Wearable in the place phase showed large variance and outliers. Overall, providing gaze-related instructions improved the capability across most devices. Figures 7 and 8 show qualitative examples of gaze behavior for each device.

(B) Analysis on Workload: To gain further insights, we conducted an additional workload analysis, described in Appendix B. Table 3 shows the means and standard deviations of raw NASA-TLX (RTLX) scores [34], used for the repeated measures analysis of variance (repeated measures ANOVA). The analysis shows that device type is statistically significant ($p < 1.0 \times 10^{-3}$), whereas the device usage order and the presence or absence of gaze-relevant instructions are not ($p \geq 0.10$). Specifically, switching from the Wearable to UMI increased workload by approximately 10 points, from UMI to Leader by approximately 20 points, and from Leader to Leader-Follower by approximately 20 points. These results align with intuition, showing that increasing embodiment constraints lead to increased workload.

(C) Analysis on Sub-indicator: Figure 4 shows the effect of embodiment emulation devices on each NASA-TLX sub-indicator. Across all sub-indicators, scores increased stepwise from Wearable to UMI, Leader, and Leader-Follower. No sub-indicator contradicted the trend in total workload described in Sec. 4.1, result (B).

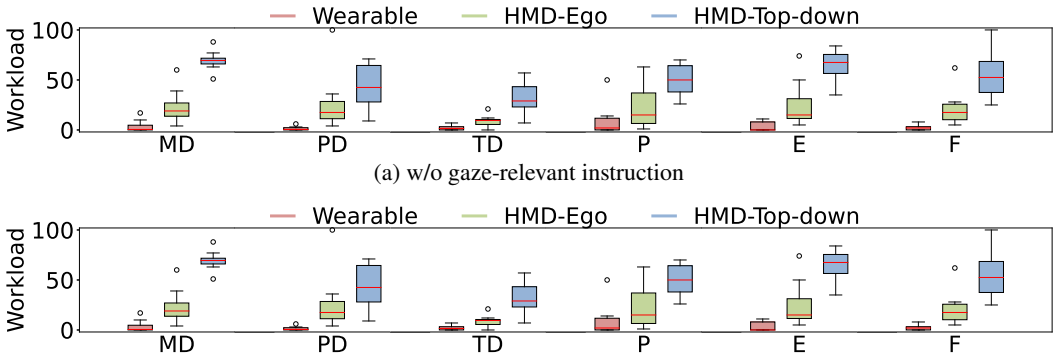
Table 3: Workload means and standard deviations used for repeated ANOVA conditioned on device order, gaze-relevant instructions (rows), and demonstration devices (columns).

Order	Inst.	Demonstration Devices							
		Wearable		UMI		Leader		Leader-Follower	
		Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Fwd.	w/o	0.25	0.29	9.17	8.47	20.13	12.54	44.54	27.41
	w	1.92	0.78	12.29	6.29	33.46	21.57	50.58	33.26
Rev.	w/o	0.54	0.98	12.83	7.21	39.08	12.01	62.63	26.56
	w	4.58	4.15	14.58	8.42	36.96	11.55	67.00	15.49



(a) w/o gaze-relevant instruction (b) w/ gaze-relevant instruction

Figure 5: Effect of visual condition emulation devices on task-relevant cue extraction capability. Capability refers to the number of trials (out of eight) in which the demonstrator fixated on the *target* or *destination* (defined in Sec. 3.2).



(a) w/o gaze-relevant instruction

(b) w/ gaze-relevant instruction

Figure 6: Effect of visual condition emulation devices on NASA-TLX sub-indicator.

Discussion: In response to RQ1, we found that embodiment emulation devices impair demonstrators' capability to extract task-relevant cues, depending on the degree of emulation. In addition, we identified three key findings: (1) The degree of embodiment emulation influence the trade-off between capturing high-quality task-relevant cues and collecting effective demonstration data. Devices with weaker embodiment emulation are better at extracting rich gaze-based task-relevant cues but tend to produce demonstration data with a larger domain gap [14]. Conversely, devices with stronger embodiment emulation extract fewer cues but generate demonstration data with a smaller domain gap [16]. To provide reliable task-relevant cues and demonstration data for a policy, our results suggest using the Wearable for cue extraction and the Leader-Follower for demonstration data collection. Although collecting data with the Wearable incurs additional costs, its lower workload burden makes it a feasible approach. (2) No significant difference in workload was observed based on the presence or absence of gaze-relevant instructions. Thus, providing gaze-related instructions is recommended whenever possible. (3) While gaze-relevant instructions improved the Leader-Follower's capability, it still did not surpass the performance of the Wearable. Although the Leader-Follower allows intuitive operation and enables anyone to demonstrate fine manipulation tasks [16], the results suggest that gaze data should ideally be collected using more natural devices with fewer embodiment constraints.

4.2 An experiment on RQ2: Investigating how visual condition differences influence the capability to extract task-relevant cues

The detailed experiment setup (participants, equipment, procedure, data collection and annotation) is described in Appendix C.

Results: *(A) Analysis on RQ2:* As previously defined, we measure the capability to extract task-relevant cues as the participant fixating on the *target* during the pick phase or the *destination* during the place phase (Sec. 3.2). Figure 5 shows the effect of visual condition emulation devices on the task-relevant cue extraction capability. Without gaze-relevant instruction (Fig. 5a), the Wearable and HMD-Ego exhibited higher capability, followed by HMD-Top-down. With gaze-relevant instruction (Fig. 5b), all devices demonstrated similar capabilities. Overall, providing gaze-related instructions

Table 4: Workload means and standard deviations used for repeated ANOVA conditioned on device order, gaze-relevant instructions (rows), and demonstration devices (columns).

Order	Inst.	Demonstration Devices					
		Wearable		HMD-Ego		HMD-Top-down	
		Mean	Std.	Mean	Std.	Mean	Std.
Fwd.	w/o	1.86	2.07	13.29	7.50	50.50	9.31
	w	8.08	7.33	25.75	14.21	60.96	18.08
Rev.	w/o	5.83	7.15	30.25	20.70	54.33	12.57
	w	15.79	8.29	36.71	19.24	51.54	8.93

improved the capability across most devices. Figures 9 and 10 show qualitative examples of gaze behavior for each device.

(B) Analysis on Workload: To gain further insights, we conducted an additional workload analysis, described in Appendix C. Table 4 shows the means and standard deviations of RTLX scores used for the repeated measures ANOVA. The analysis shows that both device type and the presence or absence of gaze-relevant instructions are statistically significant ($p < 0.05$), whereas the device usage order is not ($p \geq 0.10$). A larger effect size was observed for the device type ($\eta^2 = 0.6995$), suggesting that its impact was greater than that of the presence or absence of gaze-relevant instructions ($\eta^2 = 0.0244$). Specifically, switching from the Wearable to HMD-Ego increased workload by approximately 20 points, and from HMD-Ego to HMD-Top-down by approximately 30 points. These results align with intuition, showing that the degree of deviation from the natural visual condition leads to increased workload.

(C) Analysis on Sub-indicator: Figure 6 shows the effect of visual condition emulation devices on each NASA-TLX sub-indicator. Across all sub-indicators, scores increased stepwise from Wearable to HMD-Ego to HMD-Top-down. No sub-indicator contradicted the trend of total workload described in Sec. 4.2, result (B).

Discussion: In response to RQ2, we found that changing the viewpoint (from HMD-Ego to HMD-Top-down) impairs demonstrators' capability to extract task-relevant cues when gaze-relevant instructions are not provided. In addition, we revealed two key findings: (1) Changing the device type (from Wearable to HMD-Ego) significantly increases workload. This result suggests that HMDs should be avoided when remote viewing is not necessary. (2) No statistically significant difference in workload was observed based on the presence or absence of gaze-relevant instructions. Thus, providing gaze-related instructions is recommended whenever possible.

5 Additional Experiments: Policy-Based Comparison

Building on Sec. 4.1 and Sec. 4.2, we examined whether cognitive science-informed gaze behavior that extracts task-relevant cues improves policy robustness against environmental shifts. We evaluated performance on pick and place tasks in both in-distribution (ID) and out-of-distribution (OOD) environments (Fig. 12). The detailed setup (environments, tasks, policy, data) is in Appendix D.

Table 5 shows that the policy that uses gaze behavior from Wearable achieved comparable performance to the Oracle, which uses manually annotated gaze data. While similar performance was expected for the Baseline (non-gaze) and Wearable in the ID-Pick, the Wearable surprisingly outperformed the Baseline. This result suggests that gaze facilitates the extraction of cues from small, hard-to-see objects. Additionally, the Wearable improved 50.0 points over the Baseline in OOD-Pick

Table 5: Task success rates (%) conditioned on environment, task (rows), and gaze behavior data source (columns). Oracle uses manually annotated gaze as ground truth.

Env.	Task	Baseline	Oracle	Gaze Behavior Data Source Used by the Policy			
		(non-gaze)	(gt. gaze)	Wearable	UMI	Leader	Leader-Follower
ID	Pick	43.8	75.0	**75.0**	**75.0**	37.5	0.0
	Place	93.8	100.0	**100.0**	31.3	62.5	18.8
OOD	Pick	18.8	75.0	68.8	**75.0**	31.3	0.0
	Place	37.5	100.0	**81.3**	18.8	50.0	18.8

and 43.8 points in OOD-Place. This result suggests that cognitive science-informed gaze behavior improves policy robustness. Unexpectedly, the UMI performed worse in ID- and OOD-Place. This result reveals the limitations of applying zero-shot style gaze prediction to policies (described in Appendix. D), motivating future work on more generalizable representations.

6 Conclusion

In this study, we propose an experimental framework that systematically analyzes demonstrators' gaze behavior across a spectrum of demonstration devices. Experimental results reveal that demonstration devices with natural embodiment and visual conditions, such as the Wearable, are more effective in extracting task-relevant cues via gaze behavior. Additionally, to provide a policy with reliable task-relevant cues and demonstration data, the results suggest using the Wearable for the former and the Leader-Follower for the latter.

One promising direction for future work is to explore policy architectures that can jointly learn from gaze behavior (in the form of egocentric video) and demonstration data by extending EgoMimic [26]. Beyond extracting task-relevant cues, future research could leverage gaze to capture subgoals and apply this capability to hierarchical policy learning. We hope that our preliminary investigation of gaze behavior in imitation learning will contribute to the further advancement of the field.

References

[1] 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. $\pi 0$: A vision-language-action flow model for general robot control. *ArXiv*, abs/2410.24164, 2024.

[2] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *ArXiv*, abs/2303.04137, 2023.

[3] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016.

[4] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020.

[5] Y. Ishida, Y. Noguchi, T. Kanai, K. Shintani, and H. Bito. Robust imitation learning for mobile manipulator focusing on task-related viewpoints and regions. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2885–2892, 2024.

[6] J. Pelz, M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental brain research.*, 139:266–77, 2001.

[7] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[8] M. Land and D. Lee. Where we look when we steer. *Nature*, 369:742–744, 1994.

[9] M. F. Land and S. M. Furneaux. The knowledge base of the oculomotor system. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 352 1358:1231–1239, 1997.

[10] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–1328, 02 1999.

[11] M. M. Hayhoe, D. G. Bensinger, and D. H. Ballard. Task constraints in visual working memory. *Vision Research*, 38(1):125–137, 1998.

[12] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz. Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1), 2003.

[13] D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7:66–80, 12 1995.

[14] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[15] H. Kim, Y. Ohmura, A. Nagakubo, and Y. Kuniyoshi. Training robots without robots: Deep imitation learning for master-to-robot policy transfer. *IEEE Robotics and Automation Letters*, 8:2906–2913, 2022.

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

[17] Z. Fu, T. Z. Zhao, and C. Finn. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024.

[18] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163, 2024.

[19] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning*, 2024.

[20] 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. *ArXiv*, abs/2410.08464, 2024.

[21] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni. Eye-hand behavior in human-robot shared manipulation. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 4–13, 2018.

[22] A. Saran, E. S. Short, A. Thomaz, and S. Niekum. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*, 2019.

[23] A. L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.

[24] K. A. Turano, D. R. Geruschat, and F. H. Baker. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43(3):333–346, 2003.

[25] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. K. Gebreselasie, C. González, J. M. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolár, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbeláez, D. J. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. A. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2021.

[26] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video. In *CoRL Workshop on Learning Robot Fine and Dexterous Manipulation: Perception and Control*, 2024.

[27] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney. Predicting driver attention in critical situations. abs/1711.06406, 2018.

[28] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi. Gaze training by modulated dropout improves imitation learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7756–7761, 2019.

[29] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. Muller, J. Whritner, L. Zhang, M. Hayhoe, and D. Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6811–6820.

[30] A. Saran, R. Zhang, E. S. Short, and S. Niekum. Efficiently guiding imitation learning agents with human gaze. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, page 1109–1117, 2021.

[31] Y. Li, M. Liu, and J. M. Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:6731–6747, 2020.

[32] 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. C. 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. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. A. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022.

[33] A. Xie, L. Lee, T. Xiao, and C. Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160, 2024.

[34] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. volume 52 of *Advances in Psychology*, pages 139–183. 1988.

[35] B. Lai, M. Liu, F. Ryan, and J. M. Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132:854 – 871, 2022.

[36] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. *CVPR*, 2022.

A Limitations

Evaluation by individual emulation: We found that devices emulating (1) a robot's embodiment or (2) visual condition impair demonstrators' capability to extract task-relevant cues via gaze behavior, depending on the degree of emulation. However, devices that emulate both embodiment and visual condition have not yet been explored. While we assume that simultaneous constraints in both embodiment and visual condition impair gaze-based task-relevant cues extraction, future work should further investigate this assumption to uncover unexpected findings.

Focused Task Exploration: We conducted experiments using the pick-and-place task, which constitutes a significant portion of robot tasks. However, more complex and dexterous tasks have not yet been explored. For instance, we are particularly interested in human gaze behavior during tasks where avoiding collisions with the environment is critical, and in the cues that can be extracted from such gaze behavior.

B Detailed experimental setup: Investigating how embodiment differences influence the capability to extract task-relevant cues

Participants: We conducted an experiment to investigate how demonstration devices with different embodiments influence task-relevant cue extraction. Eight able-bodied participants (six males and two females, aged in their 20s to 40s) were recruited from within the institution. Participants were not limited to robotics researchers. To eliminate sequence effects, the participants were divided into two groups: one group followed the order (Wearable, UMI, Leader, Leader-Follower), and the other followed the reverse order (Leader-Follower, Leader, UMI, Wearable), with four participants in each group. All participants provided informed consent.

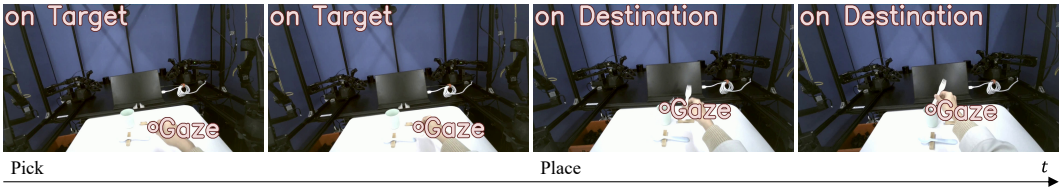
Equipment: We used Tobii Pro Glasses 3 as the eye tracker. This equipment captures eye movement and forward-facing camera video simultaneously, and shows where participants look in the video. In the Wearable setting, Tobii Pro Glasses 3 was also used as a wearable camera (Fig. 2a). For the UMI setting, we extracted and used the gripper part of ALOHA [16], as shown in Fig. 2a. For the Leader and Leader-Follower settings, we used ALOHA [16], as shown in Fig. 2a and 2b.

Procedure: Participants performed a pick-and-place task using multiple demonstration devices over up to three hours. The experiment consisted of two phases: first, using devices without gaze-related instructions; and second, with gaze-related instructions. For each device, a non-time-limited practice session was provided to reduce the effect of unfamiliarity, followed by a proficiency evaluation. Participants who passed the evaluation, indicating a certain skill level, proceeded to the actual measurements.

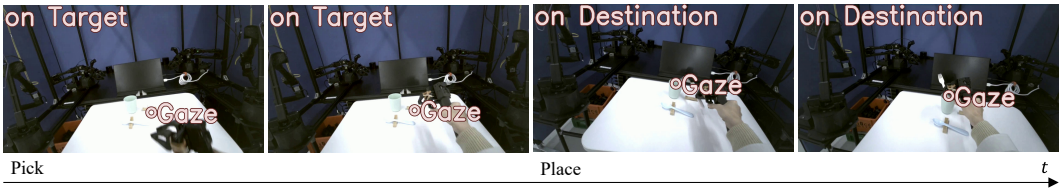
Each experimental trial followed a consistent sequence. Participants first stepped on a foot pedal to show an instruction on the display (Fig. 2a and Fig. 2b) and memorized it. They then stepped on the pedal again to clear the display and indicate the start of gaze measurement. Finally, they performed the pick-and-place task, moving the *target* to the *destination* based on the memorized instruction. Each device was used for eight such trials during the actual measurements. Table 6 lists the instructions used in the eight trials. Participants could use either left or right end-effector but

Table 6: High-level task instructions for embodiment experiment.

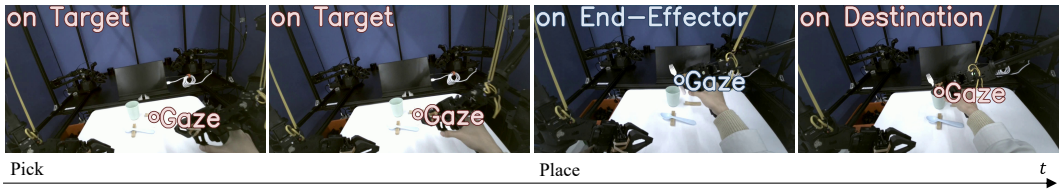
Trial	Instruction
1	Pink fork -> Green cup
2	Blue spoon -> Left knife rest
3	Pink fork -> Right knife rest
4	Blue spoon -> Green cup
5	Blue spoon -> Left knife rest
6	Pink fork -> Green cup
7	Blue spoon -> Right knife rest
8	Pink fork -> Left knife rest



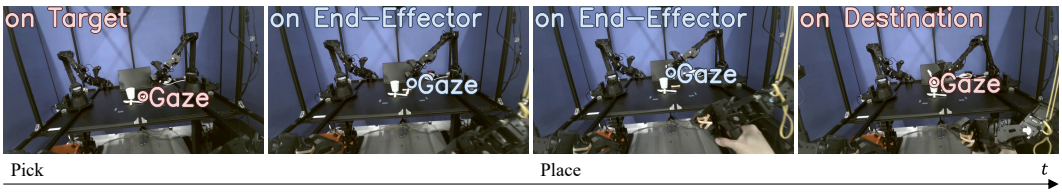
(a) Wearable



(b) UMI



(c) Leader



(d) Leader-Follower

Figure 7: Gaze sequences w/o gaze-relevant instructions in the embodiment experiment. The gaze collecting with higher degrees of embodiment emulation devices in (c) and (d) impairs the extraction of task-relevant cues (*target* and *destination*) and is instead directed toward the end-effector.

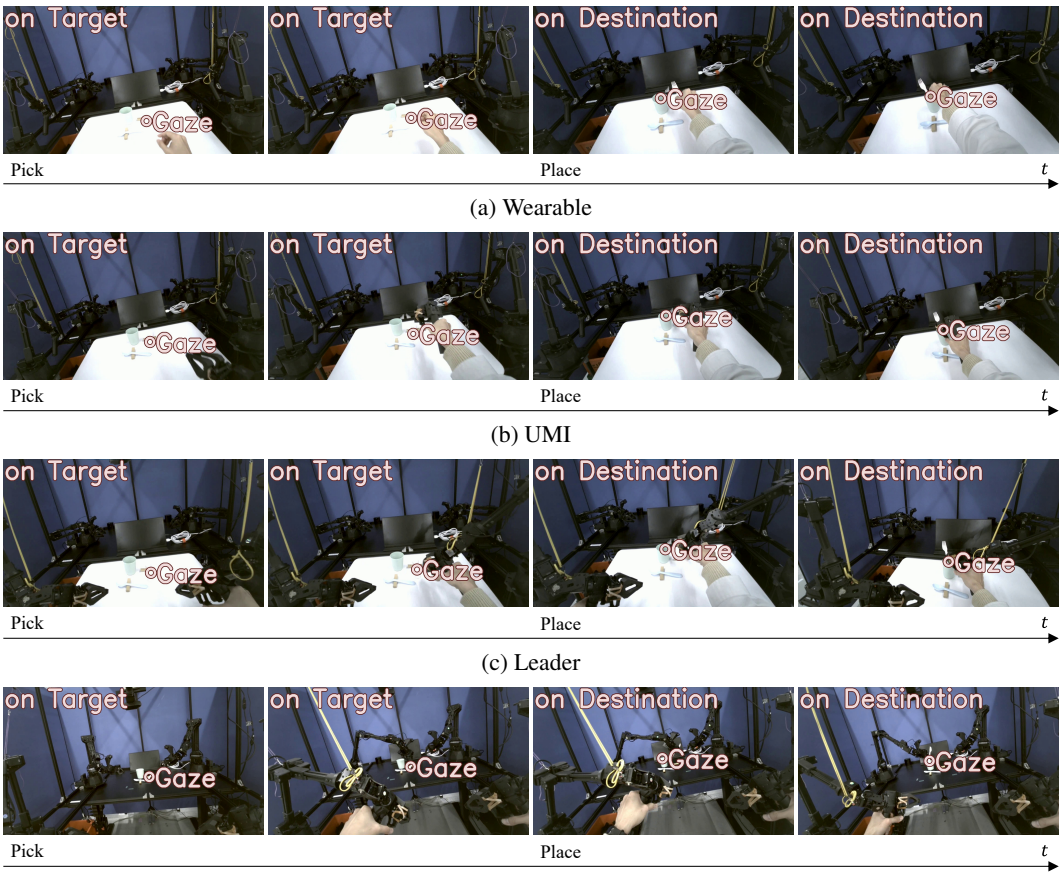
were required to use the same hand throughout a trial. No constraints were imposed on the direction or orientation when placing the *target* at the *destination*.

After all measurements, participants completed a survey assessing workload via NASA-TLX. Although not directly related to the RQ, this survey provided valuable insights. NASA-TLX measures workload based on six sub-indicators: mental demand (MD), physical demand (PD), temporal demand (TD), performance (P), effort (E), and frustration (F).

Data Collection and Annotation: As described in Sec. 3.1, data were collected and manually annotated for the *end effector*, *target*, and *destination*. The *end effector* was annotated at the point between the fingers grasping the object, the *target* at the grasped position, and the *destination* at the center of the placement area. Annotation for one participant required approximately six hours.

C Detailed experimental setup: Investigating how visual condition differences influence the capability to extract task-relevant cues

Participants: We conducted an experiment to investigate how demonstration devices with different visual conditions influence task-relevant cues extraction. Eight able-bodied participants (four males and four females, aged in their 20s to 40s) were recruited from within the institution. Participants were not limited to robotics researchers. To eliminate sequence effects, the participants were divided into two groups: one group followed the order (Wearable, HMD-Ego, HMD-Top-down), and the other followed the reverse order (HMD-Top-down, HMD-Ego, Wearable), with four participants in each group. All participants provided informed consent.



(a) Wearable

(b) UMI

(c) Leader

(d) Leader-Follower

Figure 8: Gaze sequences w/ gaze-relevant instruction in the embodiment experiment. Providing gaze-related instructions enhances the extraction of task-relevant cues (*target* and *destination*).

Equipment: We used Tobii Pro Glasses 3 and HTC VIVE Pro Eye as the eye tracker. Tobii Pro Glasses 3 captures eye movement and forward-facing camera video simultaneously, and shows where participants look in the video. HTC VIVE Pro Eye captures eye movement and video simultaneously, and shows where participants look in the video. In the Wearable setting, Tobii Pro Glasses 3 was also used as a wearable camera (Fig. 2c). For the HMD-Ego setting, we used HTC VIVE Pro Eye with ego view camera, as shown in Fig. 2c. For the HMD-Top-down setting, we used HTC VIVE Pro Eye with top-view camera, as shown in Fig. 2c.

Procedure: Participants performed a pick-and-place task using multiple demonstration devices over up to two hours. The experiment consisted of two phases: first, using the devices without gaze-related instructions; and second, with gaze-related instructions. For each device, a non-time-limited practice session was provided to reduce the effect of unfamiliarity, followed by a proficiency evalu-

Table 7: High-level task instructions for visual condition experiment.

Trial	Instruction
1	Pick up the pink fork and place it in the green cup.
2	Pick up the blue spoon and place it on the left knife rest.
3	Pick up the pink fork and place it on the right knife rest.
4	Pick up the blue spoon and place it in the green cup.
5	Pick up the blue spoon and place it on the left knife rest.
6	Pick up the pink fork and place it in the green cup.
7	Pick up the blue spoon and place it on the right knife rest.
8	Pick up the pink fork and place it on the left knife rest.

(a) Wearable

(b) HMD-Ego

(c) HMD-Top-down

Figure 9: Gaze sequences w/o gaze-relevant instructions in the visual condition experiment. The gaze collecting with top-down viewing in (c) impairs the extraction of task-relevant cues (*target* and *destination*) and is instead directed toward the end-effector.

ation. Participants who passed the evaluation, indicating a certain skill level, proceeded to the actual measurements.

Each experimental trial followed a consistent sequence. Participants first memorized the instructions read aloud by the experimenter. They then looked at a marker on the desk (Fig. 2d) to indicate the start of gaze measurement. Finally, they performed the pick-and-place task, moving the *target* to the *destination* based on the memorized instruction. Each device was used for eight such trials during the actual measurements. Table 7 shows the instructions used in the eight trials. Participants could use either left or right end-effector but were required to use the same hand throughout a trial. No constraints were imposed on the direction or orientation of placing the *target* at the *destination*.

After all measurements, participants completed a survey assessing workload via NASA-TLX. Although not directly related to the RQ, this survey provided valuable insights. NASA-TLX measures workload based on six sub-indicators: MD, PD, TD, P, E, and F.

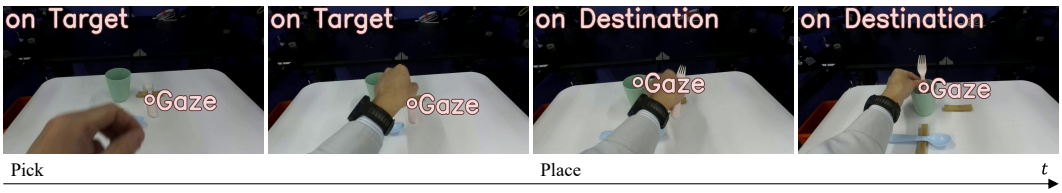
Data Collection and Annotation: As described in Sec. 3.1, data were collected and manually annotated for the *end effector*, *target*, and *destination*. The *end effector* was annotated at the point between the fingers grasping the object, the *target* at the grasped position, and the *destination* at the center of the placement area. Annotation for one participant required approximately four hours.

D Detailed experimental setup: Policy-Based Comparison

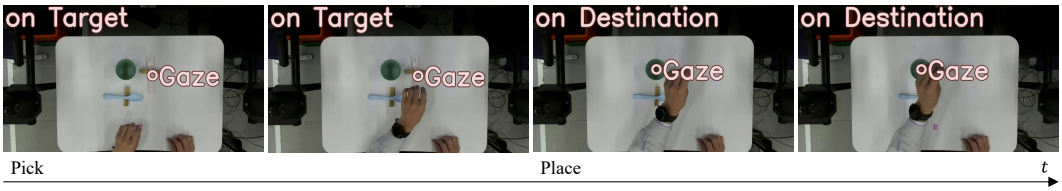
Focus: In this experiment, we compare gaze behavior in the embodiment experiment without gaze-related instruction, as outlined in Sec. 4.1. This focus is motivated by the fact that gaze behavior varies significantly depending on the degree of embodiment emulation by the demonstration devices, which might affect policy performance. Under this condition, the Wearable exhibits the most cognitive science-informed gaze behavior for extracting task-relevant cues, followed by UMI, Leader, and Leader-Follower.

Environments and Tasks: We compare these gaze behaviors based on the task "pick up the pink fork and place it on the green cup," in accordance with the embodiment experiment described in

(a) Wearable



(b) HMD-Ego

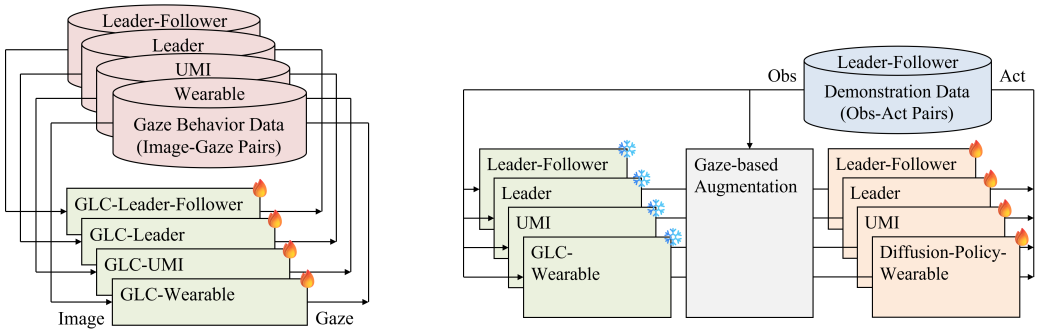


(c) HMD-Top-down

Figure 10: Gaze sequences w/ gaze-relevant instruction in the visual condition experiment. Providing gaze-related instructions enhances the extraction of task-relevant cues (*target* and *destination*).

Sec. 4.1. We used a CNN-based Diffusion Policy (DP) [2], which is commonly used in imitation learning. Note that, because DP employs a ResNet backbone, it lacks temporal modeling capabilities. As a result, the policy struggles to capture the transition of the task object from the pink fork (pick) to the green cup (place). To address this limitation, we trained separate policies for the pick and place phases and evaluated them independently. Figure 12 shows the in-distribution (ID) and distractor object-induced out-of-distribution (OOD) environments, as well as the pick and place tasks, used for evaluation. The task success rate was calculated from the results of 16 trials.

Policy and Data: Following the findings regarding trade-offs discussed in Sec. 4.1, we collected gaze behavior data (image-gaze pairs) for each demonstration device and demonstration data (observation-action pairs) for the Leader-Follower. Figure 11 illustrates the data collection and training pipeline, which consists of two phases: (1) We performed the pick-and-place task with each demonstration device and collected the corresponding image-gaze pairs over 40 episodes (Fig. 11a, top). As a model to estimate gaze behavior collected using each device, we trained egocentric gaze estimation model, Global-Local Correlation (GLC) [35], using each image-gaze pair (Fig. 11a, bottom). (2) We performed the pick-and-place task with the Leader-Follower and collected observation-action pairs over 40 episodes (Fig. 11b, top). We then estimated gaze behavior on the demonstration



(a) (b)

Figure 11: Data collection and training pipeline. (a) The egocentric gaze estimation model (GLC) is trained on gaze behavior data. (b) The DP is trained on gaze-based augmented demonstration data.

Table 8: Hyperparameters for DP [2].

Parameter	Value
Ctrl	Pos
To	2
Ta	8
Tp	16
ImgRes	1x320x240
CropRes	1x228x216
#D-Params	67 millions
#V-Params	22 millions
Lr	1e-4
WDecay	1e-6
D-Inters Train	100
D-Inters Eval	8

Table 9: Hyperparameters for GLC [35].

Parameter		Demonstration Device-dependent values			
		Wearable	**UMI**	**Leader**	**Leader-Follower**
Pick	$sampling\_rate$	2	4	9	12
	$length$	0.9	1.9	3.7	5.2
Place	$sampling\_rate$	3	5	6	15
	$length$	1.3	2.1	2.5	6.3

data using each GLC model in a zero-shot manner. As a policy, we trained a CNN-based DP using gaze-based augmented data (Fig. 11b, bottom). We applied gaze-based augmentation to enhance the policy's robustness in extracting task-relevant cues guided by gaze behavior.

The gaze-based augmentation procedure was as follows: We computed the pixel-wise distance D from the predicted gaze coordinate (x, y) using Eq. 1. A hybrid weighting strategy was used: a hard value was applied within a threshold radius r, and a soft Gaussian decay was applied beyond r, as shown in Eq. 2. The decay according to distance was controlled by the parameter σ. The resulting weight map W was then normalized. Using the normalized weight map, we performed spatially-aware data augmentation: High-weight regions retained more of the original image, while low-weight regions retained more of the images augmented using PixMix [36], a fractal pattern-based augmentation method. In this experiment, we used $r = 30.0$ and $\sigma = 100.0$.

$$D(i,j) = \sqrt{(i-x)^2 + (j-y)^2} \tag{1}$$

$$W(i,j) = \begin{cases} 1 & \text{if } D(i,j) \leq r \\ \exp\left(-\frac{(D(i,j)-r)^2}{2\sigma^2}\right) & \text{otherwise} \end{cases} \tag{2}$$

Tables 8 and 9 show the hyperparameters used for training. $sampling\_rate$ is a GLC parameter determined by Eq. 3. $length$ denotes the duration in seconds required to complete a pick or place. $frame\_rate$ refers to the frame rate of the gaze behavior data. $ratio$ determines the temporal segment within the duration $length$ from which time-series images are sampled. num_frames is a GLC parameter indicating the number of images to be sampled. In this experiment, we used $frame\_rate = 24$, $ratio = 0.8$, and $num\_frames = 8$.

$$sampling\_rate = \frac{(length \times frame\_rate \times ratio)}{num\_frames} \tag{3}$$

(a) ID-Pick (b) ID-Place (c) OOD-Pick (d) OOD-Place

Figure 12: Evaluation environments and tasks. In the ID environments, the same pink fork, green cup, and knife rest as in Sec. 4 are used. In the OOD environments, distractor objects induce the distribution shift.