Comparing Visual and Neural Representations for Imitation Learning in Robot Manipulation

Ethan Kogan
Paul G. Allen School of Computer Science and Engineering
University of Washington
dbzfan@uw.edu

Alexander Metzger
Paul G. Allen School of Computer Science and Engineering
University of Washington
metzgera@uw.edu

Abstract: Robot policies depend on how the environment is represented. We compare three representations for imitation learning: raw RGB images, segmented images, and EEG signals recorded from a human observer. Using identical demonstrations of a pick-and-place task, we train policies with each modality. Visual inputs perform best, but EEG still produces task-correlated behavior, suggesting neural observations contain weak information about task state that potentially could be leveraged to enable future accessibility interfaces for assistive robots.

Keywords: Imitation Learning, Image Segmentation, Brain-Computer Interface

1 Introduction

Robotic manipulation policies rely on representations of the environment to determine appropriate actions. In imitation learning systems, these representations form the observation space from which policies infer the current task state. The design of this representation strongly influences learning efficiency, robustness, and generalization.

Most modern robot learning approaches derive representations directly from visual perception. End-to-end visuomotor policies have demonstrated that manipulation behaviors can be learned from raw camera images without explicit perception pipelines [1]. More recent systems combine imitation learning with transformer-based architectures to learn complex manipulation tasks from demonstration using visual observations [2]. While effective, raw visual observations often contain large amounts of irrelevant information such as background clutter and lighting variation.

A common strategy is therefore to transform raw images into more structured representations that emphasize task-relevant components. Image segmentation methods such as the Segment Anything Model can provide object-level abstractions that simplify the learning problem by isolating important scene elements [3, 4]. These representations reduce visual complexity while preserving spatial structure.

A very different form of representation arises in brain–computer interface (BCI) systems, where neural signals are used to control robotic devices. EEG-based systems have enabled users with motor impairments to operate assistive robots through decoded neural activity [5]. In these systems, neural signals are typically interpreted as high-level control commands rather than as a representation of the task state. This means their ability to adapt to new accessibility contexts is limited by the low-level actions.

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

Meanwhile, humans observing a task produce neural responses correlated with visual stimuli, attention, and motor understanding. This raises a broader question: can neural signals recorded from a human observer encode enough information about a task to support robot learning?

Research Question. *How does the choice of observation representation—from raw visual perception to human neural signals—affect the ability of imitation learning systems to learn robotic manipulation tasks?*

To study this question, we compare three representations that span a spectrum from environment-centric sensing to human-centered signals: (1) raw RGB images from a robot-mounted camera, (2) segmented images that highlight task-relevant objects, and (3) EEG signals recorded from a human observer watching the task. Using a leader–follower setup with two SO-101 robotic arms [6], we collect demonstrations of a pick-and-place task while simultaneously recording visual observations and EEG signals. Policies are then trained using the same imitation learning architecture but different observation modalities.

Our experiments show that visual representations remain the most reliable for learning manipulation policies. Segmented images achieve comparable performance while sometimes improving robustness to out-of-distribution states. Policies trained directly on raw EEG signals perform substantially worse but still exhibit task-correlated behavior, suggesting that neural observations contain weak information about the task state.

Contributions

- We introduce a representation-centric comparison of imitation learning for manipulation spanning visual perception and human neural signals.
- We present a synchronized data collection pipeline that records robot demonstrations, camera observations, and EEG signals.
- We empirically compare raw images, segmented images, and EEG signals as representations for learning manipulation policies.

2 Methodology

Our goal is to compare how different observation representations affect imitation learning performance while holding all other factors constant. To isolate the effect of representation, we collect demonstrations once and derive three observation modalities from the same underlying data: raw RGB images, segmented images, and EEG signals. All policies are trained using the same architecture and hyperparameters.

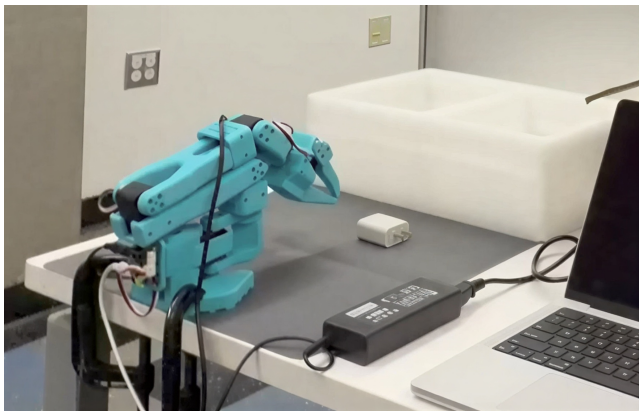

Figure 1: **Task setup.** The follower arm must grasp the block and place it into the container.

2.1 Task and Hardware Setup

Experiments are conducted using the SO-101 3D-printed robotic arms [7], which includes native compatibility with the LeRobot framework [6]. Demonstrations are collected in a leader–follower configuration: a human teleoperates the leader arm while the follower arm records joint states and observations.

The manipulation task requires the follower arm to pick up a rectangular block placed on the workspace and drop it into a container. The initial configuration is shown in Fig. 1. To reduce visual noise and ensure consistent observations, the workspace is standardized with a uniform gray surface and controlled lighting.

For neural recordings we use a Neuropawn EEG headset with eight channels positioned according to the 10–10 electrode system [8]. Channels were selected to cover regions associated with motor planning and visuomotor processing. The electrode layout is shown in Fig. 2.

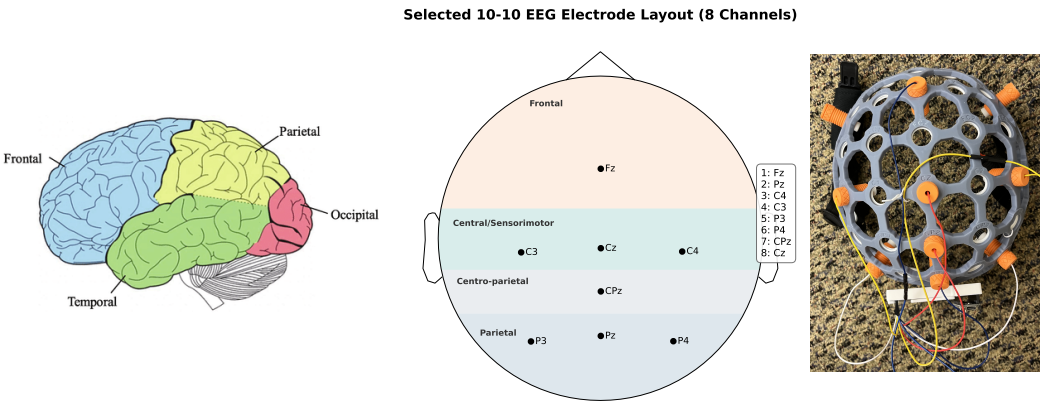

Figure 2: **EEG electrode layout.** Eight channels from the standard 10-10 system are used: Fz, Pz, C4, C3, P3, P4, CPz, and Cz. The brain diagram is from [9] and the electrode layout diagram was made with Python MNE [10].

2.2 Representation Modalities

We evaluate three observation representations spanning a spectrum from environment-centric sensing to human neural signals.

Raw Visual Observations. RGB images from the robot-mounted camera serve as the baseline representation. These images provide full scene context but include background clutter and lighting variation.

Segmented Visual Observations. To produce a structured representation, we apply FastSAM [4] segmentation to the recorded images. Segmentation highlights task-relevant objects while suppressing background information. The resulting masked images are used as observations during training and evaluation.

Neural Observations. EEG signals recorded from a human observer watching the task provide a human-centered representation. Eight-channel neural signals are sampled at 125 Hz concurrently with robot observations and appended to the policy input in place of visual data.

2.3 Data Collection

We collect 170 demonstrations of the manipulation task. During each episode, the leader arm teleoperates the follower through the full pick-and-place sequence while recording robot joint states and

actions, RGB camera observations, and synchronized EEG signals from the human observer. Data collection is implemented using custom extensions to the LeRobot [6] pipeline that integrate the BrainFlow API [11] for real-time EEG recording. Because all modalities are recorded simultaneously, the resulting datasets correspond to identical demonstrations. Segmented image datasets are generated offline by applying FastSAM [4] to the recorded camera frames. To improve robustness, demonstrations include multiple initial object configurations, varying the position and orientation of the block.

2.4 Policy Training

All policies are trained using Action Chunking with Transformers (ACT) [2]. To isolate the impact of representation, training hyperparameters are kept identical across experiments. Each model is trained for 50,000 optimization steps with a batch size of 8. The only difference between models is the observation modality. Training datasets of varying sizes (10, 50, and 170 demonstrations) are used to analyze scaling behavior.

2.5 Evaluation Procedure

Policies are evaluated on the physical robot using a binary task success metric: the robot must successfully grasp the block and place it in the container. Each trained model is tested across six initial object configurations: four in-distribution locations observed during training (middle, left, right, rotated), and two out-of-distribution locations (far left and rotated right). For the visual representations, policies operate directly on camera observations (with brief segmentation preprocessing from FastSAM). For EEG evaluation, the observer wears the EEG headset and focuses on the robot performing the task to reproduce neural activity patterns similar to those during training. The leader arm remains disconnected to avoid teleoperation interference. Success rates and qualitative failure modes are analyzed to understand how representation choice affects policy behavior.

3 Experimental Evaluation

Figure 3 compares success rates across the three representation modalities and training dataset sizes. Visual observations consistently produce the strongest policies, while neural observations provide weaker but still task-correlated signals. In addition to overall success rates, qualitative observations during evaluation reveal distinct behavioral patterns and failure modes associated with each representation.

3.1 Performance Scaling with Dataset Size

Across all representations, increasing the number of demonstrations improves performance. However, the rate of improvement differs substantially across modalities.

With only 10 demonstrations, the camera-based policy occasionally completes the manipulation sequence but is generally unreliable. It succeeds in 2 out of 4 trials when the object begins in the center location but fails entirely on other starting configurations. Qualitatively, the robot often begins the correct sequence but lacks consistency in grasp alignment and object placement. With human assistance it was able to complete the task, suggesting that it had learned the general structure of the behavior but lacked sufficient training data to reliably execute it.

The EEG-based policy trained on 10 demonstrations shows no successful task completions. However, the robot still responds to the observer's cognitive state. When the wearer focuses on the robot and the task, the arm tends to move toward the object and occasionally toward the drop location. When the wearer becomes distracted (for example, performing an unrelated task on a phone), the robot instead exhibits small exploratory motions and quickly returns to a neutral position. This behavior suggests that even at small dataset sizes the model is detecting coarse neural signals related to attention, but lacks enough information to complete the task.

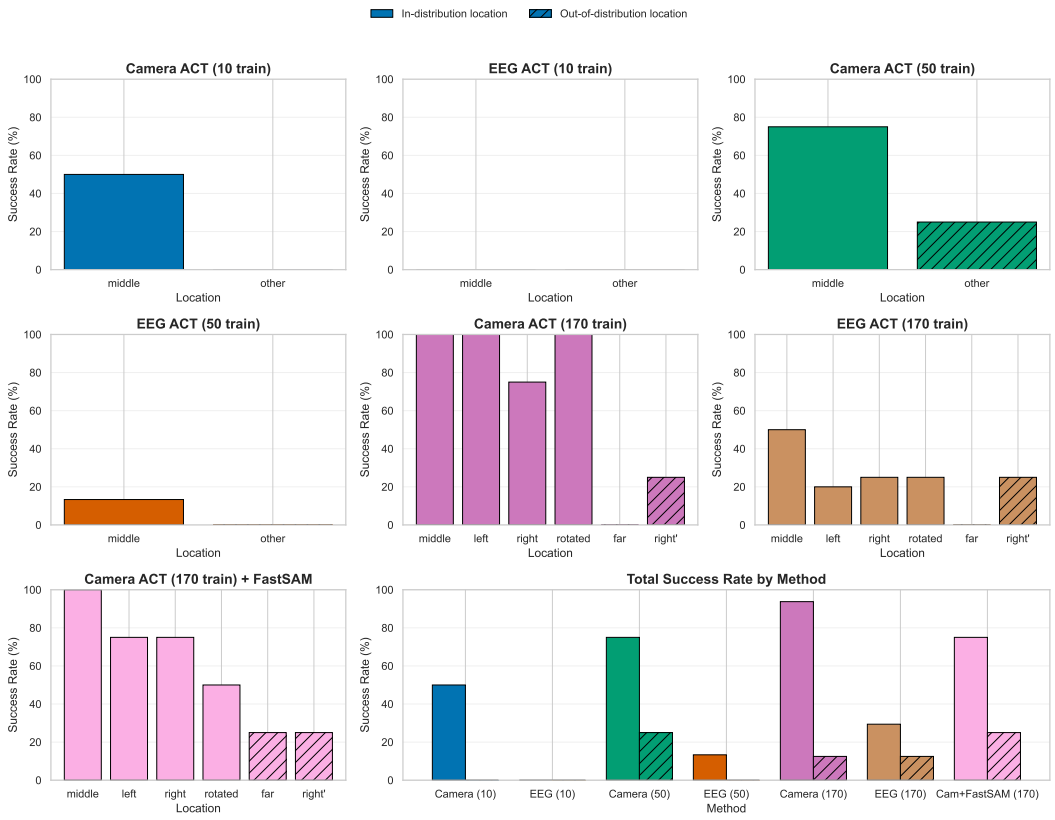

Figure 3: **Success Rate by Method, Training Data, and Charger Location**. We test each method on picking up the charger from four known positions from the training data (left, right, middle, and middle but with the charger rotated 90 degrees) and two unseen locations (far left, and rotated on the right). We can see that both the EEG and Camera policy success rates improve with the number of training samples. The camera input is overall more effective than the EEG input.

At 50 demonstrations, the camera-based policy improves significantly. It succeeds in 75% of trials when the block starts in the center position and occasionally generalizes to nearby positions not seen during training. Failures at this stage typically occur when the robot makes small positioning errors during the grasp step. In these cases the gripper opens near the object but slightly misaligned, pushing the block away and causing the subsequent grasp attempt to fail.

The EEG-based policy trained on 50 demonstrations begins to exhibit limited task-following behavior. It succeeds twice out of fifteen trials in the middle position but fails entirely on other configurations. Once the arm deviates from the expected trajectory, it often becomes unable to recover and drifts into unstable motions. As a control experiment, we trained an identical model on the same demonstrations but with both the camera and EEG inputs zeroed out. This control model was unable to pick up the block in any trials, confirming that the learned behavior in the EEG model is not simply memorizing a trajectory and that some signal is present in the neural inputs.

3.2 Large-Scale Training (170 Demonstrations)

With 170 demonstrations, the differences between representations become clearer.

The raw camera policy performs the best overall. It achieves a 94% success rate across the four in-distribution positions (middle, left, right, and rotated) and succeeds in roughly two-thirds of all trials including unseen configurations. Failures occur primarily in the out-of-distribution positions,

particularly when the block is placed far to the side of the workspace where it becomes smaller in the camera view and harder to distinguish from the background. In these cases the policy sometimes moves toward the correct region but fails to properly localize the object.

Another observed failure mode occurs during grasp alignment. Occasionally the robot opens the gripper slightly off target, pushing the block away before closing the gripper. Interestingly, the larger 170-demonstration model appears slightly less likely than the 50-demonstration model to attempt recovery after a failed grasp, suggesting that the policy may have overfit to the typical successful trajectory observed during training.

The segmented-image representation performs similarly on the center position and moderately well across other in-distribution configurations, achieving a 75% success rate overall on known locations. Its primary failure mode arises from segmentation errors: if the segmentation model fails to identify the block for even a short sequence of frames, the policy loses track of the object and the arm drifts away from the target. Despite this limitation, segmentation improves performance in some unseen conditions. In particular, the policy succeeds approximately twice as often as the raw camera policy when the block is placed in more distant or unusual positions. Qualitatively, the segmented policy is also more likely to retry grasp attempts after failure, remaining near the object rather than abandoning the task.

The EEG-based policy remains substantially less reliable than the visual policies but shows meaningful improvements with increased data. With 170 demonstrations it succeeds in approximately 31% of in-distribution trials and 24% overall. The arm consistently moves toward the correct region of the workspace and often attempts to grasp the block, but failures typically occur during fine alignment or gripping. Unlike the smaller models, the 170-demonstration EEG policy is sometimes able to attempt multiple grasp attempts if the first one fails, indicating that it has learned some temporal structure of the task.

Several qualitative behaviors suggest that the EEG representation encodes coarse task context. The robot almost always moves in the correct general direction (left, right, or center) relative to the block's location, and often correctly approaches rotated configurations. However, the policy lacks the spatial precision necessary for reliable grasp positioning. Additionally, the robot's behavior remains sensitive to the wearer's cognitive state. When the observer is focused on the task, the robot attempts the manipulation sequence, whereas distraction typically causes the arm to stop attempting the task and return to neutral motions.

Overall, these results demonstrate a clear gradient in representation quality. Direct visual sensing provides the most precise and reliable information for manipulation, structured visual abstractions trade some reliability for improved robustness, and neural signals provide weaker but still detectable information about task context and intent.

4 Discussion

Our experiments highlight how the choice of observation representation fundamentally shapes the performance and capabilities of imitation learning systems.

4.1 Environment-Centric Representations

Raw camera observations provide the most direct information about the physical environment and therefore produce the most reliable manipulation policies. These policies achieve near-perfect success on in-distribution tasks and generalize moderately to unseen configurations. This result is consistent with prior work demonstrating the effectiveness of end-to-end visuomotor learning for robotic manipulation.

Segmented visual observations represent a structured intermediate representation. By removing background information and highlighting task-relevant objects, segmentation reduces the dimensionality of the observation space. In our experiments this representation performed slightly worse

overall but demonstrated improved robustness in some out-of-distribution conditions. This suggests that structured perceptual representations may encourage policies to focus on causal task features rather than incidental visual details.

4.2 Human-Centric Representations

Neural observations represent a fundamentally different form of representation. Rather than directly sensing the environment, the policy must infer task state indirectly from the neural activity of a human observer. Despite this challenge, EEG-based policies still exhibit task-correlated behavior and measurable success rates.

The main limitation of the EEG representation is precision. Neural signals appear to encode coarse information about task progress and intent, but they do not provide the spatial resolution necessary for reliable fine-grained manipulation. This leads to common failure modes during grasp alignment and object pickup.

However, the improvement in success rates with larger datasets suggests that neural signals do contain usable information. With more sophisticated signal processing or architectures designed specifically for neural data, these representations may become significantly more effective.

4.3 Implications for Accessible Robotics

These results support a broader view of robot learning representations as lying on a spectrum between autonomous sensing and human-mediated signals. Environment-centric sensing (e.g., cameras) enables highly autonomous systems, while human-centered signals such as EEG enable forms of interaction where users may guide or influence robot behavior without direct physical control.

For accessibility applications, this spectrum is particularly important. Individuals with severe motor impairments may not be able to provide physical demonstrations or operate traditional control interfaces. Neural signals offer a potential pathway for these users to communicate intent to robotic systems, even if those signals are currently less precise than visual sensing.

Our results suggest that future assistive robots may benefit from hybrid approaches that combine environment perception with human neural or cognitive signals, allowing robots to maintain autonomy while remaining responsive to user intent.

5 Conclusion

In this work we investigated how different observation representations influence imitation learning for robotic manipulation. We compared three modalities spanning a spectrum from environment-centric sensing to human-centered neural signals: raw RGB images, segmented images, and EEG recordings from a human observer. Our results show that visual representations remain the most reliable input for learning manipulation policies, with raw images achieving the highest success rates. Structured visual abstractions produced through segmentation perform comparably while occasionally improving robustness to unseen configurations. Policies trained on neural observations perform substantially worse but still demonstrate task-correlated behavior, indicating that human neural signals contain weak but meaningful information about task state. These findings highlight a broader design space for robot learning systems in which representations range from direct environmental perception to indirect signals derived from human cognition. While neural representations alone are currently insufficient for precise manipulation, they offer a promising pathway toward accessible interfaces for assistive robotics. Future work could explore hybrid perception architectures that combine environmental sensing with neural or cognitive signals, improved neural signal processing pipelines, and larger-scale datasets. Such systems may enable assistive robots that remain autonomous while adapting to user intent through non-traditional interaction modalities.

References

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, Jan. 2016. ISSN 1532-4435.

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

[4] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang. Fast segment anything, 2023. URL https://arxiv.org/abs/2306.12156.

[5] A. Casey, H. Azhar, M. Grzes, and M. Sakel. Bci controlled robotic arm as assistance to the rehabilitation of neurologically disabled patients. *Disability and Rehabilitation: Assistive Technology*, 16(5):525–537, 2021. doi:10.1080/17483107.2019.1683239. URL https://doi.org/10.1080/17483107.2019.1683239. PMID: 31711336.

[6] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[7] R. Knight, P. Kooijmans, R. Cadene, S. Alibert, M. Aractingi, D. Aubakirova, A. Zouitine, R. Martino, S. Palma, C. Pascal, and T. Wolf. Standard Open SO-100 & SO-101 Arms. URL https://github.com/TheRobotStudio/SO-ARM100.

[8] G. E. Chatrian, E. Lettich, and P. L. Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities. *American Journal of EEG Technology*, 25(2): 83–92, 1985. doi:10.1080/00029238.1985.11080163. URL https://doi.org/10.1080/00029238.1985.11080163.

[9] W. Hu, G. Huang, L. Li, L. Zhang, Z. Zhang, and Z. Liang. Video-triggered eeg-emotion public databases and current methods: A survey. *Brain Science Advances*, 6(3):255–287, 2020. doi:10.26599/BSA.2020.9050026.

[10] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi:10.3389/fnins.2013.00267.

[11] B. Developers. Brainflow: A library intended to obtain, parse and analyze eeg, emg, ecg, and other kinds of data from biosensors. https://github.com/brainflow-dev/brainflow, 2018. Accessed: 2026-03-11.

[12] NeuroPawn Biotechnologies Inc. Neuropawn products: The neuropawn biopotential kit. https://www.neuropawn.tech/products/, 2025. Accessed: 2026-03-11.

[13] D. R. Freer and G. Yang. Mindgrasp: A new training and testing framework for motor imagery based 3-dimensional assistive robotic control. *CoRR*, abs/2003.00369, 2020. URL https://arxiv.org/abs/2003.00369.