NOIR: Neural Signal Operated Intelligent Robot for Everyday Activities

Abstract: We present Neural Signal Operated Intelligent Robot (NOIR), a general-purpose, intelligent brain-robot interface system that enables humans to command robots to perform everyday activities through brain signals. Through this interface, humans communicate their intended objects of interest and actions to the robots using electroencephalography (EEG). Our novel system demonstrates success in an expansive array of 20 challenging, everyday household activities, including cooking, cleaning, personal care, and entertainment. The effectiveness of the system is improved by its synergistic integration of robot learning algorithms, allowing for NOIR to adapt to individual users and predict their intentions. Our work enhances the way humans interact with robots, replacing traditional channels of interaction with direct, neural communication. Project website: https://sites.google.com/view/noir-corl2023

Keywords: Brain-Robot Interface; Human-Robot Interaction

1 Introduction

Brain-robot interfaces (BRIs) are a pinnacle achievement in the realm of art, science, and engineering. This aspiration, which features prominently in speculative fiction, innovative artwork, and groundbreaking scientific studies, entails creating robotic systems that operate in perfect synergy with human beings. A critical component of such systems is their ability to communicate with humans. In human-robot
collaboration and robot learning, humans communicate their intents through actions [1], gaze [2], facial
expression [3], language [4], etc. However, the prospect of direct communication through neural signals
stands out to be the most thrilling but challenging medium.

We present Neural Signal Operated Intelligent Robot (NOIR), a general-purpose, intelligent brain-robot
interface system with non-invasive electroencephalography (EEG), that enables individuals to command
robots directly through brain signals. The primary principle of this system is hierarchical shared autonomy,
where humans define the overarching high-level goals while the robot actualizes the goals through the
execution of low-level motor commands. This model leverages the unique strengths of both participants,
establishing a symbiotic relationship between human cognitive capabilities and robotic precision. Taking
advantage of the recent progress in neuroscience, robotics, and machine learning, our system distinguishes
itself by extending beyond previous attempts to make the following contributions.

First, NOIR is general-purpose in its diversity of tasks and accessibility to the general population. For task
diversity, we show that humans can accomplish an expansive array of 20 daily everyday activities, in contrast
to existing BRI systems that are typically specialized at one or a few tasks or exist solely in simulation.
Additionally, the system can be used by the general population, with a minimum amount of training.

Second, the “I” in NOIR means that our robots are intelligent and adaptive. The robots are intelligent
in the sense that they are equipped with a library of diverse skills, allowing them to perform low-level
actions without dense human supervision. Human behavioral goals can naturally be communicated,
interpreted, and executed by the robots with parameterized primitive skills, such as \texttt{Pick(obj-A)} or
\texttt{MoveTo(x,y)}. Additionally, our robots are capable of learning and predicting human intended goals
during their collaboration. We show that by leveraging the recent progress in foundation models and robot
learning from humans, we can make such a system more adaptive with limited data. We show that this
can significantly increase the efficiency of the system.

The key technical contributions of NOIR include a modular neural signal decoding pipeline for human
intentions and its integration with our robotic systems. Decoding human intended goals (e.g., “I want to pick
up the mug from the handle”) from neural signals is extremely challenging. We decompose human intention
into three components: \textit{What} object to manipulate, \textit{How} to interact with the object, and \textit{Where} to interact,
and show that such signals can be decoded from different types of neural data. These decomposed signals
naturally correspond to parameterized robot skills and can be communicated effectively to the robots.

In 20 household activities involving tabletop or mobile manipulations, three human subjects successfully
use our system to accomplish these tasks with their brain signals. We demonstrate that few-shot robot
learning from humans can significantly improve the efficiency of our system. This approach to building
intelligent robotic systems, which utilizes human brain signals for collaboration, holds immense potential
for the development of critical assistive technologies for individuals with or without disabilities and to
improve the quality of their life.

\section{Brain-Robot Interface (BRI): Background}

Since Hans Berger’s discovery of electroencephalography (EEG) in 1924, several types of devices are
developed to record human brain signals. We choose non-invasive, saline-based EEG as the recording
device due to its cost and accessibility to the general population, signal-to-noise ratio, temporal and
spatial resolutions, and types of signals that can be decoded (see Appendix 2). In EEG, brain activity
is captured by measuring the spontaneous electrical activity of the brain using electrodes placed on the
scalp. EEG-based BRI has been applied to prosthetics, wheelchairs, as well as navigation and manipulation
robots. For comprehensive reviews, see [5–8]. We utilize two types of EEG signals that are frequently
employed in BRI, namely, steady-state visually evoked potential (SSVEP) and motor imagery (MI).

SSVEP is the brain’s exogenous response to periodic external visual stimulus [9], wherein the brain
generates periodic electrical activity at the same frequency as flickering visual stimulus. The application of
SSVEP in assistive robotics often involves the usage of flickering LED lights physically affixed to different
objects [10, 11]. Attending to an object (and its attached LED light) will increase the EEG response at that
stimulus frequency, allowing the object’s identity to be inferred. Inspired by prior work [12], our system
Figure 2: NOIR has two components, a modular pipeline for decoding goals from human brain signals, and a robotic system with a library of primitive skills. The robots possess the ability to learn to predict human intended goals hence reducing the human effort required for decoding.

utilizes computer vision techniques to detect and segment objects, attach virtual flickering masks to each object, and display them to the participants for selection.

Motor Imagery (MI) differs from SSVEP due to its endogenous nature, requiring individuals to mentally simulate specific actions, such as imagining oneself waving a hand or manipulating an object. The decoded signals can be used to indicate a human’s intended way to interact with the object. This approach is widely used for rehabilitation, and for navigation tasks [13] in BRI systems. This approach often suffers from low decoding accuracy [8], which may reduce the efficiency of the system.

Much existing BRI research focuses on the fundamental problem of brain signal decoding, while several existing studies focus on how to make robots more intelligent and adaptive [14, 12, 15–18]. Inspired by this line of work, we leverage few-shot policy learning algorithms to enable robots to learn human preferences and goals. This minimizes the necessity for extensive brain signal decoding, thereby streamlining the interaction process and enhancing overall efficiency.

Our study is grounded in substantial advancements in both the field of brain signal decoding and robot learning. Currently, many existing BRI systems target only one or a few specific tasks. To the best of our knowledge, no previous work has presented an intelligent, versatile system capable of successfully executing a wide range of complex tasks, as demonstrated in our study.

3 The NOIR System

The section aims to equip the reader with an intuitive understanding before we dive into the details. The challenges we try to tackle are: 1) How do we build a general-purpose BRI system that works for a variety of tasks? 2) How do we decode relevant communication signals from human brains? 3) How do we make robots more intelligent and adaptive for more efficient collaboration? An overview of our system is shown in Fig. 2. NOIR is a hierarchical shared autonomy system in which humans and robots collaborate to complete various tasks. Humans act as planning agents to perceive, plan, and communicate behavioral goals to the robot, while robots use pre-defined primitive skills to achieve these goals.

The overarching goal of building a general-purpose BRI system is achieved by synergistically integrating two designs together. First, we propose a novel modular brain decoding pipeline for human intentions, in which the human intended goal is decomposed into three components: what, how, and where (Sec. 3.1). Second, we equip the robots with a library of parameterized primitive skills to accomplish human-specified goals (Sec. 3.2). This design enables humans and robots to collaborate to accomplish a variety of challenging, long-horizon everyday tasks. At last, we show a key feature of NOIR to allow robots to act more efficiently and to be capable of adapting to individual users, we adopt few-shot intention (skill and parameter choices) learning from humans (Sec. 3.3). Our approach utilizes recent progress in large, pre-trained foundation models and few-shot learning.

3.1 The brain: A modular decoding pipeline

We hypothesize that the key to building a general-purpose EEG decoding system is modularization. Decoding complete behavioral goals (e.g., in the form of natural language) is only feasible with expensive devices like fMRI, and with many hours of training data for each individual [19]. As shown in Fig. 3, in NOIR, we decompose human intention into three components: (a) What object to manipulate; (b) How to interact with the object; (c) Where to interact (see Figure 3). The decoding of specific user intents from EEG signals is
Figure 3: A modular pipeline for decoding human intended goals from EEG signals: (a) What object to manipulate, decoded from SSVEP signals using CCA classifiers; (b) How to interact with the object, decoded from MI signals using CSP+QDA algorithms; (c) Where to interact, decoded from MI signals. A safety mechanism that captures muscle tension from jaw clench is used to confirm or reject decoding results.

Selecting objects with steady-state visually evoked potential (SSVEP). Upon showing the task set-up on a screen, we first infer the user’s intended object using SSVEP. We make the various objects on screen flicker with different frequencies (Fig. 3a), which, when focused on by the user, evokes SSVEP. By identifying which frequency is significantly stronger in the EEG data, we may infer the frequency of the flickering visual stimulus, and hence the object that the user focuses on. As mentioned in Sec. 2, we apply modern computer vision techniques to circumvent the problem of having to physically attach LED lights, which limits the usage of SSVEP-based systems. Specifically, we use the pre-trained foundation model OWL-ViT to detect and track objects, which takes in an image and object descriptions and outputs object segmentation masks. By overlaying each mask of different flickering frequencies (6Hz, 7.5Hz, 8.57Hz, and 10Hz), and having the user focus on the desired object for 10 seconds, we are able to identify the SSVEP frequency and the attended object.

Selecting skill and parameters with motor imagery (MI). The user then chooses a skill and its parameters. Both selections are done through MI. We frame this as a k-way (k ≤ 4) MI classification problem, where, given a 5-second-long EEG signal, we aim to decode which of the k pre-decided actions the user was imagining. Unlike SSVEP, a small amount of calibration data (10-min) is required due to the distinct nature of each user’s MI signals. The four classes are: Left Hand, Right Hand, Legs, and Rest; the class names describe the body parts that users imagine using to execute some skills (e.g. pushing a pedal with feet). Upon being presented with the list of k skill options, we record a 5-second EEG signal, and use a classifier trained on the calibration data. The user then guides a cursor on the screen to the appropriate location for executing the skill. To move the cursor along the x axis, for instance, the user is prompted to imagine moving their Left hand for leftward cursor movement and to imagine Rest for rightward cursor movement. We record another five seconds of data and utilize a 2-way classifier. This process is repeated for the x, y, and z axes.

For decoding, we use only EEG channels from the visual cortex (Appendix 6) and preprocess the data with a notch filter to remove noise. We then use Canonical Correlation Analysis (CCA) for classification. We create a Canonical Reference Signal (CRS), which is a set of sin and cos waves, for each of our frequencies and their harmonics. We then use CCA to calculate the frequency whose CRS has the highest correlation with the EEG signal, and identify the object that was made to flicker at that frequency.
correlated with MI activity [24]. The classification algorithm is based on the common spatial pattern (CSP) [25–28] algorithm and quadratic discriminant analysis (QDA). Due to its simplicity, CSP+QDA is explainable and amenable to small training datasets. Contour maps of electrode contributions to the top few CSP-space principal components are shown in the middle row of Fig. 3. There are distinct concentrations around the right and left motor areas, as well as the visual cortex (which correlates with the Rest class).

Confirming or interrupting with muscle tension. Safety is a critical issue in BRI due to noisy decoding. We follow a common practice and collect electrical signals generated from facial muscle tension (Electromyography, or EMG). This signal appears when users frown or clench their jaws, indicating a negative response. This signal is strong with near-perfect decoding accuracy, and thus we use it to confirm or reject the decoded object, skill, and parameter selections. With a pre-determined threshold value obtained through the calibration stage, we can reliably detect muscle tension from 500-ms windows.

3.2 The robot: Parameterized primitive skills

Our robots must be able to solve a diverse set of manipulation tasks under the guidance of humans, which can be achieved by equipping them with a set of parameterized primitive skills. The benefits of using these skills are that they can be combined and reused across tasks. Moreover, these skills are intuitive, capture the high-level intention of humans, and can be communicated to the robots effectively. Since skill-augmented robots have shown promising results in solving long-horizon tasks, we follow recent works in robotics with parameterized skills [29–39], and augment the action space of our robots with a set of primitive skills and their parameters. Neither the human nor the agent requires knowledge of the underlying control mechanism for these skills, thus the skills can be implemented in any method as long as they are robust and adaptive to various tasks.

We use two types of robots in our experiment: A Franka Emika Panda arm for tabletop manipulation tasks, and a PAL Tiago robot for mobile manipulation tasks (see Appendix for hardware details). Skills for the Franka robot use the operational space pose controller (OSC) [40] from the Deoxys API [41]. For example, Reaching skill trajectories are generated by numerical 3D trajectory interpolation conditioned on the current robot end-effector 6D pose and target pose. Then OSC controls the robot to reach the waypoints along the trajectory orderly. The Tiago robot’s navigation skill is implemented using the ROS MoveBase package, while all other skills are implemented using MoveIt motion planning framework [42]. A complete list of skills for both robots is in Appendix 3. Later, we will show that humans and robots can work together using these skills to solve all the tasks.

3.3 Leveraging robot learning for efficient BRI

The modular decoding pipeline and the primitive skill library lay the foundation for NOIR. However, the efficiency of such a system can be further improved. During the collaboration, the robots should learn the user’s object, skill, and parameter selection preferences, hence in future trials, the robot can predict users’ intended goals and be more autonomous, hence reducing the effort required for decoding. Learning and generalization are required since the location, pose, arrangement, and instance of the objects could differ from trial to trial. Meanwhile, the learning algorithms should be sample-efficient since human data is expensive to collect.

Retrieval-based few-shot object and skill selection. In NOIR, human effort can be reduced if the robot intelligently learns to propose appropriate object-skill selections for a given state in the task. Inspired by retrieval-based imitation learning [43–45], our proposed method learns a latent state representation from observed states. Given a new state observation, it finds the most similar state in the latent space and the corresponding action. Our method is shown in Fig. 4. During task execution, we record data points that consist of images and the object-skill pairs selected by the human. The images are first encoded by a pre-trained R3M model [46] to extract useful features for robot manipulation tasks, and are then passed through several trainable, fully-connected layers. These layers are trained using contrastive learning with a triplet loss [47] that encourages the images with the same object-skill label to be embedded closer in the latent space. The learned image embeddings and object-skill labels are stored in the memory. During test time, the model retrieves the nearest data point in the latent space and suggests the object-action pair associated with that data point to the human. Details of the algorithm can be found in Appendix 7.1.

One-shot skill parameter learning. Parameter selection requires a lot of human effort as it needs precise cursor manipulation through MI. To reduce human effort, we propose a learning algorithm for predicting
parameters given an object-skill pair as an initial point for cursor control. Assuming that the user has once successfully pinpointed the precise key point to pick a mug’s handle, does this parameter need to be specified again in the future? Recent advancement in foundation models such as DINOv2 [48] allows us to find corresponding semantic key points, eliminating the need for parameter re-specification. Compared to previous works, our algorithm is one-shot [49–53] and predicts specific 2D points instead of semantic segments [54, 55]. As shown in Fig. 4, given a training image (300 × 240) and parameter choice \((x, y)\), we predict the semantically corresponding point in the test images, in which positions, orientations, instances of the target object, and contexts may vary. We utilize a pre-trained DINOv2 model to obtain semantic features [48]. We input both train and test images into the model and generate 768 patch tokens, each as a pixel-wise feature map of dimension \(75 \times 100\). We then extract a \(3 \times 3\) patch centered around the provided training parameter and search for a matching feature in the test image, using cosine similarity as the distance metric. Details of this algorithm can be found in Appendix 7.2.

4 Experiments

Tasks. NOIR can greatly benefit those who require assistance with everyday activities. We select tasks from the BEHAVIOR benchmark [56] and Activities of Daily Living [57] to capture actual human needs. The tasks are shown in Fig. 1, and consist of 16 tabletop tasks performed by the Franka robot and four mobile manipulation tasks performed by the Tiago robot. The tasks encompass various categories, including eight meal preparation tasks, six cleaning tasks, three personal care tasks, and three entertainment tasks. For systematic evaluation of task success, we provide formal definitions of these activities in the BDDL language format [56, 58], which specifies the initial and goal conditions of a task using first-order logic. Task definitions and figures can be found in Appendix 4.

Procedure. The human study conducted has received approval from Institutional Review Board number [omitted for anonymity]. Three healthy human participants (2 male, 1 female) were recruited. 15 Franka tasks are performed by all three users. Sukiyaki, four Tiago tasks, and learning tasks are performed by one user. We use the EGI NetStation EEG system, which is completely non-invasive, making almost everyone an ideal subject. Before experiments, users are familiarized with task definitions and system interfaces. During the experiment, users stay in an isolated room, remain stationary, watch the robot on a screen, and solely rely on brain signals to communicate with the robots (more details about the procedure can be found in Appendix 5).

5 Results

We seek to provide answers to the following questions through extensive experimental evaluation: 1) Is NOIR truly general-purpose, in that it allows all of our human subjects to accomplish the diverse set of everyday tasks we have proposed? 2) Does our decoding pipeline provide accurate decoding results? 3) Does our proposed robot learning and intention prediction algorithm improve NOIR’s efficiency?

System performance. Table 1 summarizes the performance based on two metrics: the number of attempts until success and the time to complete the task in successful trials. Task horizons (number of primitive skills executed) are included as a reference. Although these tasks are long-horizon and challenging, NOIR shows very encouraging results: on average, tasks can be completed with only 1.83 attempts. The reason
Video recordings of all tasks can be found in the supplemental video.

Table 1: NOIR system performance. Task horizon is the average number of primitive skills executed. # attempts indicate the average number of attempts until the first success (1 means success on the first attempt). Time indicates the task completion time in successful trials. Human time is the percentage of the total time spent by human users, this includes decision-making time and decoding time. With only a few attempts, all users can accomplish these challenging tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Object selection (What?)</th>
<th>Skill selection (How?)</th>
<th>Parameter selection (Where?)</th>
<th>Confirmation / interruption</th>
<th>Decoding Stage</th>
<th>Signal</th>
<th>Technique</th>
<th>Calibration Acc.</th>
<th>Task-Time Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MakePasta</td>
<td>SSEP</td>
<td>MI</td>
<td>MI</td>
<td>EMG</td>
<td>Task horizon</td>
<td>8.33</td>
<td>CCA (4-way)</td>
<td>0.580</td>
<td>0.812</td>
</tr>
<tr>
<td>CollectToy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td># Attempts</td>
<td>1.00</td>
<td>MI</td>
<td>0.350</td>
<td>0.422</td>
</tr>
<tr>
<td>SweepTrash</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Time (min)</td>
<td>2.00</td>
<td>MI</td>
<td>0.250</td>
<td>0.350</td>
</tr>
<tr>
<td>IronClth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Human time (%)</td>
<td>40.00</td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>OpenGift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>TicTacToe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>Sukiyaki</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>TrashDisposal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>CovidCare</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>WaterPantz</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
<tr>
<td>PetDog</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MI</td>
<td>0.150</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Table 2: Decoding accuracy at different stages of the experiment.

Decoding accuracy. A key to our system’s success is the accuracy in decoding brain signals. Table 2 summarizes the decoding accuracy of different stages. We find that CCA on SSVEP produces a high accuracy of 81.2%, meaning that object selection is mostly accurate. As for CSP + QDA on MI for parameter selection, the 2-way classification model performs at 73.9% accuracy, which is consistent with current literature [24]. The 4-way skill-selection classification models perform at about 42.2% accuracy. Though this may not seem high, it is competitive considering inconsistencies attributed to long task duration (hence the discrepancy between calibration and task-time accuracies). A potential solution is to have a longer calibration phase or more frequent calibrations. Our calibration time is only 10 minutes, which is significantly shorter compared to the duration of typical MI calibration and training sessions by several orders of magnitude [59]. More calibration provides more data for training more robust classifiers, and allows human users to practice more which typically yields stronger brain signals. Overall, the decoding accuracy is satisfactory, and with the safety mechanism, there has been no single instance of task failure caused by incorrect decoding.

Object and skill selection results. We then answer the third question: Does our proposed robot learning algorithm improve NOIR’s efficiency? First, we evaluate object and skill selection learning. We collect a dataset offline with 15 training samples for each object-skill pair in the MakePasta task. We compare our algorithm with multiple baselines. Given an image, a prediction is considered correct if both the object and the skill are predicted correctly. Results are shown in Table 3. While a simple image classification model using ResNet [60] achieves an average accuracy of 0.31, our method with a pre-trained ResNet backbone achieves significantly higher accuracy at 0.73, highlighting the importance of contrastive learning and retrieval-based learning. Using R3M as the feature extractor further improves the performance to 0.94. The generalization ability of the algorithm is tested on the same MakePasta task. For instance-level generalization, 20 different types of pasta are used; for context generalization, we randomly select and place 20 task-irrelevant objects in the background. Results are shown in Table 3. In all variations, our
model achieves accuracy over 93%, meaning that the human can skip the skill and object selection 93% of the time, significantly reducing the amount of time and effort required.

We further test our algorithm during actual task execution (Fig. 5). A human user completes the task with and without object-skill prediction two times each. With object and skill learning, the average time required for each object-skill selection is reduced by 60% from 45.7 seconds to 18.1 seconds. More details about the experiments and visualization of learned representation can be found in Appendix 7.1.

**One-shot parameter learning results.** First, using our pre-collected dataset (see Appendix 7.2), we compare our algorithm against multiple baselines. The MSE values of the predictions are shown in Table 4. Random sample shows the average error when randomly predicting points in the 2D space. Sample on objects randomly predicts a point on objects and not on the background; the object masks here are detected with the Segment Anything Model (SAM) [61]. For Pixel similarity, we employ the cosine similarity and sliding window techniques used in our algorithm, but on raw images without using DINOv2 features. All of the baselines are drastically outperformed by our algorithm. Second, our one-shot learning method demonstrates robust generalization capability, as tested on the respective dataset; table 4 presents the results. The low prediction error means that users spend much less effort in controlling the cursor to move to the desired position.

We further demonstrate the effectiveness of the parameter learning algorithm in actual task execution for SetTable, quantified in terms of saved human effort in controlling the cursor movement (Fig. 5). Without learning, the cursor starts at the chosen object or the center of the screen. The predicted result is used as the starting location for cursor control which led to a considerable decrease in cursor movement, with the mean distance reduced by 41%. These findings highlight the potential of parameter learning in improving efficiency and reducing human effort. More results and visualizations can be found in Appendix 7.2.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE↓</th>
<th>Generalization</th>
<th>MSE↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>175.8±29.7</td>
<td>5.6±6.0</td>
<td>23.8±5.5</td>
</tr>
<tr>
<td>Sample on objects</td>
<td>137.2±55.7</td>
<td>12.0±11.7</td>
<td>62.5±22.2</td>
</tr>
<tr>
<td>Pixel similarity</td>
<td>45.9±50.1</td>
<td>16.4±22.2</td>
<td>26.8±62.5</td>
</tr>
<tr>
<td>Ours (R3M)</td>
<td>15.8±23.8</td>
<td>26.8±62.5</td>
<td>0.09±0.02</td>
</tr>
</tbody>
</table>

Table 3: Object-skill learning results. Our method is highly accurate and robust.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE↓</th>
<th>Generalization</th>
<th>MSE↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>15.8±23.8</td>
<td>26.8±62.5</td>
<td>0.09±0.02</td>
</tr>
<tr>
<td>Classification (ResNet)</td>
<td>0.31±0.11</td>
<td>0.94±0.04</td>
<td>0.04±0.02</td>
</tr>
<tr>
<td>Ours (ResNet)</td>
<td>0.73±0.09</td>
<td>0.93±0.02</td>
<td>0.02±0.02</td>
</tr>
<tr>
<td>Ours (R3M)</td>
<td>0.94±0.04</td>
<td>0.98±0.02</td>
<td>0.01±0.02</td>
</tr>
</tbody>
</table>

Table 4: One-shot parameter learning results. Our method is highly accurate and generalizes well.

6 Conclusion, Limitations, and Ethical Concerns

In this work, we presented a general-purpose, intelligent BRI system, NOIR. This system empowers human users to control a robot to accomplish a diverse, challenging set of real-world activities using brain signals. NOIR enables human intention prediction through few-shot learning, thereby facilitating a more efficient collaborative interaction. NOIR holds a significant potential to augment human capabilities and enable critical assistive technology for individuals who require everyday support.

NOIR represents a pioneering effort in the field, unveiling immense potential opportunities while simultaneously raising questions about its limitations and potential ethical risks. These critical considerations necessitate a thorough discussion, which we address in a dedicated section, “Questions and Answers about NOIR” (Appendix 1). The decoding speed, as it currently stands, restricts tasks to those devoid of time-sensitive interactions. However, advancements in the field of neural signal decoding, particularly those predicated on machine learning methods, hold promise for alleviating this concern. Furthermore, the compilation of a comprehensive library of primitive skills presents a long-term challenge in robotics, necessitating additional exploration and development. Nonetheless, we maintain that once a robust set of skills is successfully established, human users will indeed be capable of applying these existing skills to complete new tasks.
References


