NOIR 2.0: Neural Signal Operated Intelligent Robots for Everyday Activities

Tasha Kim[1], **Yingke Wang**[2], **Hanvit Cho**[3], **Alex Hodges**[2]

[1]Institute for Computational and Mathematical Engineering
[2]Department of Computer Science
[3]Department of Mechanical Engineering
{tashakim, yingkewang, hvcho74, alexh555}@stanford.edu

Stanford University

Abstract: Neural Signal Operated Intelligent Robots (NOIR) system is a versatile brain-robot interface that allows humans to control robots for daily tasks using their brain signals. This interface utilizes electroencephalography (EEG) to translate human intentions regarding specific objects and desired actions directly into commands that robots can execute. We present NOIR 2.0, an enhanced version of NOIR. NOIR 2.0 includes faster and more accurate brain decoding algorithms, which reduce task completion time by 46%. NOIR 2.0 uses few-shot robot learning algorithms to adapt to individual users and predict their intentions. The new learning algorithms leverage foundation models for more sample-efficient learning and adaptation (15 demos vs. a single demo), significantly reducing overall human time by 65%.

Keywords: Brain-Robot Interface; Human-Robot Interaction

1 Introduction

Brain-robot interfaces (BRIs) represent a major milestone in the fields of art, science, and engineering. The Neural Signal Operated Intelligent Robots (NOIR) [1], unveiled in 2023, is a versatile, intelligent BRI system that employs non-invasive electroencephalography (EEG). The system operates on the concept of hierarchical shared autonomy, where humans set high-level objectives, and the robot carries out these objectives through the execution of detailed motor commands. At the time of its introduction, NOIR demonstrated its general-purpose nature by being able to handle a variety of tasks (20 everyday activities) and showing broad accessibility, as it requires minimal training for use by the general public. Moreover, NOIR is adaptive and intelligent, equipped with a broad set of skills that enable it to autonomously perform low-level actions. Human intentions are conveyed, interpreted, and executed by the robots through *parameterized primitive skills*, such as `Pick(obj-A)` or `MoveTo(x,y)`. Additionally, NOIR can learn and adapt to human goals throughout the course of their collaboration.

NOIR is built on a modular neural signal decoding pipeline. Decoding human intentions (e.g., "grasp the mug by the handle") from neural signals is highly complex. Therefore, we break down human intentions into three components: *What* (the object to interact with), *How* (the manner of interaction), and *Where* (the specific location of interaction), demonstrating that these elements can be extracted from various neural data types. These signals, once decoded, map naturally to the robot's parameterized skills and can be effectively communicated to the robots.

However, there is yet much to be improved for NOIR. First, the decoding time and effort were considerably high. Performing the tasks (4-15 primitive skills) took 3 to 43 minutes, and 55%-85% of the time was spent on the decoding side. Additionally, the decoding accuracy at test time, especially

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

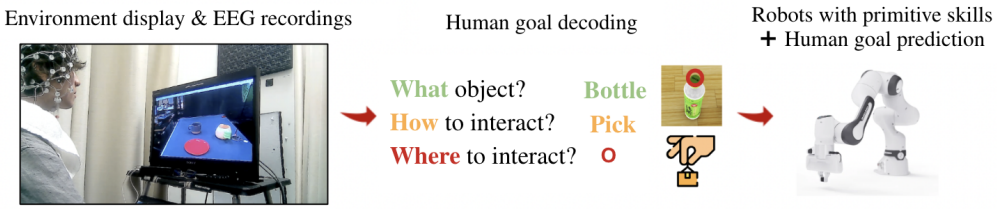

Figure 1: NOIR 2.0 system overview. Following NOIR, NOIR 2.0 implements a modular pipeline for decoding goals from human brain signals, and a robotic system with a library of primitive skills. While minimizing the effort needed for decoding, the robot system is capable of learning to anticipate the goals that humans intend to achieve.

for skill selection, was relatively low (42% for 4-way classification, 74% for 2-way classification). By leveraging the recent progress in neural decoding [2], we show that both the decoding time and accuracy can be significantly improved.

Second, once a human participant has successfully performed the task multiple times, NOIR was able to use its retrieval-based few-shot object and skill selection algorithm to predict human intention, which was shown to save decoding time by 60%. However, the algorithm was based on a pre-trained R3M model [3], which requires 15 training trajectories to predict human intention successfully. Although this number does not seem high for typical robot learning research, it is considered impractical for BRIs, especially in clinical trials. We show that by leveraging the recent progress in large, pre-trained vision-language models, we can accurately predict human intention with only one trajectory, making NOIR 2.0 more effective and practical.

2 The NOIR 2.0 System

Fig. 1 shows a schematic representation of the system. In this setup, humans function as planning agents who perceive and communicate behavioral objectives to the robot. The robot, equipped with pre-defined primitive skills then executes these objectives.

2.1 The Brain: A modular decoding framework

We opted for a non-invasive, saline-based EEG system that records the brain's spontaneous electrical activity through electrodes positioned on the scalp. EEG-based BRIs have been utilized in various applications, including prosthetics, wheelchairs, and robots designed for navigation and manipulation [4–7]. Our approach leverages two commonly used EEG signal types in BRIs: steady-state visually evoked potential (SSVEP) and motor imagery (MI).

SSVEP is the brain's response to external visual stimuli presented at regular intervals [8]. When an individual focuses on a flickering object, the EEG response at the frequency of that stimulus becomes stronger, enabling identification of the object. In contrast, MI is endogenous, requiring the user to mentally simulate specific actions, such as imagining how to manipulate an object. The decoded signals from MI provide insight into how a person intends to interact with the object. As depicted in Fig. 2, we break down human intention into three parts: (a) *What* object is being manipulated; (b) *How* to engage with the object; and (c) *Where* to engage with it.

Object selection via steady-state visually evoked potential (SSVEP). Once the task setup is displayed on a screen, we first determine the user's target object: each object on the screen is made to flicker at distinct frequencies (Fig. 2), and when the user focuses on one, it generates an SSVEP response [8]. By detecting which frequency shows a stronger presence in the EEG data, we can identify the flickering visual stimulus and thus which object the user is focusing on. We utilize OWL-ViT [9] for object detection and tracking, which processes images along with object descriptions to generate segmentation masks. By superimposing each mask with the corresponding flicker frequencies ($6Hz$, $7.5Hz, 8.57Hz$, and $10Hz$[10, 11]), we can accurately identify the chosen object when the user concentrates on it for 10 seconds. For this, we only process signals from the visual cortex and apply a notch filter to the data. Next, we employ canonical correlation analysis

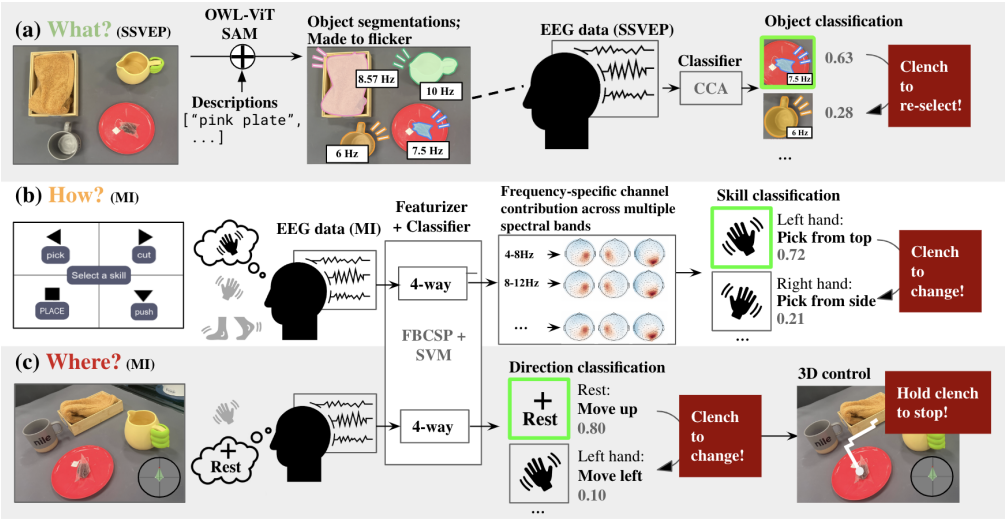

Figure 2: A structured framework for interpreting human objectives from EEG signals that include: (a) *What* object to manipulate, determined by SSVEP signals using CCA classifiers; (b) *How* to engage with the object; and (c) *Where* to interact with the object, deciphered through MI signals using FBCSP+SVM algorithms. A safety mechanism also monitors muscle tension from the jaw to either confirm or reject the decoding results.

(CCA) for classification [12]. We generate a canonical reference signal (CRS), which consists of $\sin$ and $\cos$ waves corresponding to each frequency and its harmonics. CCA is then used to determine which CRS correlates most strongly with the EEG signal, allowing us to find the object that flickered at the corresponding frequency.

Skills and parameters selection via motor imagery (MI). The user selects a skill and its parameters, which we present as a k-way ($k \leq 4$) MI classification problem. The goal is to identify which of the k pre-determined actions the user envisions. Unlike SSVEP, brief calibration data (10 min) is required due to the unique MI signal patterns present for each individual. The four categories are: `Left Hand`, `Right Hand`, `Legs`, and `Rest`, corresponding to the body parts the user imagines moving to perform certain actions (e.g. pressing a pedal with their feet). After displaying the skill options, we capture a 3-second EEG recording and apply a classifier trained on the calibration data to interpret the signals. The user then controls a cursor on the display to the desired skill execution point. To move the cursor continuously along the xy plane, the user is prompted to imagine moving their `Left` or `Right Hands` for leftward or rightward movement; `Legs` or `Rest` for downward or upward movement, respectively. This method is also used for moving the cursor vertically along the z axis, enabling control in three-dimensional space.

For the decoding process, we specifically utilize EEG channels located near the brain regions associated with motor imagery. The data undergoes band-pass filtering within the $8Hz$ to $30Hz$ range to capture the μ-band and β-band frequencies, which are known to be lined with MI activity [13]. The classification algorithm is based on the filter-bank common spatial pattern (FBCSP) [14] algorithm coupled with a support vector machine (SVM) classifier. Due to its simplicity, FBCSP+SVM is explainable and amenable to smaller training datasets. Specifically, the FBCSP algorithm identifies unique spatial patterns across different frequencies, highlighting concentrated activity over the left and right motor regions as well as the visual cortex, the latter being associated with the `Rest` class. The multi-band nature of FBCSP allows for a more comprehensive analysis of these spatial patterns across different spectral ranges.

In Fig. 3, the evolution of parameter selection methodology from NOIR to NOIR 2.0 is illustrated. In the NOIR system, parameter selection is conducted through binary classification using the CSP algorithm, where the direction is decoded then the cursor first moves along the x-axis until the user gives a cue to stop; followed by the y-axis then the z. However, in NOIR 2.0, the process is refined to

Figure 3: Example cursor trajectories during parameter selection on a graphical user interface: (a) shows how the mouse cursor moves using binary control in NOIR. (b) shows how the mouse cursor moves in all four directions through continuous, closed-loop control in NOIR 2.0.

enable simultaneous and continuous selection of the x and y directions through 4-way classification. The enhancement leverages MI decoding to facilitate real-time cursor movement, offering a more seamless user experience. Moreover, by introducing a closed-loop feedback mechanism, NOIR 2.0 allows users to reject or re-select, and actively adapt to new cursor positions.

Confirmation or interruption through muscle tension. As a safety measure, we incorporate a widely used method of capturing electrical signals produced in facial muscle tension (i.e. electromyography, or EMG). These signals are triggered whenever the user frowns, flinches, or clenches their jaw. Since EMG gives strong and highly accurate signals, we use them to confirm or reject selections of objects, skills, or parameters. Using a threshold value established during the calibration phase, NOIR 2.0 reliably detects muscle tension using variance-based threshold filters within a short 500-ms window. To enable real-time interruption signaling during human control time compared to NOIR, NOIR 2.0 filters EEG artifacts that are not generated by the brain such as blinking, lateral eye movement, respiration, and pulse. In addition to this, NOIR 2.0 further utilizes spatial filtering, frequency and duration analysis of the signals, and targeted channel selection to more accurately and robustly identify signals to reject.

2.2 The Robot: Parameterized primitive skills

Like NOIR, NOIR 2.0 equips the robots with a collection of parameterized primitive skills [15–26]. For our experiments, we use the Franka Emika Panda arm to perform tabletop manipulation tasks. The skills for the Franka robot rely on the operational space pose controller (OSC) [27], integrated through the Deoxys API [28]. For instance, the `Reaching` skill involves generating trajectories via numerical 3D interpolation, based on the robot's current 6D end-effector pose and the target pose. The OSC then directs the robot to follow the trajectory by moving through each waypoint in sequence.

2.3 Leveraging efficient BRI through robot learning

While performing tasks, the robots need to learn the user's preferences for object, skill, and parameter selections. This allows the robots to predict the user's intended goals in future trials, making them more autonomous and reducing the need for decoding. It is crucial for the system to learn and generalize effectively since factors like the location, orientation, arrangement, and instances of objects vary across different trials. Additionally, the learning algorithms must be highly sample-efficient, since gathering human data is costly and time-consuming in this setting. The overall design of the learning algorithm is shown in Fig. 4.

Retrieval-based few-shot object and skill selection. Human effort can be reduced if the robot intelligently learns to propose the appropriate object-skill selection for a given state in the task. NOIR uses retrieval-based imitation learning [29–31], which learns a latent state representation (pre-trained R3M model [32] plus a few trainable layers) from observed states. Given a new state observation, it finds the most similar state in the latent space and the corresponding action. However, such a method still requires a handful of demonstration data to be effective (around 15 for NOIR), which is impractical in BRI research. With recent progress in large, pre-trained vision language

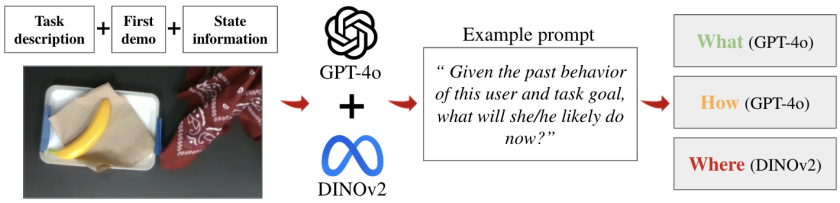

Figure 4: Overview of the robot learning method for NOIR 2.0.

models (VLMs) such as GPT-4o, we show that this can be done with a single training example. We design the prompts for the vision-language model (VLM) to assist the following two parts (I) and (II):

(I) State understanding and mapping.

Given a sequence of annotated human demonstration state images, the VLM is tasked with interpreting the spatial relationships between the robotic gripper and objects on the table at each state. It must also comprehend the state of these objects. Putting the task description to use, the model maps each state to relevant objects and their associated skills, while recognizing the logical progression between the states. *Sample prompt:*

> The gripper view image captures what is held in the robot's gripper during the current state. ... The top-view image captures the spatial relationships and status of objects on the table during the current state. ... I will give you a sequence of all the states in the current task, arranged in order. Please do the following:
>
> 1. Fully understand the object types in the gripper view images and pair them with the '`obj`' text annotations. ...
> 2. Fully understand the objects in the top-view images, the spatial relationships between these objects, and the current status of each object, pairing them with the '`skill`' text annotations. ...
>
> After that, I will send you two query images. ...

(II) State inference and task retrieval.

Given a query image of the robot's current state, the VLM is expected to infer the robot's current state. It then retrieves the subsequent state by identifying the relevant object and the task to be performed, informed by the prior human demonstration. Finally, it returns the index of the inferred current state. *Sample prompt:*

> Identify the state index from the sequence that best matches this new query pair. Then, return the '`obj`' and '`skill`' information of the next state (i.e., `index + 1`). ... Rank your answer by the order of possibility. ...

One-shot skill parameter learning. Selecting parameters requires significant human effort because it involves precise cursor control through continuous MI. NOIR 2.0 retains the design of NOIR by using a learning algorithm to predict parameters based on an object-skill pair. With DINOv2 [33], we can locate corresponding semantic key points, removing the need for repeatedly specifying parameters. Given a top-view training image (360×240) and a chosen parameter $(x,\ y)$, the model predicts the semantically equivalent point in the top-view test images, even when the target object's position, orientation, instance, or context vary. With a side-view training image (360×240) and a chosen parameter (z), the model first predicts the semantically equivalent point A in the side-view test image, then retrieves the point B on the z-axis calculated based on the predicted $(x,\ y)$ that gives the minimal distance from A. We convert the pixel coordinates of B into the 3D coordinates that the robot's primitive skill should take in.

Task Name	Time (min.)			Human Time (min.)		
	NOIR	NOIR 2.0	NOIR 2.0+Learning	NOIR	NOIR 2.0	NOIR 2.0+Learning
WipeSpill	14.74	9.12	5.46	11.65	5.12	3.15
OpenBasket	15.90	6.79	5.80	13.04	2.60	1.52
PourTea	13.53	8.90	12.60	11.25	6.55	7.87
Avg. Time Reduced (%)	-	**43.82**	**45.97**	-	**60.30**	**65.11**

Table 1: NOIR vs. NOIR 2.0 system performance comparison. Time (min.) indicates the total task completion time for successful trials. Human time (min.) refers to the total time spent by the user, which includes both decision-making and decoding time. Time reduced in both is measured for NOIR 2.0 and NOIR 2.0 with robot learning (NOIR 2.0+Learning) against NOIR.

We use a pre-trained DINOv2 model to extract semantic features [33]. Both the training and test images are processed by the model to generate 768 patch tokens, forming a pixel-wise feature map of 75×100. We then take a 3×3 patch centered on the provided training parameter and identify the matching feature in the test image, using cosine similarity as the distance measure.

3 Experiments

Tasks. We focus on three tabletop tasks described in NOIR: WipeSpill, OpenBasket, and PourTea [1]. For systematic evaluation of the task success, we provide formal definitions of these activities in the BDDL language format [34, 35], which specify the initial and goal conditions of a task using first-order logic (FOL).

Our human study was approved by the Institutional Review Board (IRB). One healthy participant completed all three tasks above. We employed the EGI NetStation EEG system, which is entirely non-invasive, and makes it suitable for a wide range of participants. Prior to the experiments, the user was introduced to task descriptions and the system interface. During the experiment, the participant stayed in a secluded room, remained still, watched the robot's actions through a screen, and used only brain signals to interact and communicate with the robot.

4 Results

System performance. Table 1 summarizes the performance based on two metrics: the (a) total time taken to complete a task (Time), and (b) human time spent on the task (Human Time). Note that for all tasks, the task horizon (average number of primitive skills executed) fell in the range 4-6, and the average number of attempts until the first success was 1-2 (1 means success on the first attempt). If the participant encountered a state during task execution from which recovery was not possible, we would reset the environment, and allow them to start the task anew.

For NOIR, the average task completion time was 14.72 minutes across the tasks, and the time humans spent on decision-making and decoding was relatively long (81.28% of total time). With NOIR 2.0, we saw a marked performance improvement in average task completion time at 8.27 minutes, with a 60.30% decrease in human time spent. Coupled with robot learning, the task completion time was reduced by 45.97% in total, and human time spent was reduced by 65.11%.

Decoding accuracy. A key to the success of a BRI system is in the accuracy of decoding brain signals. Tables 2 and 3 compares NOIR vs. NOIR 2.0 by summarizing the decoding accuracy at different stages of the pipeline. Our results indicate that using CCA on SSVEP yields an accuracy rate of 88% at task time, which shows that object selection is predominantly precise. The improvement in SSVEP accuracy in NOIR 2.0 is attributed to color and contrast optimizations of the masks and background. As for FBCSP+SVM on MI for parameter selection, the 4-way skill-selection classification models increased the accuracy of skill selection from 42% to 61% at task-time. While the accuracy might not appear high, this is quite competitive given the variations caused by long task duration and differences in setting observed at calibration and task times.

Object and skill selection results. We then turn to the question: Does our new robot learning algorithm further improve the efficiency of NOIR 2.0? We assess this by evaluating learning in the

Decoding Stage	Signal	Technique	Calibration Acc.	Task-Time Acc.
Object selection (What?)	SSVEP	CCA (4-way)	–	0.81
Skill selection (How?)	MI	CSP+QDA (4-way)	0.58	0.42
Confirmation / interruption	EMG	Thresholding (2-way)	1.0	1.0

Table 2: NOIR decoding accuracy at different stages of the experiment.

Decoding Stage	Signal	Technique	Calibration Acc.	Task-Time Acc.
Object selection (What?)	SSVEP	CCA (4-way)	–	**0.88**
Skill selection (How?)	MI	FBCSP+SVM (4-way)	**0.64**	**0.61**
Confirmation / interruption	EMG	Thresholding (2-way)	1.0	1.0

Table 3: NOIR 2.0 decoding accuracy at different stages of the experiment.

object and skill selection stage. When an image is presented, a prediction is deemed accurate if and only if it correctly identifies the object and the associated skill. Results are shown in Table 4.

Model	Task Prediction	Offline Acc.	Task-Time Acc.
GPT-4o	Object / skill selection	0.94	0.83
DINOv2	Parameter selection	0.79	0.67

Table 4: Offline and task-time accuracy for robot learning.

In offline experiments, our algorithm demonstrates an accuracy of 94% for the VLM object and task proposal and 79% for the DINO parameter prediction. In online task-time experiments, our algorithm achieves an accuracy of over 83% for the VLM object and task proposal and 79% for the dino parameter prediction. Note the slight decrease in accuracies between offline to online demonstrations primarily stemmed from minor object shifts caused by the gripper during execution and small offsets introduced by the top-view camera that does not perfectly align in parallel to the table surface.

What this means is that the user can skip object and skill selection 83% of the time, and can skip parameter selection over 79% of the time during a task. This significantly reduces the time and effort required from the user. Hence we see that with object and skill learning, the average time of human involvement is reduced by 12% from 4.76 to 4.18 minutes.

5 Discussion and Ethical Concerns

The improved decoding framework of NOIR 2.0 substantially increased the skill selection accuracy at task time. The transition from binary to 4-way classification with continuous, closed-loop control allows for real-time adjustments and re-selections, enhancing user control in 3D space. By integrating retrieval-based few-shot object and skill selection with one-shot skill parameter learning, NOIR 2.0 showed significant performance improvements. Through pre-trained vision-language models for state understanding and task retrieval, we reduced the amount of demonstration data required from 15 examples to 1 user demo and enabled more effective learning. With parameter prediction capability, we addressed challenges in object placement and generalizability. These changes contribute considerably to reducing overall task completion time and effort in performing everyday activities.

It is also important to acknowledge existing challenges in accurately decoding extracortical EEG data, as well as the subject-variability of demonstrable signals. User fatigue in concentrating for long periods of time on long-horizon tasks may affect the quality of extracted signals which translates to a decrease in system performance. The EMG-based interruption system using facial muscle tension continues to provide crucial and reliable safety features for NOIR 2.0, allowing users to maintain control over the system. It is imperative to ensure the robustness and safety of end-to-end brain-robot systems like NOIR, especially when they are operated in dynamic environments or human spaces.

6 Conclusion

In this paper, we introduced NOIR 2.0, an enhanced version of the NOIR BRI. Through increased speed and accuracy when decoding brain activity and incorporating one-shot learning to predict human intent, NOIR 2.0 facilitates more efficient human-robot collaboration. By leveraging the algorithm's ability to adapt to individual users and their intentions real-time, NOIR 2.0 reduces cognitive burden and increases the system's adaptability This enables everyday users to manipulate a robot to perform various real-world tasks using brain signals with little to no training. We believe that NOIR 2.0 holds a significant potential to augment human capabilities and to serve as a critical component of assistive technology. Its intuitive interface and natural interactive paradigm make it especially valuable for individuals who may require daily physical assistance. Integration of these technologies opens up new avenues for future research in collaborative human-robotics and the expansion of BRIs into other areas of human activity.

References

[1] R. Zhang, S. Lee, M. Hwang, A. Hiranaka, C. Wang, W. Ai, J. J. R. Tan, S. Gupta, Y. Hao, G. Levine, et al. Noir: Neural signal operated intelligent robots for everyday activities. In *Conference on Robot Learning*, pages 1737–1760. PMLR, 2023.

[2] Z. Y. Chin, K. K. Ang, C. Wang, C. Guan, and H. Zhang. Multi-class filter bank common spatial pattern for four-class motor imagery BCI. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 571–574, 2009. doi:10.1109/IEMBS.2009.5332383.

[3] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. URL https://arxiv.org/abs/2203.12601.

[4] M. Aljalal, S. Ibrahim, R. Djemal, and W. Ko. Comprehensive review on brain-controlled mobile robots and robotic arms based on electroencephalography signals. *Intelligent Service Robotics*, 13:539–563, 2020.

[5] L. F. Nicolas-Alonso and J. Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12(2): 1211–1279, 2012.

[6] L. Bi, X.-A. Fan, and Y. Liu. Eeg-based brain-controlled mobile robots: a survey. *IEEE transactions on human-machine systems*, 43(2):161–176, 2013.

[7] N. M. Krishnan, M. Mariappan, K. Muthukaruppan, M. H. A. Hijazi, and W. W. Kitt. Electroencephalography (eeg) based control in assistive mobile robots: A review. In *IOP Conference Series: Materials Science and Engineering*, volume 121, page 012017. IOP Publishing, 2016.

[8] E. D. Adrian and B. H. Matthews. The berger rhythm: potential changes from the occipital lobes in man. *Brain*, 57(4):355–385, 1934.

[9] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby. Simple open-vocabulary object detection with vision transformers, 2022.

[10] D. Zhu, J. Bieger, G. G. Molina, and R. M. Aarts. A survey of stimulation methods used in ssvep-based bcis. *Computational intelligence and neuroscience*, 2010:1–12, 2010.

[11] R. Kuś, A. Duszyk, P. Milanowski, M. Łabecki, M. Bierzyńska, Z. Radzikowska, M. Michalska, J. Żygierewicz, P. Suffczyński, and P. J. Durka. On the quantification of ssvep frequency responses in human eeg in realistic bci conditions. *PloS one*, 8(10):e77536, 2013.

[12] L. Shao, L. Zhang, A. N. Belkacem, Y. Zhang, X. Chen, J. Li, and H. Liu. Eeg-controlled wall-crawling cleaning robot using ssvep-based brain-computer interface, 2020.

[13] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren. Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges. *Sensors*, 19(6):1423, 2019.

[14] Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki. Multi-class filter bank common spatial pattern for four-class motor imagery bci. *Frontiers in Neuroscience*, 13:1275, 2019. doi:10.3389/fnins.2019.01275.

[15] R. Chitnis, T. Silver, J. B. Tenenbaum, T. Lozano-Perez, and L. P. Kaelbling. Learning neuro-symbolic relational transition models for bilevel planning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4166–4173. IEEE, 2022.

[16] S. Nasiriany, H. Liu, and Y. Zhu. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7477–7484. IEEE, 2022.

[17] Y. Zhu, J. Tremblay, S. Birchfield, and Y. Zhu. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. IEEE, 2021.

[18] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[19] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. *arXiv preprint arXiv:2209.05451*, 2022.

[20] W. Liu, C. Paxton, T. Hermans, and D. Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6322–6329. IEEE, 2022.

[21] D. Xu, A. Mandlekar, R. Martín-Martín, Y. Zhu, S. Savarese, and L. Fei-Fei. Deep affordance foresight: Planning through what can be done in the future. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6206–6213. IEEE, 2021.

[22] C. Wang, D. Xu, and L. Fei-Fei. Generalizable task planning through representation pretraining. *IEEE Robotics and Automation Letters*, 7(3):8299–8306, 2022.

[23] S. Cheng and D. Xu. Guided skill learning and abstraction for long-horizon manipulation. *arXiv preprint arXiv:2210.12631*, 2022.

[24] C. Agia, T. Migimatsu, J. Wu, and J. Bohg. Taps: Task-agnostic policy sequencing. *arXiv preprint arXiv:2210.12250*, 2022.

[25] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *6th Annual Conference on Robot Learning*, 2022.

[26] A. Hiranaka, M. Hwang, S. Lee, C. Wang, L. Fei-Fei, J. Wu, and R. Zhang. Primitive skill-based robot learning from human evaluative feedback. *arXiv preprint arXiv:2307.15801*, 2023.

[27] O. Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal on Robotics and Automation*, 3(1):43–53, 1987.

[28] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Object-centric imitation learning for vision-based robot manipulation. In *Conference on Robot Learning*, pages 1199–1210. PMLR, 2023.

[29] E. Mansimov and K. Cho. Simple nearest neighbor policy method for continuous control tasks, 2018. URL `https://openreview.net/forum?id=ByL48G-AW`.

[30] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *6th Annual Conference on Robot Learning*, 2022.

[31] M. Du, S. Nair, D. Sadigh, and C. Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets, 2023.

[32] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2021.

[33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[34] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[35] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pages 477–490. PMLR, 2022.