

Robotic Grasping and Placement Controlled by EEG-Based Hybrid Visual and Motor Imagery

Yichang Liu¹, Tianyu Wang^{2,5}, Ziyi Ye^{3,4*}, Yawei Li⁶, Yu-Gang Jiang^{3,4}, Shouyan Wang^{1*}, Yanwei Fu^{2,4,5}

Abstract—We present a framework that integrates EEG-based visual and motor imagery (VI/MI) with robotic control to enable real-time, intention-driven grasping and placement. Motivated by the promise of BCI-driven robotics to enhance human-robot interaction, this system bridges neural signals with physical control by deploying offline-pretrained decoders in a zero-shot manner within an online streaming pipeline. This establishes a dual-channel intent interface that translates visual intent into robotic actions, with VI identifying objects for grasping and MI determining placement poses, enabling intuitive control over both what to grasp and where to place. The system operates solely on EEG via a cue-free imagery protocol, achieving integration and online validation. Implemented on a Base robotic platform and evaluated across diverse scenarios, including occluded targets or varying participant postures, the system achieves online decoding accuracies of 40.23% (VI) and 62.59% (MI), with an end-to-end task success rate of 20.88%. These results demonstrate that high-level visual cognition can be decoded in real time and translated into executable robot commands, bridging the gap between neural signals and physical interaction, and validating the flexibility of a purely imagery-based BCI paradigm for practical human-robot collaboration.

I. INTRODUCTION

Recently, robotic technologies have achieved remarkable progress in end-to-end grasping and manipulation [1], [2], [3]. Modern robots, empowered by deep learning and advanced control algorithms, can recognize objects, plan complex motions, and interact with real-world environments [4], [5], [6]. However, they still struggle to interpret and execute commands derived from human *high-level* intentions [7], [8], such as natural language instructions, context-dependent requests, and multimodal cues. In contrast, over the past two decades, *low-level* robotic capabilities have matured significantly, enabling reliable perception, motion, object manipulation, and basic action execution. Consequently, bridging these robust low-level skills with high-level cognitive processes that reflect human intent remains a central challenge.

To address this challenge, studies have explored combining cognitive signals with robotic control. Early approaches used explicit cues, such as voice commands, hand gestures, or eye

tracking, to infer user intent and guide robot actions [9], [10]. More recently, brain-computer interfaces (BCIs) have been integrated directly, allowing robots to interpret a user’s brain activity related to object recognition or movement intentions, without requiring physical or verbal input.

However, the integration of brain signals with robotic actions remains in its early stages and faces significant challenges. Most existing BCI-robot control systems rely on constrained and explicit control signals such as event-related potentials (e.g., P300), steady-state visually evoked potentials (SSVEP) [11], or motor imagery (MI) [12]. While effective in laboratory environments, these methods are often restricted to predefined commands or simple actions (e.g., “move left” or “move right”), making them ill-suited for generating complex, intuitive, and natural-intention-driven commands for real-world physical interaction. This gap between a user’s cognitive intent and the robot’s executable actions remains a major challenge for the widespread adoption of BCI technology [13], [14].

Recent advances in brain decoding, particularly with electroencephalography (EEG), have increasingly targeted higher-order cognitive states such as visual perception and imagery [15], [16], [17]. EEG, as a non-invasive and portable modality, shows promise in interpreting complex brain activity. Nevertheless, most studies have focused on passive observation and offline analysis, without directly translating these rich cognitive signals into real-time robotic actions. As a result, the potential of high-level visual cognition for controlling robotic systems remains largely underexplored.

In this study, we present an end-to-end framework that integrates high-level visual cognition with robotic control for a pick-and-place task. As shown in Figure 1, our **EEG-based cognitive control pipeline** first collects both visual imagery (VI) and motor imagery (MI) EEG data from five participants in an offline setup to pretrain the corresponding decoders. The selected decoders are then integrated into the robotic control system, where VI selects the target object for grasping and MI specifies the placement pose. The offline evaluation achieves average accuracies of 44.11% for VI and 76.53% for MI, respectively. Motivated by the offline results, we deployed the VI/MI decoders in a zero-shot online setting on a dual-channel EEG system for a pick-and-place platform, achieving online accuracies of 40.23% (VI) and 62.59% (MI) with an end-to-end task success rate of 20.88%. This **online application** demonstrates that VI and MI can convey **higher-level intentions** directly usable for task execution, and the system remains effective in scenarios such as control in supine position without direct sight of the physical objects.

This work was supported by STI 2030 Major Projects (2021ZD0200407), the National Natural Science Foundation of China (62521004), the National Key Research and Development Program of China (2022YFC2405100), the Shanghai Pujiang Talents Program (Grant 24PJ006), and the AI for Science Foundation of Fudan University.

Project Page: <https://yichangliu1224.github.io/EEG-Grasptbot/>

All authors are affiliated with ¹Institute of Science and Technology for Brain-inspired Intelligence (ISTBI), Fudan University; ²School of Data Science, Fudan University; ³Institute of Trustworthy Embodied AI, Fudan University; ⁴Shanghai Key Laboratory of Multimodal Embodied AI; ⁵Shanghai Innovation Institute; ⁶ETH Zürich.

* indicates the corresponding authors

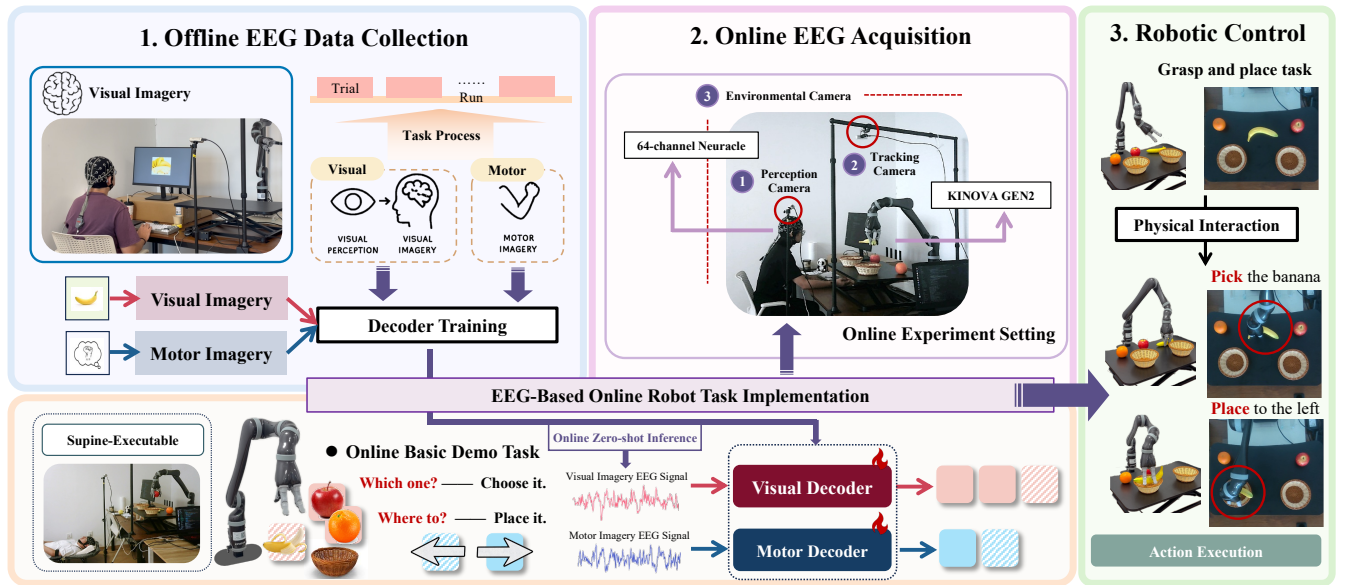


Fig. 1: **High-level cognitive control pipeline for EEG-based robotic manipulation.** The framework consists of three main components: (1) **Offline EEG data collection:** visual imagery (VI) and motor imagery (MI) EEG data are collected and labeled to train visual and motor decoders. (2) **Online EEG acquisition:** A dual-channel system is deployed in which VI-EEG is decoded into grasping intentions and MI-EEG determines placement positions, enabling real-time task-level command generation. (3) **Robotic control:** the trained decoders drive a robotic arm to perform the grasp and place task in real-world, demonstrating the seamless mapping from high-level visual cognition to physical manipulation.

Our main contributions are:

- 1) We propose and empirically validate a high-level cognitive control paradigm as a natural-intention-driven BCI grounded in **visual imagery (VI)**, translating VI-EEG signals directly into task-level robotic commands for pick-and-place manipulation.
- 2) We develop a **real-time framework** that bridges brain signals and robotic actions via a dual-channel intent interface: VI-EEG decodes object intent and MI-EEG infers target poses, achieving a seamless mapping from high-level visual cognition to physical manipulation.
- 3) We perform **end-to-end system integration** and conduct online system validation, implementing functional demonstration on a robotic platform, which achieves precise real-world grasping and placement, and showcases advanced interaction capabilities.

II. RELATED WORKS

A. Visual Decoding from Brain Signals

Visual decoding from brain signals has become a rapidly advancing field, with significant strides made in translating neural activity into interpretable visual information [18], [19]. Early work relied on fMRI due to its high spatial resolution, enabling reconstruction of simple and natural images from voxel patterns, and coupled with the recent integration of latent diffusion models improving semantic fidelity [20], [21], [17]. However, its inherent cost, immobility, and low temporal resolution limit its suitability for real-time BCI [22]. In contrast, EEG/MEG offer unparalleled temporal resolution in the millisecond range, with EEG being particularly suitable for real-time and online BCI

systems [23]. Over the years, EEG-based decoding methods have evolved from basic CNNs and temporal networks to more complex architectures [24], including attention-based and graph-augmented models [25], [26]. Some models also incorporate alignment with large-scale vision and multimodal datasets [27], [28]. However, most existing studies primarily focus on offline classification and passive observation [29], [30], often failing to connect decoded visual semantics to actionable tasks in real-time scenarios. As a result, bridging high-level visual imagery with actionable robotic commands remains an ongoing challenge.

B. Robotics Control using EEG

Prevailing BCI-robot paradigms typically rely on P300, SSVEP, or MI signals. While these methods are reliable in controlled settings, they are limited by predefined commands and often require the user to focus on flickering stimuli [31], [32], [33], which diminishes naturalness and can lead to user fatigue. Consequently, these approaches struggle to facilitate intuitive, task-level interactions in more complex and dynamic environments. Even in more naturalistic settings, grounding implicit *high-level* intent into physically feasible robot actions remains challenging [7], [8]. This motivates more direct, action-oriented intent interfaces. For real-time BCI applications, approaches such as LDA, SVMs, and shallow neural networks are often preferred due to their simplicity [34], [35]. However, even more advanced techniques, including pretrained neural networks, are typically restricted to offline processing. Therefore, decoding zero-shot visual imagery EEG into robotic grasping and placement actions is vital for bridging the gap between laboratory-based

BCI systems and real-world applications, allowing for more intuitive and natural interactions in diverse environments.

III. DATA COLLECTION AND SYSTEM CONFIGURATION

A. Participants

Five healthy graduate students (2 females; age 23–30 years; $M=25.4$, $SD=2.97$) from a university participated in this study. All were right-handed with normal or corrected-to-normal vision and hearing, and none reported neurological disorders. Two participants had prior experience with EEG experiments. Written informed consent was obtained and the protocol was approved by the appropriate ethics committee¹. Each participant received approximately \$42 USD per experimental session. All five participants completed the offline experiments, and four further engaged in the online testings.

B. Task Paradigms

Offline data collection comprised visual perception (VP), visual imagery (VI), and motor imagery (MI) tasks. This study involved a single offline session, with continuous EEG signal recording throughout the experiment.

Visual Perception (VP)& Visual Imagery(VI): Stimuli were drawn from the public *Fruit Images for Object Detection* dataset², including three easily recognizable fruit categories (apple, banana, and orange). Each trial began with a 1-second fixation period marked by a central cross, followed by a 2-second presentation of the target image (perception phase). Participants were then instructed to press the space bar to perform eyes-closed visual imagery of the presented object for 5 seconds (imagery phase), a duration supported by previous studies on visual imagery timing [16], [36]. The imagery period ended with a 300 ms auditory tone (PsychoPy default pitch “A”, approximately 440 Hz). After each trial, participants reported their attentional engagement on a 10-point scale (0 = not at all focused, 9 = fully focused). The collected average attention level for each subject in each trial of the VI task was 8.25 ± 1.02 (range: 0-9), indicating a high attention during the study.

The VI task comprised 10 runs, each including 3 blocks of 10 randomized trials. Short breaks (10 sec) were provided between blocks, and participants rested for 1–2 min between runs. Each run lasted approximately 6–7 min (30 trials), yielding 300 trials in total (~90 min).

Motor Imagery (MI): The MI protocol followed the BCI Competition IV-2b design [12]. Each trial started with a 1-second fixation period with the center cross, followed by a 1.25-second directional arrow cue (left or right). A short auditory warning tone (70 ms, approximately 440 Hz) was presented 1 second after cue onset to prompt imagery initiation. Participants then performed kinesthetic motor imagery of the cued hand for 4 seconds, and a 200 ms auditory tone signaled the end of the trial. Each trial was followed by a 1.5-second pause and a 1.5-second ITI, which allowed the

EEG signal to recover and reduced temporal overlap between consecutive trials.

Each run consisted of 20 trials with balanced, randomly intermixed left/right cues. Participants completed 5 runs (100 trials in total), with each run lasting approximately 3 minutes. The entire MI session lasted 20–30 minutes.

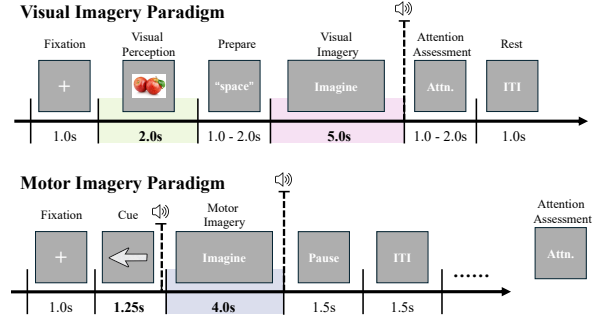


Fig. 2: Visual Imagery and Motor Imagery Paradigms.

The visual imagery and motor imagery paradigms are shown in Figure 2. For each task, the phase triggers were recorded by PsychoPy and transmitted to the EEG system, marking the start and end of every phase. The total duration of all offline sessions was approximately 2 hours.

C. Signal Acquisition and Preprocessing

EEG was recorded with a 64-channel Neuracle system (59 scalp EEG, 2 mastoid, 2 EOG, 1 ECG, 1000 Hz sampling; 24-bit A/D resolution; $\pm 0.003\%$ accuracy; analog input range ± 5 V) on the NeuroHUB platform, keeping electrode impedances below 10 k Ω . Stimuli were presented on a 24.5-inch, 100 Hz monitor at a 50–70 cm viewing distance. Preprocessing was performed in MNE-Python, including standard procedures such as band-pass filtering, re-referencing to linked mastoids (HEOR and HEOL), artifact removal using Independent Component Analysis (ICA), and artifact identification based on correlations with recorded signals (threshold > 0.95). Three band-pass versions were created for comparison (0.5–40 Hz, 0.5–60 Hz, 0.5–100 Hz), and a 50 Hz notch filter was applied to suppress line noise. After preprocessing, we collect two sets of offline data (VI and MI) for each participant, with three band-pass filtered versions to enable the exploration of different high-frequency (e.g., high gamma) components in subsequent analyses.

D. Robotics Platform

We employed a KINOVA GEN2 robot as the embodied platform for systematic interaction. Two Intel RealSense D435 cameras served as the primary sensors of the robotic perception system. The camera positioned directly in front of the workstation acted as the main sensor and was extrinsically calibrated in an eye-to-hand setup to support both the perception system and model inference. An additional camera, mounted vertically above the workstation, provided a top-down view during online inference, enabling execution monitoring and tracking of object interaction. Furthermore, we installed an XWF-1080P6 RGB camera outside the workstation to capture global information for the EEG-based robotic grasping system.

¹Ethics approval was granted by the Ethics Committee of Fudan University, approval number FE255001.

²Aliyun Tianchi: <https://tianchi.aliyun.com/dataset/93848>

IV. METHODOLOGY AND EVALUATION

Prior to deploying the online system, we trained and validated our EEG decoding models using offline datasets to establish a performance baseline.

A. Model Selection

This study refers to the model selections and experimental setups in EEG-ImageNet [37] with specific modifications tailored to the task of decoding distinct intentions for real-time robotic control driven by brain signals. The models encompass both traditional machine learning algorithms, including support vector machine (SVM), random forest, K-nearest neighbors (KNN), decision tree (DT), and ridge regression, as well as deep learning models such as multi-layer perceptron (MLP), EEGNet [38], and RGNN [39]. MLP is a fundamental feedforward neural network, while EEGNet is a lightweight convolutional neural network specifically designed for EEG signals, and RGNN is a recurrent graph neural network that effectively captures the spatio-temporal dependencies within EEG data.

B. Feature Extraction

Considering the cognitive processes involved in visual and motor intention processing, we removed 10% of the signal from both the beginning and end of each EEG segment to ensure the neural signals were entirely confined to the task-relevant period. The remaining signal segments were split into 500 ms sliding windows with minimal overlap. The segmented signals were used as features for EEGNet which requires time-domain signals as input. For other models, **Differential Entropy (DE)** [40] are extracted as the input feature, which estimate $P(f)$ for each frequency band using the following formula:

$$P(f) = \lim_{T \rightarrow \infty} \frac{1}{T} |X_T(f)|^2, \quad (1)$$

where f denotes the (continuous) frequency (Hz), $P(f)$ represents the Power Spectral Density (PSD) at f , and $X_T(f)$ is the Fourier transform of the truncated signal $x(t)$ over the interval $[0, T]$. Then, DE is computed as

$$DE = - \int_{\mathcal{F}} P(f) \log(P(f)) df, \quad (2)$$

where \mathcal{F} is the frequency range of interest.

In addition to the traditional frequency bands (δ (0.5-4 Hz), θ (4-8 Hz), α (8-13 Hz), and β (13-30 Hz)), we specifically computed three new γ bands: (30-40 Hz), (30-60 Hz), and (30-100 Hz), tailored to different filtering strategies. This customization was designed to more effectively capture individual brain activity states for intention decoding. For each EEG signal segment, we obtained differential entropy values for each electrode and frequency band.

TABLE I: **Offline classification results for different models in VP, VI and MI tasks.** The mean decoding accuracy across all subjects is reported for each model in different frequency conditions across three tasks. **Bolded values** indicate the highest accuracy in each row, while underlined values denote the second highest accuracy to report competitive alternatives for model selection.

Subject	EEGNet	MLP	RGNN	SVM	DT
Visual Perception					
00	0.4844	<u>0.4767</u>	0.4622	0.4067	0.4033
01	0.4911	<u>0.4667</u>	0.4278	0.4211	0.3711
02	0.4678	<u>0.4522</u>	0.4155	0.3833	0.3667
03	0.4978	<u>0.4533</u>	0.4322	0.4011	0.3911
04	0.4911	<u>0.4889</u>	0.4633	0.4633	0.3978
Average	0.4864	<u>0.4676</u>	0.4402	0.4151	0.3860
Visual Imagery					
00	0.3979	0.4702	<u>0.4451</u>	0.4014	0.3694
01	0.3688	0.4069	<u>0.3903</u>	0.3694	0.3549
02	0.3639	0.4173	<u>0.3868</u>	0.3639	0.3451
03	0.3736	0.4757	<u>0.4507</u>	0.3826	0.4111
04	0.3694	0.4354	<u>0.4299</u>	0.4028	0.3521
Average	0.3747	0.4411	<u>0.4206</u>	0.3840	0.3665
Motor Imagery					
00	0.7967	<u>0.6983</u>	0.6567	0.5850	0.5250
01	0.6933	<u>0.6783</u>	0.6567	0.6317	0.5550
02	<u>0.7250</u>	0.7267	0.7117	0.6867	0.5933
03	0.7850	<u>0.6917</u>	0.6700	0.6183	0.5850
04	0.8267	<u>0.8167</u>	0.7800	0.7017	0.6450
Average	0.7653	<u>0.7223</u>	0.6950	0.6447	0.5807

C. Experimental Setup

The MLP was implemented with two hidden layers, utilizing the ReLU activation function. Both the EEGNet and RGNN models were implemented strictly according to the original architectures. To ensure model generalizability and accurate evaluation, each participant's dataset was partitioned using stratified sampling into an 80% training set and a 20% test set, which maintained a consistent class label distribution. Experiments were conducted on a server equipped with an NVIDIA A6000 GPU, leveraging Python 3.11.11 and PyTorch 2.6.0. Each model was trained for 1000 epochs.

D. Experimental Results

a) *Offline Performance:* The offline data results for five subjects under different models are summarized in Table I and Figure 3, representing the mean decoding accuracy across all subjects. The MLP model achieved the highest average accuracy in the offline-VI task, with an accuracy of 44.11%, while EEGNet achieved an average accuracy of 76.53% in the offline-MI task. This performance disparity can be attributed to the fundamental differences in brain signal generation. Motor imagery tasks are known to elicit strong and distinct event-related desynchronization/synchronization (ERD/ERS) patterns in the μ and β frequency bands, which are more readily detectable and classifiable from EEG signals [41], [42]. In contrast, the neural activity associated with visual imagery is more distributed and complex [43], [44]. Given that our objective

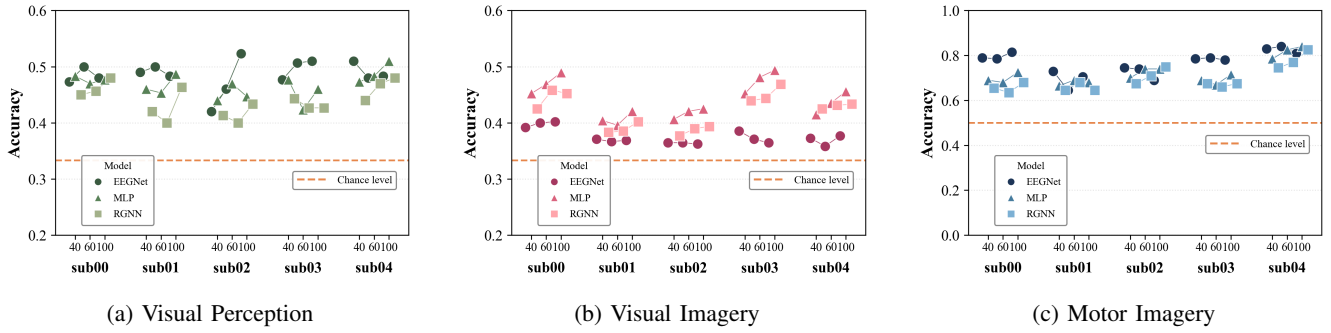


Fig. 3: **Offline results for three tasks: (a) Visual Perception, (b) Visual Imagery, and (c) Motor Imagery.** Points show per-subject accuracies across models and frequency settings; the dashed line denotes chance level (VP/VI: 33.3%, MI: 50%). MI consistently outperforms VP/VI, and all tasks exceed chance.

for the visual imagery task was to achieve object-level differentiation for fine-grained, real-world items, the obtained accuracy is a promising initial result.

b) Task Comparison: As shown in Figure 4, a comparison between the VI and MI tasks revealed significant differences in performance. For the VI task, decoding accuracy for perception was consistently higher than for imagery. This finding is in line with previous studies, suggesting that direct visual stimuli evoke a stronger and more distinct neural response compared to imagined visual stimuli [16], [36].

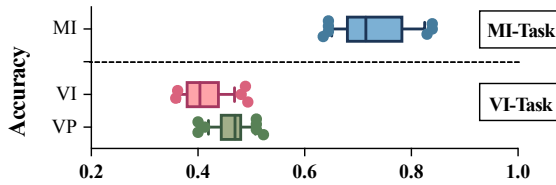


Fig. 4: **Comparison of decoding accuracies for VI and MI.** Motor imagery achieves higher accuracy than the visual EEG tasks overall, and within the visual paradigm, VP slightly outperforms VI.

While the accuracy for the VI task may appear modest, it is crucial to consider the specific context of our experimental setup. At the system level, this level of accuracy makes the task of grasping specific fruits using EEG signals feasible and effective. Despite the lower direct visual decoding accuracy in the VI task, our results consistently surpass chance levels, indicating the potential of using EEG-based system to effectively support fruit-grasping tasks.

c) Average Evoked Responses: Figure 5 shows the average event-related potentials (ERPs) for Sub00 during the visual imagery task, illustrating how brain activity evolves in response to different stimuli. Brain activity showed clear temporal variation, with ERP differences across fruits indicating that visual imagery tasks can modulate neural responses to distinct stimuli and support classification. However, despite this variation, a fairly consistent ERP pattern emerges across signals for all three fruit conditions. This uniformity of the brain’s response across categories also indicates the inherent challenge in discriminating between conditions, emphasizing the complexity of interpreting EEG signals in this context.

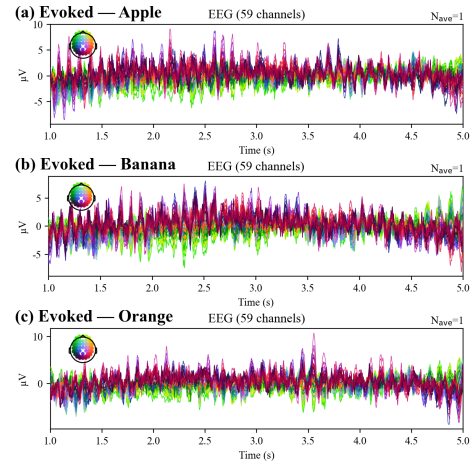


Fig. 5: **Evoked EEG responses to fruit visual imagery.** Brain activity patterns during VI for three stimuli in Sub00 (59 channels): (a) Apple, (b) Banana, (c) Orange. Responses vary over time yet remain broadly similar across categories, reflecting the fine-grained nature of VI decoding.

V. ONLINE DEMONSTRATION

Based on the collected offline data and pre-trained model, we conducted online experiments for both VI and MI tasks. The system was deployed within a real-time data transmission pipeline with EEG streamed from the NeuroHUB platform to the robotic control system. All online experiments were conducted on a laptop running Ubuntu 20.04, equipped with an Intel i7-7700HQ CPU and an NVIDIA GeForce GTX 1070 GPU.

A. Online System Workflow

In the **basic demonstration** scenario, the task was to recognize different fruits and place them in the correct positions. This demonstration highlights the system’s ability to translate mental imagery into actionable robotic commands for the selected object and the intended position. The online test for our base demonstration followed a structured, multi-stage workflow to ensure reliable human–robot interaction.

Preparation and Task Initiation. Each trial began with a beep prompt. Participants performed an online-VI task for 15 seconds by mentally visualizing the target object, followed

TABLE II: **Online system results. Top:** Base demo performance per subject on online VI/MI tasks across model/frequency settings (e.g., 63.64 (RGNN/40) denotes accuracy using the RGNN model at 40 Hz). For each subject, we report the top-2 online accuracies and the overall mean. **Bolded values** indicate the row maximum, while **underlined values mark** the second highest. **Middle:** Overall online system accuracy, with online VI/MI accuracy representing the mean top-2 accuracies across all subjects. **Bottom:** Overall system runtime statistics.

Online Base Demonstration results								
Subject	VI-online Task			MI-online Task				
	Top1 Acc. (%)	Top2 Acc. (%)	Overall Avg. (%)	Top1 Acc. (%)	Top2 Acc. (%)	Overall Avg. (%)		
01	63.64 (RGNN/40)	44.44 (RGNN/60)	47.14	75.00 (RGNN/60)	60.00 (MLP/40)	61.67		
02	50.00 (EEGNet/100)	39.13 (MLP/100)	<u>40.82</u>	65.22 (MLP/100)	60.00 (MLP/40)	58.41		
03	50.00 (MLP/60)	33.67 (RGNN/60)	38.89	60.00 (RGNN/100)	53.85 (RGNN/60)	54.62		
04	37.50 (RGNN/100)	33.67 (MLP/60)	34.72	66.67 (MLP/60)	60.00 (MLP/40)	<u>60.00</u>		
Online System Accuracy								
Online VI avg Acc. (%)	Query success rate (%)	Online MI avg Acc. (%)	Place success rate (%)	System Accuracy (%)				
40.23	76.11	62.59	100	20.88				
System operation times (s)								
Prepare	VI Task	VI Data Proc.	VI Infer	MI Task	MI Data Proc.	MI Infer	Robot Exec.	System Total
6.010	15.000	0.639	8.191	15.000	0.524	8.000	54.872	107.266

by an online-MI task for 15 seconds by imagining left or right hand movements, consistent with the offline protocol.

Neural Decoding and Inference. Real-time EEG was streamed to pretrained VI and MI decoders. VI decoding identified the target object, while MI decoding determined the placement position, providing the robot with the information needed to compute the optimal interaction location.

Action Execution. The system integrated VI (object) and MI (position) results to guide grasp-and-place actions. Based on these inferences, the manipulator grasped the target object and placed it at the inferred location.

By converting neural intentions into robotic actions, the system effectively bridged mental imagery with physical execution, demonstrating reliable real-time interaction between human cognition and the robotic platform.

B. System Performance and Results

The online system results were evaluated with four participants in a base demonstration scenario. Each participant completed at least 50 iterations of the *Online System Workflow* to validate the entire system. Following the offline accuracy hierarchy, we tested each model-frequency configuration (model/Hz), running at least 4-5 trials per configuration. A model was not considered effective for online application if its results showed no change. Models that performed best in the offline setting did not necessarily yield the same success rate in the online scenario, indicating that factors such as model stability and other conditions also play a role in real-time performance. After initial evaluation of all configurations, models with relatively higher success rates were selected for at least 15 online trials to assess the system’s overall performance.

Table II presents the online system results in detail.

a) Detailed Online Performance Analysis: Upon completing experiments with the four participants, we selected the top-2 online accuracies for both the VI and MI tasks,

along with their corresponding model/frequency configurations. Among all participants, Sub01 demonstrated the best overall online performance, with average accuracies of **47.14%** for the online-VI task and **61.67%** for the online-MI task. Online inference used a 10-second window (3–13s) within the 15-second task. This temporal constraint, combined with less refined nature of real-time preprocessing, largely explains the reduced accuracies compared to offline results. For Sub03 and Sub04, certain VI conditions (model/Hz) exhibited systematic biases (e.g., apples misclassified as bananas, bananas as oranges) and were therefore excluded from the final analysis. Overall, the mean online accuracy across all subjects, including configurations not shown in the top-2, was 40.23% for VI and 62.59% for MI, reflecting the impact of inter-subject variability, physiological state, and the inherent characteristics of EEG signals in the online setting. Notably, among the top-performing configurations, MLP/100 and RGNN/60 consistently performed well, suggesting promise for real-time use.

b) Overall System and Operational Performance: From the system perspective, the online system accuracy was **20.88%** (Table II, middle). This composite metric integrates visual recognition accuracy, motor position recognition accuracy, and the robotic arm’s object acquisition and placement success rates. While the overall score is modest, it indicates that the primary bottleneck is converting noisy EEG signals into reliable commands for robotic execution. Notably, the placement success rate was 100%, showing that once the system correctly identified the placement position, execution was consistently successful. The average runtime per demonstration was **107.266 s** (Table II, bottom). VI and MI inference take about 8 s each per task, with average preprocessing times of 0.639 s (VI) and 0.524 s (MI), totaling about 17 s for EEG decoding (~16% of the runtime). In contrast, robot execution accounts for the largest portion (54.872 s), indicating that timing is mainly constrained by

system-level operations rather than the EEG pipeline. Taken together, these findings indicate that strengthening EEG decoding robustness and optimizing task design are the most effective levers for overall reliability and performance.

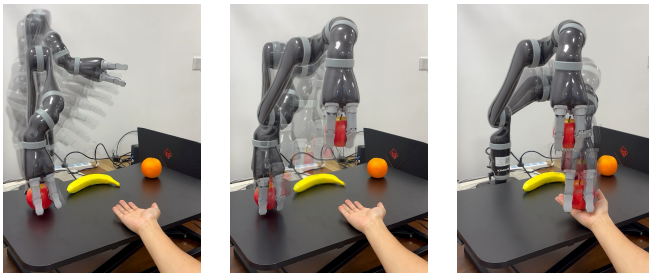
TABLE III: Grasp Success Rates for Different Fruits

Fruit Catalogue	Success Rate (%)
Apple	89.83
Banana	55.84
Orange	84.44
Overall Success Rate	76.11

c) Robotic Grasping and System Reliability: The average grasp query success rate was **76.11%**, as shown in Table III, which includes the success rates for different objects in the fruit catalogue. Specifically, apples and oranges had high success rates of **89.83%** and **84.44%**, respectively, while the success rate for bananas was comparatively lower at 55.84%. These discrepancies can be attributed to small execution and pose estimation inaccuracies (e.g., a depth-related Z error of about $\pm 2\%$ for the RealSense D435 at a 2 m range, and slight deviations in 6D grasp pose computation). Elongated, non-uniform objects such as bananas are more sensitive to such deviations. Despite these challenges, the system maintained a reasonable grasping performance overall.



(a) Hidden Object Interaction



(b) Direct Human-Robot Interaction

Fig. 6: Online Demos for extended application scenarios. (a) **Hidden Object Interaction:** Using only VI-EEG, the subject guides the system to identify and reveal a hidden object, revealing an ability to decode mental imagery without visual cues. (b) **Direct Human-Robot Interaction:** By imagining the object and controlling hand movements through Visual Imagery and Motor Imagery, the subject interacts seamlessly with the robot, which places the object directly into their hand.

C. Extended Application Scenarios

Beyond the base demonstration, we conducted a series of online tests in various scenarios to further validate the

system’s robustness and adaptability. A **key strength of our system is its posture-independent operation.** Tasks rely solely on mental imagery without additional visual or environmental cues, allowing the system to function in any environment or orientation. In addition to validating the system in a seated posture, **we also successfully completed online tests with subjects in a supine position**, as shown in the lower left corner of Figure 1.

To further assess the system’s performance in environments with no visual input, we designed specific experiments to test its ability to decode neural signals and interact with objects. Demonstrations are presented in Figure 6.

Hidden Object Interaction: EEG Decoding Without Visual Input. The Hidden Object Interaction task was designed to evaluate the system’s ability to decode EEG-derived mental imagery and perform object manipulation without any visual information. In this experiment, various objects were placed on a table and concealed from the subject’s view with lids. Without knowing the type or location of the objects, the subject relied solely on mental imagery to guide the system. EEG signals corresponding to the subject’s visual imagination were decoded to identify the intended object, and the robot then manipulated and revealed the object that matched the subject’s mental imagery.

Direct Human-Robot Interaction: Object Manipulation with Visual and Motor Imagery. The Direct Human-Robot Interaction task aimed to evaluate the system’s ability to support seamless, intuitive interaction between humans and robots. In this scenario, the subject first visualized a target by VI, and the system decoded the EEG signals to identify the intended object. The subject then performed MI by imagining left- or right-hand movements while extending the corresponding hand. The system decoded these motor signals to determine the intended hand position, enabling the robot to grasp the target and place it directly into the subject’s hand. This task demonstrated the integration of visual and motor imagery for reliable, real-time human-robot interaction.

VI. CONCLUSIONS

In this study, we present a high-level cognitive control paradigm for natural intention-driven BCI, where VI-EEG decodes grasp intentions and MI-EEG infers target placement positions. By combining offline pretraining with a real-time framework, we achieved seamless mapping from visual cognition to robotic actions, enabling accurate grasping and reliable placement of real-world objects in online scenarios. Although the system accuracy can still be improved, our results demonstrate the feasibility of leveraging higher-order cognitive states for intuitive brain–robot interaction. Current limitations mainly include the intrinsic constraints of EEG, insufficient feature separability of VI-EEG, and the need for improved long-term robustness and generalizability, which point to promising directions for future work.

REFERENCES

- [1] H. Zhang, J. Tang, S. Sun, and X. Lan, “Robotic grasping from classical to modern: A survey,” *arXiv preprint arXiv:2202.03631*, 2022.

- [2] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.
- [3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.
- [4] T. Wang, H. Lin, J. Yu, and Y. Fu, "Polaris: Open-ended interactive robotic manipulation via syn2real visual grounding and large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9676–9683.
- [5] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," *arXiv preprint arXiv:2308.07931*, 2023.
- [6] J. Björck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.
- [7] 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," *arXiv preprint arXiv:2311.01454*, 2023.
- [8] Y. Zhang, N. Rajabi, F. Taleb, A. Matvienko, Y. Ma, M. Björkman, and D. Kragic, "Mind meets robots: a review of eeg-based brain-robot interaction systems," *International Journal of Human-Computer Interaction*, pp. 1–32, 2025.
- [9] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8476–8482.
- [10] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 863–869.
- [11] F. Beverina, G. Palmas, S. Silvoni, F. Piccione, S. Giove *et al.*, "User adaptive bcis: Ssvep and p300 based interfaces," *PsychNology J.*, vol. 1, no. 4, pp. 331–354, 2003.
- [12] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 4, pp. 473–482, 2007.
- [13] K. Värbu, N. Muhammad, and Y. Muhammad, "Past, present, and future of eeg-based bci applications," *Sensors*, vol. 22, no. 9, p. 3331, 2022.
- [14] P. Tai, P. Ding, F. Wang, A. Gong, T. Li, L. Zhao, L. Su, and Y. Fu, "Brain-computer interface paradigms and neural coding," *Frontiers in neuroscience*, vol. 17, p. 1345961, 2024.
- [15] H. Wilson, X. Chen, M. Golbabaee, M. J. Proulx, and E. O'Neill, "Feasibility of decoding visual information from eeg," *Brain-computer interfaces*, vol. 11, no. 1-2, pp. 33–60, 2024.
- [16] J. Kilmarx, I. Tashev, J. d. R. Millán, J. Sulzer, and J. Lewis-Peacock, "Evaluating the feasibility of visual imagery for an eeg-based brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 2209–2219, 2024.
- [17] J. Gao, Y. Liu, B. Yang, J. Feng, and Y. Fu, "Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing," *arXiv preprint arXiv:2503.06940*, 2025.
- [18] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.
- [19] Y.-T. Lan, K. Ren, Y. Wang, W.-L. Zheng, D. Li, B.-L. Lu, and L. Qiu, "Seeing through the brain: image reconstruction of visual perception from human brain signals," *arXiv preprint arXiv:2308.02510*, 2023.
- [20] J. Huo, Y. Wang, Y. Wang, X. Qian, C. Li, Y. Fu, and J. Feng, "Neuro-pictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 56–73.
- [21] J. Gao, Y. Fu, Y. Fu, Y. Wang, X. Qian, and J. Feng, "Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [22] L. Yang and Z. Wang, "Applications and advances of combined fmri-fnirs techniques in brain functional research," *Frontiers in Neurology*, vol. 16, p. 1542075, 2025.
- [23] B. He, H. Yuan, J. Meng, and S. Gao, "Brain-computer interfaces," in *Neural engineering*. Springer, 2020, pp. 131–183.
- [24] W. Li, P. Zhao, C. Xu, Y. Hou, W. Jiang, and A. Song, "Deep learning for eeg-based visual classification and reconstruction: Panorama, trends, challenges and opportunities," *IEEE Transactions on Biomedical Engineering*, 2025.
- [25] S. Bagchi and D. R. Bathula, "Eeg-convtransformer for single-trial eeg-based visual stimulus classification," *Pattern Recognition*, vol. 129, p. 108757, 2022.
- [26] J. Kalafatovich, M. Lee, and S.-W. Lee, "Learning spatiotemporal graph representations for visual perception using eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 97–108, 2022.
- [27] C. Du, K. Fu, J. Li, and H. He, "Decoding visual neural representations by multimodal learning of brain-visual-linguistic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10760–10777, 2023.
- [28] Y. Song, B. Liu, X. Li, N. Shi, Y. Wang, and X. Gao, "Decoding natural images from eeg for object recognition," *arXiv preprint arXiv:2308.13234*, 2023.
- [29] H. Guo, "Eegvision: reconstructing vision from human brain signals," *Appl Math Nonlinear Sci*, vol. 9, no. 1, p. 2, 2024.
- [30] D. Li, C. Wei, S. Li, J. Zou, H. Qin, and Q. Liu, "Visual decoding and reconstruction via eeg embeddings with guided diffusion," *arXiv preprint arXiv:2403.07721*, 2024.
- [31] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of eeg-based brain-computer interface paradigms," *Journal of neural engineering*, vol. 16, no. 1, p. 011001, 2019.
- [32] M. Kim, M.-S. Choi, G.-R. Jang, J.-H. Bae, and H.-S. Park, "Eeg-controlled tele-grasping for undefined objects," *Frontiers in Neuro-robotics*, vol. 17, p. 1293878, 2023.
- [33] I. Hameed, D. M. Khan, S. M. Ahmed, S. S. Aftab, and H. Fazal, "Enhancing motor imagery eeg signal decoding through machine learning: A systematic review of recent progress," *Computers in Biology and Medicine*, vol. 185, p. 109534, 2025.
- [34] E. M. Dos Santos, R. San-Martin, and F. J. Fraga, "Comparison of subject-independent and subject-specific eeg-based bci using lda and svm classifiers," *Medical & biological engineering & computing*, vol. 61, no. 3, pp. 835–845, 2023.
- [35] I. Carrara, B. Aristimunha, M.-C. Corsi, R. Y. de Camargo, S. Chevalier, and T. Papadopoulos, "Geometric neural network based on phase space for bci-eeg decoding," *Journal of Neural Engineering*, vol. 22, no. 1, p. 016049, 2025.
- [36] S. Lee, S. Jang, and S. C. Jun, "Exploring the ability to classify visual perception and visual imagery eeg data: Toward an intuitive bci system," *Electronics*, vol. 11, no. 17, p. 2706, 2022.
- [37] S. Zhu, Z. Ye, Q. Ai, and Y. Liu, "Eeg-imagenet: An electroencephalogram dataset and benchmarks with image visual stimuli of multi-granularity labels," *arXiv preprint arXiv:2406.07151*, 2024.
- [38] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
- [39] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2020.
- [40] J. Wang and M. Wang, "Review of the emotional feature extraction and classification using eeg signals," *Cognitive robotics*, vol. 1, pp. 29–40, 2021.
- [41] B. Yang, F. Rong, Y. Xie, D. Li, J. Zhang, F. Li, G. Shi, and X. Gao, "A multi-day and high-quality eeg dataset for motor imagery brain-computer interface," *Scientific Data*, vol. 12, no. 1, p. 488, 2025.
- [42] T. Zhu, H. Tang, L. Jiang, Y. Li, S. Li, and Z. Wu, "A study of motor imagery eeg classification based on feature fusion and attentional mechanisms," *Frontiers in Human Neuroscience*, vol. 19, p. 1611229, 2025.
- [43] E. L. Holm, D. F. Slezak, and E. Tagliazucchi, "Contribution of low-level image statistics to eeg decoding of semantic content in multivariate and univariate models with feature optimization," *NeuroImage*, vol. 293, p. 120626, 2024.
- [44] T. Orima and I. Motoyoshi, "Spatiotemporal cortical dynamics for visual scene processing as revealed by eeg decoding," *Frontiers in Neuroscience*, vol. 17, p. 1167719, 2023.