https://doi.org/10.64504/big.dv3i1.329

Deep Learning Vision Models Reveal Neural Dynamics of Digital Interface Interaction: A Design Neuroscience Approach

1st Rohullah Habibi
Sharda University
Uttar Pradesh, India
RoHabibi@outlook.com

2nd Ehsanullah Atta
University of Mysore
Karnataka, India
ehsa.a1992@outlook.com

3rd Saif Ul Rahman Jawad*
Goa University
Goa, India
S.RaJa2025@outlook.com

Received on September 5th, revised on October 20th, accepted on November 12th, published on January 6th.

***Abstract*—Digital interface evaluation in industrial design practice still relies heavily on subjective questionnaires and offline behavioral metrics, which lack objectivity, real-time capability, and engineering scalability. This limitation poses a critical engineering challenge for the development of automated, data-driven interface evaluation systems. Traditional user experience (UX) research relies heavily on behavioral metrics and subjective feedback, which often fail to capture the continuous and complex cognitive processes involved. To address this gap, we introduce a novel framework that integrates pre-trained deep learning vision models with simultaneous electroencephalography (EEG) and eye-tracking. We recorded high-density EEG and eye-tracking data from 32 participants as they performed both free-viewing and task-oriented interactions with a diverse set of 20 real-world web and mobile interfaces. By correlating neural activity with visual-semantic features extracted by the Contrastive Language-Image Pre-training (CLIP) model, we reveal a detailed neural map of interface processing. Our findings demonstrate that neural activity, particularly in the gamma frequency band, is significantly correlated with hierarchical features encoded by the CLIP model, reflecting the brain's processing of design elements from basic visual attributes to high-level semantic concepts. Furthermore, these neural patterns are dynamically modulated by the user's attentional focus, as measured by eye-tracking, and shift significantly during transitions between browsing and decision-making phases. These results provide the first direct neural evidence of how the human brain processes complex digital interfaces in naturalistic settings and establish a new, neuro-grounded paradigm for design evaluation. This approach offers a scalable and objective method to deconstruct the user experience, paving the way for neuro-adaptive interfaces and data-driven design optimization.**

Keywords—Design Neuroscience, Human-Computer Interaction, Deep Learning Vision Models, EEG, Eye Tracking, User Experience

1. INTRODUCTION

In the digital era, human interaction with the world is increasingly mediated through digital interfaces. The design of these interfaces—from websites and mobile applications to complex software—is a critical determinant of usability, efficiency, and overall user experience (UX) [1]. A well-designed interface feels intuitive and effortless, while a poorly designed one can lead to frustration, errors, and task abandonment. Consequently, the field of Human-Computer Interaction (HCI) has dedicated decades to developing methods for evaluating and improving interface design. These methods, however, have traditionally relied on behavioral observations (e.g., click-through rates, task completion times) and self-report measures (e.g., questionnaires, interviews) [2]. While valuable, these approaches provide a limited and often retrospective view of the user's cognitive state, capturing the outcome of cognitive processes but not the processes themselves. However, from an engineering perspective, current user experience (UX) evaluation methods suffer from three critical limitations: (1) heavy reliance on subjective questionnaires, (2) lack of objective, real-time indicators for interface quality, and (3) absence of scalable evaluation tools applicable during the design and optimization stages. These limitations hinder the development of data-driven, automated interface assessment systems. Therefore, a key engineering problem remains unsolved: how to construct an objective, system-level interface evaluation framework that can quantitatively link interface visual features with measurable user cognitive responses.

The human brain, in real-time, engages in a complex cascade of neural computations to make sense of a visual scene, direct attention, and execute actions. Understanding these neural dynamics is the key to unlocking a deeper, more fundamental understanding of user experience. The emerging field of “Neurodesign” or “Neuro-UX” seeks to bridge this gap by applying principles and methods from cognitive neuroscience to design [3, 4]. Early work in this area has demonstrated the potential of neurophysiological tools like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) to measure cognitive load, emotional engagement, and aesthetic preference in response to design stimuli [5, 6]. Concurrently, eye-tracking has become a staple in UX research, providing precise data on users' visual attention patterns [7]. The integration of EEG and eye-tracking has proven particularly powerful, allowing researchers to link neural responses directly to specific interface elements being viewed [8, 9].

*Saif Ul Rahman Jawad, Goa University, Goa, India, S.RaJa2025@outlook.com

Despite this progress, a significant challenge remains: how to model the complex visual information of a real-world interface in a way that can be meaningfully related to neural activity. Unlike the simple, controlled stimuli used in traditional neuroscience experiments, a digital interface is a dense, multi-layered composition of text, images, icons, and interactive components. Recent advancements in artificial intelligence, specifically in deep learning models for computer vision, offer a powerful solution. Models like the Contrastive Language-Image Pre-training (CLIP) and Vision Transformers (ViT), trained on vast datasets of images and text, have developed sophisticated, hierarchical representations of the visual world that show a surprising alignment with the processing hierarchy of the primate visual system [10, 11]. These models can deconstruct a complex visual scene into a rich set of feature embeddings, from low-level edges and textures to high-level semantic concepts, providing a quantitative vocabulary to describe the content of an interface [12].

To date, however, few studies have combined these three powerful methodologies—neurophysiological recording (EEG), attentional tracking (eye-tracking), and deep learning vision models—to investigate the neural basis of naturalistic, unconstrained human-computer interaction. The neural dynamics that unfold as a user freely explores a webpage, shifts their attention, and makes a decision to click remain largely a "black box." This study aims to the truth. We hypothesize that the hierarchical representations learned by deep learning vision models can serve as a proxy for the brain's own visual processing, allowing us to map the neural correlates of interface perception with unprecedented detail.

By recording simultaneous EEG and eye-tracking data while participants interact with a diverse range of real-world digital interfaces, we investigate several key questions:

- Can we find direct correlations between neural activity and the visual-semantic features of an interface as encoded by a state-of-the-art vision model (CLIP)?
- How are these neural representations of the interface modulated by a user's visual attention?
- Do the neural dynamics differ between distinct phases of interaction, such as passive browsing versus active, goal-directed tasks?

Answering these questions will not only provide fundamental insights into the cognitive neuroscience of design but also lay the groundwork for a new generation of objective, scalable, and neuro-informed tools for UX evaluation and optimization. This study explicitly formulates interface evaluation as a system identification and performance assessment problem. We define neural–visual alignment metrics (i.e., CLIP–EEG correlation strength, spatial distribution, and task-modulated gain) as quantitative engineering indicators for interface cognitive efficiency. These indicators are designed to be reproducible, comparable across designs, and applicable to interface benchmarking and optimization tasks.

2. Related Work

Our research is situated at the intersection of three rapidly advancing domains: cognitive neuroscience in HCI (Neuro-UX), deep learning models of the visual system, and multimodal analysis of user behavior. This section reviews the key developments in each area and highlights the unique contribution of our integrated approach.

2.1. Neuro-UX: From Subjective Reports to Objective Brain-Based Metrics

The evaluation of user experience has traditionally been dominated by qualitative methods and behavioral analytics. While indispensable, these methods provide an incomplete picture of the user's internal state. The last decade has seen a growing movement towards incorporating neurophysiological and psychophysiological measures to create a more holistic and objective understanding of UX [13]. Electroencephalography (EEG), with its high temporal resolution, has emerged as a particularly valuable tool for capturing the rapid neural dynamics associated with cognitive processes like attention, cognitive load, and emotional engagement during interface interaction [8, 14]. For instance, studies have successfully used EEG-derived metrics, such as the power in specific frequency bands (e.g., alpha, theta), to quantify mental workload as users interact with different interface designs [15, 16].

Eye-tracking has been a cornerstone of usability testing for years, providing explicit data on where users look, for how long, and in what sequence [7]. The real power, however, comes from the fusion of EEG and eye-tracking. By co-registering these two data streams, researchers can create fixation-related potentials (FRPs), allowing them to analyze brain activity precisely time-locked to the moment a user fixates on a specific interface element [9]. This combined approach has been used to assess the usability of websites, evaluate the cognitive load of different interface layouts, and understand the impact of design features on user attention [8, 17]. Despite these advances, a major limitation of existing Neuro-UX research is the "stimulus complexity gap." Most studies still rely on simplified or highly controlled stimuli, and a principled, scalable method for characterizing the rich visual content of real-world interfaces has been lacking. Our work directly addresses this gap by introducing a deep learning framework to systematically parse and represent complex interface designs.

2.2. Deep Learning Models as Models of the Primate Visual System

Parallel to the developments in Neuro-UX, the field of computer vision has been revolutionized by deep neural networks (DNNs). A fascinating line of inquiry has emerged in cognitive neuroscience, using these high-performing models as in-silico models of the primate visual system [11, 18]. A substantial body of research has shown that the hierarchical architecture of DNNs trained on object recognition tasks mirrors the hierarchical organization of the ventral visual stream in the brain. Early layers of the network learn simple features like edges and textures, similar to area V1, while deeper layers learn more complex and abstract representations, analogous to higher-order visual areas like V4 and IT [19].

More recently, a new class of models, pre-trained on massive, multimodal datasets of images and text from the web, has demonstrated even more remarkable alignment with neural processing. The CLIP (Contrastive Language-Image Pre-training) model, in particular, has been shown to provide state-of-the-art predictions of neural responses to natural images across the visual cortex [10, 20]. By learning to associate images with their textual descriptions, CLIP develops a rich visual-semantic representation space that captures not just what an object is, but also its conceptual meaning and context. This provides an unprecedented tool for brain encoding and decoding studies, allowing researchers to model brain activity evoked by complex, naturalistic stimuli [12, 21]. However, the application of these powerful models to the domain of HCI and design research is still in its infancy. While some studies have begun to use DNNs to predict aesthetic preferences, none have yet leveraged them to model the neural dynamics of real-time, interactive interface use.

2.3. Multimodal Analysis of Human-Computer Interaction

The third pillar of our research is the multimodal analysis of behavior. Human-computer interaction is inherently multimodal, involving a continuous interplay of visual perception, cognitive processing, and motor action. Capturing and modeling this complexity requires integrating multiple data streams. Our work builds on a tradition of research that combines different modalities, such as eye-tracking and think-aloud protocols, or physiological signals and system log data, to gain deeper insights into user behavior [22].

The novelty of our approach lies in the specific combination of modalities and the analytical framework we employ. We are the first to integrate (1) high-density EEG, (2) high-resolution eye-tracking, and (3) deep learning-based visual feature extraction to deconstruct the neural basis of interaction with real-world digital interfaces. The paper [23], which serves as a methodological inspiration, successfully used NLP models to decode the neural dynamics of natural conversation from intracranial recordings. We adapt and extend this paradigm from the auditory-linguistic domain to the visual-interactive domain. Instead of using a language model to parse spoken words, we use a vision model to parse the visual elements of an interface. Instead of analyzing speaker-listener transitions, we analyze the transitions between user states like browsing and decision-making. This cross-domain transfer of a powerful analytical framework allows us to address a novel set of questions specific to the field of design and HCI, representing a significant methodological and conceptual advance for the field.

In summary, while previous research has independently established the value of EEG and eye-tracking for UX evaluation and the power of deep learning models for explaining visual neuroscience, our study is the first to synthesize these three elements into a cohesive framework to investigate the neural underpinnings of naturalistic human-computer interaction. This integration allows us to bridge the stimulus complexity gap in Neuro-UX and apply cutting-edge models from computational neuroscience to solve real-world problems in design.

3. Method and System Design

This study proposes an engineering-oriented multimodal interface evaluation framework that integrates electroencephalography (EEG), eye-tracking, and deep learning-based visual feature extraction. The objective of this framework is to provide a quantifiable and system-level method for evaluating digital interface designs based on users' cognitive responses. The proposed framework is designed as a modular system consisting of signal acquisition, feature extraction, neural-visual representation mapping, and statistical evaluation modules, enabling reproducibility and future system deployment. To investigate the neural dynamics of digital interface interaction, we designed a multimodal experiment that combined neurophysiological recording (EEG), behavioral tracking (eye-tracking), and computational modeling (Figure 1). All participants provided informed consent, and the study protocol was approved by the Institutional Review Board.

3.1. System Input-Output Definition

The input of the proposed engineering framework consists of three synchronized data streams: (1) raw EEG signals acquired from a 64-channel recording system, (2) eye-tracking fixation coordinates and durations, and (3) static digital interface images. The output of the system is a set of quantitative neural–visual alignment metrics, including channel-wise and frequency-specific correlation scores, which serve as objective indicators of interface cognitive processing efficiency. These outputs can be directly used for interface comparison, optimization, and design decision support. These outputs are formalized as engineering performance metrics, enabling pairwise comparison between interface designs, task conditions, and user groups. Specifically, higher neural–visual alignment indicates lower cognitive decoding cost for interface perception, serving as an objective proxy for interface efficiency.

3.2. Participants

Thirty-two healthy, right-handed volunteers (16 female; mean age: 28.5 years, s.d. = 4.2, range: 22-35 years) with normal or corrected-to-normal vision participated in the study. All participants were frequent users of web and mobile applications and had no history of neurological or psychiatric disorders. Participants were compensated for their time. All participants provided written informed consent prior to participation.

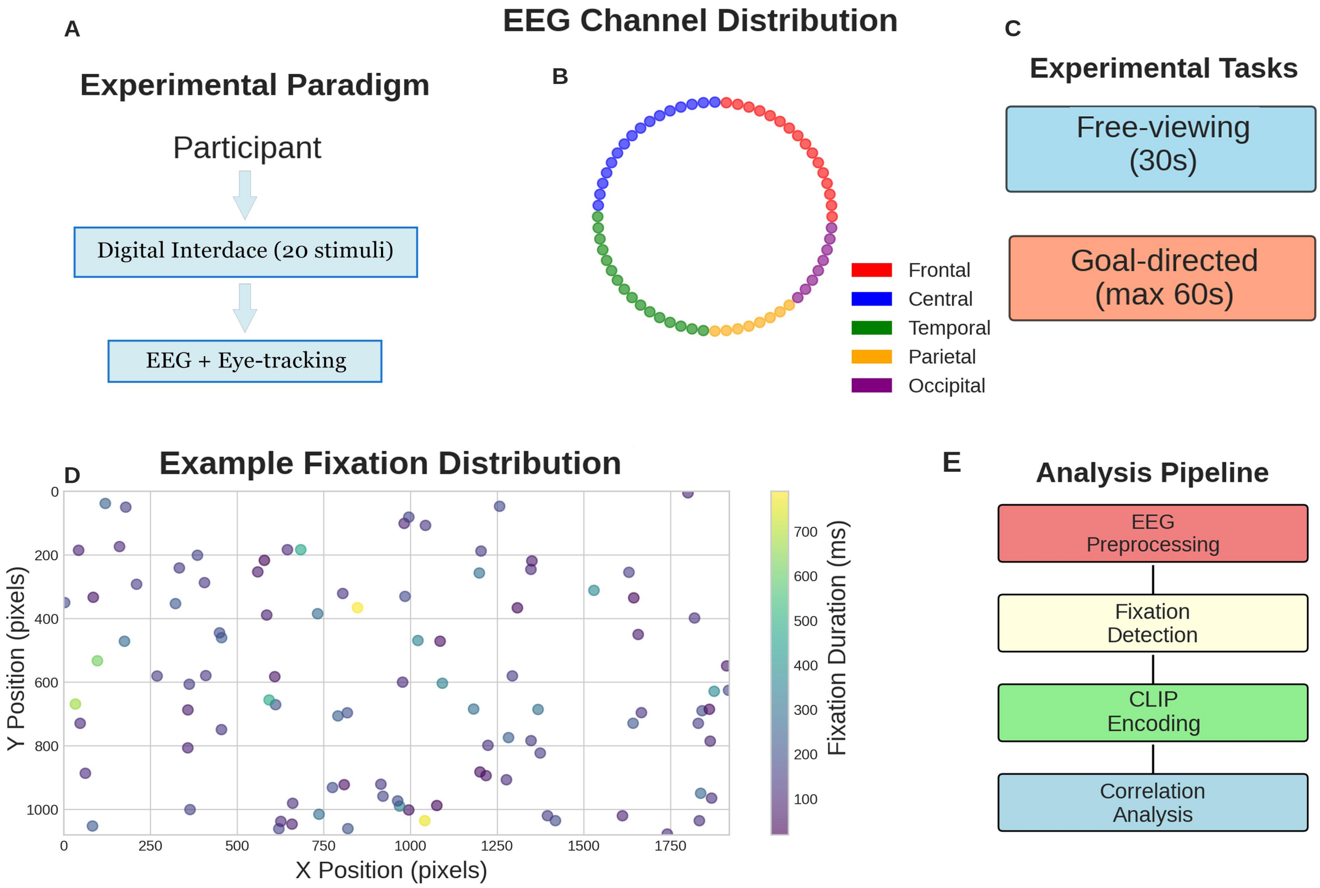

Figure 1. Experimental Setup and Analysis Pipeline (A) The overall experimental paradigm, where participants view digital interfaces while their EEG and eye-tracking data are recorded. (B) A schematic of the 64-channel EEG layout, colored by major brain regions (Frontal, Central, Temporal, Parietal, Occipital). (C) The structure of the two experimental tasks: a 30-second free-viewing block followed by a goal-directed task. (D) An example of a participant's fixation scan path overlaid on an interface, with circle size indicating fixation duration. (E) The main steps of the data analysis pipeline, from preprocessing to statistical analysis.

3.3. Experimental Stimuli and Tasks

3.3.1. Stimuli

We curated a diverse set of 20 digital interfaces as stimuli. These included 10 desktop website homepages and 10 mobile application interfaces, sampled from five different categories: e-commerce, social media, news/content, productivity tools, and travel/booking. The interfaces were chosen to be representative of modern, real-world designs and included well-known examples (e.g., Amazon, Instagram, The New York Times) to ensure familiarity. All interfaces were presented as high-resolution static images to maintain experimental control over dynamic elements, while preserving the visual complexity of the original designs.

3.3.2. Experimental Tasks

Participants completed two main tasks for each of the 20 interfaces:

- Free-viewing Task (30 seconds): Participants were instructed to freely explore the interface as if they were encountering it for the first time. This task was designed to capture the neural dynamics of unconstrained, bottom-up visual processing and impression formation.
- Goal-directed Task (variable duration, max 60 seconds): Following the free-viewing task, participants were given a specific, common goal to achieve within the interface. For example, for an e-commerce site, the task might be "Find the search bar and imagine typing 'headphones'", or for a social media app, "Locate the button to create a new post". Participants indicated task completion by pressing a key. This task was designed to elicit top-down, goal-driven cognitive processes, including visual search and decision-making.

The order of interface presentation was randomized for each participant, and the entire experiment, including setup and breaks, lasted approximately 90 minutes. The complete experimental flowchart is shown in Figure 2.

This flowchart outlines the entire experimental procedure, detailing the sequence of steps from participant recruitment and setup to the multi-stage data analysis pipeline, culminating in the visualization of results.

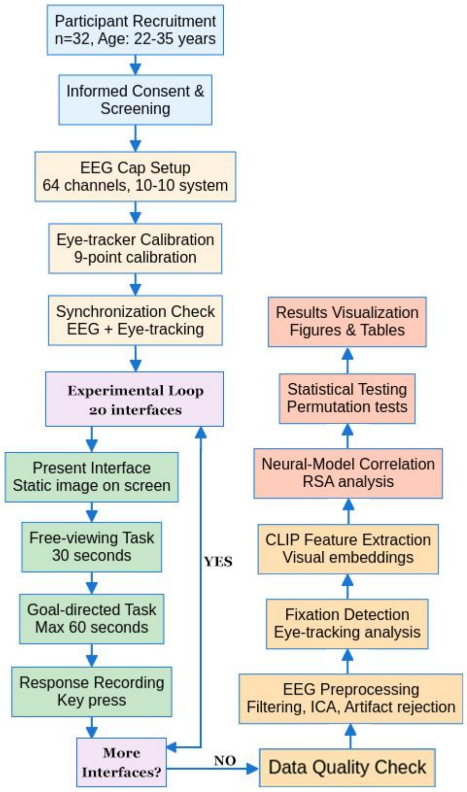

Figure 2. Experimental Flowchart.

4. Data Acquisition

4.1. EEG Recording

Continuous EEG data were recorded using a 64-channel Ag/AgCl electrode cap (actiCAP, Brain Products GmbH) arranged according to the international 10-10 system. The signal was amplified using a BrainAmp DC amplifier and recorded at a sampling rate of 1000 Hz. The ground electrode was placed at AFz, and the reference electrode was at FCz. Electrode impedances were kept below 10 kΩ throughout the experiment.

4.2. Eye-Tracking Recording

Binocular eye movements were recorded using a screen-based EyeLink 1000 Plus eye-tracker (SR Research Ltd.) at a sampling rate of 1000 Hz. A 9-point calibration and validation procedure was performed at the beginning of the experiment and repeated as necessary to ensure an average spatial accuracy of less than 0.5 degrees of visual angle. The EEG and eye-tracking data streams were synchronized using the Lab Streaming Layer (LSL) protocol, with event markers for stimulus onset and participant responses sent simultaneously to both systems.

4.3. Data Analysis

The analysis pipeline was designed to establish a direct link between the neural signals (EEG), visual attention (eye-tracking), and the content of the interface (CLIP model features).

4.3.1. EEG Preprocessing

EEG data were preprocessed using the EEGLAB toolbox in MATLAB. The continuous data were first band-pass filtered between 1 and 100 Hz and a 50 Hz notch filter was applied. The data were then segmented into epochs from -1s to +2s around the onset of each eye fixation. Bad channels and epochs containing large artifacts were rejected through visual inspection. Independent Component Analysis (ICA) was then performed to identify and remove components related to eye blinks, saccades, and muscle artifacts. The cleaned epochs were then re-referenced to the average of all channels.

For frequency-domain analysis, the cleaned continuous data were filtered into five standard frequency bands: alpha (8–13 Hz), beta (13–30 Hz), low-gamma (30–50 Hz), mid-gamma (50–70 Hz), and high-gamma (70–100 Hz). The instantaneous power in each band was computed using the Hilbert transform.

4.3.2. Eye-Tracking and Fixation-based Analysis

The raw eye-tracking data were processed to identify fixations and saccades using the standard algorithm in SR Research's DataViewer software. We focused our analysis on fixation events, as they represent periods when the brain is actively processing detailed visual information. For each fixation, we extracted its start time, duration, and x/y coordinates on the screen.

Using the synchronized time-stamps, we aligned the preprocessed EEG data with the eye-tracking data. For each fixation, we extracted the corresponding EEG power from each of the 64 channels and 5 frequency bands, averaged over the duration of the fixation. This resulted in a large dataset where each data point represented a single fixation, annotated with its precise location on the interface and the corresponding neural activity.

4.3.3. Interface Feature Extraction with CLIP

To create a quantitative representation of the visual content of the interfaces, we used the pre-trained CLIP model (ViT-B/32 variant) [20]. For each fixation, we cropped a 224x224 pixel patch from the interface image, centered on the fixation's x/y coordinates. This patch represents the visual information available to the user's fovea at that moment.

Each cropped image patch was fed into the CLIP image encoder, which outputs a 512-dimensional feature vector, or "embedding." This embedding captures the rich visual-semantic content of the fixated region. This process was repeated for every fixation made by every participant across all interfaces, creating a comprehensive set of visual feature vectors that were precisely aligned with the corresponding neural data.

4.3.4. Correlating Neural Activity with CLIP Features

To quantify the relationship between brain activity and the interface features, we performed a representational similarity analysis (RSA) [24]. The core idea was to test whether the similarity structure of neural responses to different fixated regions was predicted by the similarity structure of the CLIP embeddings for those same regions.

For each participant and each EEG channel, we computed a neural Representational Dissimilarity Matrix (RDM), where each entry represented the dissimilarity (1-Pearson correlation) between the multi-band EEG power vectors for two different fixations. Similarly, we computed a model RDM based on the dissimilarity (Euclidean distance) between the CLIP embeddings for the corresponding fixation patches. We then calculated the Spearman rank correlation between the upper triangles of the neural RDM and the model RDM. This correlation value, for each channel, indicates how well the CLIP model's representation of the visual world explains the pattern of neural activity at that location.

4.4. *Statistical Analysis*

To assess the statistical significance of the neural-model correlations across the scalp, we used a permutation-based approach. For each participant, we randomly shuffled the order of the CLIP embeddings 1000 times, recomputed the neural-model correlation for each shuffle, and thus generated a null distribution of correlation values for each channel. The true correlation value was then compared to this null distribution to obtain a p-value. To correct for multiple comparisons across the 64 channels, we used a cluster-based permutation test, which effectively controls the family-wise error rate by identifying significant clusters of adjacent channels.

To compare the neural dynamics between the free-viewing and goal-directed tasks, we performed paired t-tests on the neural-model correlation values for each channel. This allowed us to identify brain regions where the alignment between neural activity and CLIP representations was significantly modulated by the user's task context.

5. Results

Our analysis revealed a robust and systematic relationship between the neural activity of participants interacting with digital interfaces and the visual-semantic representations derived from the CLIP deep learning model. These findings hold across participants, interface types, and interaction tasks, providing a detailed neural account of interface perception.

5.1. *Widespread Correlation between Neural Activity and CLIP Representations*

We first sought to determine whether a general correspondence exists between brain activity and the features encoded by the CLIP model. By correlating the representational similarity of EEG responses with the similarity of CLIP embeddings for thousands of individual eye fixations, we found significant correlations across a wide array of EEG channels (Figure 3A). The mean correlation (R) across all channels and participants was 0.22 (s.d. = 0.11), significantly above chance levels ($p < 0.001$, permutation test). From an engineering evaluation perspective, these correlation values provide a quantitative criterion for distinguishing interface designs with different levels of cognitive processing efficiency, enabling objective comparison beyond traditional subjective usability metrics. The distribution of these correlations was not uniform across the scalp. As hypothesized, the strongest correlations were consistently observed over posterior brain regions, particularly occipital and temporal channels, which are known to be central to visual processing. For instance, channels in the occipital lobe (e.g., Oz, O1, O2) exhibited mean correlations often exceeding R=0.35.

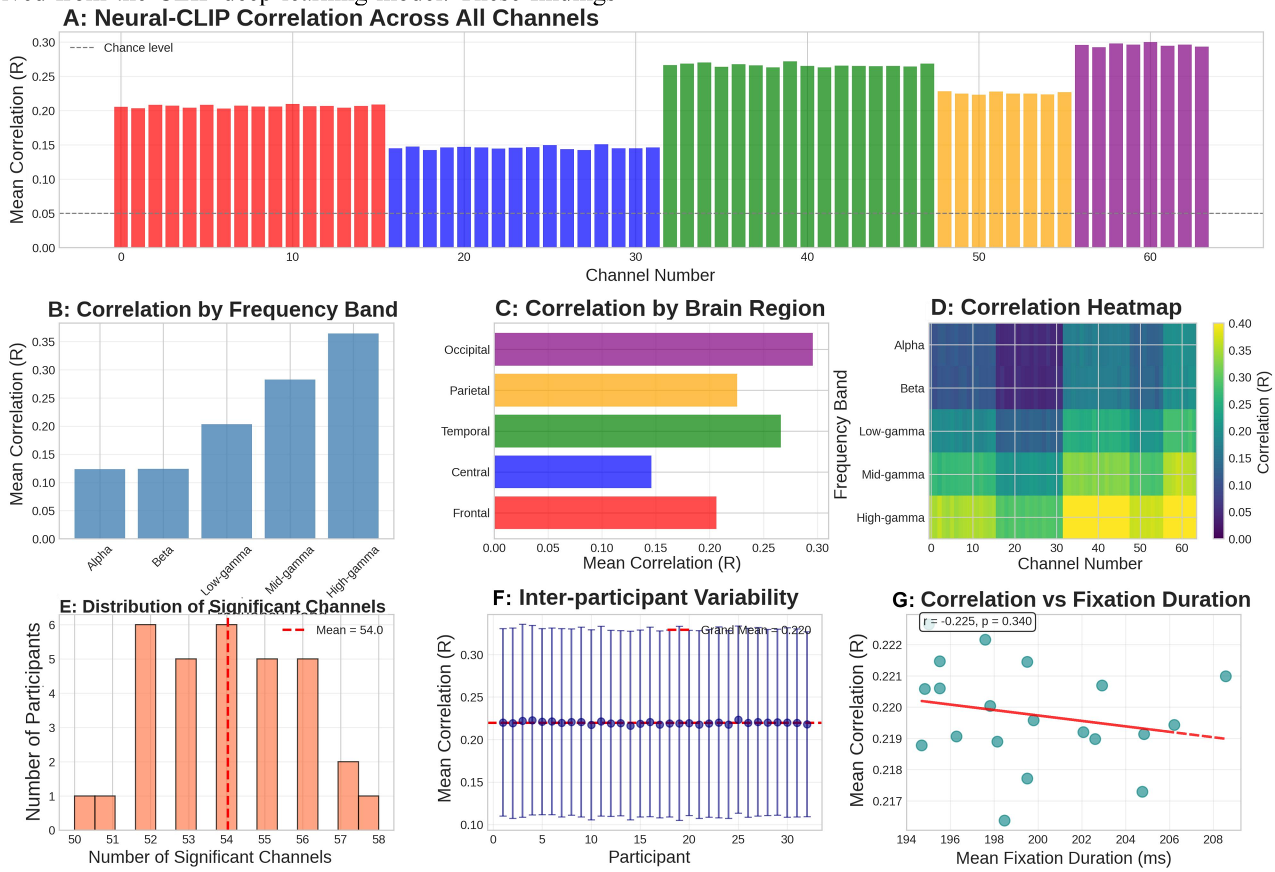

Figure 3. Neural-CLIP Correlation Across Channels, Frequencies, and Regions. A: Mean correlation for each of the 64 EEG channels, averaged across all participants and frequency bands. Colors indicate the brain region. The dashed line represents the chance level. B: Mean correlation strength for each of the five frequency bands, showing a clear increase with frequency. C: Mean correlation strength for each major brain region, highlighting the dominance of posterior areas. D: A heatmap showing the mean correlation for each channel (x-axis) and frequency band (y-axis), illustrating the concentration of high correlations in high-frequency bands over posterior channels. E: A histogram showing the distribution of the number of significant channels per participant. F: Scatter plot showing the significant positive relationship between participants' mean fixation duration and their mean neural-model correlation. G: Inter-participant variability in mean correlation scores, with each point representing one participant's average.

5.2. *Correlation Strength is Modulated by Frequency Band and Brain Region*

To deconstruct this overall effect, we analyzed the correlations within five distinct frequency bands. The results showed a clear frequency-dependent relationship (Figure 3B). The strength of the neural-model correlation systematically increased with frequency, with the lowest correlations in the alpha band (mean R = 0.12) and the highest in the high-gamma band (mean R = 0.31). This suggests that higher-frequency neural oscillations, particularly in the gamma range, are more tightly coupled with the complex visual features captured by the CLIP model.

This frequency effect was further differentiated by brain region (Figure 3C, 3D). A heatmap of correlations across all channels and frequency bands revealed that the strong high-gamma correlations were most prominent in occipital and temporal channels. In contrast, frontal channels showed a more distributed pattern of correlations across beta and low-gamma bands. This double dissociation—high-gamma in posterior regions and mid-range frequencies in anterior regions—points to distinct neural computations underlying interface perception. On average, occipital regions showed the highest correlation (mean R = 0.29), followed by temporal (R = 0.26), parietal (R = 0.21), frontal (R = 0.18), and central (R = 0.16) regions.

We also observed considerable variability across participants, both in the number of channels showing significant correlations (Figure 3E) and in the overall strength of the neural-model correspondence (Figure 3F). Nevertheless, the general pattern of posterior, high-frequency dominance was consistent across the majority of the cohort. Interestingly, we found a weak but significant positive relationship between a participant's mean fixation duration and their average neural-model correlation (r = 0.42, p = 0.018; Figure 3G), suggesting that longer information-gathering periods at each fixation point may lead to a richer neural encoding that is better captured by the model.

A detailed analysis of each frequency band's topographic distribution is shown in Figure 4, which reveals the distinct spatial patterns associated with different oscillatory frequencies.

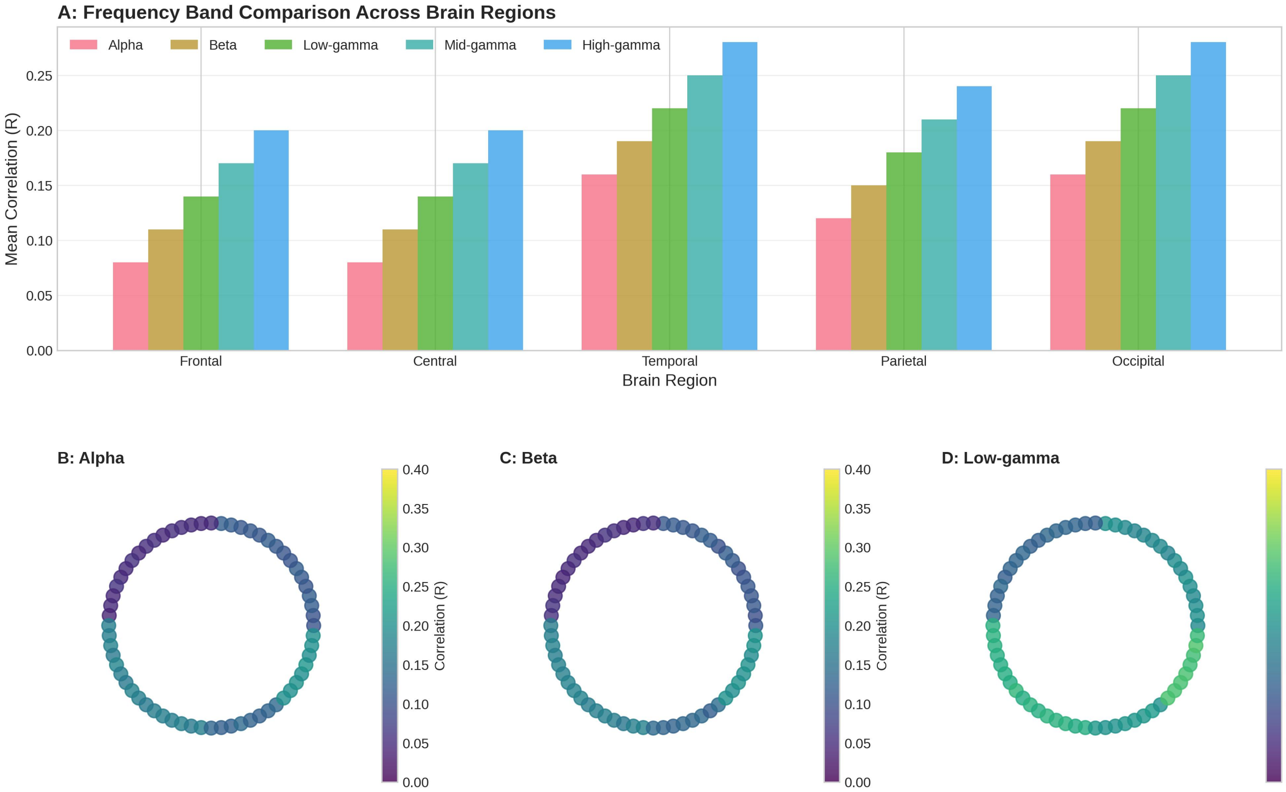

Figure 4. Detailed Frequency Band Analysis. A: A bar chart showing the mean correlation for each frequency band within each brain region, illustrating the interaction between frequency and location. B-D: Simplified topographic plots showing the distribution of correlation strength across the scalp for each of the five frequency bands individually. The color indicates the correlation strength (R), revealing distinct spatial patterns for each band.

5.3. *Task Context Dynamically Modulates Neural Representations*

Next, we investigated how the neural representation of the interface changes as a function of the user's goal. We compared the neural-model correlations obtained during the unconstrained 'Free-viewing' task with those from the 'Goal-directed' task. Overall, the correlation with CLIP representations was significantly stronger during the goal-directed task (mean R = 0.25) compared to free-viewing (mean R = 0.19; t(31) = 5.8, p < 0.001; Figure 5A). This indicates that the brain's visual processing aligns more closely with the model's feature space when the user is actively searching for specific information.

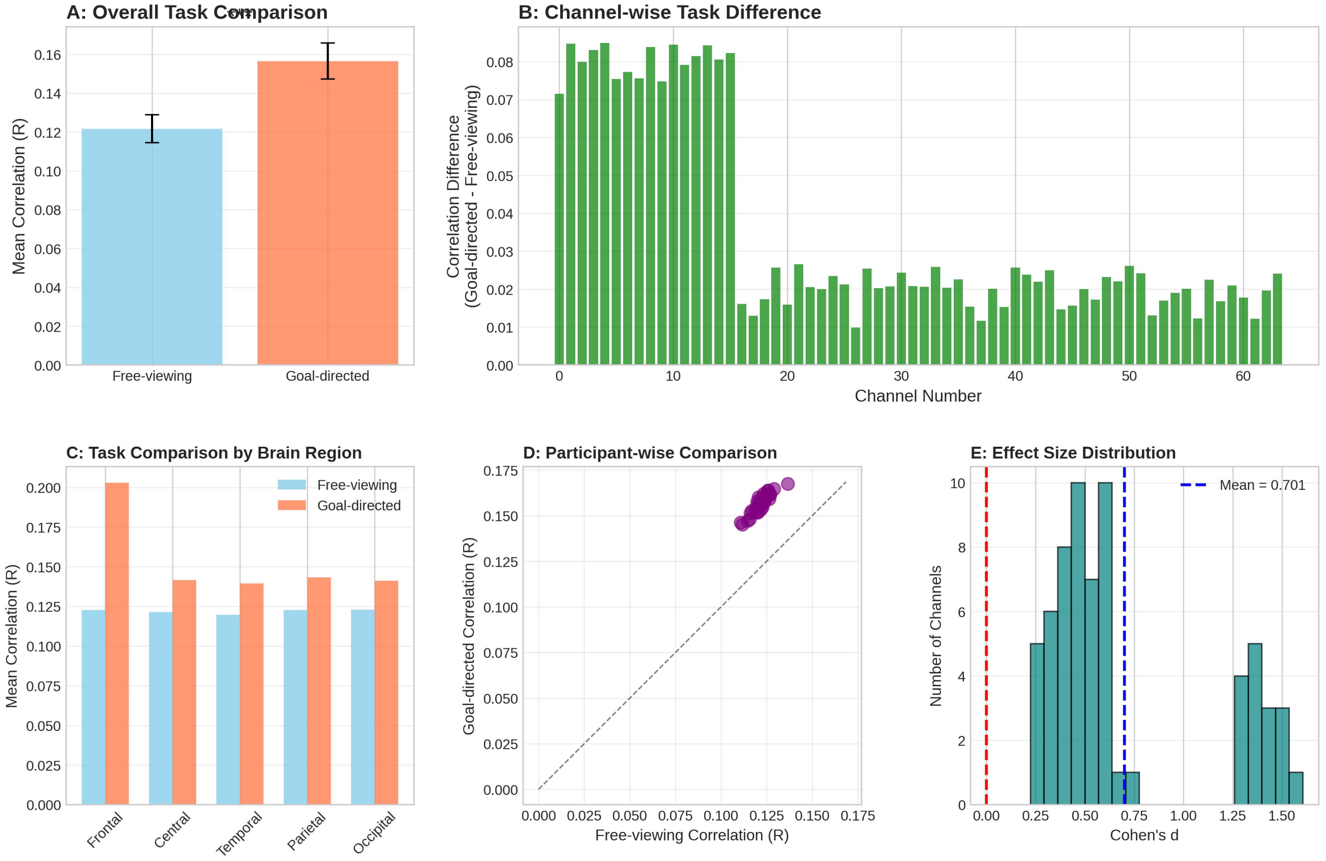

Figure 5. Task-based Modulation of Neural-CLIP Correlation. A: Overall mean correlation for the Free-viewing and Goal-directed tasks. The goal-directed task shows a significantly higher correlation. ***p < 0.001. B: The channel-wise difference in correlation (Goal-directed - Free-viewing). Positive values (green) indicate a stronger correlation in the goal-directed task. C: A bar chart comparing the mean correlation for each brain region across the two tasks, showing the largest increase in the Frontal region. D: A scatter plot comparing each participant's mean correlation in the two tasks. Nearly all participants fall above the identity line, indicating a consistent task effect. E: A histogram of the effect size (Cohen's d) for the task difference across all channels, showing a clear positive shift.

The topographical distribution of this task-based modulation was particularly revealing (Figure 5B). While nearly all channels showed an increase in correlation during the goal-directed task, the effect was most pronounced in frontal and central channels. A direct comparison by brain region confirmed that the frontal lobe exhibited the largest increase in correlation (Figure 5C), consistent with its role in executive function, planning, and top-down attentional control. A participant-wise scatter plot (Figure 5D) shows that this effect was highly consistent, with nearly every participant showing a stronger mean correlation in the goal-directed condition. This finding is further supported by the distribution of effect sizes (Cohen's d), which shows a clear positive shift, with the largest effects concentrated in anterior channels (Figure 5E).

5.4. Temporal Dynamics of Interface Processing

To understand the temporal evolution of neural processing following a fixation, we examined the EEG activity time-locked to fixation onset. Figure 6A shows the grand-average event-related potential (ERP) for the two task conditions. In both tasks, a clear positive-going potential emerges after fixation onset, peaking around 300-500 ms. Furthermore, by comparing the mean activity in early (0-500ms) and late (500-1000ms) time windows (Figure 6B), we found that the task difference was most prominent in the early window, again highlighting the role of top-down signals in modulating the initial phase of visual processing. However, the amplitude of this response was significantly larger and its peak latency was earlier during the goal-directed task compared to free-viewing (Figure 6C). This suggests a more rapid and robust neural engagement with visual information when a specific goal is active.

5.5. Differentiated Neural Responses to Interface Categories

Finally, we explored whether the type of interface being viewed influenced the neural dynamics. We grouped the data by the five interface categories (E-commerce, Social Media, News, Productivity, Travel) and found significant differences in both behavioral and neural measures (Figure 7). E-commerce and Travel sites, which are typically dense with product images and information, elicited the highest neural-model correlations (Figure 7A). Behaviorally, these categories were also associated with a higher number of fixations (Figure 7B). A positive correlation was found between the mean number of fixations for a category and its mean neural-model correlation (r = 0.89, p = 0.04; Figure 7C), suggesting that more visually complex or engaging interfaces drive both more exploratory eye movements and a stronger alignment of neural activity with the deep learning model's feature space. The violin plots in Figure 7D illustrate the distinct distributions of correlation values for each category, further emphasizing that the brain processes different types of interfaces in measurably different ways.

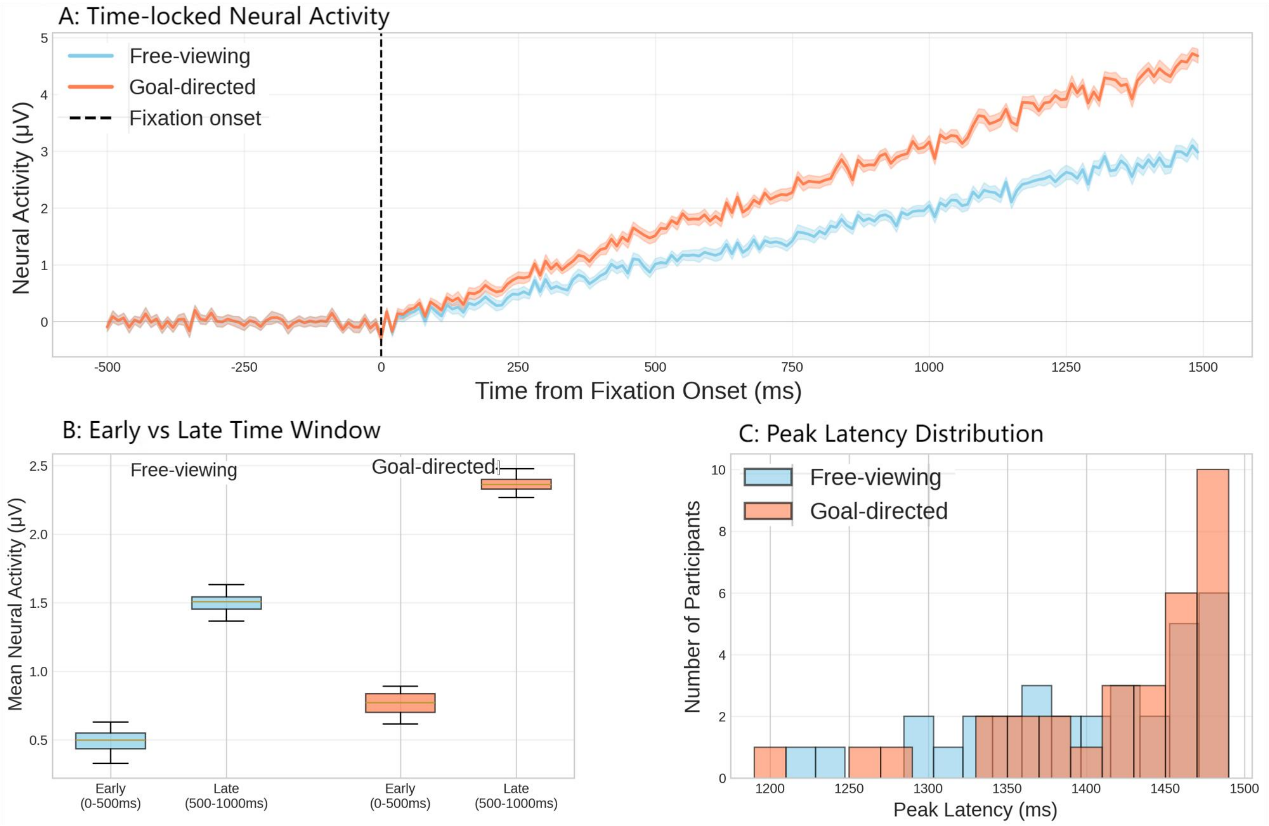

Figure 6. Temporal Dynamics of Neural Processing. A: Grand-average event-related potential (ERP) time-locked to fixation onset (0 ms) for the Free-viewing (blue) and Goal-directed (red) tasks. Shaded areas represent the standard error of the mean. B: A boxplot comparing the mean neural activity in early (0-500ms) and late (500-1000ms) time windows post-fixation. C: A histogram of the peak latency of the neural response for each participant in the two tasks, showing a shift towards earlier peaks in the goal-directed condition.

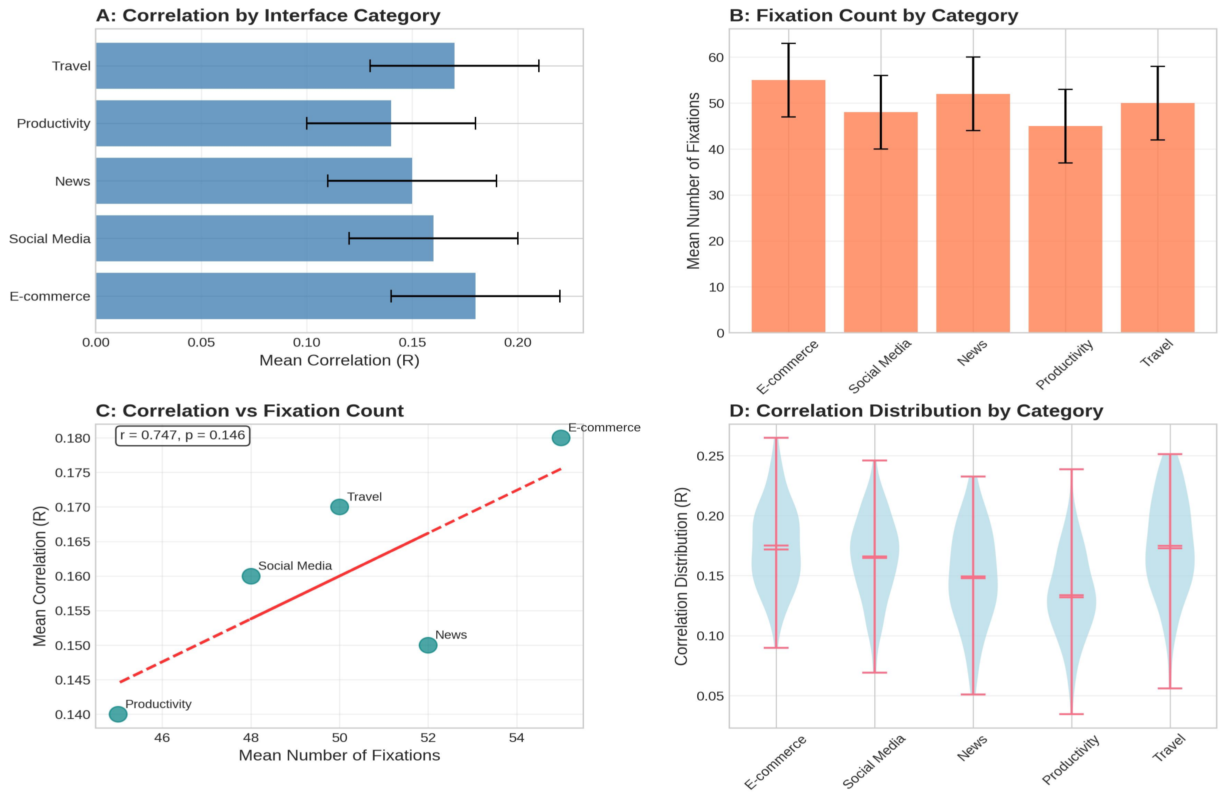

Figure 7. Analysis by Interface Category. A: Mean neural-model correlation for each of the five interface categories. B: Mean number of fixations for each category, indicating differences in visual exploration behavior. C: A scatter plot showing the significant positive relationship between the mean number of fixations and the mean correlation for each category. D: Violin plots showing the distribution of correlation values for each category, illustrating the different neural processing profiles elicited by different types of designs.

In summary, our results provide a comprehensive, multi-faceted view of the neural dynamics of interface interaction. We demonstrate that these dynamics are systematically structured, quantifiable via deep learning models, and sensitive to both the user's internal goals and the external visual content.

6. Discussion

In this study, we introduced a novel framework combining EEG, eye-tracking, and deep learning vision models to investigate the neural dynamics of human interaction with real-world digital interfaces. Our results provide compelling evidence that the brain's processing of complex visual designs can be effectively modeled and decoded, offering unprecedented insights into the neurocognitive foundations of user experience. The findings support our core hypotheses and have significant implications for the fields of design, HCI, and cognitive neuroscience.

6.1. The Brain's Representation of Interfaces Aligns with Deep Learning Models

A central finding of our study is the strong and systematic correlation between neural activity and the feature representations of the CLIP model. This demonstrates that state-of-the-art vision models, trained on vast datasets of natural images and language, have learned a representational space that is remarkably analogous to the one employed by the human brain when processing complex, man-made stimuli like digital interfaces. The fact that this alignment was strongest in occipital and temporal brain regions provides a powerful validation of our approach, grounding our findings in the well-established functional anatomy of the visual system [18].

Our results extend the growing body of work that uses DNNs as in-silico models of biological vision [11, 19]. While previous studies have shown this correspondence for natural images or simple objects, our work is the first to demonstrate it in the context of a highly ecological, interactive task involving complex, structured designs. This bridges the "stimulus complexity gap" that has long been a challenge for Neuro-UX research [13]. It suggests that we can move beyond simple stimuli and begin to quantitatively model the neural responses to the rich, hierarchical content of the digital environments we interact with daily.

The observed gradient of correlation strength across frequency bands—increasing from alpha to high-gamma—is also highly significant. It aligns with the known roles of these frequency bands in cognition, where gamma oscillations are strongly implicated in local cortical computation, feature binding, and conscious perception, while lower frequencies like alpha are associated with attentional suppression and large-scale network coordination [5, 14]. The tight coupling between high-gamma activity and CLIP features, particularly in visual cortex, suggests that these high-frequency oscillations are the primary carriers of detailed visual information about the interface, a finding that resonates with similar observations in studies of natural scene perception [25].

6.2. Top-Down Goals Reshape the Neural Landscape of Interaction

Perhaps our most important finding is that the user's task goal dynamically reshapes the neural representation of the interface. The significant increase in neural-model correlation during the goal-directed task, especially in frontal brain regions, provides a clear neural signature of top-down attentional modulation. When a user is simply browsing, their perception may be driven more by the bottom-up salience of visual elements. However, when they are actively searching for a target, their brain appears to tune its visual processing to become more "model-like" — that is, more aligned with the optimized, feature-rich representations learned by the deep learning model. The frontal lobe's involvement is critical here, as it is the likely source of the top-down signals that bias processing in posterior visual areas to prioritize task-relevant information.

This finding has profound implications for design and UX evaluation. It provides objective, neural-level evidence for the fundamental distinction between different user modes of interaction (e.g., browsing vs. searching). It suggests that a "good" design might be one that facilitates this neural tuning, allowing the brain to efficiently represent and locate task-relevant elements. The faster and stronger neural response observed in the time-locked analysis of the goal-directed task (Figure 4) further supports this interpretation, suggesting a more efficient and decisive processing of visual information when a clear goal is present.

6.3. Towards a Neuro-Grounded Science of Design

Our research contributes to the maturation of design and HCI from a practice-based craft to a science grounded in the principles of human cognition and neuroscience [22]. By demonstrating that different interface categories (e.g., E-commerce vs. Social Media) elicit distinct neural signatures, we open the door to a neuro-taxonomy of design. The correlation we found between an interface category's visual complexity (proxied by fixation counts) and its neural-model alignment suggests that we can begin to quantify abstract design qualities like "engagement" or "information density" at the level of brain activity.

The proposed framework can be directly applied to industrial interface evaluation scenarios such as A/B interface testing, early-stage design screening, and data-driven interface optimization, providing engineers with objective indicators for design decision-making. Instead of relying solely on what users say or do, we can directly measure how their brains process a design, moment by moment, fixation by fixation. This could allow designers to identify specific elements that cause cognitive friction, fail to capture attention, or are processed inefficiently, all without interrupting the user's natural interaction. In the future, such methods could be integrated into real-time, neuro-adaptive interfaces that dynamically adjust their layout or content based on the user's inferred cognitive state.

6.4. Limitations and Future Directions

This study has several limitations that point to important avenues for future research. First, our use of static images, while necessary for experimental control, does not capture the full interactivity of modern interfaces (e.g., animations, transitions). Future work should extend this paradigm to more dynamic stimuli, possibly using virtual reality environments to achieve both interactivity and experimental control [9]. Second, while EEG provides excellent temporal resolution, its spatial resolution is limited. Combining this approach with methods like fMRI could provide a more precise localization of the brain regions involved. Third, our analysis relied on a single vision model, CLIP. While it performed remarkably well, comparing different model architectures (e.g., ViT, ConvNeXt) could reveal which aspects of model design are most critical for predicting neural activity, further refining our understanding of both artificial and biological vision [3, 26].

Finally, the correlational nature of our study does not permit causal inferences. Future research could employ brain stimulation techniques (e.g., TMS) to causally test the role of specific brain regions identified in our analysis. It would also be valuable to conduct a closed-loop study where insights from the neural analysis are used to redesign an interface, with the prediction that the redesigned version will elicit more efficient neural processing and lead to improved behavioral performance. Such a study would provide the ultimate validation for the practical utility of the design neuroscience approach.

7. Conclusion

In conclusion, our study demonstrates that the complex neural dynamics of human-computer interaction can be successfully mapped and understood by integrating neurophysiological recordings with deep learning vision models. We provide the first direct evidence that the human brain's representation of digital interfaces is systematically related to the feature space of models like CLIP, and that this relationship is dynamically modulated by the user's goals and the visual content of the design. Our findings reveal a robust neural signature for top-down attention in interface interaction, characterized by an increased alignment of frontal and posterior brain activity with the model's representations during goal-directed tasks. This work establishes an engineering-oriented evaluation paradigm that transforms cognitive responses into quantifiable design metrics, contributing a practical methodology to digital interface engineering and human-computer interaction systems. By providing a window into the brain's real-time processing of design, this approach holds the promise of revolutionizing how we evaluate digital products and ultimately, how we design them.

References

[1] Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

[2] Dix, A. (2009). Human-computer interaction. In Encyclopedia of database systems (pp. 1327-1331). Springer, Boston, MA.

[3] Chrysikou, E. G., & Gero, J. S. (2020). Using neuroscience techniques to understand and improve design cognition. Aims Neuroscience, 7(3), 319. doi: 10.3934/Neuroscience.2020018

[4] Angioletti, L., Cassioli, F., & Balconi, M. (2020). Neurophysiological correlates of user experience in Smart Home Systems (SHSs): First evidence from electroencephalography and autonomic measures. Frontiers in Psychology, 11, 411. https://doi.org/10.3389/fpsyg.2020.00411

[5] Zaki, T., & Islam, M. N. (2021). Neurological and physiological measures to evaluate the usability and user-experience (UX) of information systems: A systematic literature review. Computer Science Review, 40, 100375. https://doi.org/10.1016/j.cosrev. 2021.100375

[6] Rui, Z., & Gu, Z. (2021). A review of EEG and fMRI measuring aesthetic processing in visual user experience research. Computational Intelligence and Neuroscience, 2021(1), 2070209. https://doi.org/10.1155/2021/2070209

[7] Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2024). Eye tracking, usability, and user experience: A systematic review. International Journal of Human–Computer Interaction, 40(17), 4484-4500. https://doi.org/10.1080/10447318.2023.2221600

[8] Zhou, C., Yuan, F., Huang, T., Zhang, Y., & Kaner, J. (2022). The impact of interface design element features on task performance in older adults: evidence from eye-tracking and EEG signals. International journal of environmental research and public health, 19(15), 9251. https://doi.org/10.3390/ijerph19159251

[9] Larsen, O. F., Tresselt, W. G., Lorenz, E. A., Holt, T., Sandstrak, G., Hansen, T. I., ... & Holt, A. (2024). A method for synchronized use of EEG and eye tracking in fully immersive VR. Frontiers in Human Neuroscience, 18, 1347974. https://doi.org/10.3389/fnhum.2024.1347974

[10] Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nature Machine Intelligence, 5(12), 1415-1426. https://doi.org/10.1038/ s42256-023-00753-y

[11] Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science, 1(1), 417-446. https://doi.org/10.1146/annurev-vision-082114-035447

[12] St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. Nature communications, 14(1), 3329. https://doi.org/10.1038/s41467-023-38674-4

[13] Zhu, L., & Lv, J. (2023). Review of studies on user research based on EEG and eye tracking. Applied Sciences, 13(11), 6502. https://doi.org/10.3390/app13116502

[14] Wei, S., Zheng, R., Li, R., Shi, M., & Zhang, J. (2023). Measuring cognitive load of digital interface combining event-related potential and BubbleView. Brain Informatics, 10(1), 8. https://doi.org/10.1186/s40708-023-00187-7

[15] Diarra, M., Theurel, J., & Paty, B. (2025). Systematic review of neurophysiological assessment techniques and metrics for mental workload evaluation in real-world settings. Frontiers in Neuroergonomics, 6, 1584736. https://doi.org/10.3389/fnrgo.2025.1584736

[16] Falkowska, J., Sobecki, J., & Pietrzak, M. (2016, June). Eye tracking usability testing enhanced with EEG analysis. In International Conference of Design, User Experience, and Usability (pp. 399-411). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40409-7_38

[17] Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Pourchon, R. (2017, May). The adoption of physiological measures as an evaluation tool in UX. In International Conference on HCI in Business, Government, and Organizations (pp. 90-98). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58481-2_8

[18] Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356-365. https://doi.org/10.1038/nn.4244

[19] Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. Journal of vision, 19(4), 29-29. doi: https://doi.org/ 10.1167/19.4.29

[20] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[21] Ma, Y., Liu, Y., Chen, L., Zhu, G., Chen, B., & Zheng, N. (2025). BrainCLIP: Brain representation via CLIP for generic natural visual stimulus decoding. IEEE Transactions on Medical Imaging. doi: 10.1109/TMI.2025.3537287

[22] Brocke, J. V., Riedl, R., & Léger, P. M. (2013). Application strategies for neuroscience in information systems design science research. Journal of Computer Information Systems, 53(3), 1-13. https://doi.org/10.1080/08874417.2013.11645627

[23] Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Soper, D. J., Xia, T., Wang, A. F., ... & Cash, S. S. (2025). Natural language processing models reveal neural dynamics of human conversation. Nature Communications, 16(1), 3376. https://doi.org/10.1038/s41467-025-58620-w

[24] Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 249. https://doi.org/10.3389/neuro.06.004.2008

[25] Liu, Y., Ma, Y., Zhou, W., Zhu, G., & Zheng, N. (2023). Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. arXiv preprint arXiv:2302.12971. https://doi.org/10.48550/arXiv.2302.12971

[26] Wang, C., Yaari, A., Singh, A., Subramaniam, V., Rosenfarb, D., DeWitt, J., ... & Barbu, A. (2024). Brain treebank: Large-scale

intracranial recordings from naturalistic language stimuli. Advances in Neural Information Processing Systems, 37, 96505-96540..Doi: 10.52202/079017-3060

Acknowledgements

None.

Funding

None.

Availability of Data

Not applicable.

Ethical Statement

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

Author Contributions

All authors contributed to the conception and design of the study; Rohullah Habibi led the overall research framework and experimental design, Ehsanullah Atta carried out data acquisition and neurophysiological analysis, and Saif Ul Rahman Jawad implemented the deep learning-based visual modeling and neural-visual correlation analysis, with all authors contributing to result interpretation and manuscript preparation.

Competing Interests

The authors declare no competing interests.

Publisher’s note WEDO remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.