ERP Markers of Visual and Semantic Processing in AI-Generated Images: From Perception to Meaning

Teodora Mitrevska
LMU Munich
Munich, Germany
teodora.mitrevska@ifi.lmu.de

Francesco Chiossi
LMU Munich
Munich, Germany
francesco.chiossi@um.ifi.lmu.de

Sven Mayer
LMU Munich
Munich, Germany
Munich Center for Machine Learning (MCML)
Munich, Germany
info@sven-mayer.com

Abstract

Perceptual similarity assessment plays an important role in processing visual information, which is often employed in Human-AI interaction tasks such as object recognition or content generation. It is important to understand how humans perceive and evaluate visual similarity to iteratively generate outputs that meet the users' expectations better and better. By leveraging physiological signals, systems can rely on users' EEG responses to support the similarity assessment process. We conducted a study (N=20), presenting diverse AI-generated images as stimuli and evaluating their semantic similarity to a target image while recording event-related potentials (ERPs). Our results show that the P2 and N400 component distinguishes medium, and high similarity of images, while the low similarity of images did not show a significant impact. Thus, we demonstrate that ERPs allow us to assess the users' perceived visual similarity to support rapid interactions with human-AI systems.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Generative AI, Human-AI Interaction, Physiological Computing, Event-Related Potentials, Semantic Processing

ACM Reference Format:
Teodora Mitrevska, Francesco Chiossi, and Sven Mayer. 2025. ERP Markers of Visual and Semantic Processing in AI-Generated Images: From Perception to Meaning. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3719907

1 Introduction

In generative AI systems, there is an increasing need to understand user intent and preferences to create personalized and relevant content [42]. This need is especially important in generative applications with complex stimuli, such as image generation [49] or visual recognition [17], where the AI must match outputs with user semantic expectations [51]. While today's models are proficient in generating photorealistic imagery, it remains a challenge to manipulate and embed semantic features of interest to the user.

Here, traditional feedback mechanisms in AI systems primarily rely on explicit user input, such as clicks, textual input, or even using other example images [19]. Although advanced, these methods still prove to be uncertain [24], inaccurate, and time-consuming [38, 51]. This issue is particularly challenging in tasks of categorization or evaluation, when users are asked to determine the relevance or similarity of complex visual content [25, 26]. Such explicit, subjective feedback can introduce biases and may not accurately represent the expected and implicit user perception. Thus, AI systems may struggle to accurately interpret and respond to user intent [2], limiting their support and productivity for generative applications based on users' feedback. Previous work explored multimodal input to inform AI outputs, particularly in systems requiring human interaction for decision-making or information retrieval [2]. Gwizdka et al. [21]) investigated the temporal dynamics of eye-tracking and EEG during reading and relevance decisions, highlighting that combining these inputs can effectively capture user attention and decision-making processes. In another study, Huang et al. [23] explored EEG correlates of visual perception using rapid serial visual presentation paradigms, showing that EEG signals can track relevance and attention across quickly changing visual stimuli. In our work, we explored new methods of capturing and utilizing implicit user feedback to enhance the adaptability of AI systems by employing ERPs to assess the perceived similarity of AI-generated images.

We conducted an experiment with 20 participants, where each participant was presented with a series of AI-generated images, including target images and deviants. The participants evaluated the similarity of these images to a predefined target image. During this process, we recorded the participants' Event-Related Potential (ERP) responses, focusing on the P2 and N400 components. The P2 was analyzed to examine early attentional and perceptual processes involved in comparing visual features of the images, while the N400 was investigated to explore potential semantic or conceptual processes underlying the evaluation of similarity. Our goal was to understand how these electrophysiological responses relate to image perception and to assess the feasibility of predicting perceived similarity based on these ERP components.

We demonstrate that the P2 and N400 component discriminates between medium and high levels of perceived semantic similarity.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI EA '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719907

This work set a foundation for employing ERP components as a metric for the assessment of semantic similarity between thematically diverse photorealistic images. Additionally, we give implications for future work in employing ERP components as a form of implicit feedback in classification tasks for AI systems.

2 Related Work

This section explores implicit interactions with generative AI. Further on, it examines electrophysiological correlates of semantic similarity as a reliable signal for implicit feedback in interactions with such systems.

2.1 Extended Interactions Beyond Textual Input

Results achieved from a collaboration with generative AI can be impressive. From different artistic styles to photorealistic images, text-to-image generation can be an extremely powerful co-creator for humans. However, in terms of communication, this is a process limited to typing on a keyboard [6, 8]. By taking part in it, the human is limited by the proficiency of their motor skills, their level of literacy, input control, and expressivity [14, 33, 37, 38]. A promising approach involves using EEG to capture neural responses to visual stimuli. This method integrates humans into the loop by translating natural brain activity into implicit feedback. Here, AI systems can rely on physiological correlates of users' states and adapt without explicit user input. This allows AI systems to adjust their outputs in real time, aligning with users' underlying cognitive and perceptual responses. Torre-Ortiz et al. [46] demonstrate that brain relevance feedback can effectively control a generative model, showing that implicit signals can guide the creation process. A further study by Torre-Ortiz et al. [47] reveals that images generated from brain feedback closely match the goals of study subjects and are comparable to those produced with manual feedback.

We investigate ERP components that reflect early perceptual processing and semantic similarity, proposing them as a method to evaluate the divergence between users' semantic expectations and the content of images. This approach could serve as implicit feedback for generative models, enabling these systems to align their outputs with user expectations better. Future applications may include real-time evaluation and input for generative tasks, supporting personalized content creation and recommendations.

2.2 Electrophysiological Correlates of Semantic Similarity

ERPs have been extensively utilized in human-computer interaction (HCI) research to examine how the brain processes visual stimuli [50]. These time-locked brain responses provide insights into how users perceive [12] and process visual stimuli [10], as well as how they identify similarity [22]. When comparing semantic similarity between images, the P2 and N400 components are expected to be elicited within paradigms involving visual stimulus processing.

The P2 component is typically observed around 200 ms after stimulus onset, is associated with early attentional processes and the detection of visual features. It reflects the initial selection and categorization of visual stimuli, indicating how quickly and efficiently the brain can identify relevant features in the visual field [13, 41]. P2 is linked to visual differences like color and spatial arrangements [7], detecting visual features rather than semantic [15], and not to vary with semantic content [40], reinforcing that P2 is focused on basic visual attributes.

The N400 component, typically occurring between 300-500 ms after stimulus onset, is a well-established marker of semantic processing and integration of meaning [9, 43]. In linguistic contexts, the N400 is elicited when a word is semantically incongruent with the preceding context, indicating a processing challenge or expectation violation [29]. This ERP component's relevance extends beyond verbal stimuli to non-verbal contexts, demonstrating sensitivity to semantic incongruities found in linguistic and visual stimuli [27]. The transition from linguistic to visual stimuli highlights the N400's general role in semantic processing. Nigam et al. [36] showed that N400 is consistently elicited by semantic anomalies, regardless of whether the stimuli are words or pictures, emphasizing its broad applicability in detecting semantic incongruence.

This suggests that the evaluation of AI-generated images might show similar ERP patterns, with greater similarity to the target image producing distinct ERP responses, particularly in the N400 component. While the N400 is attuned to the semantic content and relevance of stimuli, the P2 component, which is mainly involved in early visual processing, may not be influenced by how similar or different AI-generated images are from a target picture.

3 User Study

Prior research highlights the importance of incorporating user input into AI systems to improve outcomes, especially in co-creation tasks where human preferences shape the creative process [34, 35]. In this study, we focused on using electrophysiological signals as a means to evaluate users' visual and semantic processing of AI-generated images. We conducted an experiment with 20 participants, presenting them with AI-generated images comprising a target image alongside deviant and similar stimuli. Participants evaluated the similarity of each stimulus to the target image using a slider scale. Drawing from previous work, we formulate the following research question:

RQ: Can ERP responses reliably reflect the perceived semantic of AI-generated images to a target one?

3.1 Stimuli Generation

The stimulus sets were generated using Midjourney (Version 6.1). The generated stimuli consist of everyday scenes, incorporating different objects, food, animals, etc. In total, we generated 220 images, which allowed us to create 20 sets. Every set consists of one target image and 10 stimuli images, including instances with different similarity levels to the target image, one instance of the target image itself, and two unrelated instances. Each set of images followed a different theme, such as a tennis court, garden, and living room. We used multiple semantic topics to enhance the richness and diversity of the image stimuli, allowing for more robust ERP responses [32]. Lu et al. [32] suggests that diverse semantic information allows for stable ERP amplitudes and accurately reflects real-world scenarios where multiple, often complex, factors influence attention allocation and decision-making. Similarly, Ullsperger and Grune

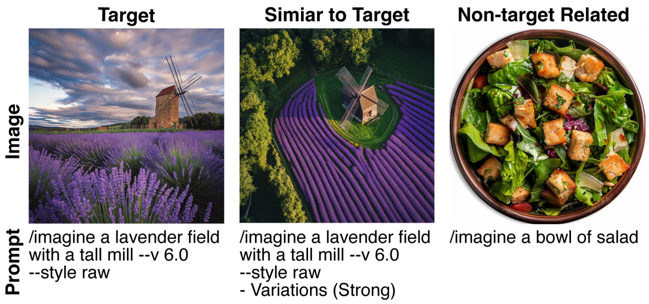


Figure 1: Generation of stimuli. A showcase of the stimuli generation process, depicting the prompt (right) used to generate an image (left). Left: the first image represents a target image, whereas the second one is a stimuli image. Middle: non-target image but semantically related to the target image. To ensure similarity, we used a prompt to create a strong variation of the target image. Right: a non-target related image, semantically unrelated to the target image.

[48] demonstrated that the use of multi-dimensional stimuli increases ERP amplitude, highlighting the role of complex semantic content in engaging relevance processing. We illustrated the stimuli generation in Figure 1.

3.2 Apparatus

We built the study application using PsychoPy v2024.2.1. We used a VIEWPixx Liquid Crystal Display (LCD) with a refresh rate of 120 Hz and a 1920 × 1080 pixels resolution of 23.6 inches. We used an EyeLink 1000 Plus to track the user's gaze behavior, measuring 60cm of the eye-to-screen distance. The images were displayed centrally on the screen, at 2.69° by 2.65° area of visual angle. EEG data were collected using 64 Ag-AgCl pin-type passive electrodes embedded in a water-based EEG cap (R-Net, BrainProducts GmbH, Germany). Electrodes were positioned according to the 10–20 system at the following locations: Fp1, Fz, F3, F7, F9, FC5, FC1, C3, T7, CP5, CP1, Pz, P3, P7, P9, O1, Oz, O2, P10, P8, P4, CP2, CP6, T8, C4, Cz, FC2, FC6, F10, F8, F4, Fp2, AF7, AF3, AFz, F1, F5, FT7, FC3, C1, C5, TP7, CP3, P1, P5, PO7, PO3, Iz, POz, PO4, PO8, P6, P2, CPz, CP4, TP8, C6, C2, FC4, FT8, F6, F2, AF4, and AF8. EEG signals were acquired using two LiveAmp amplifiers at a sampling rate of 500 Hz. Electrode impedances were maintained at or below 20 kΩ. FCz was used as the online reference, while AFz was the ground electrode. We employed the Lab Streaming Layer (LSL) framework to synchronize and record physiological data and experiment events.

3.3 Procedure

Upon arrival, we briefed the participants about the study and asked them to provide informed written consent. We asked them to fill out a questionnaire consisting of demographic information and AI literacy questionnaire [5]. In the demographics, we noted the participants' gender, age, level of education and employment status. We explained the protocol in detail and presented participants with a paper-based example of the task they would do later. We asked the participants to wear the EEG headset and sit on a chair in a comfortable position in front of a monitor, providing them with a mouse and keyboard as a feedback input method. We ensured the ergonomic comfort of the participants by adjusting the table, chair, or headrest height to accommodate each individual. Each participant placed their head on a head and chin rest and sat 60 cm from the screen. As next, we calibrated the eye tracker and began the experiment. In total, participants rated the similarity of 200 image pairs in succession. This was a recurring process, where after one iteration (1 target image, ten image stimuli), a new target image was introduced, followed by a new set of 10 image stimuli, and we randomized the order of the target images. The participants were shown a target image, followed by a test image stimuli. For each image stimulus, we asked participants to assess the similarity to the target image via the following statement: "This image has a very high similarity to the target image." As an input method, we provided a 101-point slider, ranging end values from strongly disagree(lowest similarity) to strongly agree(highest similarity). Participants could take a break after each similarity rating and continue when ready.

3.3.1 Trial Structure. At the beginning of the task, we present a target image to the participants. The duration of the presentation was relative, as the participants could proceed to the next step when they were feeling ready by pressing "space." In the next step, the participants are presented with a black fixation cross (+) with a given duration of 1000ms, positioned centrally on the screen. The next step presented an inter-stimulus interval (ISI) of 1000ms with an additional random offset of 250ms, 500ms, or 750ms to reset the neural and attentional reserve and avoid fatigue effects [52]. For 3000ms, an image stimulus is presented to the participants. Finally, we presented them with a slider to rate the similarity; once a value was chosen, they moved to the potential break. An optional break was offered to the participants, which they can take or continue with the task with a button press [1]. After each break, we re-calibrated the eye tracker to ensure high-quality data.

3.4 Participants

We recruited 27 volunteers through the institutional email list and convenience sampling methods. We excluded eleven participants from the analysis due to unsatisfactory EEG data quality, as identified by RANSAC, which revealed that the electrodes of interest compromised the data reliability for those participants. This totaled 16 participants (9 females, 7 males, none diverse). 12 participants were aged 18-24, and 4 of them were aged 25-34, and one in the age range 45-54. Additionally, we surveyed the participants using the Meta AI Literacy Questionnaire [5]. The participants reported an overall score (M = 5.6, SD = 1.5), encompassing the conditions AI literacy (M = 6.3, SD = 1.4), Create AI (M = 2.6, SD = 3.0), AI Self Efficacy (M = 4.8, SD = 2.1), AI Self Competency (M = 6.4, SD = 1.7).

Exclusion criteria for the recruitment included a medical history of psychological or neurological disorders, color blindness, and visual impairments. The local ethics committee approved the study, qualifying as fast-track approval since the participants were not subjected to any risk (e.g., deception, stress beyond normal levels, recording of sensitive information). The study averaged 1.5 hours, which we compensated with 12 EUR per hour.

3.5 EEG Preprocessing & ERP Analysis

We first automatically detected bad or outlier channels via the random sample consensus (RANSAC) method [3] using spherical splines for estimating scalp potential based on algorithms proposed by Perrin [39]. We then applied a notch filter at 50 Hz and band-passed the signal between 1–15 Hz to remove high and low-frequency noise. The signal was re-referenced to the common average reference (CAR). Independent Component Analysis (ICA) for artifact detection and correction was performed using the extended Infomax algorithm [30]. ICA components' labeling and rejection process was automated using the MNE plugin "ICLabel" [31]. We segmented continuous signals from 200 ms before to 1000 ms after image display onset, using a 200 ms pre-stimulus baseline for correction. The P2 component was quantified as the average positive peak amplitude in the 150–275 ms window. This analysis focused on electrodes Fp1, Fp2, F7, and F8, which are known to be sensitive to early perceptual processing in the frontal regions, as supported by previous research [41]. The N400 component was measured as the average negative peak amplitude in the 350–550 ms window, centered around the typical N400 peak latency observed in the grand average waveforms. For the N400 analysis, electrodes Fz, Cz, C3, C4, F3, and F4 were used [20, 43].

4 Results

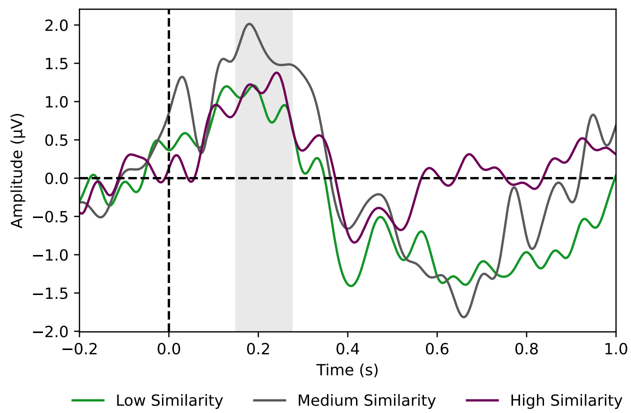
In the following, we present the results of our evaluation. We employed a Linear Mixed Model (LMM) to investigate P2 and N400 peak amplitude differences.

4.1 P2 Amplitude

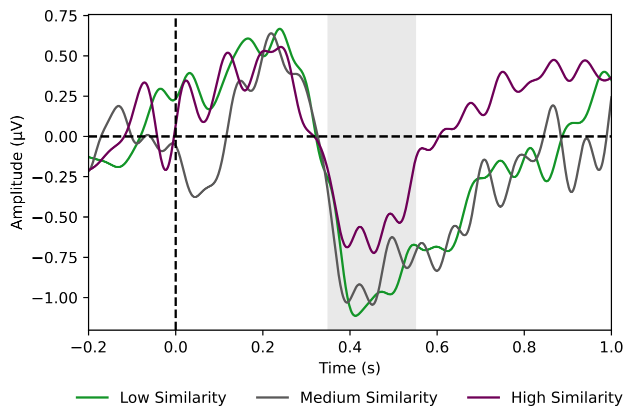
We fitted a linear mixed to predict the P2 amplitude based on Semantic Similarity. The total explanatory power of the model was substantial, with a conditional R^2 of .48, indicating that the model explained 48% of the variance in peak amplitude. The variance explained by the fixed effects alone was moderate, with a marginal R^2 of .10. The intercept, corresponding to the High condition, was statistically significant at 4.18 (95% CI $[2.17, 6.18]$, $t(43) = 4.21$, $p < .001$). The effect of the Medium condition was statistically significant and positive, with $\beta = 2.23$, 95% CI $[.08, 4.39]$, $t(43) = 2.09$, $p = .042$. The standardized effect size was $\beta_{\text{std}} = .55$, 95% CI $[.02, 1.07]$, indicating a moderate positive impact on amplitude. The effect of the Low condition was statistically non-significant and negative, with $\beta = -.81$, 95% CI $[-2.97, 1.34]$, $t(43) = -.76$, $p = .450$. The standardized effect size was $\beta_{\text{std}} = -.20$, 95% CI $[-.72, .33]$, suggesting a small and non-significant impact on amplitude.

4.2 N400 Amplitude

The total explanatory power of the model was substantial, with a conditional R^2 of .42, indicating that the model explained 42% of the variance in peak amplitude. The variance explained by the fixed effects alone was moderate, with a marginal R^2 of .09. The intercept, corresponding to the High Similarity condition, was statistically significant at -3.09 (95% CI $[-3.80, -2.37]$, $t(43) = -8.71$, $p < .001$). The effect of the Medium Similarity condition was statistically significant and negative, with $\beta = -1.06$, 95% CI $[-1.87, -.26]$, $t(43) = -2.67$, $p = .011$. The standardized effect size was $\beta_{\text{std}} = -.73$, 95% CI $[-1.28, -.18]$, indicating a moderate-to-strong negative impact on Amplitude. The effect of the Low Similarity condition was statistically non-significant and negative, with $\beta = -.42$, 95% CI $[-1.23, .38]$, $t(43) = -1.06$, $p = .295$. The standardized effect size was $\beta_{\text{std}} = -.29$, 95% CI $[-.84, .26]$, suggesting a small and non-significant impact on amplitude.



(a) P2 Grand Average

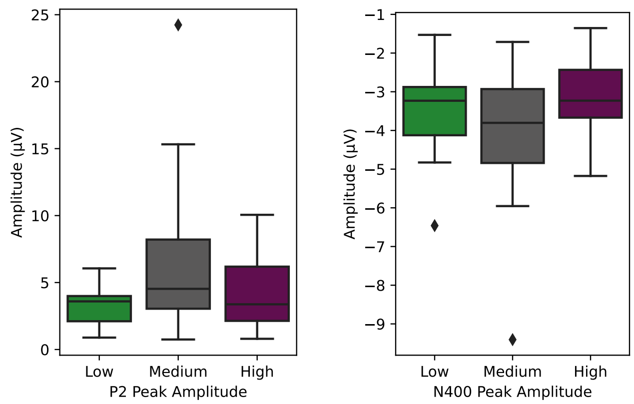


(b) N400 Grand Average

Figure 2: Grand Average event-locked to image display onset. Data reflect the results obtained from frontocentral ROI for each Image Similarity condition. (a) The P2 component (150-275 ms) shows a statistically significant differences between Medium and High. (b) the N400 component (350-550 ms) exhibits significant modulation, with greater positivity for high similarity conditions in AI-generated images.

5 Discussion

We explored how the human brain processes AI-generated images at different levels of semantic similarity, providing the groundwork for the development of implicit human-AI interaction systems. ERP responses reveal distinct neural mechanisms that underlie both perceptual and semantic processing of artificial visual content. These findings significantly impact user experience with generative AI systems and developing more effective implicit feedback mechanisms.



(a) P2 Peak Amplitude **(b) N400 peak Amplitude**

Figure 3: (a) P2 Peak Amplitude and (b) N400 Peak Amplitude. For both, we found significant differences between the Medium similarity images and the High similarity images.

5.1 P2 and N400 Components Reveal Distinct Processing Stages in AI-Generated Image Perception

The increasing use of AI-generated imagery requires a better understanding of how humans process these artificial outputs. We examined whether ERPs, particularly the P2 and N400 components, could serve as implicit markers of semantic similarity perception in AI-generated images. The results confirmed a dissociation between early perceptual processing (P2) and later semantic integration (N400).

Regarding the P2, which underlies early visual attention and feature detection, revealed an interesting pattern across semantic similarity conditions. We found no differences between high and low similarity, but we found an increased positivtiy for medisum similarity images. This stability across high and low conditions confirms that early perceptual processing operates independently of semantic meaning, with the brain similarly processing visually dissimilar content at this early stage (150-250ms post-stimulus). The increase for the Mediun similarity conditition suggests these ambiguous images may recruit additional attentional resources during early processing, possibly reflecting the brain's attempt to resolve perceptual uncertainty when confronted with partially matching features. This U-shaped response pattern aligns with predictive coding theories, where completely novel stimuli are quickly categorized as "different," without requiring additional processing. Overall, the relative stability of P2 amplitudes across conditions supports previous research demonstrating that P2 primarily responds to physical features rather than semantic relationships [7, 15, 40, 41].

This finding aligns with predictive coding theories, which propose that the brain continuously generates and updates predictions about incoming stimuli. When a stimulus closely matches an expected category (high similarity) or is clearly incongruent (low similarity), prediction error is minimized, requiring less processing effort. However, medium-similarity images create greater uncertainty, leading to increased prediction error and, consequently, a stronger N400 response as the brain engages in additional processing to resolve the ambiguity. This is consistent with research showing that ambiguous or partially matching stimuli require greater cognitive effort for semantic integration, as the brain must reconcile competing interpretations before reaching a stable representation [18, 28, 45].

5.2 ERP as Future Implicit Feedback Mechanisms for Generative AI Systems

The integration of ERPs into generative AI systems opens possibilities towards implicit and diverse feedback. Previous work that explored the use of ERPs as a form of implicit feedback Studies done on a certain type of stimuli report

The P2 and N400 components provide insights into perceptual and semantic processing, capturing the user's immediate reaction based on which they form their opinion [47]. These findings pose them as promising real-time indicators of user engagement or comprehension when interacting with AI-generated outputs [16, 46, 47]. The increased N400 response can be employed to detect semantic incongruities in AI outputs, while altered P2 amplitudes could reflect perceptual mismatches. This dissociation allows us to track at which stage potential issues with AI-generated content arise - whether at basic visual perception or deeper semantic understanding. Further on, these findings suggest that our brain processes AI images similarly to natural images, with distinct stages for visual features and meaning. This validates that humans can meaningfully engage with AI-generated content while giving a method to measure on how they process such content.

5.3 Limitations & Future Work

While our study provides first insights into utilizing ERPs as an implicit technique for assessing semantic similarity in AI-generated images, we address several limitations that should be addressed in future work. Firstly, our analysis was conducted entirely offline, meaning we only processed the data after all trials had been collected. This limits the applicability of our findings to real-time systems, which needs to be explored. Our work provides a stepping stone for future research that extends to the development of systems that leverage machine learning models that could leverage neurophysiological data to detect semantic similarity or dissimilarity between images in real time. The trial-based setup of this study also fails to fully capture a natural interaction between humans and, for e.g., an AI-driven image generator. For instance, an AI system that is fully capable of implicitly interpreting perceptual visual similarity could create a path for studying iterative interactions that involve real-time feedback. Future work should explore how such interactions influence ERP responses and how neural data could guide AI systems to adapt or generate semantically meaningful content in real time [4, 11, 44].

Second, although we conducted an experiment where diverse visual stimuli were presented, we did not include images with faces. Faces are strongly associated with the N170 ERP component, which reflects early face-sensitive perceptual processes; however, we did not inspect them in this study. The lack of facial stimuli limits the generalization of our findings for scenarios that are relevant to the assessment of visuals that include faces. Future studies should include face stimuli to explore how facial features might interact with semantic similarity processing, as this could reveal different neural activity. Additionally, incorporating a comparison with non-AI-generated stimuli could have resulted in differences in perception or confirmed the reliability of our results. Last but not least, incorporating eye-tracking data into the experimental design could provide insights on how visual attention contributes to semantic similarity assessments. Eye-tracking could help identify specific visual features (e.g., regions of interest) that create differences in ERP responses, such as the N400. By combining eye-tracking with EEG, future research could achieve a more detailed understanding of the interaction between visual attention and semantic processing.

6 Conclusion

Our findings demonstrate that ERPs can serve as reliable markers of how humans process and evaluate AI-generated images, with P2 and N400 components reflecting distinct aspects of visual and semantic processing. P2 and N400 responses indicate perceptual processing across Medium and High similarity, providing a potential implicit feedback mechanism for AI systems. This work establishes a foundation for integrating electrophysiological measures into human-AI interaction, offering a path toward more intuitive and user-aligned content generation.

7 Open Science

We encourage readers to reproduce and expand upon our findings. To facilitate this, our experimental setup, datasets, and analysis scripts are openly accessible on the Open Science Framework at the following link: https://osf.io/b89f6/.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161.

References

[1] Atsunori Ariga and Alejandro Lleras. 2011. Brief and rare mental "breaks" keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition* 118, 3 (2011), 439–443. doi:10.1016/j.cognition.2010.12.007

[2] Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. 2023. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet* 15, 8 (2023), 260. doi:10.3390/fi15080260

[3] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. 2015. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in neuroinformatics* 9 (2015), 16. doi:10.3389/fninf.2015.00016

[4] Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. 2001. Classifying Single Trial EEG: Towards Brain Computer Interfacing. In *Advances in Neural Information Processing Systems*, Vol. 14. MIT Press, Cambridge, MA, USA. https://proceedings.neurips.cc/paper_files/paper/2001/file/2d579dc29360d8bbfbb4aa541de5afa9-Paper.pdf

[5] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100014. doi:10.1016/j.chbah.2023.100014

[6] Cheng Chen, Sangwook Lee, Eunchae Jang, and S. Shyam Sundar. 2024. Is Your Prompt Detailed Enough? Exploring the Effects of Prompt Coaching on Users' Perceptions, Engagement, and Trust in Text-to-Image Generative AI Tools. In *TAS '24* (Austin, TX, USA). Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. doi:10.1145/3686038.3686060

[7] Qingfei Chen, Xiuling Liang, Peng Li, Chun Ye, Fuhong Li, Yi Lei, and Hong Li. 2015. The processing of perceptual similarity with different features or spatial relations as revealed by P2/P300 amplitude. *International Journal of Psychophysiology* 95, 3 (2015), 379–387. doi:10.1016/j.ijpsycho.2015.01.009

[8] Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. 2024. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA, 7727–7736. doi:10.48550/arXiv.2310.08129

[9] Francesco Chiossi. 2017. Developing a N400 Brain Computer Interface Based on Semantic Expectancy. In *Proceedings of the Doctoral Consortium, Posters and Demos at CHItaly 2017 (DCPD@CHItaly 2017)*. CEUR Workshop Proceedings, Aachen, Germany, 66–77.

[10] Francesco Chiossi, Uwe Gruenefeld, Baosheng James Hou, Joshua Newn, Changkun Ou, Rulu Liao, Robin Welsch, and Sven Mayer. 2024. Understanding the Impact of the Reality-Virtuality Continuum on Visual Search Using Fixation-Related Potentials and Eye Tracking Features. *Proc. ACM Hum.-Comput. Interact.* 8, MHCI, Article 281 (Sept. 2024), 33 pages. doi:10.1145/3676528

[11] Francesco Chiossi, Changkun Ou, Felix Putze, and Sven Mayer. 2024. Detecting Internal and External Attention in Virtual Reality: A Comparative Analysis of EEG Classification Methods. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*. Association for Computing Machinery, New York, NY, USA, 381–395. doi:10.1145/3701571.3701579

[12] Francesco Chiossi, Ines Trautmannsheimer, Changkun Ou, Uwe Gruenefeld, and Sven Mayer. 2024. Searching Across Realities: Investigating ERPs and Eye-Tracking Correlates of Visual Search in Mixed Reality. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (2024), 6997–7007. doi:10.1109/TVCG.2024.3456172

[13] Francesco Chiossi, Yannick Weiss, Thomas Steinbrecher, Christian Mai, and Thomas Kosch. 2024. Mind the Visual Discomfort: Assessing Event-Related Potentials as Indicators for Visual Strain in Head-Mounted Displays. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, New York, NY, USA, 150–159. doi:10.1109/ISMAR62088.2024.00029

[14] Neo Chung. 2021. Human in the Loop for Machine Creativity. *9th AAAI Conference on Human Computation and Crowdsourcing* (10 2021). doi:10.48550/arXiv.2110.03569

[15] Julien Dampuré, Abdelmoumin Benraiss, and Nicolas Vibert. 2015. Task-dependent modulation of word processing mechanisms during modified visual search tasks. *Quarterly Journal of Experimental Psychology* 69, 6 (2015), 1145–1163. doi:10.1080/17470218.2015.107088

[16] Carlos de la Torre-Ortiz and Tuukka Ruotsalo. 2024. Perceptual Visual Similarity from EEG: Prediction and Image Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 11146–11155. doi:10.1145/3664647.3685508

[17] Carlos de la Torre-Ortiz, Michiel Spapé, and Tuukka Ruotsalo. 2023. The P3 indexes the distance between perceived and target image. *Psychophysiology* 60, 5 (March 2023), e14225. doi:10.1111/psyp.14225

[18] Karl Friston. 2005. A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, 1456 (2005), 815–836. doi:10.1098/rstb.2005.1622

[19] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA, 2414–2423. doi:10.1109/CVPR.2016.265

[20] Gisela Govaart, Antonio Schettino, Saskia Helbling, David M. A. Mehler, William Xiang Quan Ngiam, David Moreau, Francesco Chiossi, Anthony Zanesco, Fang Yu, Johannes Algermissen, Marta Topor, Jose Pérez-Navarro, Greta Häberle, Melanie S. Schreiner, Katrin Rothmaler, Guiomar Niso, and Mariella Paul. 2022. EEG ERP Preregistration Template. *MetaArXiv* (2022). doi:10.31222/osf.io/4nvpt Preprint.

[21] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. *Journal of the Association for Information Science and Technology* 68, 10 (2017), 2299–2312. doi:10.1002/asi.23904

[22] Jeff P Hamm, Blake W Johnson, and Ian J Kirk. 2002. Comparison of the N300 and N400 ERPs to picture stimuli in congruent and incongruent contexts. *Clinical neurophysiology* 113, 8 (2002), 1339–1350. doi:10.1016/S1388-2457(02)00161-X

[23] Yonghong Huang, Kenneth E. Hild, Misha Pavel, Santosh Mathan, and Deniz Erdogmus. 2012. Neural correlates of visual perception in rapid serial visual presentation paradigms. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, New York, NY, USA, 1–6. doi:10.1109/MLSP.2012.6349766

[24] Kevin Jasberg and Sergej Sizov. 2020. Human uncertainty in explicit user feedback and its impact on the comparative evaluations of accurate prediction and personalisation. *Behaviour & Information Technology* 39, 5 (2020), 544–577. doi:10.1080/0144929X.2019.1604804

[25] Deok-Hwan Kim, Chin-Wan Chung, and Kobus Barnard. 2005. Relevance feedback using adaptive clustering for image similarity retrieval. *Journal of Systems and Software* 78, 1 (2005), 9–23. doi:10.1016/j.jss.2005.02.005

[26] Yan Kong, Weiming Dong, Xing Mei, Chongyang Ma, Tong-Yee Lee, Siwei Lyu, Feiyue Huang, and Xiaopeng Zhang. 2016. Measuring and Predicting Visual Importance of Similar Objects. *IEEE Transactions on Visualization and Computer Graphics* 22, 12 (2016), 2564–2578. doi:10.1109/TVCG.2016.2515614

[27] Marta Kutas and Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* 62 (2011), 621–647. doi:10.1146/annurev.psych.093008.131123

[28] Marta Kutas and Kara D. Federmeier. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology* 62 (2011), 621–647. doi:10.1146/annurev.psych.093008.131123

[29] Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 4427 (1980), 203–205. doi:10.1126/science.7350657

[30] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. 1999. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation* 11, 2 (1999), 417–441. doi:10.1162/089976699300016719

[31] Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. 2022. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software* 7, 76 (2022), 4484. doi:10.21105/joss.04484

[32] Zhaohua Lu, Qi Li, Ning Gao, and Jingjing Yang. 2020. Time-varying networks of ERPs in P300-speller paradigms based on spatially and semantically congruent audiovisual bimodality. *Journal of Neural Engineering* 17, 4 (2020), 046015. doi:10.1088/1741-2552/aba07f

[33] Teodora Mitrevska and Sven Mayer. 2024. Physiological Signals as Implicit Multimodal Input in Human-AI Interactions During Creative Tasks. In *GenAICHI: CHI 2024 Workshop on Generative AI and HCI* (2024-01-01). 3. https://sven-mayer.com/wp-content/uploads/2024/07/mitrevska2024physiological.pdf

[34] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Manning.

[35] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (01 Apr 2023), 3005–3054. doi:10.1007/s10462-022-10246-w

[36] Ashok Nigam, James E. Hoffman, and Robert F. Simons. 1992. N400 to semantic anomalies in pictures and words. *Journal of Cognitive Neuroscience* 4, 1 (1992), 15–22.

[37] Jonas Oppenlaender. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* (2023), 1–14. doi:10.1080/0144929X.2023.2286532

[38] Jonas Oppenlaender, Hannah Johnston, Johanna Silvennoinen, and Helena Barranha. 2024. Artworks Reimagined: Exploring Human-AI Co-Creation through Body Prompting. *arXiv preprint arXiv:2408.05476* (2024). doi:10.48550/arXiv.2408.05476

[39] François Perrin, Jacques Pernier, Olivier Bertrand, and Jean Francois Echallier. 1989. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology* 72, 2 (1989), 184–187. doi:10.1016/0013-4694(89)90180-6

[40] Svetlana Pinet and Nazbanou Nozari. 2023. Different electrophysiological signatures of similarity-induced and Stroop-like interference in language production. *Journal of Cognitive Neuroscience* 35, 8 (2023), 1329–1349. doi:10.1162/jocn_a_02014

[41] Geoffrey F Potts. 2004. An ERP index of task relevance evaluation of visual stimuli. *Brain and cognition* 56, 1 (2004), 5–13. doi:10.1016/j.bandc.2004.03.006

[42] Hamed Nabizadeh Rafsanjani and Amir Hossein Nabizadeh. 2023. Towards human-centered artificial intelligence (AI) in architecture, engineering, and construction (AEC) industry. *Computers in Human Behavior Reports* (2023), 100319. doi:10.1016/j.chbr.2023.100319

[43] Anđela Šoškić, Vojislav Jovanović, Suzy J Styles, Emily S Kappenman, and Vanja Ković. 2022. How to do better N400 studies: reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review* 32, 3 (2022), 577–600. doi:10.1007/s11065-021-09513-4

[44] Andrew X Stewart, Antje Nuthmann, and Guido Sanguinetti. 2014. Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of neuroscience methods* 228 (2014), 1–14. doi:10.1016/j.jneumeth.2014.02.014

[45] Christopher Summerfield and Floris P. de Lange. 2014. Expectation in Perceptual Decision Making: Neural and Computational Mechanisms. *Nature Reviews Neuroscience* 15, 11 (2014), 745–756. doi:10.1038/nrn3838

[46] Carlos de la Torre-Ortiz, Michiel M. Spapé, Lauri Kangassalo, and Tuukka Ruotsalo. 2020. Brain Relevance Feedback for Interactive Image Generation. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1060–1070. doi:10.1145/3379337.3415821

[47] Carlos de la Torre-Ortiz, Michiel M. Spapé, Niklas Ravaja, and Tuukka Ruotsalo. 2024. Cross-Subject EEG Feedback for Implicit Image Generation. *IEEE Transactions on Cybernetics* (2024), 1–0. doi:10.1109/TCYB.2024.3406159

[48] Peter Ullsperger and Kerstin Grune. 1995. Processing of multi-dimensional stimuli: P300 component of the event-related brain potential during mental comparison of compound digits. *Biological psychology* 40, 1-2 (1995), 17–31. doi:10.1016/0301-0511(95)05110-4

[49] Kailas Vodrahalli and James Zou. 2023. Artwhisperer: A dataset for characterizing human-ai interactions in artistic creations. *arXiv preprint arXiv:2306.08141* (2023), 31. doi:10.48550/arXiv.2306.08141

[50] Philipp Wolf and Timo Götzelmann. 2023. A Systematic Review of Interaction Approaches based on Visually Evoked Potentials. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '23)*. Association for Computing Machinery, New York, NY, USA, 396–401. doi:10.1145/3594806.3594862

[51] Melvin Wong, Yew-Soon Ong, Abhishek Gupta, Kavitesh Kumar Bali, and Caishun Chen. 2023. Prompt Evolution for Generative AI: A Classifier-Guided Approach. In *2023 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, New York, NY, USA, 226–229. doi:10.1109/CAI54212.2023.00105

[52] Geoffrey F Woodman and Steven J Luck. 1999. Electrophysiological measurement of rapid shifts of attention during visual search. *Nature* 400, 6747 (1999), 867–869. doi:10.1038/23698