bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Temporal misalignment in scene perception: Divergent representations of locomotive action affordances in human brain responses and DNNs

Clemens G. Bartnik[1*], Evelyne I.C. Fraats[2], Iris I. A. Groen [1,2]

[1] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[2] Psychology Research Institute, University of Amsterdam, Amsterdam, The Netherlands
[*] c.g.bartnik@uva.nl

Abstract

The human visual system processes scenes with remarkable speed, enabling the extraction of essential information to navigate our surroundings in a single glance. To elucidate how the brain transforms visual inputs into neural representations of navigationally relevant information, we collected electroencephalography (EEG) responses to diverse indoor and outdoor scenes along with behavioral annotations of locomotive action affordances (e.g., walking, cycling), object annotations, and low-level image features to model distinct types of scene information. Using representational similarity analysis, we examined the neural representation of locomotive action affordances over time, their co-localization within scene-selective cortex, and their computational alignment with deep neural networks (DNNs). Our results show that locomotive action affordance representations emerge within 200 ms of visual processing, showing unique contributions to EEG responses at temporally distinct time-points from objects and low-level properties. Spatiotemporal fusion with functional magnetic resonance imaging (fMRI) recordings in scene-selective brain regions reveals that both the parahippocampal (PPA) and occipital place region (OPA), but not the medial place region (MPA), contribute to locomotive action affordance representations, with a distinct temporal hierarchy between them. While DNNs exhibit good predictivity of early EEG responses, they primarily capture low-level features and show limited alignment with affordance processing. These findings reveal a temporally distinct neural representation of action affordances and highlight a limitation of current DNNs in modeling affordance perception.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Humans effortlessly navigate dynamic environments, where changes can occur in milliseconds—such as cycling through dense morning traffic. This highlights the brain's remarkable speed in processing visual scenes (Fei-Fei et al., 2007; Potter, 1975; Thorpe et al., 1996), allowing it to capture relevant information to guide navigational actions in a glance (Greene and Oliva, 2009). The process of identifying potential relevant interactions with the environment is broadly known as *affordance perception* (Gibson, 1977). Although the presence of scene-selective cortical regions is well established (Bartnik and Groen, 2023; Epstein and Baker, 2019; Dilks et al., 2022), and growing evidence highlights their sensitivity to navigational affordances (Bonner and Epstein, 2017; Dwivedi et al., 2021b; Bartnik et al., 2024), most prior research has relied on fMRI measurements. As a result, the temporal dynamics of navigational affordance perception and the cascade of underlying neural computations remain poorly understood.

Studies using time-resolved brain measurements during scene perception indicate that scene feature representations emerge as early as 100 ms after image onset, seemingly following a temporal hierarchy. Low-level features, such as global scene properties (e.g., clutter level, scene size, and overall spatial layout), are extracted between 90–150 ms (Groen et al., 2013; Ramkumar et al., 2016; Cichy et al., 2017; Hansen et al., 2018). Open vs. closed scene discrimination (Lowe et al., 2018) and manmade vs. natural scene categorization (Groen et al., 2013; Harel et al., 2016) also emerge around that time. Greene and Hansen (2020) showed that information about objects contained in the scenes is processed around 175–225 ms, while correlations with behavioral tasks, especially affordance-based scene sorting, appear at later stages. Recent studies also show that spatial structure is processed between 90–125 ms, semantic content between 140–175 ms (Mononen et al., 2025), while navigational affordances may be processed even later, around 300 ms post-onset (Dwivedi et al., 2024).

This sequence suggests that affordance perception might build upon previously extracted scene features. However, isolating individual contributions is challenging due to feature interdependencies (Groen et al., 2017; Malcolm et al., 2016), and it has been suggested that scene processing may not follow a straightforward temporal cascade from low- to high-level scene properties (Groen et al., 2017; Ramkumar et al., 2016; Greene and Hansen, 2020). Indeed, one study found that neural representations of local features diagnostic of navigational affordances emerge as early as 134 ms after image onset (Harel et al., 2022), and perceived affordances strongly shape scene categorization (Greene et al., 2014), inherently affecting how the environment itself is perceived (Djebbara et al., 2019). This view of affordances as a 'visual primitive' is supported by evidence that affordance perception is largely task-independent (Bonner and Epstein, 2017; Bartnik and Groen, 2023), though prior knowledge may still influence scene perception (Djebbara et al., 2019; Naveilhan et al., 2024).

When navigational affordance representations emerge in visual processing may also depend on how they are operationalized. Most existing studies focus on indoor environments, defining affordances as possible pathways through space (Bonner and Epstein, 2017; Dwivedi et al., 2024; Harel et al., 2022). An alternative approach considers locomotive action affordances, emphasizing the identification of specific *types* of actions required to navigate an environment (Bartnik et al., 2024). In a prior fMRI study, we found representation of such affordances in the scene-selective regions Parahippocampal Place Area (PPA) and Occipital Place Area (OPA). These representations were not explained away by low-level features or object labels, and were poorly captured by features learned by modern DNNs (Bartnik et al., 2024). Given these findings, we hypothesized that temporally-resolved neural measurements will also reveal unique locomotive action affordances representations that do not overlap with other visual features or DNN features. Such measures furthermore allow determining the temporal onset of locomotive affordance representations, which could help elucidate the temporal dynamics of affordance perception.

To test this hypothesis, we collected human electroencephalography (EEG) responses to a diverse set of images as used in our prior study (Bartnik et al., 2024), and related these responses to behavioral annotations, fMRI data, and DNN feature activations. Our results confirm that locomotive action affordances form a unique representational space not only in fMRI but also in EEG signals; moreover, we find that they are

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

processed within 200 ms but later than low-level image features or object representations, are temporally aligned with scene-selective regions, with a processing hierarchy showing earlier responses in OPA compared to PPA. While DNNs exhibit strong representational alignment with EEG signals, their representations align more closely with low-level features, suggesting they may be insufficient for accurately modeling the neural processing of locomotive action affordances.

Results

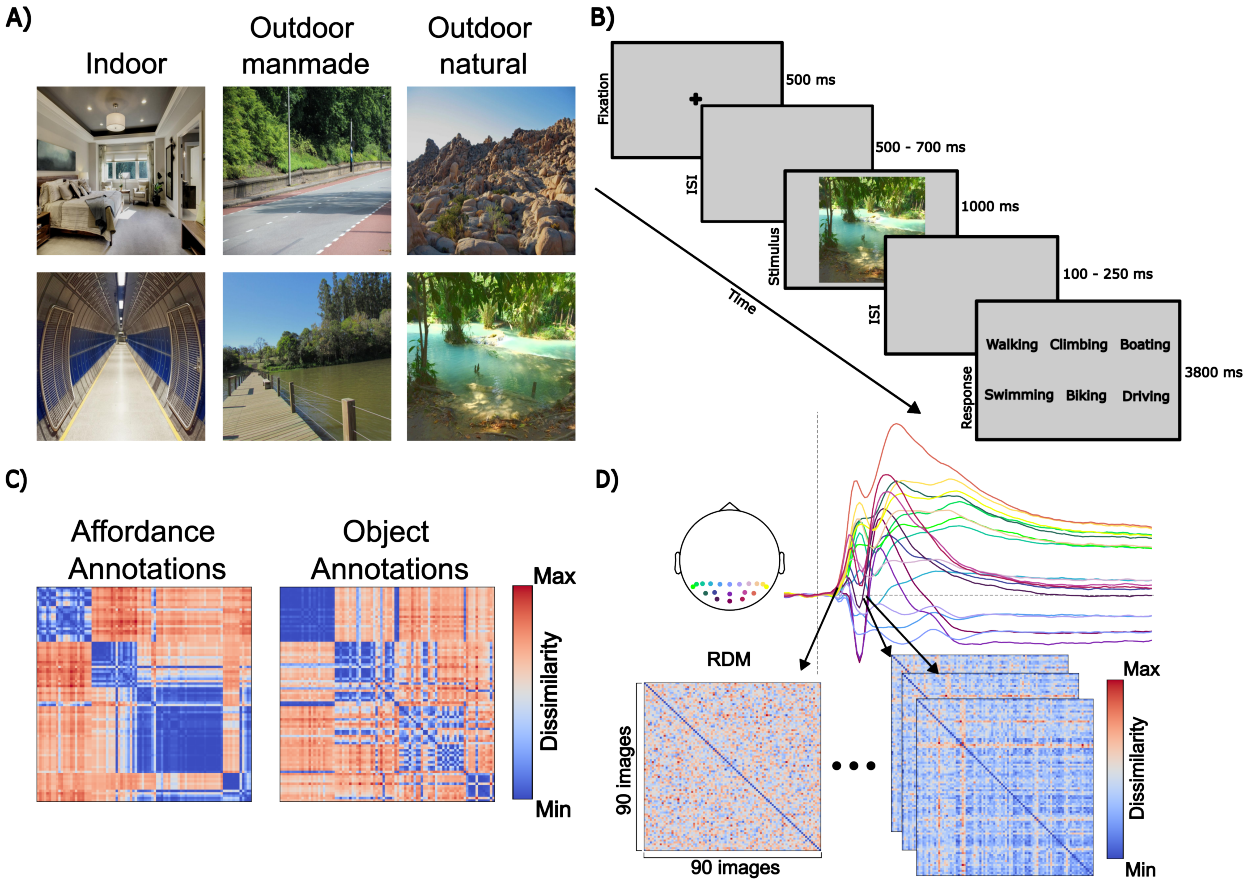


Figure 1: Overview: (A) Example scenes from our stimulus set (collected from Flickr containing (CC0 1.0, CC BY 2.0 and CC BY-SA 4.0 licensed). (B) Example trial sequence of the EEG experiment from the locomotive affordance annotation task. Participants additionally performed an object annotation and a fixation task, and in each task run responded to only subset of the trials. (C) RDMs from locomotive action affordance (left) and object annotations (right), with red indicating high and blue low dissimilarity. (D) Single-image evoked responses from 19 occipital electrodes, with RDMs computed from ERP response amplitudes every 7.86 ms (-100 to 1000 ms relative to image onset).

We recorded EEG responses as human participants viewed 90 real-world scene images spanning indoor, outdoor man-made, and outdoor natural environments (**Fig. 1A**). All participants completed three task blocks: (1) categorizing scenes based on six locomotive action affordances (**Fig. 1B**), (2) identifying objects, and (3) reporting the color of a fixation cross (control task). To quantify perceived locomotive action affordance and object representations, we computed pairwise dissimilarities between the obtained behavioral annotations, generating representational dissimilarity matrices (RDMs) for both affordance and object representational spaces (**Fig. 1C**). In addition, we computed EEG RDMs based on pairwise dissimilarities between event-related potentials (ERP) amplitudes across posterior electrodes recorded for each scene, at each time point between -100ms to +1000ms relative to stimulus onset (**Fig. 1D**). Similar to (Bartnik

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al., 2024), task-based analyses revealed no robust effects of task instructions on neural responses (see **Supplementary Figure 1**), so all results reported below are based on task-averaged ERPs.

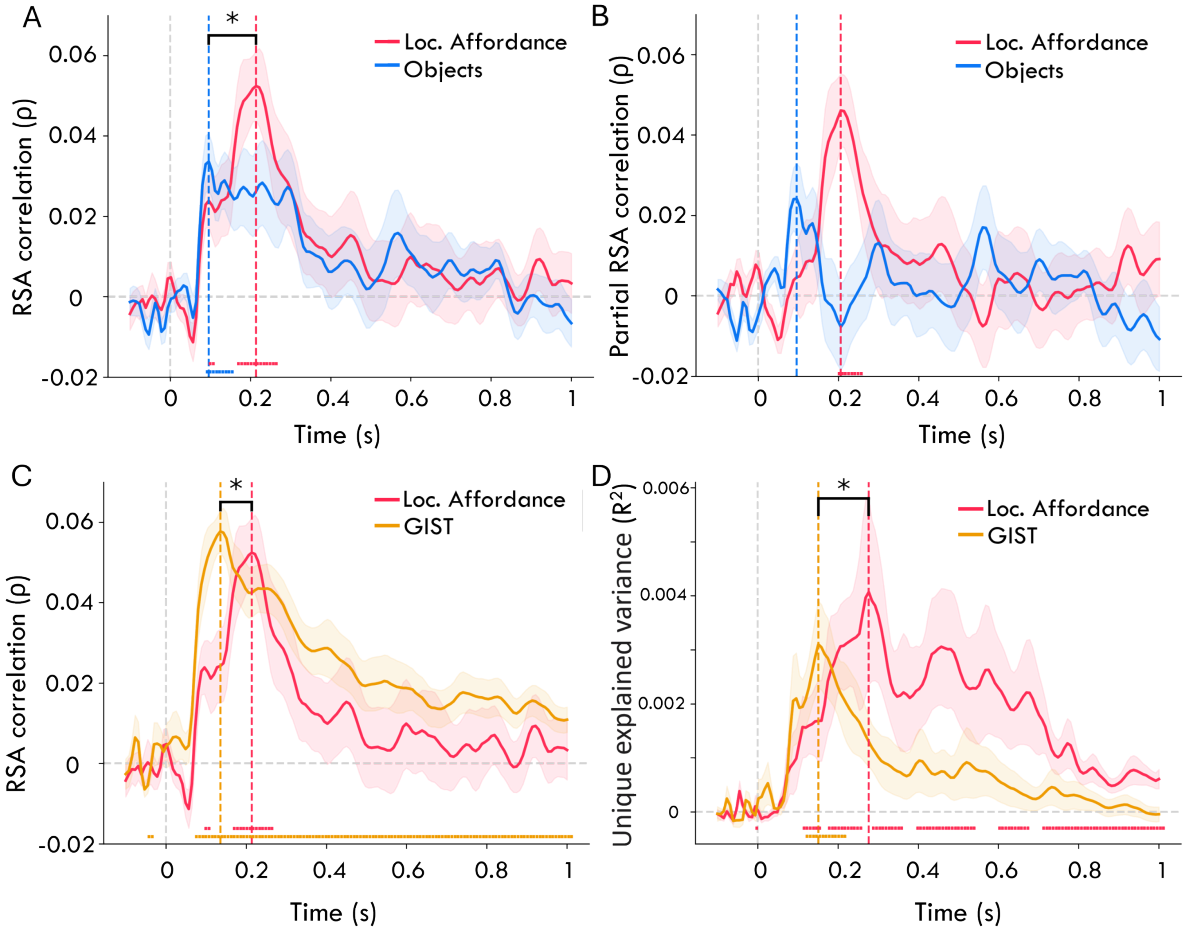


Figure 2: (A) Time-resolved, across-subject averaged RDM correlations between behavioral annotations and ERPs; shaded areas indicate SEM. Significant time points (fdr-corrected) are marked with dots, and vertical dashed lines highlight peak correlation times. Asterisks indicate a significant (α <0.05) different in peak correlation times. (B) Partial correlations between action affordance and ERP RDMs, controlling for object representations, and vice versa. (C) Time-resolved correlations between ERP and GIST RDMs vs. locomotive action affordance RDMs. (D) Unique variance explained by action affordance annotations and GIST features.

Unique locomotive action affordance representations emerge around 200 ms

To investigate when locomotive action affordance representations emerge during visual processing, we correlated the task-averaged ERP RDMs per time point with the behavioral RDMs for both locomotive action affordance and objects. The resulting correlation time courses (**Fig. 2A**) show that object representations emerged relatively early, around 100ms after image onset, while action affordance representations emerged later, peaking shortly after 210ms. A paired-sample t-test comparing peak correlation times across participants confirmed a significant delay in the processing of locomotive action affordances relative to objects (t(17) = 2.40, p = 0.028).

While these results provide initial evidence of a temporally distinct representation of locomotive action affordances relative to objects, the two behavioral RDMs also show considerable inter-correlation (ρ = 0.60) (a similar correlation was found in behavior obtained in a larger-scale online experiment and in the fMRI

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

experiment reported in Bartnik et al. (2024); see **Supplementary Figure 2A** for an overview). To better understand the individual contributions of each type of representation, we computed partial correlations, thereby removing the shared contribution of the object representational RDM from the action affordance RDM, and vice versa. Controlling for object annotations reduces the correlation with action affordance at early time-points but maintains the distinct peak around 200 ms (**Fig. 2B**). This peak, along with subsequent time points, remained significantly above zero (all $t > 3.59$, all $p < 0.002$; fdr-corrected), indicating processing of unique locomotive action affordance-related information. In contrast, controlling for locomotive action affordances rendered the object correlation peak at 100 ms insignificant, and also reduced correlations with object representations at later time points, revealing the opposite pattern as observed for action affordances.

To take into account potential overlap with other image features than objects, we also compared locomotive action affordance representations to the GIST model (Oliva and Torralba, 2001), a well-established image-computable model for low-level features known to correlate with early EEG signals (Greene and Hansen, 2020; Groen et al., 2017). As shown in **Fig. 2C**, correlations with GIST model RDMs emerge significantly earlier (peak at 135ms) than the affordance representations (one-sample t-test: $t(17) = 2.39$, $p = 0.022$). Partial correlation analysis produced nearly identical curves (**Supplementary Figure 3**), reflecting the fact that these representational spaces are not highly correlated ($\rho = 0.08$) to another. Interestingly, while **Fig. 2C** suggests a stronger correlation with GIST throughout the ERP time course, a variance partitioning analysis provided a clear distinction of each representational space's unique contribution to the EEG signal (**Fig. 2D**), with GIST features and affordances both accounting for unique variance between 150 and 200 ms, but affordances continuing to explain unique variance at later time points.

Overall, these results show that locomotive action affordances are represented in EEG responses and are processed later in visual processing than objects or GIST features.

Scene-selective regions match the time course of locomotive action affordance representations

In our prior fMRI study (Bartnik et al., 2024) we demonstrated that scene-selective regions OPA and PPA, but not the MPA, automatically extract representations of locomotive action affordances. Our recordings of EEG responses to the same image set enables spatiotemporal fusion (Cichy et al., 2019) to assess each region's temporal correspondence with the ERP responses.

For this, we constructed average RDMs by calculating pairwise correlation distances of multi-voxel activity patterns from PPA, OPA, and MPA (see Bartnik et al. (2024) for further details on the fMRI analysis). Next, we correlated these three fMRI ROI RDMs with the ERP RDMs, separately for each time point. **Fig. 3A** shows that all three scene-selective regions exhibit significant correlations with ERP responses around 90ms after stimulus onset (one-sample t-tests against 0, fdr-corrected for multiple comparisons). OPA peaks first at approximately 90ms, followed by MPA at 150ms. PPA shows an initial peak at 90ms but exhibits additional peaks at 150ms and 210ms.

Because all three regions are known to be involved in scene perception, they likely exhibit a substantial covariance. To identify their unique correspondence to the EEG signal, we again used variance partitioning to determine how much unique variance each scene-selective region's RDM explains in the ERPs. **Fig. 3B** shows that indeed all three regions contain overlapping representations, indicated by the shared explained variance peaking around 150 ms. Nevertheless, each region also explains unique variance, starting from around 100 ms, with OPA and MPA contributing the most unique variance around this time. In contrast, PPA accounts for more unique variance at later time points, particularly after 200 ms.

While analyzing unique variance helps disentangle the individual contributions of each scene-selective region, the significant time windows remain highly overlapping. Indeed, despite the apparent differences in the time points of the highest average unique explained variance, a paired-samples t-test comparing the peak times for OPA and PPA showed no significant difference ($t(17) = -0.83$, $p = 0.417$). However, a more focused comparison using partial correlations between OPA and PPA only reveals distinct temporal dynam-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ics (**Fig. 3C**): OPA exhibits a unique partial correlation peak around 90ms post-image onset (all $t > 3.62$, all $p < 0.002$; fdr-corrected) indicating its distinct contribution at this early stage. Conversely, PPA's early correlations disappear when controlling for OPA, with significant peaks emerging only later, between 180–250ms post-onset (all $t > 2.71$, all $p < 0.015$; fdr-corrected). These non-overlapping time windows suggest that OPA processes information earlier, while PPA becomes engaged at later stages.

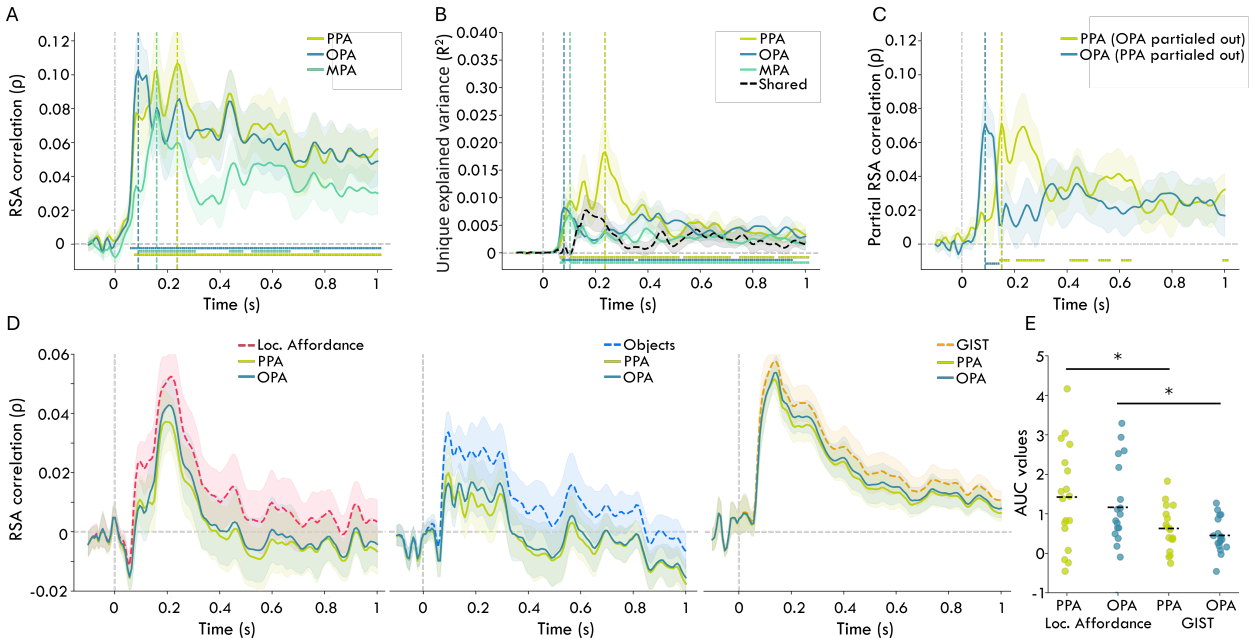


Figure 3: (A) Time-resolved, across-subject averaged RDM correlations between scene-selective ROIs and ERP RDMs, with shaded areas indicating SEM. Significant time points (fdr-corrected) are marked with dots, and vertical dashed lines indicate peak correlation times. (B) Variance partitioning of RDM correlations (C) Partial RDM correlations between PPA and ERP RDMs, controlling for OPA representations, and vice versa. (D) Left: RDM correlations between locomotive action affordances and ERP RDMs, with correlations after controlling for PPA and OPA. Middle: Objects Right: GIST (E) AUC values of difference curves from (D), subtracting partial correlations from full correlations.

Thus far, we demonstrated temporal alignment between scene-selective ROI representations and EEG responses. But to what extent does this temporal alignment reflect the processing of action affordances, versus other image properties? To answer this question, we examined how correlations of affordance representations with ERPs change when controlling for OPA and PPA representations. In **Fig. 3D**, the first panel shows that partialing out either OPA or PPA RDMs reduces locomotive action affordance correlations with ERPs, suggesting both regions contribute to the affordance representations. The second panel shows a similar pattern for objects, with correlations reducing when partialing out both regions. In contrast, the GIST model correlations with ERPs remain largely unchanged when controlling for OPA and PPA (final panel), indicating minimal correspondence of these regions with the representation of GIST features in the ERPs.

To quantify these effects, we computed difference curves by subtracting the correlation time courses obtained by partialing out the fMRI ROIs (**Fig. 3D**) from the original correlation time courses, and then calculated the area under the curve (AUC) (see Methods), reflecting the extent of shared variance with the different feature representations. The resulting AUC values (**Fig. 3E**) were significantly larger when partialing scene-selective regions out of affordance representations in ERPs, compared to the GIST representations (PPA $t(17) = 4.656$, $p < 0.001$); OPA $t(17) = 4.310$, $p < 0.001$). This shows that both regions indeed contribute to the ERP variance related to affordance processing, but not to GIST-related variance.

Overall, these results show that scene-selective regions account for variance in ERP signals, suggesting a processing hierarchy where OPA processes information earlier than PPA. Additionally, these regions ap-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pear to primarily share variance related to high-level features such as locomotive action affordances and object representations, rather than low-level features captured by the GIST model.

Pre-trained DNNs correlate well with ERPs but show alignment with low-level features rather than locomotive action affordances

DNNs serve as effective models of human visual processing (Kietzmann et al., 2019), but their ability to encode features relevant to navigational affordances remains unclear. While prior research suggests they capture affordance-related information in the form of pathways (Bonner and Epstein, 2018), we found stronger alignment with object representations (Bartnik et al., 2024). Here, we assess how well a variety of DNNs capture navigational affordances by examining their alignment with emerging visual representations in ERP responses over time. Specifically, we extracted feature activations for each of 90 scenes separately for each layer and DNN model, which we then used to create RDMs that we correlated with the ERP RDMs at each time point (see Materials and Methods for further details).

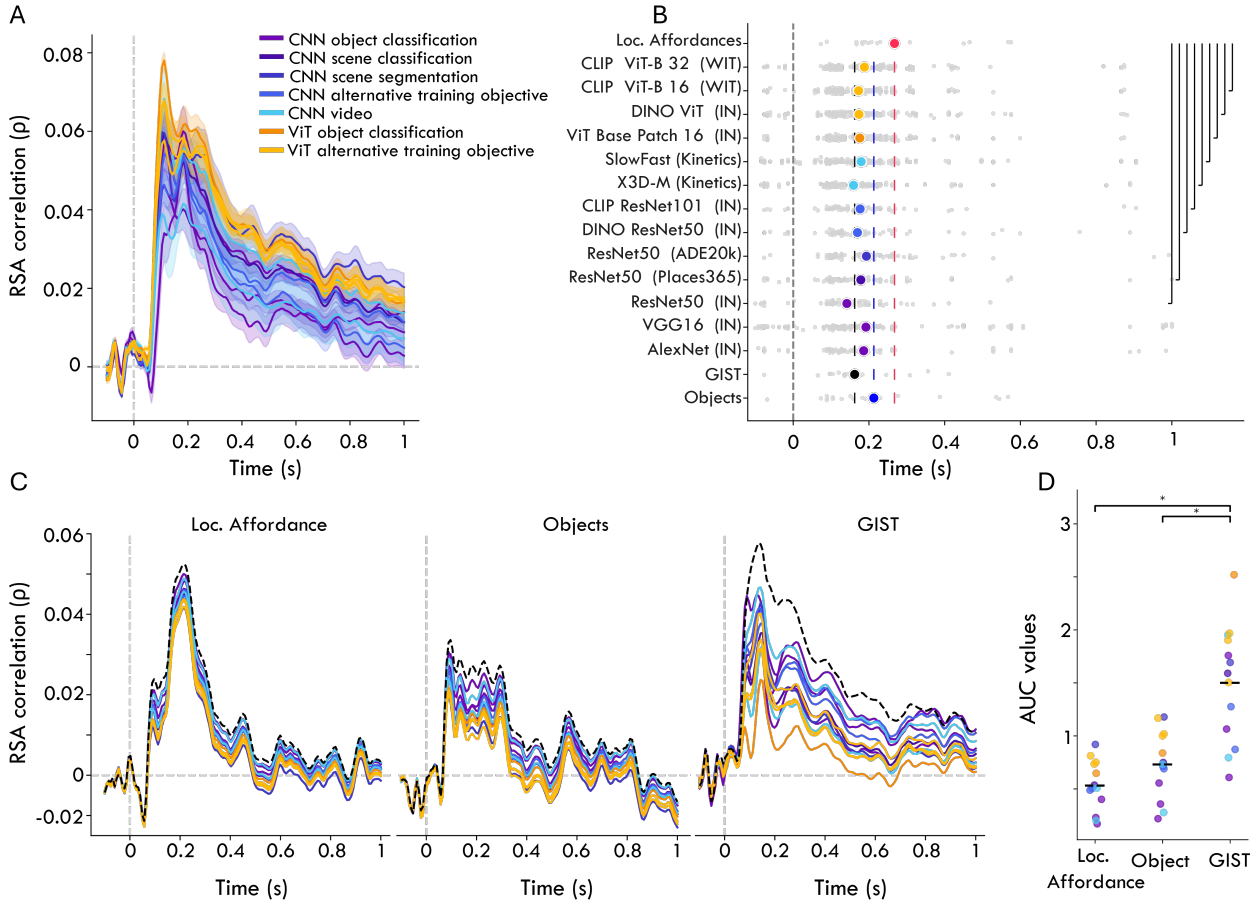


Figure 4: (A) Time-resolved, across subject RDM correlations between ERP RDMs and DNN layer activation RDMs (averaged across subjects and layers), with shaded areas indicating SEM across layers. (B) Peak correlation time points for DNNs and the affordance, object and GIST representational spaces, with gray dots representing individual layer correlations and colored dots showing the average. Non-DNN representational space peak averages are marked with colored lines, and significant differences from action affordance annotations (fdr-corrected) are highlighted with black lines. (C) Time-resolved partial RDM correlations between feature models and ERP RDMs, controlling for DNN activations. First panel: Locomotive action affordance correlations, second: object correlations, third: GIST correlations. (D) AUC values of difference curves from (C), subtracting partial correlations from full correlations.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 4A confirms the general notion that DNNs are good models capturing representations in visual cortex, as all tested DNN models, when averaging correlations across layers, display substantial RSA correlations with the ERPs starting around 100 ms (see **Supplementary Figure 4** for layer-specific results). Further, we observe that Transformer architectures (ViT) and models trained on image-text pairs (CLIP) show somewhat higher correlations than classic CNNs trained on object recognition, indicating higher representational alignment with brain activity.

To what extent are these DNNs able to capture the neural representation of locomotive action affordances in the EEG signal? To examine this, we first compared the peak correlation times for activations extracted from each individual DNN layer, with the average peak time across subjects for each of the three representational spaces we tested so far (affordances, objects and GIST). DNN model groups show no consistent ordering of peak timing, with all models showing a peak correlation in between the maxima for the GIST model and the object features; none of the models shows a clear temporal co-localization with the peak for affordances (**Fig. 4B**).

This peak comparison provides a first indication that DNN features primarily align with earlier visual processing stages that reflect low-level feature processing, rather than affordance processing. However, this peak misalignment does not provide a complete picture; RSA correlations with DNN features extend throughout the ERP epoch, and could thus still capture affordance representations at later stages. To compare representations across the entire ERP time course, we performed an analogous partial correlation analysis to the spatiotemporal fusion section above. We reasoned that a significant reduction in correlation between ERPs and affordances when partialing out DNN activations would indicate representational overlap. We indeed observe a slight reduction when partialing out DNN activations from the locomotive action affordance correlations with the ERPs (**Fig. 4C**, left panel). ViTs and CLIP models induce the largest reductions, consistent with their overall stronger alignment with ERPs. Object correlations with ERPs show a similar, slightly stronger reduction, especially in early time points (**Fig. 4C**, middle panel). In comparison, GIST model correlations exhibit a much larger reduction when controlling for DNN activations, indicating more overlap of DNN representations with GIST features in ERPs than with affordances or objects (**Fig. 4C**, right panel).

To quantify these reductions, we again calculated the AUC between the original and the partial correlations. **Fig. 4D** shows the lowest AUC values for the action affordance model, followed by object correlations. The highest AUC difference scores are found for the GIST model, which are significantly higher than the affordance model (Mann-Whitney $U = 8.0$, $p = 0.0001$, corrected $p = 0.0015$) and the object model ($U = 20.0$, $p = 0.0010$, corrected $p = 0.0003$), suggesting a greater alignment of DNN representations with low-level features represented by the GIST model than with affordance- and object-related information in ERPs.

Overall, these results suggest that, although DNNs exhibit strong representational alignment with EEG responses, their representations correspond more closely to low-level features than to locomotive action affordances. This indicates that current DNN models may be insufficient for accurately modeling the neural processing of locomotive action affordances.

Discussion

In this study, we examined the temporal dynamics of locomotive action affordance perception, investigating whether representations of such affordances emerge distinctly from other scene features, and how well they align with neural responses in scene-selective regions and layer activations of DNNs.

Our results revealed that locomotive action affordances form a distinct representation, explaining ERP variability between 175 and 250 ms post-stimulus. These action affordances appear to reflect comparatively higher-level features, processed in a distinct time window after low-level GIST and object features. Scene-selective regions OPA, PPA and MPA all three explain variability in ERPs, with results indicating a potential temporal processing hierarchy: OPA processes information earlier, while PPA reaches peak explanatory

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

power around 200 ms. Interestingly, partial correlation analyses reveals that both regions account for more ERP variability shared with affordances and objects than with GIST-based low-level features, indicating a role in processing high-level features. In contrast, while DNNs pre-trained on visual tasks also align with ERP responses, they primarily captured low-level visual feature-related variance in the ERPs, showing limited alignment with representations of locomotive action affordances.

Prior research on the temporal dynamics of navigational affordance perception has predominantly focused on synthetic (Harel et al., 2022; Djebbara et al., 2019) and exclusively indoor real-world environments(Dwivedi et al., 2024), typically defining affordances as walkable pathways. In contrast, our study adopts the paradigm of Bartnik et al. (2024), expanding the operationalization of navigational affordances to include multiple locomotive actions across both indoor and outdoor scenes. The timing of affordance perception is debated, with some studies suggesting it occurs relatively late, around 300 ms (Dwivedi et al., 2024) therefore likely building on previously extracted features. In contrast, our findings indicate that locomotive action affordances emerge earlier, around 200 ms, consistent with other studies (Harel et al., 2022; Djebbara et al., 2019). While our definition of affordances is broader than in previous studies, it remains more specific than all affordances in a scene. Nonetheless, our findings align with (Greene and Hansen, 2020), who reported that a much broader set of 227 scene functions peaked in neural correlation between 176 and 200 ms. In addition, the fact that our affordance representations explained unique variance in EEG responses independent of objects or GIST features suggests that affordance perception may occur in parallel, rather than using other scene features as building blocks.

By using the same stimuli as (Bartnik et al., 2024) and replicating key findings in the temporal domain, our study provided the opportunity to integrate fMRI and EEG measurements by spatiotemporal fusion (Cichy et al., 2016), offering a more complete picture of how locomotive action affordances are processed in the brain. Our results confirm that both OPA and PPA show correspondence with ERP responses to scenes, but crucially, they do so at different time points, revealing a temporal hierarchy between them. OPA, which has been linked to scene layout representation and navigational affordance processing (Bonner and Epstein, 2017; Park and Park, 2020; Epstein and Baker, 2019), exhibits an early peak around 100 ms, consistent with its role in rapid invariant encoding of scene layout (Henriksson et al., 2019). In contrast, PPA, associated with encoding spatial layout, objects, and texture (Epstein and Kanwisher, 1998; Walther et al., 2009; Harel et al., 2013; Bastin et al., 2013), reaches peak correlation around 200 ms, overlapping with the later stage of locomotive action affordance processing.

These distinct temporal dynamics may explain why previous studies that defined navigational affordances as walkable pathways (Bonner and Epstein, 2017), relying primarily on spatial structure processed around 90–125 ms (Mononen et al., 2025), predominantly identified OPA rather than PPA. In contrast, locomotive action affordances likely engage additional scene features (e.g. textures) processed in PPA (Henriksson et al., 2019). By comparing to different representational feature spaces, we further show that these regions are not fully explained by low-level features. While GIST features are extracted early (100–150 ms) (Ramkumar et al., 2016) and significantly influence fMRI activity in PPA and OPA (Watson et al., 2014, 2017; Rice et al., 2014; Bartnik et al., 2024), our results indicate that they primarily account for early ERPs but do not overlap with affordance representations in these regions.

DNNs are widely used to model representations in visual cortex (Kietzmann et al., 2019; Storrs and Kriegeskorte, 2019), with their feature activations mapping onto various brain regions (Khaligh-Razavi and Kriegeskorte, 2014; Kriegeskorte, 2015; Guclu and van Gerven, 2015) and effectively capturing the temporal dynamics of visual perception (Cichy et al., 2017; Bankson et al., 2018). However, most studies have focused on object perception, leaving DNNs ability to encode navigational affordances less explored. Our results confirm that DNN representations correlate strongly with ERPs (Gifford et al., 2022), peaking around 100 ms, suggesting they effectively capture early visual processing, but do not fully capture affordance-related processing at later time points. Moreover, unlike other prior work suggesting that a DNN's training objective influences its alignment with brain representations (Dwivedi et al., 2021b,a), we find no clear benefit of models trained on e.g. scene segmentation or video, as opposed to object recognition, in explaining ERP variance to scene images. Although CLIP models trained on image-text pairs exhibit one of the strongest correlations with

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ERP responses, the same correlation was obtained with ViT Base Patch 16, which was trained on regular object labels, arguing against a strong benefit of language supervision on brain alignment (Wang et al., 2023; Doerig et al., 2022).

It is important to acknowledge several limitations of our study, which also highlight key directions for future research. While we extend the operationalization of affordances by using six locomotive action affordances, a key definition of affordances is that they should reflect the relations between environmental properties and the perceiver's action capabilities (Gibson, 1977; Rietveld and Kiverstein, 2014). Here more immersive tasks and naturalistic inputs (Gregorians and Spiers, 2022; Zhang and Gallant, 2020; Djebbara et al., 2021) are needed to account for the embodiment that ties a particular set of scene, object, and material properties together to an individual observer's affordances. A lack of embodiment could also explain DNNs inability to capture locomotive affordances. Here, a way forward could be to test DNNs that are trained on human affordance labels, or to explore vision-language-action models from robotics such as (Kim et al., 2024) which are trained in more embodied contexts.

Finally, our locomotive action affordance annotations showed considerable correlation with object annotations, making it challenging to fully disentangle their distinct contributions to the temporal dynamics of scene perception and pinpoint the diagnostic features for affordance perception (Groen et al., 2017). While our results reveal the presence of independent affordance-related variance in the EEG signals through partial correlations and variance partitioning, a more fine-grained representation of affordance-related features, such as spatial layout and textures, could help clarify how the brain processes locomotive affordances over time. Furthermore, expanding the spatiotemporal fusion analysis beyond scene-selective regions to include other brain areas could provide a more comprehensive understanding of the full temporal hierarchy underlying affordance perception, especially since prior searchlight results show that locomotive affordance representations extended into early and mid-level visual regions (Bartnik et al., 2024).

Overall, our findings demonstrate that locomotive action affordances form distinct neural representations that emerge after low-level and object features, while current DNNs fail to adequately capture these affordance-related processes.

Materials and Methods

Participants

Twenty healthy volunteers (6 males; age 18–27, M = 21.45, SD = 2.43) participated in the study, all with normal or corrected vision and providing informed consent. They received research credits or monetary compensation. Two participants were excluded due to incomplete data. The study was approved by the Ethical Committee of the institution.

Stimuli

The stimuli consisted of 90 high-resolution (1024×1024 pixels) color photographs, sourced from a copyright-free image database (Flickr) previously used in a neuroimaging study (Bartnik et al., 2024). Each image was captured from a human-scale, eye-level perspective and depicted an everyday scene without humans or prominent central objects, and belonged to one of three environmental categories: indoor, outdoor-natural, or outdoor-man-made (see **Fig. 1A**).

Experimental design and procedure

The experiment consisted of three blocks, each corresponding to a different task (action, object, or fixation) which were presented in a counterbalanced order. In each block, participants viewed the same 90 images, each repeated six times. Once per image per block, a response screen appeared after the presentation of the image, prompting participants to press one of six keyboard buttons corresponding to task-specific labels:

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

action affordances (walking, biking, driving, swimming, boating, climbing), contained objects (building, plant, water, furniture, road, stones), or the color of the fixation cross (blue, red, orange, purple, yellow, cyan). Image order and response option order were randomized separately for each participant. The response screen appeared randomly after one of the six repetitions, resulting in six task-specific repetitions and 18 total repetitions per image across tasks. Before each block, participants completed a training session with 15 practice images not included in the main experiment.

Stimuli (1024x1024 px) were presented using PsychoPy (v3.2.4) on a 2560×1440 px screen (59.6 × 33.6 cm) in a controlled lighting environment. Participants sat 70 cm from the screen, with images spanning 20° of their visual field. A fixation cross was displayed at the start (8 s) and end (12 s) of the experiment. Each trial followed this sequence: fixation cross (500 ms), blank grey screen (500–750 ms), stimulus image with a randomly colored fixation cross (1000 ms), and another blank grey screen (100–250 ms). For response trials, a response screen appeared (max 3800 ms), followed by a jittered inter-stimulus interval (ISI) of 2000–4000 ms. Non-response trials proceeded directly to the ISI. Breaks were allowed after every 90 images and between task blocks.

EEG acquisition and pre-processing

EEG data was recorded using a Biosemi 64-channel Active Two EEG system (www.biosemi.com) with a 10-20 layout at a sampling frequency of 2048.0 Hz. To capture more vision-related activity, two frontal electrodes (F5 & F6) were repositioned posteriorly to the left and right of Iz (renamed I1 & I2). Eye movements were monitored with electro-oculograms (EOGs). Preprocessing for the purpose of computing event-related responses (ERPs) was done in Python, MNE (Gramfort et al., 2013) and included the following steps.

High-pass filter at 0.1 Hz (6 dB/octave); a low-pass filter at 30 Hz (6 dB/octave) (one-pass, zero-phase, non-causal bandpass filter, hamming window, 0.0194 passband ripple, 53 dB stopband attenuation, -6 dB cut-off frequency); two notch filters (zero-phase) at 50 Hz and 60Hz; epoch segmentation in -100 to 1000 ms from stimulus onset, downsampled to 128 Hz for computational efficiency; baseline correction between -100 and 0 ms; ocular correction using the EOG electrodes (Gratton et al., 1983)); conversion to Current Source Density responses (Perrin et al., 1987). Artifacts were rejected using maximal allowed amplitudes of -75 and +75 µV in the Oz channel. This led to a rejection of 2.04% of the epochs with (mean = 5.5%, SD = 0.45% images per participant and task).

Trials with the same image were averaged across all repetitions to obtain non-task-specific ERPs for each image and each subject. For task-specific ERPs, only the image repetitions within each block corresponding to the same task were averaged. Following previous studies that demonstrated that navigational affordance processing happens in the visual cortex Bonner and Epstein (2017); Harel et al. (2022); Dwivedi et al. (2024), we selected 19 posterior and occipital channels for further data analysis (P1, P3, P5, P7, P9, PO7, PO3, O1, O2 Oz, POz, Pz, P2, P4, P6, P8, P10, PO4, PO8).

RDM construction

Trial-averaged ERP responses were used to construct representational dissimilarity matrices (RDMs) for each subject, based on pairwise Pearson correlation distances between ERP amplitude values across 19 electrodes. This was done in a time-resolved manner at 7.86 ms intervals from -100 to 1000 ms relative to image onset (**Fig. 1D**). RDMs were computed using the Python version of the RSA toolbox (Nili et al., 2014). To smooth the data, we applied a sliding window averaging method, averaging across consecutive 39.29 ms windows while maintaining dimensional consistency through padding. The effect of different smoothing window sizes is illustrated in **Supplementary Figure 2B**. Given that previous studies (e.g., Walther et al. (2016)) use various distance metrics, we multiple metrics and selected correlation distance as the most appropriate (**Supplementary Figure 2C**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Behavioral annotations

RDMs were calculated for each participant based on the responses they provided in the affordance and objects tasks, indicating the possibility for six locomotive actions and the object annotations. RDMs were calculated using Euclidean distance metric. These were then averaged across participants to create group-level behavioral RDMs.

fMRI data

fMRI-based RDMs were constructed from multivoxel activation patterns of 20 separate participants viewing the same 90 stimuli (see Bartnik et al. (2024) for acquisition and processing details). Voxel response patterns were limited to include only voxels from three scene-selective regions of interest (ROIs): PPA, OPA, and MPA, which were identified using a separate standard six-category localizer procedure. The resulting voxel patterns were averaged across runs and used to compute subject-specific RDMs for each ROI using Pearson correlation distances between each pair of images.

DNN feature activations

We extracted layer activations from several pre-trained deep neural network (DNN) models using the Net2Brain Python package `https://github.com/cvai-roig-lab/Net2Brain` (Bersch et al., 2022). We adopted the same model and layer selection as described in (Bartnik et al., 2024). We used the layers pre-selected by Net2Brain, and computed RDMs by standardizing the features within each layer (mean removal and scaling to unit variance) and computing pairwise distances between the flattened feature activations in each layer using the Pearson correlation distance.

Representational similarity analysis

To compare the EEG responses with the behavioral annotations, fMRI responses and DNN layer feature representations we used representational similarity analysis (RSA) (Kriegeskorte, 2008). Model RDMs (behavioral annotations, fMRI response and DNN layer features) were compared to the ERP RDMs for each individual participant and at each time point by computing Spearman's ρ as proposed by the RSA toolbox (Kriegeskorte, 2008; Nili et al., 2014) (lower triangle, excluding the diagonal; Ritchie et al. (2017)). At each time point, one-sample t-tests were conducted to determine whether the average correlation across participants differed significantly from zero. P-values were corrected for multiple comparisons using fdr correction at alpha level 0.05. To compare peak correlation time points, paired t-tests were used, while independent t-tests were applied for DNN comparisons. All comparisons were corrected using fdr correction.

Variance partitioning and partial correlations

We used variance partitioning via regression to identify the unique and shared explained variance of each source. Therefore, we calculated the difference between the variance explained when all model RDMs were included as independent variables and the variance explained when all except the current model RDM were used as independent variables. This analysis was performed for each participant and time point separately. We then averaged the unique explained variance across participants and calculated the standard error of the mean. We also computed partial correlations to account for the influence of different model RDMs. Therefore we computed the Spearman correlations while setting various model RDMs as covariates.

For the analyses aimed at quantifying the impact of controlling for fMRI RDMs or DNN RDMs on the representational similarity between the affordance, object and GIST RDMs (referred to as 'feature spaces') and EEG responses, we computed difference curves, whereby we subtracted the partial Spearman correlation time courses from the original feature space correlation time courses. We then computed the Area under the Curve (AUC) for these difference curves using the trapezoidal rule for the full ERP epoch (-100ms to 1000ms). AUC values were compared using paired-sample t-tests for the fMRI ROI comparisons, and Mann-Whitney *U* Test for the DNN comparisons.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data and code availability

EEG data, behavioral annotations, and analysis code are available on OSF `https://osf.io/v3rcq/` and GitHub `https://github.com/cgbartnik/EEG---temporal-dynamics-of-affordance-perception.git`.

Acknowledgements

This work was supported by an VENI grant (VI.Veni.194030) from the Netherlands Organisation for Scientific Research to IIAG.

Author contributions

C.G.B, I.I.A.G experimental design; C.G.B, E.I.C.F performed experiments; C.G.B analyzed the data and prepared code for final submission; C.G.B and I.I.A.G wrote the manuscript; I.I.A.G provided funding.

References

Bankson, B., Hebart, M., Groen, I. I. A., and Baker, C. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, 178:172–182.

Bartnik, C. G. and Groen, I. I. A. (2023). Visual Perception in the Human Brain: How the Brain Perceives and Understands Real-World Scenes. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press.

Bartnik, C. G., Sartzetaki, C., Puigseslloses Sanchez, A., Molenkamp, E., Bommer, S., Vuksic, N., and Groen, I. I. A. (2024). Distinct representation of locomotive action affordances in human behavior, brains and deep neural networks.

Bastin, J., Committeri, G., Kahane, P., Galati, G., Minotti, L., Lachaux, J., and Berthoz, A. (2013). Timing of posterior parahippocampal gyrus activity reveals multiple scene processing stages. *Human Brain Mapping*, 34(6):1357–1370.

Bersch, D., Dwivedi, K., Vilas, M., Cichy, R. M., and Roig, G. (2022). Net2Brain: A Toolbox to compare artificial vision models with human brain responses.

Bonner, M. F. and Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798.

Bonner, M. F. and Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*, 14(4):e1006111.

Cichy, R. M., Khosla, A., Pantazis, D., and Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J., and Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24.

Cichy, R. M., Pantazis, D., and Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579.

Dilks, D. D., Kamps, F. S., and Persichetti, A. S. (2022). Three cortical scene systems and their development. *Trends in Cognitive Sciences*, 26(2):117–127.

Djebbara, Z., Fich, L. B., and Gramann, K. (2021). The brain dynamics of architectural affordances during transition. *Scientific Reports*, 11(1):2796.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Djebbara, Z., Fich, L. B., Petrini, L., and Gramann, K. (2019). Sensorimotor brain dynamics reflect architectural affordances. *Proceedings of the National Academy of Sciences*, 116(29):14769–14778.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., and Charest, I. (2022). Visual representations in the human brain are aligned with large language models. Version Number: 2.

Dwivedi, K., Bonner, M. F., Cichy, R. M., and Roig, G. (2021a). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, 17(8):e1009267.

Dwivedi, K., Cichy, R. M., and Roig, G. (2021b). Unraveling representations in scene-selective brain regions using scene-parsing deep neural networks. *Journal of Cognitive Neuroscience*, 33(10):2032–2043.

Dwivedi, K., Sadiya, S., Balode, M. P., Roig, G., and Cichy, R. M. (2024). Visual features are processed before navigational affordances in the human brain. *Scientific Reports*, 14(1):5573.

Epstein, R. A. and Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, 5(1):373–397.

Epstein, R. A. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2):67–82.

Gifford, A. T., Dwivedi, K., Roig, G., and Cichy, R. M. (2022). A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. S. (2013). MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13.

Gratton, G., Coles, M. G. H., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55(4):468–484.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., and Fei-Fei, L. (2014). Affordances provide a fundamental categorization principle for visual scenes. arXiv:1411.5340 [cs, q-bio].

Greene, M. R. and Hansen, B. C. (2020). Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. *The Journal of Neuroscience*, 40(27):5283–5299.

Greene, M. R. and Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472.

Gregorians, L. and Spiers, H. J. (2022). Affordances for Spatial Navigation. In Djebbara, Z., editor, *Affordances in Everyday Life*, pages 99–112. Springer International Publishing, Cham.

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., and Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, 33(48):18814–18824.

Groen, I. I. A., Silson, E. H., and Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160102.

Guclu, U. and van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hansen, N. E., Noesen, B. T., Nador, J. D., and Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, 114:168–180.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., and Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro*, 3(5):ENEURO.0139–16.2016.

Harel, A., Kravitz, D. J., and Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, 23(4):947–957.

Harel, A., Nador, J. D., Bonner, M. F., and Epstein, R. A. (2022). Early electrophysiological markers of navigational affordances in scenes. *Journal of Cognitive Neuroscience*, 34(3):397–410.

Henriksson, L., Mur, M., and Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, 103(1):161–171.e3.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., and Finn, C. (2024). OpenVLA: An Open-Source Vision-Language-Action Model. _eprint: 2406.09246.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

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.

Lowe, M. X., Rajsic, J., Ferber, S., and Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex*, 106:275–287.

Malcolm, G. L., Groen, I. I. A., and Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, 20(11):843–856.

Mononen, R., Saarela, T., Vallinoja, J., Olkkonen, M., and Henriksson, L. (2025). Cortical encoding of spatial structure and semantic content in 3D natural scenes. *The Journal of Neuroscience*, page e2157232024.

Naveilhan, C., Saulay-Carret, M., Zory, R., and Ramanoël, S. (2024). Spatial Contextual Information Modulates Affordance Processing and Early Electrophysiological Markers of Scene Perception. *Journal of Cognitive Neuroscience*, 36(10):2084–2099.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4):e1003553.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Park, J. and Park, S. (2020). Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. *The Journal of Neuroscience*, 40(18):3621–3630.

Perrin, F., Bertrand, O., and Pernier, J. (1987). Scalp Current Density Mapping: Value and Estimation from Potential Data. *IEEE Transactions on Biomedical Engineering*, BME-34(4):283–288.

Potter, M. C. (1975). Meaning in visual search. *Science*, 187(4180):965–966.

Ramkumar, P., Hansen, B. C., Pannasch, S., and Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study. *NeuroImage*, 134:295–304.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rice, G. E., Watson, D. M., Hartley, T., and Andrews, T. J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience*, 34(26):8837–8844. Publisher: Soc Neuroscience.

Rietveld, E. and Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, 26(4):325–352.

Ritchie, J. B., Bracci, S., and Op De Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, 148:197–200.

Storrs, K. R. and Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. Publisher: arXiv Version Number: 1.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200.

Walther, D. B., Caddigan, E., Fei-Fei, L., and Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, 29(34):10573–10581.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., and 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.

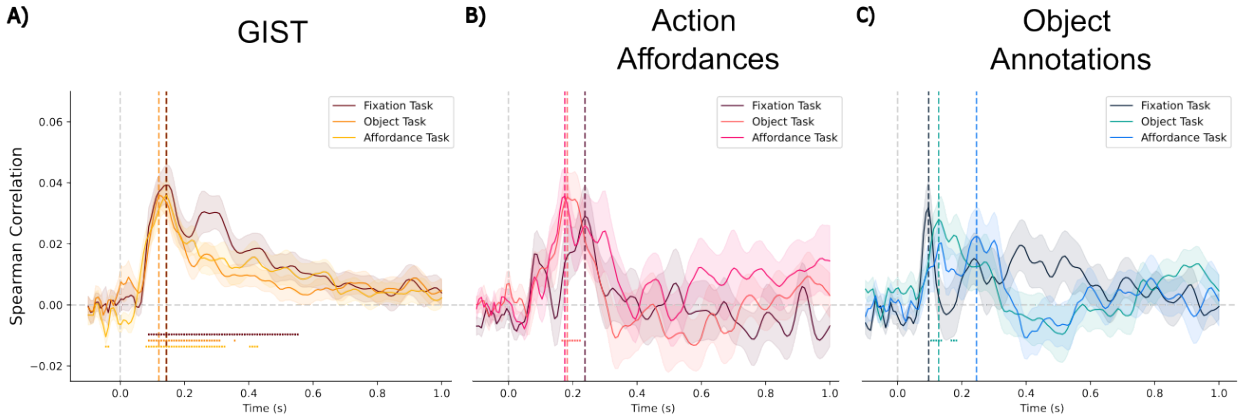
Watson, D. M., Andrews, T. J., and Hartley, T. (2017). A data driven approach to understanding the organization of high-level visual cortex. *Scientific Reports*, 7(1):3596.

Watson, D. M., Hartley, T., and Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *NeuroImage*, 99:402–410.

Zhang, T. and Gallant, J. L. (2020). A naturalistic navigation task reveals rich distributed representations of information across the human cerebral cortex. *Journal of Vision*, 20(11):462.

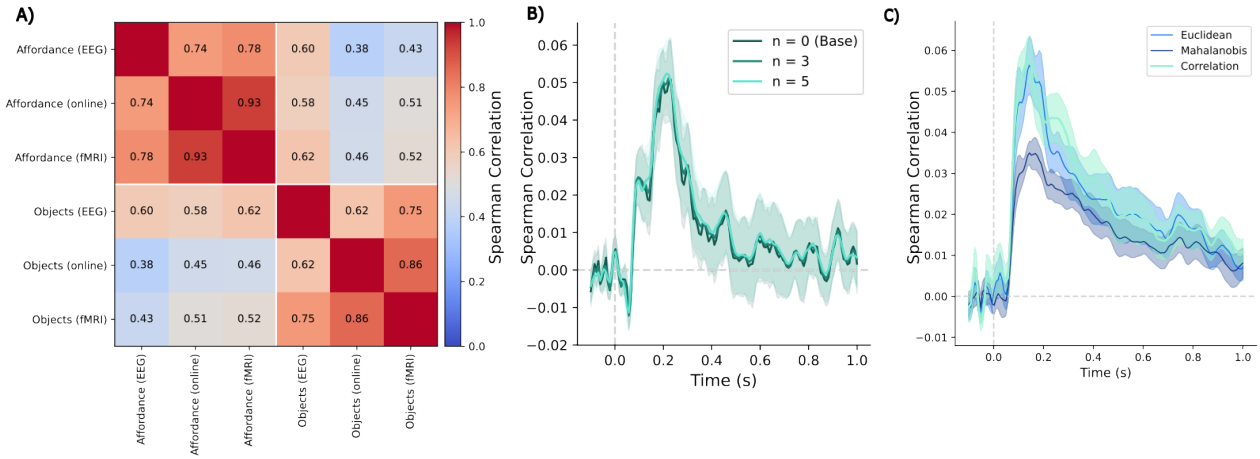
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary information



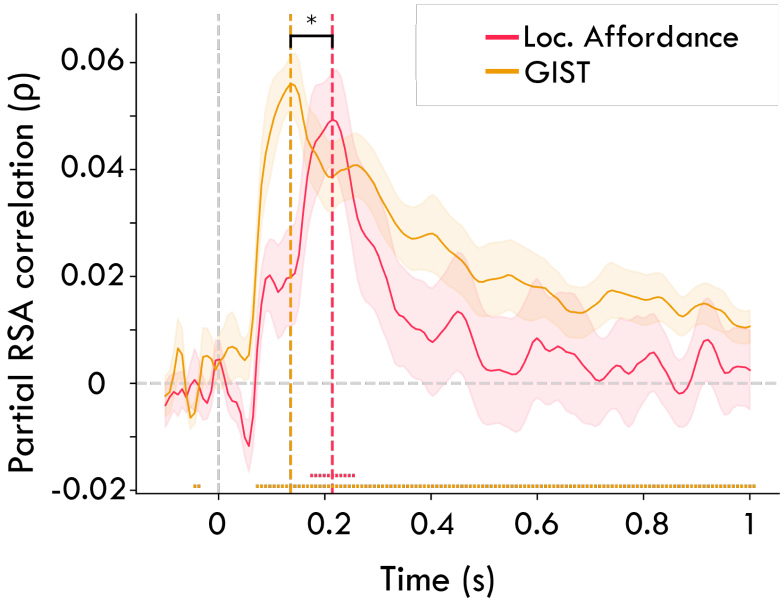
Supplementary Figure 1: Task-specific correlations between ERP RDMs and model RDMs. (A) Correlations between RDMs, derived from ERP patterns in the 19 occipital electrodes for the three specific tasks, and the GIST model using a 256 by 256 pixel image resolution. The affordance task is shown in the lightest color, the object task in the middle color, and the fixation task in the darkest color. The lines represent the average RDM correlation across participants, while the shaded areas indicate the standard error of the mean (SEM). Colored dots mark significant results from a one-sample t-test against zero ($p<0.05$), corrected for multiple comparisons using fdr correction. **(B)** Correlations for the action affordance annotations, using the same visual elements as in (A). **(C)** Correlations with the object annotations, following the same format as in (A).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



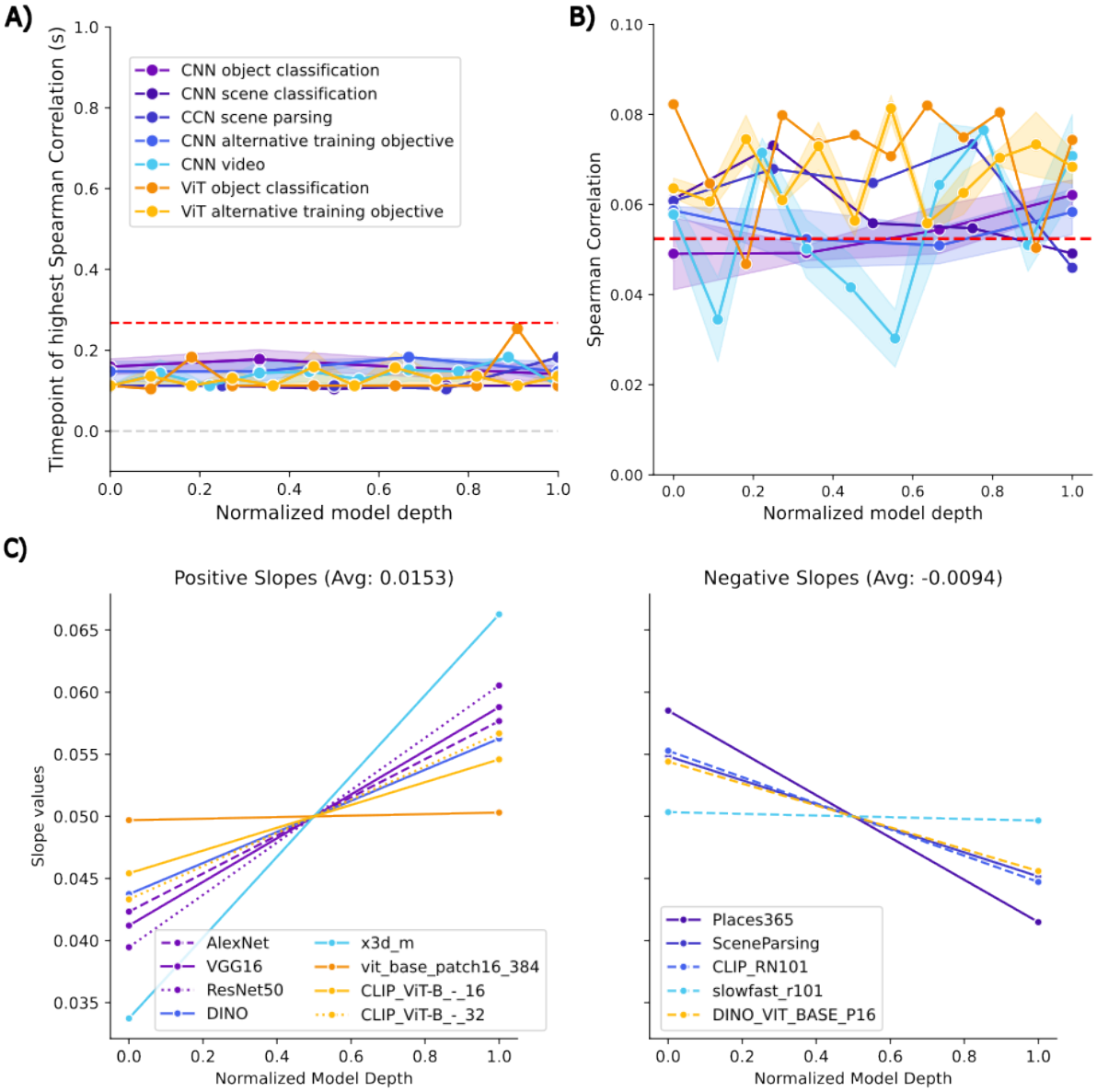
Supplementary Figure 2: Comparison to other behavioral datasets, effect of smoothing, and comparisons of distance metrics for computing EEG RDMs. **(A)** Comparison (Spearman correlation) of behavioral annotations for action affordances and objects obtained during the EEG experiment to annotations done by participants in a prior online study and fMRI experiment (Bartnik et al., 2024). **(B)** Sliding window averaging of *n* timesteps to smooth the curves. Spearman correlation of the average ERP response with the GIST RDM. **(C)** Correlation with the GIST RDM for different distance metrics used to calculate the ERP RDMs. We decided to use the correlation distance metric.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 3: Partial correlation on GIST and ERP correlations. Partial correlations between GIST and ERP RDMs, controlling for locomotive action affordance representations, and vice versa.; shaded areas indicate SEM. Significant time points (fdr-corrected) are marked with dots, and vertical dashed lines highlight peak correlation times.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.14.642994; this version posted March 15, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 4: Impact of DNN model depth on correlations with the ERPs. **(A)** Highest correlating time points by normalized model depth for each group of model. The red dashed line indicates the highest average time point for the locomotive action affordance RDM. Lines indicate the mean correlation for each model group with the shaded area indicating SEM of the mean for each group if multiple models were contained. If models in each group had different numbers of layers we reduced the numbers to the lowest common number of layers. The grey dashed line at zero indicated the image onset. **(B)** Highest RDM correlation by normalized model depth and model group. Same elements as in (A). **(C)** Regression lines for the RDM correlations for each DNN model. In the first Panel we show the positive slopes and in the second panel we show the negative slopes.