Where do I go? Decoding temporal neural dynamics of scene processing and visuospatial memory interactions using CNNs

Clément Naveilhan[1,*], Raphaël Zory[1,2], Stephen Ramanoël[1,3]

[1] Université Côte d'Azur, LAMHESS, Nice, France

[2] Institut Universitaire de France (IUF), Paris, France

[3] Sorbonne Université, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, F-75012 Paris, France

*CA, Corresponding Author: Clément Naveilhan (clement.naveilhan@univ-cotedazur.fr)

Abstract

Visual scene perception enables rapid interpretation of the surrounding environment by integrating multiple visual features related to task demands and context, which is essential for goal-directed behavior. In the present work, we investigated the temporal neural dynamics underlying the interaction between the processing of visual features (*i.e.*, bottom-up processes) and contextual knowledge (*i.e.*, top-down processes) during scene perception. We analyzed EEG data from 30 participants performing scene memory and visuospatial memory tasks in which we manipulated the number of navigational affordances available (*i.e.*, the number of open doors) while controlling for similar low-level visual features across tasks. We used convolutional neural networks (CNN) coupled with gradient-weighted class activation mapping (Grad-CAM) to assess the main channels and time points underlying neural processing for each task. We found that early occipitoparietal activity (50-250 ms post-stimulus) contributed most to the classification of several aspects of visual perception, including scene color, navigational affordances, and spatial memory content. In addition, we showed that the CNN successfully trained to detect affordances during scene perception was unable to detect the same affordances in the spatial memory task after learning, whereas a similarly trained and tested model for detecting wall color was able to generalize across tasks. Taken together, these results reveal an early common window of integration for scene and visuospatial memory information, with a specific and immediate influence of newly acquired spatial knowledge on early neural correlates of scene perception.

Keywords: Visual perception, Scene, Spatial Memory, Temporal Dynamic, EEG, Grad-CAM.

Introduction

Visual scene perception is a critical cognitive function that enables individuals to rapidly encode, interpret, and interact with their environment. This process requires the integration of low-, mid-, and high-level visual information (*e.g.*, object features, spatial configurations, and landmarks; Bartnik & Groen, 2023) to support efficient and adaptive behavior in dynamic settings such as natural environments (Epstein & Baker, 2019). The complexity of visual perception becomes even more pronounced in everyday life when we consider the variety of goals, tasks, contexts, or prior knowledge that modulate scene processing and its neural correlates (Bar, 2009; Bar et al., 2006; Kay et al., 2023; Malcolm et al., 2016; Nau et al., 2024; Ritchie et al., 2024). Therefore, a deeper understanding of the neural dynamics underlying scene perception, especially the interaction between bottom-up (*e.g.*, visual features) and top-down (*e.g.*, prior knowledge) information, is essential for elucidating the mechanisms of visual cognition.

In this context, the processing of navigational affordances available in visual scenes represents a promising theoretical and methodological framework (Bartnik et al., 2024; Djebbara et al., 2019; Gregorians & Spiers, 2022; Naveilhan et al., 2024). Initially proposed to be automatically extracted solely by bottom-up mechanisms during scene perception (Bonner & Epstein, 2017; Harel et al., 2022), recent findings also suggest the influence of contextual information on scene processing at both behavioral and neural levels (Aminoff & Tarr, 2021; Choi et al., 2020; Naveilhan et al., 2024). For example, Naveilhan et al. (2024) reported that increasing the number of navigational affordances (*i.e.*, open doors in a room) selectively decreased participants' accuracy in retrieving the position of a previously learned goal. These behavioral findings were complemented by neural data using event-related potentials (ERPs) recorded from occipitoparietal electrodes spanning scene-selective regions (SSRs), specifically the occipital place area (OPA), which plays a critical role in encoding pathways for movement in scenes (Bonner & Epstein, 2017; Harel et al., 2016, 2022; Kaiser et al., 2020; Kamps et al., 2024). These findings suggest that prior spatial knowledge modulates early neural markers of visual processing, specifically the P2 component (Harel et al., 2016, 2022; Kaiser et al., 2020). However, the restriction of previous analyses to ERPs over occipitoparietal electrodes may have limited the ability to investigate the complex global dynamics of visual information processing involved in visuospatial tasks (Groen et al., 2017; Kravitz et al., 2011).

Recent advances in signal processing and supervised learning with neural networks now make it possible to consider multiple brain regions that support complex dynamics for visual perception. Notably, Orima and Motoyoshi (2023) trained a convolutional neural network (EEGNet; Lawhern et al., 2018) to classify both scene categories and their global properties (*e.g.*, naturalness, openness, roughness) based on EEG data. By applying gradient-weighted class activation mapping (Grad-CAM; Selvaraju et al., 2017), the authors reported that early EEG signals from occipital electrodes were crucial for classifying openness, whereas later signals from frontal electrodes were crucial for determining naturalness and scene categories. These findings suggest that different properties of visual scenes are processed in different brain regions and at different times. Dwivedi et al. (2024) further explored this complex

sequencing of scene perception using representational similarity analysis to compare neural responses across occipital electrodes with computational models of 2D, 3D, semantic features, and navigational affordances. Their results showed that visual features are processed earlier in the temporal sequence (130-170 ms post-stimulus) than navigational affordances (around 300 ms), suggesting hierarchical processing in line with the findings of the Orima group. However, these results also seem to be partially inconsistent with those of Harel et al. (2022), who showed that navigational affordances were processed earlier (around 230 ms). A possible explanation is based on the fact that in previous protocols participants were passively presented with the scene (Bonner & Epstein, 2017; Harel et al., 2022), whereas in Dwivedi et al. (2024) participants had explicitly engaged in navigational affordance processing. These differences in task engagement, contextual information, and the visual features of the stimuli used could influence the temporal dynamics of navigational affordance extraction, as previously reported only at the behavioral level (Naveilhan et al., 2024). However, the temporal dynamics of the neural correlates underlying the interaction between scene information processing and prior contextual information remain elusive and represent a major challenge for current research in visual cognition (Nau et al., 2024; Ritchie et al., 2024).

To address this issue, we performed novel analyses using a convolutional neural network (CNN) on a previously acquired dataset, obtained during scene perception and visual spatial memory tasks in which we manipulated the number of available affordances while controlling for similar low-level visual features across tasks (Naveilhan et al., 2024). We included all 64 available electrodes to investigate the global dynamics of the expected interaction between task-relevant top-down (*i.e.*, prior spatial knowledge) and bottom-up information processing (*i.e.*, visual features). The aim of the present study is to provide new insights into the temporal neural dynamics underlying task-specific visuospatial representations by leveraging the complexity of CNNs to disentangle the intricate structure of EEG data. Specifically, based on previous work on visual object processing for action (Enge et al., 2023), we hypothesize that top-down information modulates the early neural signature of scene perception, including neural correlates of visual scene features. In this sense, we expect that processing of prior spatial knowledge will modulate early occipitoparietal activity (around 200 ms), allowing the CNN to accurately decode the task participants are performing. We also propose that this early time window contains specific information about the previously learned goal position. Furthermore, we hypothesize that learned spatial information will modulate the early neural signature of affordance processing despite identical visual features. In other words, a CNN trained to decode the navigational affordances available when participants have no prior spatial knowledge will not be able to decode the affordance information after participants have learned the goal position, even when the visual scenes are the same.

Method

Participants

In the present study, we conducted a reanalysis of a dataset from a previously published study (Naveilhan et al., 2024). Electroencephalographic (EEG) data were collected from 30 young adults (mean age = 24.31 years, SD = 0.65, range = 19-31; 16 females). Sample size was determined a priori using G*Power (version 3.1.9.7; Faul et al., 2007) based on a previous EEG study of scene perception (Naveilhan et al., 2023), which indicated that a minimum of 23 participants was required to achieve statistical power of 0.95 at an alpha level of 0.05. The study was approved by the local ethics committee (CERNI-UCA opinion no. 2021-050), and all participants gave informed consent prior to participation. They were all right-handed, had normal or corrected-to-normal vision, and had no history of neurological or cognitive disorders.

Stimuli and procedure

The experiment used visual stimuli developed with Unity Engine (v2019.2.0.0f1) and presented on an iiyama ProLite B2791HSU monitor (1920×1080 resolution, 30-83 Hz) positioned 60 cm from the participants. Stimulus presentation was controlled by PsychoPy (v2022.13) running on a Dell Precision 7560 workstation with an Intel Xeon W-11955 processor. The stimuli consisted of images of simple rectangular rooms, each containing either a door or a gray rectangle on three visible walls. These designs were adapted from previous studies (Bonner & Epstein, 2017; Harel et al., 2022). To avoid potential associations between door locations and wall colors, seven different door configurations and seven wall color variations were used, resulting in 49 unique stimuli. All stimuli are publicly available in the OSF repository, and a subset is presented in **Figure 1**.

Participants completed two tasks: a scene memory task and a spatial memory task, each organized into six randomized blocks. Within each block, the order of stimulus presentation was pseudorandomized to maintain participant engagement. In the scene memory task, participants performed a 1-back task in which they identified whether the current scene matched the previous scene based on wall color only, pressing a designated key upon detection of a match. This task consisted of 64 images with four door configurations presented in two different environments, each repeated eight times. The spatial memory task required participants to first passively navigate through rooms to memorize the path to a hidden goal that could be located to the left, center, or right. Each wall color corresponded to a specific goal direction. Participants were then presented with the images and had to indicate the direction of the goal by pressing a key on a USB keyboard as quickly and accurately as possible. Although the same images were used as in the scene memory task, participants now had access to spatial contextual information because they knew that a goal was hidden behind one of the doors. The spatial memory task consistently followed the scene memory task within each block to prevent prior knowledge from influencing performance during the scene memory phase.

For each task, images were displayed for 1 second, followed by a fixation cross lasting between 0.5 and 1 second, and auditory feedback for responses. Over the course of the experiment, participants viewed a total of 1,536 images (768 per task), and each session lasted approximately 55 minutes.

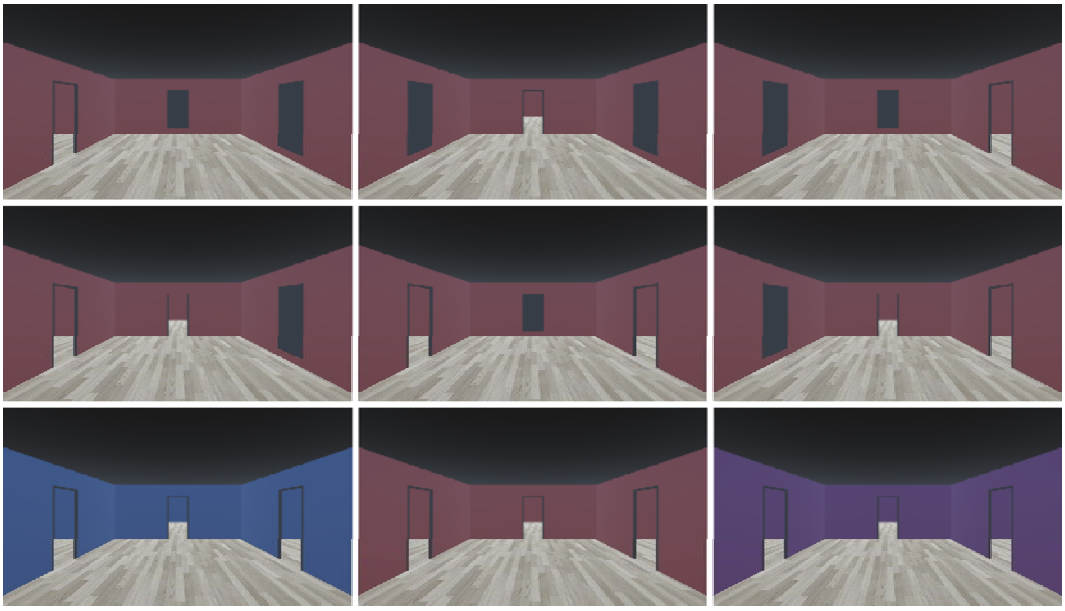


Figure 1. Presentation of a subset of the stimuli used, showing the 7 possible affordance configurations and 3 of the 8 wall colors included in the protocol. The complete set of stimuli is available in the OSF repository.

EEG preprocessing

EEG data were acquired at a sampling rate of 500 Hz using a 64-channel Waveguard™ cap with active wet electrodes, connected to an eego™ mylab amplifier (ANT Neuro) and digitized at 24-bit resolution. The reference electrode was positioned at CPz, and the ground electrode at AFz. Electrode impedances were maintained below 10 kΩ, with the majority under 5 kΩ. EEG recordings were synchronized with stimulus presentations using LabRecorder in the Lab Streaming Layer (Kothe et al., 2024).

Since preprocessing steps have been proposed to play an important role in the decoding capabilities of EEGNet (Kessler et al., 2024), we opted for an open source and fully reproducible pipeline implemented in MATLAB (R2024a), namely, the BeMobil pipeline (Klug et al., 2022) using EEGLAB v2024.2 (Delorme & Makeig, 2004). Non-experimental segments and highly artifacted portions of the data were manually excluded and the signals were downsampled to 250 Hz. Line noise was automatically detected and removed using Zapline Plus (Klug & Kloosterman, 2022). Bad electrodes were detected based on a correlation threshold of 0.8 and a maximum downtime of 0.6, and then interpolated using spherical interpolation methods, with an average of 3.10 ± 2.58 electrodes removed per subject. The data were referenced to the common average. Additional artifact cleaning was performed in the time domain using Artifact Subspace Reconstruction (ASR) with a cutoff

threshold of 20 (Chang et al., 2020), resulting in the removal of an average of 7.51 ± 3.52% of data points.

Subsequently, the data were high-pass filtered at 1.5 Hz and AMICA decomposition was applied with 10 rejection iterations and a sigma threshold of 3 (Klug et al., 2024; Klug & Gramann, 2021). Dipole fitting was performed using DipFit, and independent components were classified using ICLabel with default parameters (Pion-Tonachini et al., 2019). Components identified as muscle, line noise, eye, or heart artifacts were excluded, and only those classified as brain or other with residual variance less than 15% were retained (Delorme et al., 2012). The data were then band-pass filtered between 0.3 and 50 Hz, and epochs from 1 second before to 2 seconds after stimulus onset were extracted. Epochs containing artifacts larger than 100 µV were excluded, resulting in an average of 1,461 ± 90.02 epochs per subject, balanced across conditions.

EEGNet and Grad-CAM procedure

Data were exported from MATLAB to Python for subsequent analyses using an EEG-based classification framework implemented with PyTorch (Paszke et al., 2019). Analyses were performed on an NVIDIA RTX 4090 GPU with 24 GB of VRAM using CUDA 12.6. Each dataset was normalized and balanced to ensure equal class representation. Our methodology relied on the EEGNet architecture (Lawhern et al., 2018), which incorporates convolutional layers, batch normalization (Ioffe & Szegedy, 2015), exponential linear unit (ELU) activation functions, and optimized dropout layers (Srivastava et al., 2014) to mitigate overfitting. By capturing both spatial patterns across electrodes and temporal dynamics within single trials, EEGNet is well suited to handle the complexity of the present dataset while maintaining computational efficiency for single trial classification. Optuna's hyperband pruner (Akiba et al., 2019; Li et al., 2018), which uses a tree-structured Parzen estimator, was applied for optimal parameter selection, including learning rate and dropout rate, with early stopping based on validation accuracy. To address class imbalances and improve model generalization, class-weighted cross-entropy loss was used (Cui et al., 2019). After training the model on 80% of the full pooled dataset with early stopping, performance was evaluated using 5-fold cross-validation for each individual subject. Classification accuracy was then compared to chance levels using a binomial cumulative distribution approach. For each analysis, a binomial distribution was computed using MATLAB's binoinv function, taking into account the predefined significance level ($\alpha = 0.05$), the number of predictions (the average number of classifications made per participant), and the number of classes (2 for task predictions, 7 for affordances, and 8 for wall color). The binomial method was used to determine statistical significance, as it is comparable in reliability to permutation testing for datasets with more than 100 trials, but does not have the extensive computational requirements of permutation testing (Combrisson & Jerbi, 2015; Vahid et al., 2020). We then used 100,000 iterations of bootstrap resampling to calculate the mean difference between the model accuracy and a random sample drawn from the generated binomial distribution. From the bootstrap distribution of the mean differences between accuracy of the model and chance, we derived a 95% confidence interval and a *p-value*, which are presented in the Results section. Finally, Grad-CAM (Selvaraju et al., 2017) was used to visualize the contributions to classification by

focusing on activations from the second convolutional block (Orima & Motoyoshi, 2023). Specifically, gradients of predicted values were computed with respect to feature map activations, resulting in two-dimensional localization maps (electrodes and time points). These maps were processed through ReLU layers, normalized, and averaged across participants to identify the key EEG channels and time points that most significantly contributed to the classifications, as illustrated in the Results section. To ensure the robustness and generalizability of our findings, all analyses were repeated five times, and the median results were reported. Detailed information about the different models architectures and example code are available in the study's OSF repository. We decided to chose hyperparameter selection using Bayesian optimization and HyperBand (Falkner et al., 2018), which addresses the main limitation of hyperparameter selection highlighted previously (Orima & Motoyoshi, 2023), but still relies on some hyperparameters that may introduce small inconsistencies in test-retest reliability. However, for each analysis, hyperparameter selection, training, and testing were performed five times, and the accuracy of the model for all analyses was consistent across the five repetitions.

Results

Early EEG signal is modulated by visuospatial memory information

In the first analysis, we merged the entire dataset, performed hyperparameter optimization and then trained the model on the balanced data to classify the tasks participants were performing (*i.e.*, scene memory or spatial memory). For each subject, we then performed 5-fold cross-validation and used Grad-CAM on the second layer of the trained CNN (**Figure 2**) to identify the time points and electrodes that contributed most to classifying the task participants were performing (**Figure 2**).

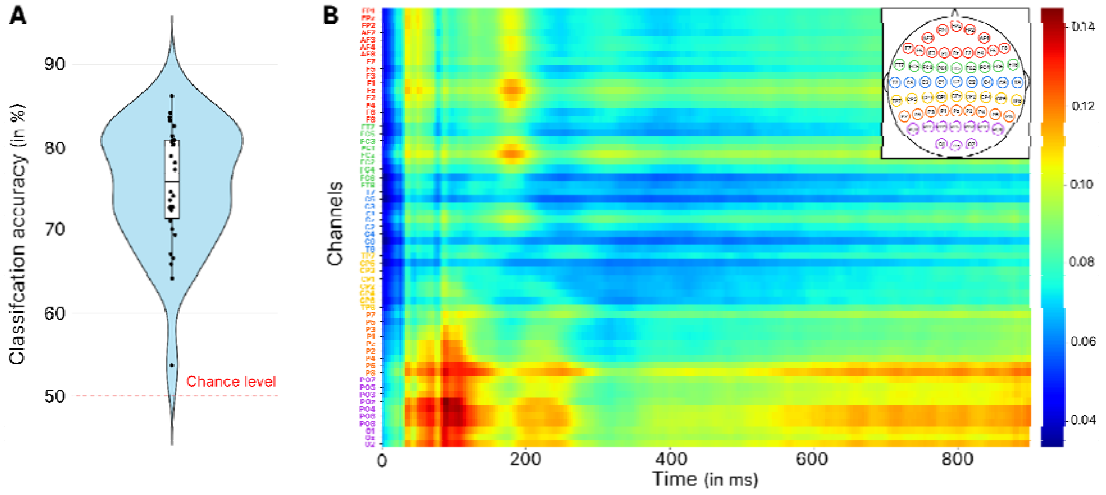


Figure 2. Results of the EEGNet classification and Grad-CAM analysis, for the CNN trained on the entire merged dataset to classify the task being performed. **A.** Violin plot showing the accuracy of the model. **B.** Grad-CAM results highlighting the time-channel points in the second CNN layer that contributed most to the classification. The most contributive points are in red, and the colorbar represents arbitrary units. The channels are arranged sequentially from frontal to occipital regions.

Looking at the accuracy of the model (**Figure 2.A**), we find an average classification accuracy of 75.39 ± 7.28%. Bootstrapping analysis using the binomial distribution showed that the accuracy of the model was better than the 50% chance level (95% confidence interval of the mean difference from chance: CI [19.40, 31.45], $p < 0.0001$). This result suggests that learning contextual information modified the neural correlates of scene perception allowing the network to classify the task. In order to identify which time points and EEG channels contributed most, we then looked at the Grad-CAM results (**Figure 2.B**). These results highlight that the activity of a cluster of occipitoparietal electrodes (P8/POz/O1/O2/P6/PO4/PO6/PO7/PO8/Oz) between 50 and 180ms after stimulus presentation contributed the most to task classification, a result that may reflect differences in visual stimuli processing. The same electrodes also seemed to be involved in a second section of data between 200 and 250ms. Finally, later activity appeared after 600ms, which also enabled the classification of the task, reflecting the onset of motor activity to produce the response.

In a second step, we extracted the EEG activity of participants involved in the spatial memory task (*i.e.*, when they had to retrieve the target). We selected only trials in which participants were presented with 2 open doors and in which they successfully retrieved the position of the goal (93.47 ± 0.58% of trials per participant). This ensured that the network had sufficient trials, and used only the EEG activity associated with goal direction information as support for the classification.

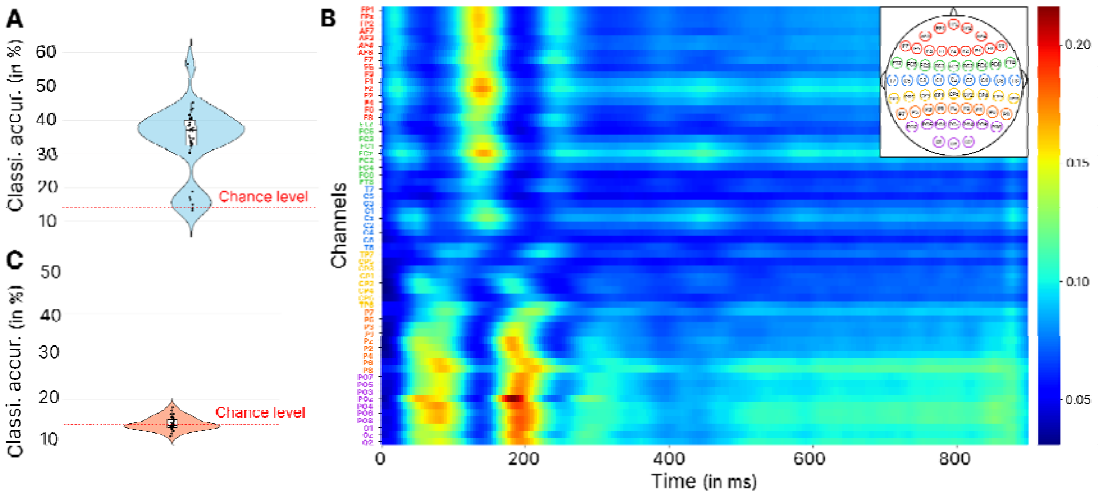


Figure 3. Outcomes of the EEGNet classification and Grad-CAM analysis, for the CNN trained on the dataset with the 'two affordances' conditions to classify the goal position. **A.** Violin plot displaying the model's accuracy. **B.** Grad-CAM results indicating the time-channel points in the second CNN layer, with those making the highest contribution to the classification presented in red.

The accuracy of the model (**Figure 3.A**) indicated that on average the network classified 45.41 ± 5.37 % of the tested trials correctly, performing better than the 33% expected by chance (95% CI: [3.74, 13.63], $p = 0.0004$). The Grad-CAM results (**Figure 3.B**) highlighted activity in frontal electrodes between 180 and 200ms as well as activity in occipitoparietal electrodes between 200 and 220ms as the most contributive to the goal position identification.

These results suggest that early EEG activity also contains information regarding the position of the goal, irrespective of the visual features of the scene. As a control, we also performed the same analysis to detect the position of the goal but in the condition with three affordances and similarly found that the model accurately classified the position of the goal (Mean classification accuracy = 48.67 ± 6.54 %; 95% CI: [6.95, 16.89], $p < 0.0001$).

Prior spatial knowledge modulates early neural activity related to affordance processing

Here, we considered the EEG activity of participants performing the scene memory task and trained the model to detect the affordances within the scene. We also tested the model on data from the spatial memory task to assess whether the information supporting decoding was modulated by learned spatial contextual information. (**Figure 4**).

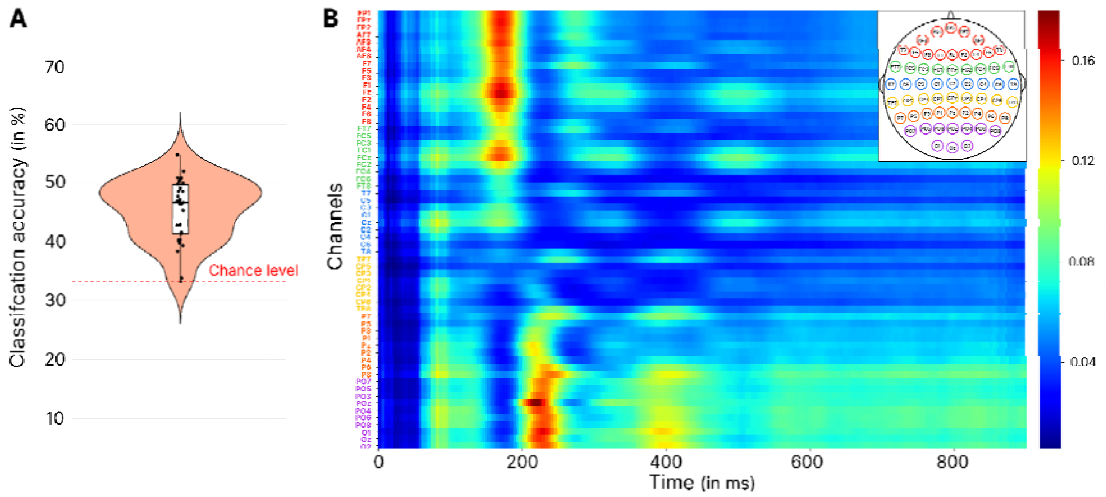


Figure 4. Results of the EEGNet classification and Grad-CAM analysis, for the CNN trained on the data of participants performing the scene memory task to detect the position of the affordances. **A.** Violin plot showing the accuracy of the model during the scene memory task. **B.** Grad-CAM results highlighting the time-channel points in the second CNN layer that contribute most to the classification. The most contributive points are in red, and the colorbar corresponds to arbitrary units. **C.** Violin plot of the accuracy for the model tested on the data of participants performing the spatial memory task.

The model tested on the scene memory task (**Figure 4.A**) showed an accuracy of 35.80 ± 10.45%, and bootstrap testing indicated that it performed better than the 14.29% chance level (95% CI: [10.88, 23.68], $p < 0.0001$). However, when we tested the model on EEG data from the spatial memory task (**Figure 4.C**), the classification accuracy dropped to 13.59 ± 1.78% and was no longer higher than chance level (95% CI: [-7.41, 3.11], $p = 0.77$). These results suggest that the information from neural activity enabling the classification of the position of affordances is modulated by spatial learning. The results of the Grad-CAM analysis (**Figure 4.B**) indicated that the most contributive time points for the classification were once again located in the occipitoparietal region around 200 ms, as previously suggested by modulation of the P2 component in ERP analyses (Harel et al., 2022). Interestingly, earlier occipitoparietal activity (50–100 ms) and frontal activity (150–180 ms) were also involved to

a lesser extent in the classification, replicating previous findings on the neural correlates of low-visual feature processing such as the openness of visual scenes (Greene & Oliva, 2009; Hansen et al., 2018; Orima & Motoyoshi, 2023).

Finally, to test the possibility that our results could simply be explained by task differences, we conducted a supplementary analysis using a CNN trained to classify the color of the wall, another scene feature that was present in both tasks. Similar to the previous affordance processing analysis we trained a CNN on EEG data from participants performing the scene memory task and then tested it on unseen data from the same task as well as on data from the spatial memory task to assess the generalization of the information supporting classification (**Figure 5**). For the scene memory task, the CNN achieved an average accuracy of 34.34 ± 11.18%, significantly better than the 12.5% chance level (95% CI: [13.75, 25.87], $p < 0.0001$). However, when tested on data from the spatial memory task, the accuracy was on average 18.75 ± 2.09%, which was statistically higher than chance level (95% CI: [0.31, 9.65], $p = 0.019$), suggesting that the EEG information allowing the classification of wall color generalizes to some extent across tasks.

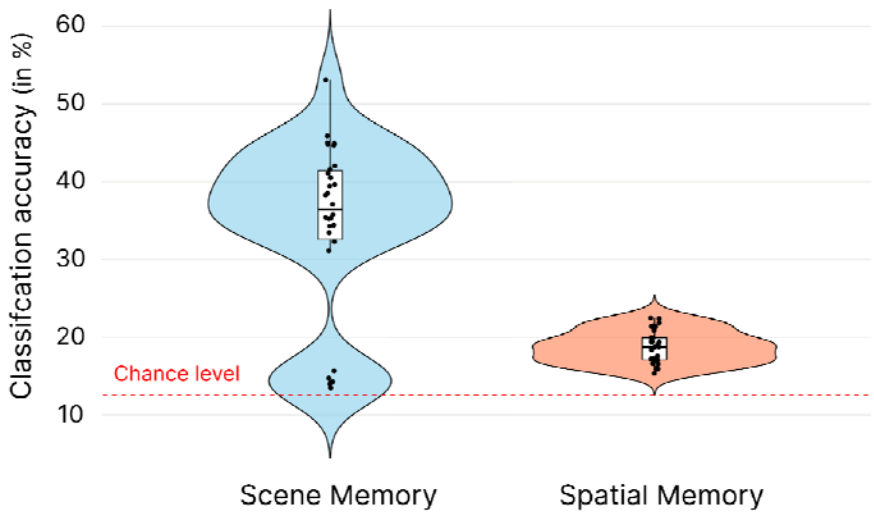


Figure 5. Accuracy of the CNN, trained on the data from scene memory task to detect the color of the wall (chance level at 12.5%), in the scene memory or the spatial memory task.

Discussion

In this study, we investigated the temporal neural dynamics underlying visual perception during scene processing and its interaction with visuospatial memory information. Our results demonstrate that a CNN trained on EEG data can accurately classify several aspects of visual perception, including the color of the wall within the scene, navigational affordances, and the location of a previously learned target, primarily based on early occipitoparietal brain activity (between 100-230 ms). To further support the potential influence of top-down information

suggested by this common time window of integration, we showed that a model trained on EEG data to decode affordance-related information in a scene memory task failed to generalize to the task in which participants had learned the position of a goal. In contrast, another model trained to detect wall color performed significantly above chance in the same context. Taken together, these results highlight the strong influence of top-down information related to prior spatial knowledge on early occipitoparietal neural activity during scene perception.

Navigational affordances and visuospatial memory processing share a common early integration time window

The Grad-CAM analyses highlighted occipitoparietal activity around 200 ms as the most important contributor to the classification of navigational affordances within the scene, consistent with previous findings by Harel et al. (2022). In their study, the authors also used a multivariate approach and suggested that the location of affordances may be represented even earlier, around 130 ms. In the present study, we did not distinguish between the position and number of affordances but found that activity around both 100 ms and 200 ms contributed to the classification of affordances, with the latter time window playing a more important role. This early extraction of the position of available pathways for movement may reflect the brain's ability to rapidly construct high-level representations of visual scenes based on categorical distinctions and global spatial properties (Orima & Motoyoshi, 2021, 2023). In support of this point, studies combining MEG measurements with deep neural network analyses have demonstrated a hierarchical progression of scene representations, integrating information from low-level features to higher-order scene features within 200 ms (Cichy et al., 2017). These findings suggest that the processing of affordances from visual scenes occurs as early as other perceptual processes, bridging the gap between scene perception (*e.g.*, encoding the number of affordances) and action (Djebbara et al., 2019; Harel et al., 2022).

Interestingly, our results suggest that higher-level information, such as spatial contextual information about the target location, is also represented in this early occipitoparietal activity. This common integration time window is consistent with several lines of evidence for top-down influences on early visual processing during scene and object perception (Enge et al., 2023; Klink et al., 2023; Steel et al., 2021, 2023, 2024). For example, Enge et al. (2023) demonstrated that prior knowledge of an object's function shapes early cortical markers at around 200 ms and influences higher-order visual perception (Greene & Hansen, 2020; Groen et al., 2018). Similarly, Klink et al. (2023) reported that scene familiarity modulates early EEG activity, suggesting a role for learned scene context in shaping early visual processing. The current results support this interpretation by highlighting the critical role of visuospatial memory-derived information in shaping early visual markers of scene processing. Furthermore, these results provide novel insights into a common early time window for integrating both bottom-up and top-down information, emphasizing the dynamic interplay between scene perception and prior visuospatial knowledge.

Visuospatial memory immediately influences the early neural activity of affordance extraction

To investigate the influence of top-down modulation on early neural correlates of scene perception, we examined how prior learning of the target location affects the classification abilities of the CNN in decoding affordance information. To test this, we trained a model that accurately decoded affordances in the scene memory task, confirming that the extraction of navigational affordances is represented in early occipitoparietal activity (Dwivedi et al., 2024; Harel et al., 2022). However, this model failed to detect affordances after participants learned the path to the goal in the spatial memory task, even though the presented images were exactly the same. This suggests that the neural correlates of navigational affordances, including their number and location, are modulated by the learning of spatial contextual information about these affordances. Thus, top-down information related to prior spatial knowledge may interact with the automatic encoding of navigational affordances. This finding complements previous behavioral findings by Naveilhan et al. (2024), which showed that increasing the number of affordances decreased participants' accuracy in the spatial memory task only. Critically, this finding does not simply reflect task differences, as highlighted by a control analysis showing significant generalization between scene memory and spatial memory tasks of a model similarly trained to decode wall colors. This lack of modulation observed for wall color features is consistent with previous findings by Hansen et al. (2018), which showed that processing of low-level visual information is rapid and minimally affected by observer-based goals. Thus, our results suggest a targeted modulation of neural activity tied to scene-specific information influenced by learning, rather than a broader, nonspecific modulation.

These findings on the interaction between scene feature integration and prior spatial information during navigational affordance processing illustrate the dynamic and complex nature of visual scene perception. From a broader perspective, this is consistent with the predictive coding framework for efficient perception (Peelen et al., 2024; Rao & Ballard, 1999) and the ecological basis of affordances (Gibson, 1979). This notion emphasizes the intrinsic connection between action and perception, highlighting the role of perception in guiding action, and is consistent with the notion of active inference, which views action and perception as integrated processes that work to minimize prediction error (Friston, 2010; Friston et al., 2017; Kaplan & Friston, 2018; Pezzulo & Cisek, 2016). Consistent with this theoretical framework, our results suggest that spatial memory content modulates the early neural correlates of navigational affordance processing to support efficient, goal-directed visual information processing. Finally, it is important to consider an alternative explanation for the modulation of affordance processing that arises from the findings of Castelhano & Witherspoon (2016). They demonstrated that knowledge of an object's function significantly increases search efficiency by directing attention to functionally relevant areas within a scene (Bouwkamp et al., 2024). In our paradigm, participants may have explored the visual scene differently or considered other navigational options (*i.e.*, open doors that did not lead to the goal). Future studies using eye-tracking or specific experimental manipulations, such as modifying the eccentricity of navigational affordances, could help disentangle the possibilities and shed light on this potential confounding factor.

Conclusion

In conclusion, this work supports the notion that during scene perception neural processes associated with behaviorally relevant tasks, such as retrieving the position of a goal hidden behind a door, share a common time window of integration—approximately 100-250 ms over occipitoparietal regions—with processes representing visual information, such as navigational affordances. The present results also demonstrate an immediate influence of the newly learned spatial knowledge on the early neural activity associated with scene processing, modulating even the first stages of visual scene perception. Thus, knowing where to go may shape what you see.

Data availability

The complete set of data, stimuli generated and code used for preprocessing and CNN is available on the OSF repository : https://osf.io/5cqxy/?view_only=2957dac1a1304c2db6e1fb3b056c5008.

Author Contributions

Clément Naveilhan: Conceptualization; Formal analysis; Investigation; Methodology; Writing—Original draft; Writing—Review & editing. **Raphaël Zory:** Funding acquisition; Writing—Review & editing. **Stephen Ramanoël**: Conceptualization; Funding Acquisition; Methodology; Project administration; Supervision, Writing—Review & editing.

Acknowledgements

This research was made possible by the generous participation of volunteer participants, to whom the authors are sincerely grateful. We also thank Catherine Buchanan for her careful reading of the manuscript and her feedback.

This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01 and, in particular, by the interdisciplinary Institute for Modeling in Neuroscience and Cognition (NeuroMod) of Université Côte d'Azur.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna□: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623□2631. https://doi.org/10.1145/3292500.3330701

Aminoff, E. M., & Tarr, M. J. (2021). Functional Context Affects Scene Processing. *Journal of Cognitive Neuroscience*, *33*(5), 933□945. https://doi.org/10.1162/jocn_a_01694

Bar, M. (2009). The proactive brain□: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1235□1243. https://doi.org/10.1098/rstb.2008.0310

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449□454. https://doi.org/10.1073/pnas.0507062103

Bartnik, C. G., & 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*. https://doi.org/10.1093/acrefore/9780190264086.013.437

Bartnik, C. G., Vukšić, N., Bommer, S., & Groen, I. I. A. (2024). Unveiling task-dependent action affordance representations□: Insights from scene-selective cortex and deep neural networks. *Journal of Vision*, *24*(10), 897. https://doi.org/10.1167/jov.24.10.897

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *114*, 4793□4798. https://doi.org/10.1073/pnas.1618228114

Bouwkamp, F. G., de Lange, F. P., & Spaak, E. (2024). Spatial Predictive Context Speeds Up Visual Search by Biasing Local Attentional Competition. *Journal of Cognitive Neuroscience*, 1□15. https://doi.org/10.1162/jocn_a_02254

Castelhano, M. S., & Witherspoon, R. L. (2016). How You Use It Matters□: Object Function Guides Attention During Visual Search in Scenes. *Psychological Science*, *27*(5), 606□621. https://doi.org/10.1177/0956797616629130

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2020). Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-Channel EEG Recordings. *IEEE Transactions on Biomedical Engineering*, *67*(4), 1114□1121. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2019.2930186

Choi, B., McCloskey, M., & Park, S. (2020). Representing navigational affordance based on high-level knowledge of scenes. *Journal of Vision*, *20*(11), 646. https://doi.org/10.1167/jov.20.11.646

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346□358. https://doi.org/10.1016/j.neuroimage.2016.03.063

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance□: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, *250*, 126□136. https://doi.org/10.1016/j.jneumeth.2015.01.010

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). *Class-Balanced Loss Based on Effective Number of Samples*. 9268□9277. https://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html

Delorme, A., & Makeig, S. (2004). EEGLAB□: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9□21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., & Makeig, S. (2012). Independent EEG Sources Are Dipolar. *PLOS ONE*, *7*(2), e30135. https://doi.org/10.1371/journal.pone.0030135

Djebbara, Z., Fich, L. B., Petrini, L., & Gramann, K. (2019). Sensorimotor brain dynamics reflect architectural affordances. *Proceedings of the National Academy of Sciences*, *116*(29), 14769□14778. https://doi.org/10.1073/pnas.1900648116

Dwivedi, K., Sadiya, S., Balode, M. P., Roig, G., & Cichy, R. M. (2024). Visual features are processed before navigational affordances in the human brain. *Scientific Reports*, *14*, 5573. https://doi.org/10.1038/s41598-024-55652-y

Enge, A., Süß, F., & Rahman, R. A. (2023). Instant Effects of Semantic Information on Visual Perception. *Journal of Neuroscience*, *43*(26), 4896–4906. https://doi.org/10.1523/JNEUROSCI.2038-22.2023

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*, 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

Falkner, S., Klein, A., & Hutter, F. (2018). BOHB: Robust and Efficient Hyperparameter Optimization at Scale. *Proceedings of the 35th International Conference on Machine Learning*, 1437–1446. https://proceedings.mlr.press/v80/falkner18a.html

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), Article 2. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press. https://doi.org/10.4324/9781315740218

Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. *Journal of Neuroscience*, *40*(27), 5283–5299. https://doi.org/10.1523/JNEUROSCI.2088-19.2020

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Gregorians, L., & Spiers, H. J. (2022). Affordances for Spatial Navigation. In Z. Djebbara (Éd.), *Affordances in Everyday Life: A Multidisciplinary Collection of Essays* (p. 99–112). Springer International Publishing. https://doi.org/10.1007/978-3-031-08629-8_10

Groen, I. I. A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLOS Computational Biology*, *14*(12), e1006690. https://doi.org/10.1371/journal.pcbi.1006690

Groen, I. I. A., Silson, E. H., & 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. https://doi.org/10.1098/rstb.2016.0102

Hansen, N. E., Noesen, B. T., Nador, J. D., & Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, *114*, 168–180. https://doi.org/10.1016/j.neuropsychologia.2018.04.040

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multi-faceted EEG investigation. *eNeuro*. https://doi.org/10.1523/ENEURO.0139-16.2016

Harel, A., Nador, J. D., Bonner, M. F., & Epstein, R. A. (2022). Early Electrophysiological Markers of Navigational Affordances in Scenes. *Journal of Cognitive Neuroscience*, *34*(3), 397–410. https://doi.org/10.1162/jocn_a_01810

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (arXiv:1502.03167). arXiv. http://arxiv.org/abs/1502.03167

Kaiser, D., Häberle, G., & Cichy, R. M. (2020). Cortical sensitivity to natural scene structure. *Human Brain Mapping*, *41*(5), 1286–1295. https://doi.org/10.1002/hbm.24875

Kamps, F. S., Chen, E. M., Kanwisher, N., & Saxe, R. (2024). Representation of navigational affordances and ego-motion in the occipital place area. *Imaging Neuroscience*. https://doi.org/10.1162/imag_a_00424

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323–343. https://doi.org/10.1007/s00422-018-0753-2

Kay, K., Bonnen, K., Denison, R. N., Arcaro, M. J., & Barack, D. L. (2023). Tasks and their role in visual neuroscience. *Neuron*, *111*(11), 1697–1713. https://doi.org/10.1016/j.neuron.2023.03.022

Kessler, R., Enge, A., & Skeide, M. A. (2024). *How EEG preprocessing shapes decoding performance* (arXiv:2410.14453). arXiv. https://doi.org/10.48550/arXiv.2410.14453

Klink, H., Kaiser, D., Stecher, R., Ambrus, G. G., & Kovács, G. (2023). Your place or mine? The neural dynamics of personally familiar scene recognition suggests category independent familiarity encoding. *Cerebral Cortex*, *33*(24), 11634–11645. https://doi.org/10.1093/cercor/bhad397

Klug, M., Berg, T., & Gramann, K. (2024). Optimizing EEG ICA decomposition with data cleaning in stationary and mobile experiments. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-64919-3

Klug, M., & Gramann, K. (2021). Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience*, *54*(12), 8406–8420. https://doi.org/10.1111/ejn.14992

Klug, M., Jeung, S., Wunderlich, A., Gehrke, L., Protzak, J., Djebbara, Z., Argubi-Wollesen, A., Wollesen, B., & Gramann, K. (2022). *The BeMoBIL Pipeline for automated analyses of multimodal mobile brain and body imaging data* (p. 2022.09.29.510051). bioRxiv. https://doi.org/10.1101/2022.09.29.510051

Klug, M., & Kloosterman, N. A. (2022). Zapline-plus—: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG. *Human Brain Mapping*, *43*(9), 2743–2758. https://doi.org/10.1002/hbm.25832

Kothe, C., Shirazi, S., Stenner, T., Medine, D., Boulay, C., Crivich, M., Mullen, T., Delorme, A., & Makeig, S. (2024). *The Lab Streaming Layer for Synchronized Multimodal Recording*. https://doi.org/10.1101/2024.02.13.580071

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), Article 4. https://doi.org/10.1038/nrn3008

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet—: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband—: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, *18*, 1–52.

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856. https://doi.org/10.1016/j.tics.2016.09.003

Nau, M., Schmid, A. C., Kaplan, S. M., Baker, C. I., & Kravitz, D. J. (2024). Centering cognitive neuroscience on task demands and generalization. *Nature Neuroscience*, *27*(9), 1656–1667. https://doi.org/10.1038/s41593-024-01711-6

Naveilhan, C., Delaux, A., Durteste, M., Lebrun, J., Raphaël, Z., Arleo, A., & Ramanoel, S. (2023). *Age-related differences in electrophysiological correlates of visuospatial reorientation*. https://doi.org/10.1101/2023.11.22.568209

Naveilhan, C., Saulay-Carret, M., Zory, R., & Ramanoël, S. (2024). Spatial Contextual Information Modulates Affordance Processing and Early Electrophysiological Markers of Scene Perception. *Journal of Cognitive Neuroscience*, *36*(10), 2084–2099. https://doi.org/10.1162/jocn_a_02223

Orima, T., & Motoyoshi, I. (2021). Analysis and Synthesis of Natural Texture Perception From Visual Evoked Potentials. *Frontiers in Neuroscience*, *15*, 698940. https://doi.org/10.3389/fnins.2021.698940

Orima, T., & Motoyoshi, I. (2023). Spatiotemporal cortical dynamics for visual scene processing as revealed by EEG decoding. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1167719

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch—: An Imperative

Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), Article 1. https://doi.org/10.1038/s44159-023-00254-0

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape□: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Sciences*, *20*(6), 414□424. https://doi.org/10.1016/j.tics.2016.03.013

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel□: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181□197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex□: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79□87. https://doi.org/10.1038/4580

Ritchie, J., Wardle, S., Vaziri-Pashkam, M., Kravitz, D., & Baker, C. (2024). *Rethinking category-selectivity in human visual cortex*. https://doi.org/10.48550/arXiv.2411.08251

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM□: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618□626. https://doi.org/10.1109/ICCV.2017.74

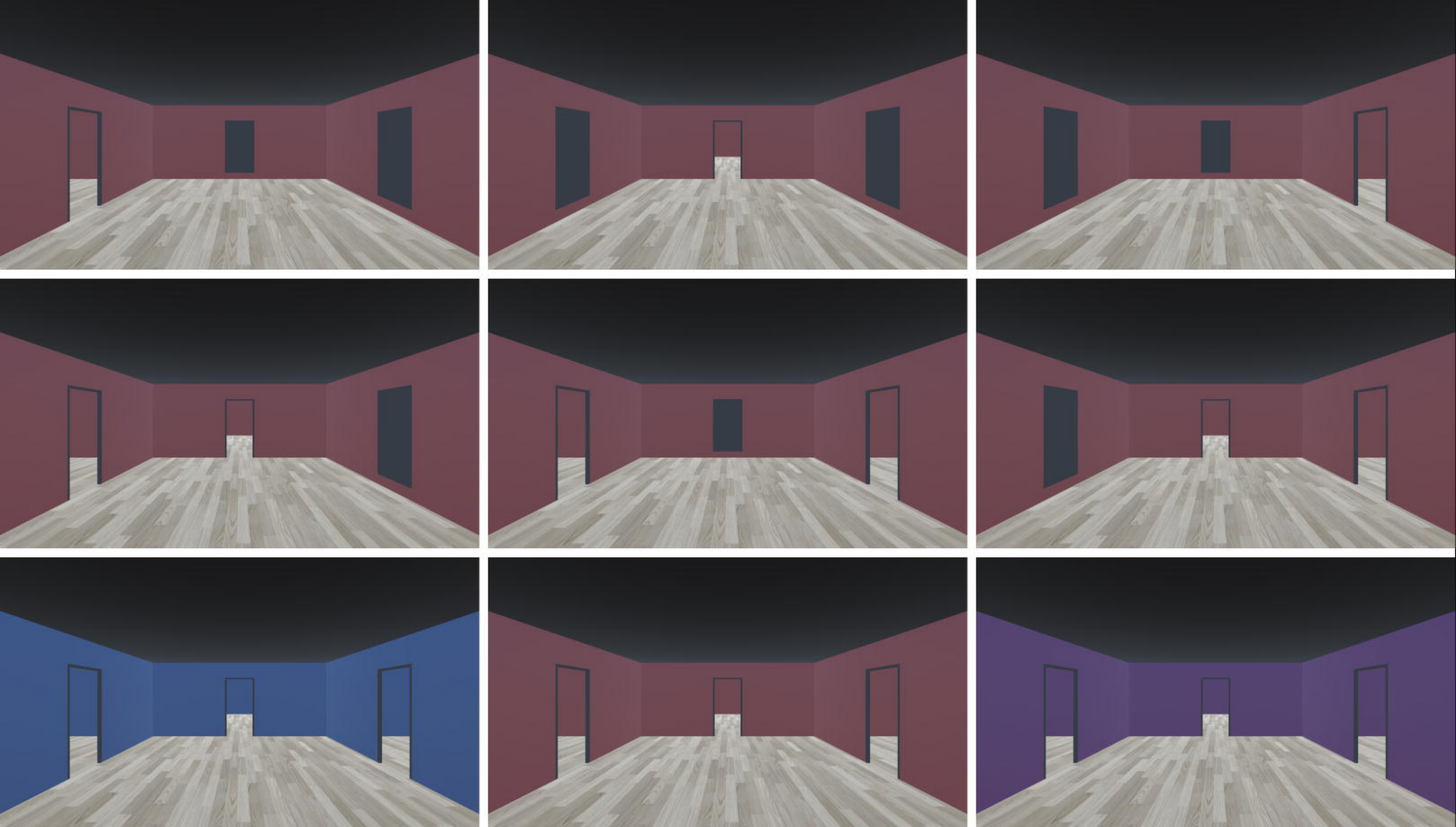
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout□: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929□1958.

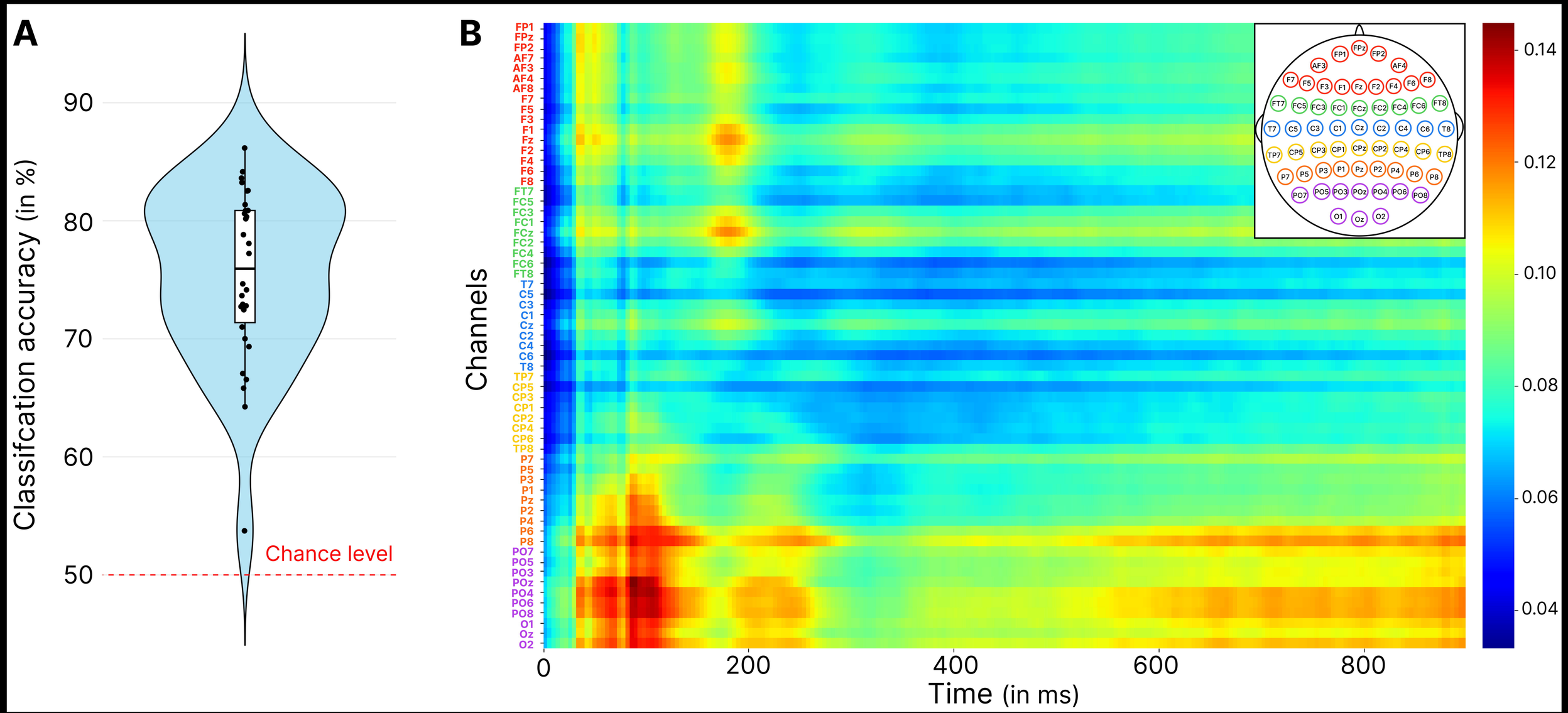
Steel, A., Billings, M. M., Silson, E. H., & Robertson, C. E. (2021). A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-22848-z

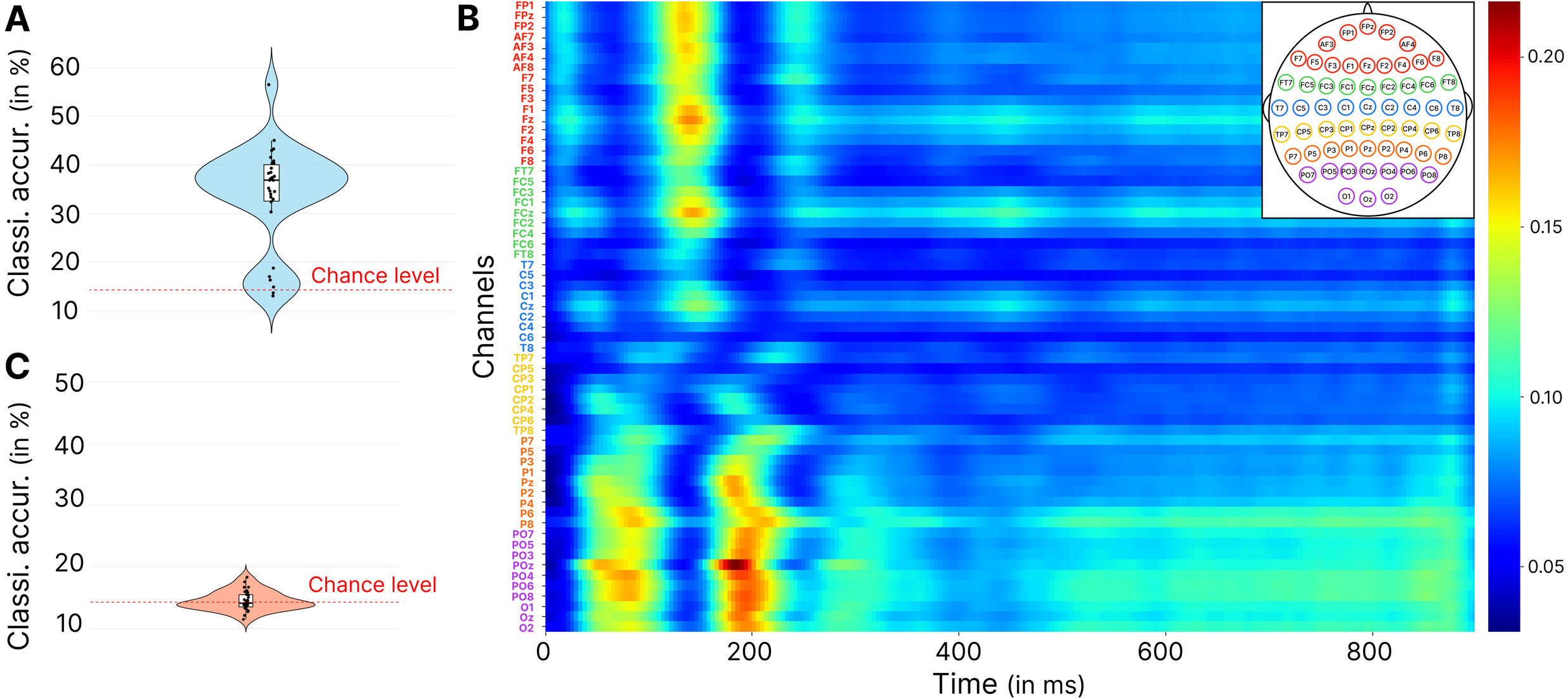
Steel, A., Garcia, B. D., Goyal, K., Mynick, A., & Robertson, C. E. (2023). Scene Perception and Visuospatial Memory Converge at the Anterior Edge of Visually Responsive Cortex. *Journal of Neuroscience*, *43*(31), 5723□5737. https://doi.org/10.1523/JNEUROSCI.2043-22.2023

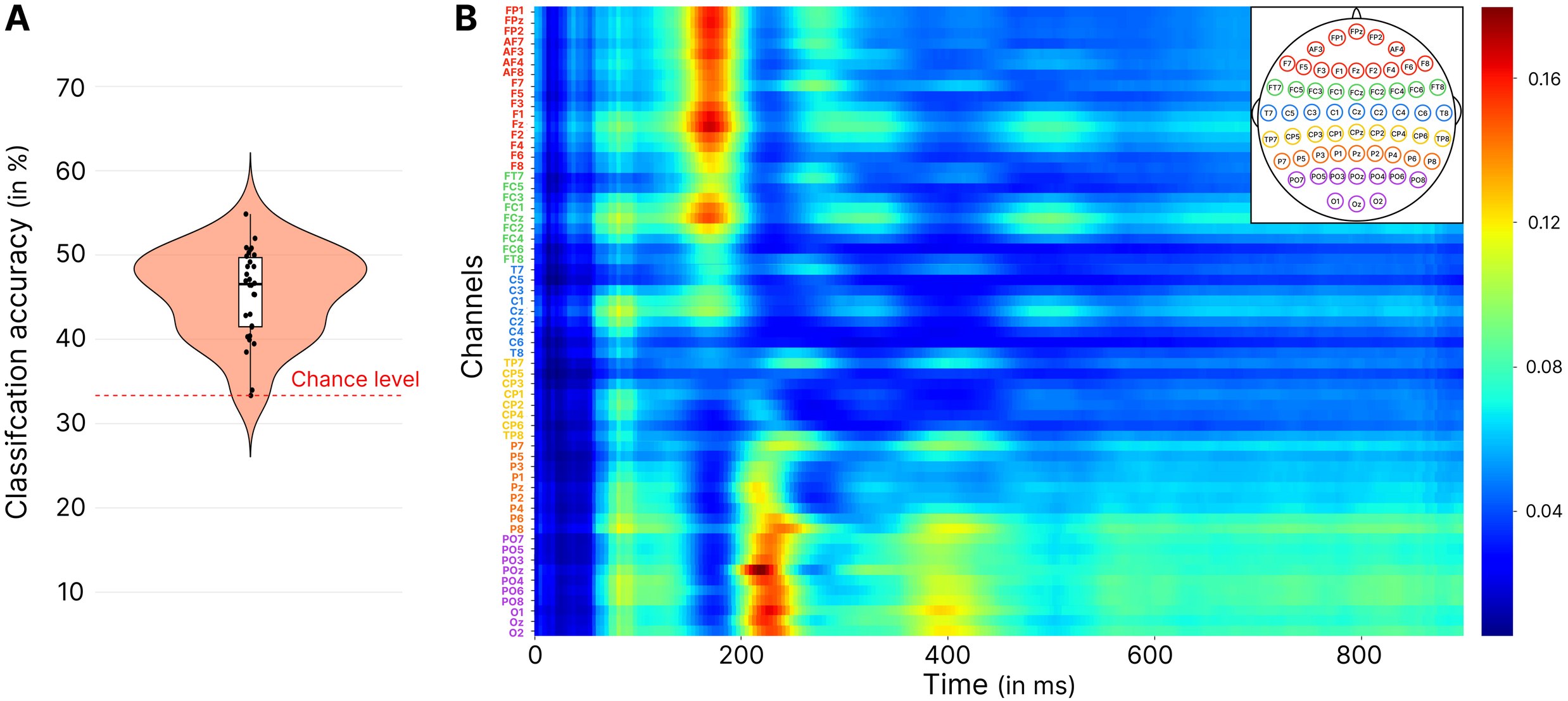
Steel, A., Silson, E. H., Garcia, B. D., & Robertson, C. E. (2024). A retinotopic code structures the interaction between perception and memory systems. *Nature Neuroscience*, *27*(2), 339□347. https://doi.org/10.1038/s41593-023-01512-3

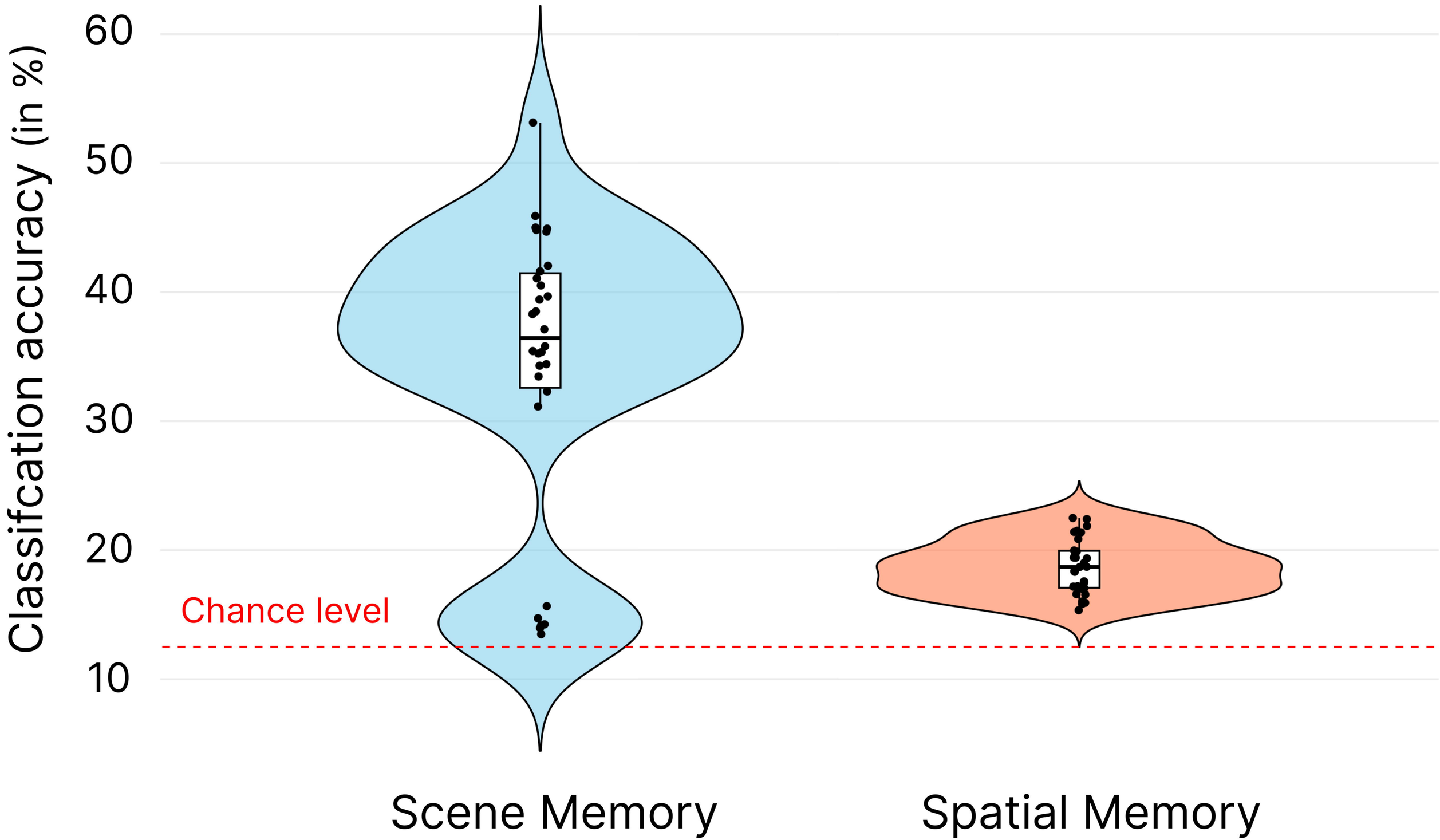
Vahid, A., Mückschel, M., Stober, S., Stock, A.-K., & Beste, C. (2020). Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control. *Communications Biology*, *3*(1), 1□11. https://doi.org/10.1038/s42003-020-0846-z




A
Classifcation accuracy (in %)
90
80
70
60
50
Chance level
B
Channels
FP1
FPz
FP2
AF7
AF3
AF4
AF8
F7
F5
F3
F1
Fz
F2
F4
F6
F8
FT7
FC5
FC3
FC1
FCz
FC2
FC4
FC6
FT8
T7
C5
C3
C1
Cz
C2
C4
C6
T8
TP7
CP5
CP3
CP1
CPz
CP2
CP4
CP6
TP8
P7
P5
P3
P1
Pz
P2
P4
P6
P8
PO7
PO5
PO3
POz
PO4
PO6
PO8
O1
Oz
O2
0
200
400
600
800
Time (in ms)
0.14
0.12
0.10
0.08
0.06
0.04


A
Classi. accur. (in %)
60
50
40
30
20
10
Chance level
C
Classi. accur. (in %)
50
40
30
20
10
Chance level
B
Channels
FP1
FPz
FP2
AF7
AF3
AF4
AF8
F7
F5
F3
F1
Fz
F2
F4
F6
F8
FT7
FC5
FC3
FC1
FCz
FC2
FC4
FC6
FT8
T7
C5
C3
C1
Cz
C2
C4
C6
T8
TP7
CP5
CP3
CP1
CP2
CP4
CP6
TP8
P7
P5
P3
P1
Pz
P2
P4
P6
P8
PO7
PO5
PO3
POz
PO4
PO6
PO8
O1
Oz
O2
0
200
400
600
800
Time (in ms)
0.20
0.15
0.10
0.05


A
Classifcation accuracy (in %)
70
60
50
40
30
20
10
Chance level
B
Channels
FP1
FPz
FP2
AF7
AF3
AF4
AF8
F7
F5
F3
F1
Fz
F2
F4
F6
F8
FT7
FC5
FC3
FC1
FCz
FC2
FC4
FC6
FT8
T7
C5
C3
C1
Cz
C2
C4
C6
T8
TP7
CP5
CP3
CP1
CP2
CP4
CP6
TP8
P7
P5
P3
P1
Pz
P2
P4
P6
P8
PO7
PO5
PO3
POz
PO4
PO6
PO8
O1
Oz
O2
0
200
400
600
800
Time (in ms)
0.16
0.12
0.08
0.04


Classifcation accuracy (in %)
60
50
40
30
20
10
Chance level
Scene Memory
Spatial Memory