communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-026-09685-w

Achieving more human brain-like vision via human EEG representational alignment

Zitong Lu [1,2] ✉, Yile Wang[3] & Julie D. Golomb [1]

Despite advancements in artificial intelligence, object recognition models still lag behind in emulating visual information processing in human brains. Recent studies have highlighted the potential of using neural data to mimic brain processing; however, these often rely on invasive neural recordings from non-human subjects, leaving a critical gap in understanding human visual perception. Addressing this gap, we present, 'Re(presentational)AI(ignment)net', a vision model aligned with human brain activity based on non-invasive EEG, demonstrating a significantly higher similarity to human brain representations. Our innovative image-to-brain multi-layer encoding framework advances human neural alignment by optimizing multiple model layers and enabling the model to efficiently learn and mimic the human brain's visual representational patterns across object categories and different modalities. Our findings demonstrate that ReAlnets exhibit stronger alignment with human brain representations than traditional computer vision models, achieving an average similarity improvement of approximately 3% and a maximum relative improvement ratio reaching up to 40%. This alignment framework takes an important step toward bridging the gap between artificial and human vision and achieving more brain-like artificial intelligence systems.

While current vision models in artificial intelligence (AI) have achieved remarkable advancements, they still fall short of capturing the full complexity and adaptability of the human brain's information processing. Deep convolutional neural networks (DCNNs) now rival human performance in object recognition[1], and many studies have identified representational similarities between the hierarchical structures of DCNNs and the ventral visual stream[2–6]. However, aligning DCNNs with human neural representations, though promising, remains an area with significant potential for further exploration. Enhancing the similarity between visual models and the human brain has become a critical concern for both computer scientists and neuroscientists. From a computer vision perspective, brain-inspired models often exhibit greater robustness and generalization, which are essential for achieving brain-like intelligence. Meanwhile, from a cognitive neuroscience perspective, models that more closely resemble brain representations can provide valuable insights into the mechanisms of human visual processing.

Conventional approaches—increasing model depth and layer count—have struggled to emulate the complexity of human visual processing[7]. Researchers have proposed various bio-inspired strategies to leverage the understanding of the human brain to enhance current AI vision models, including altering the architecture of the model (adding recurrent structures[4,8–11], dual-pathway models[12–16], topographic constraints[17–20] or feedback pathways)[21] and changing the training task (using self-supervised training[22,23] or 3D task models)[24]. However, an important question remains: Can we directly use human neural activity to align artificial neural networks (ANNs) in object recognition, thereby achieving more human brain-like vision models?

Several previous studies have begun to explore the integration of neural data into machine learning, particularly deep learning models, to enable these models to learn biologically inspired representations from neural data. One common strategy is to introduce a similarity loss during training to increase the representational alignment between models and neural activity, often derived from mouse V1 or monkey V1 and IT regions[25–28]. Another strategy from ref. 29 integrates an additional task that uses an encoding module to predict monkey V1 neural activity. Both similarity-based methods and multi-task frameworks have been shown to produce more brain-like representations and improve model robustness. However, the key challenge of these neural alignment studies is they only align single brain region or single model layer. Previous studies typically focused on aligning only a single layer of a CNN with a specific brain region, such as V1 or IT,

[1]Department of Psychology, The Ohio State University, Columbus, OH, USA. [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Neuroscience, The University of Texas at Dallas, Richardson, TX, USA. ✉e-mail: zitonglu@mit.edu

without a clear understanding of how multiple layers correspond to different brain regions. This oversimplification can lead to misalignment and inaccuracies. Moreover, most of these studies depend on invasive neural recordings from animals instead of human neural data, which limits the direct applicability of findings to human visual processing. Human noninvasive recordings, such as fMRI and EEG, often have lower data quality compared to invasive animal recordings, making it more challenging for models to learn human brain representations effectively. Early attempts involved applying human fMRI signals as additional inputs to machine learning classifiers, such as SVMs and CNNs, resulting in improved category classification performance without altering the internal feature representations of the models themselves[30–32]. While more recent research has directly optimized CNN internal representations align with human fMRI data for video emotion recognition[33], going beyond earlier approaches that only incorporated fMRI features into classification tasks[30], this approach has been limited to a relatively simple six-category emotion classification task. It remains unclear whether it could scale effectively to more complex domains like object recognition, which involves a far greater diversity of categories (e.g., 1000 classes in ImageNet-trained models). Also, it is unclear whether we can apply human neural data to optimize ANNs to achieve more human brain-like internal model representations.

To address these limitations, our study proposes a novel approach that employs an encoding-based framework for multi-layer alignment. This framework goes beyond simple similarity by training an additional encoding module to predict human neural activity, thereby enabling the model to autonomously extract complex visual features. Our approach leverages human neural data to achieve more effective alignment with human brain representations in object recognition tasks.

To bridge the gap between AI vision and human vision, we introduce Re(presentational)Al(ignment)net framework (hereafter, the ReAlnet framework), a method for effectively aligning vision models with human brain representations obtained from noninvasive EEG recordings. EEG was chosen for its high temporal resolution and the capacity to collect a large number of trials through rapid successive stimulus presentation, making it a cost-effective and scalable option. Our novel encoding-based multi-layer alignment framework effectively allows neural networks to learn human brain representations, enabling the creation of personalized vision models tailored to individual neural data. In this ReAlnet framework, each individualized ReAlnet refers to one EEG-aligned model instance trained on a single subject's data, while the plural ReAlnets refers to the set of ten such individualized models. Here, we define "direct alignment" as modifying a model's internal representational structure via a learning objective that explicitly incorporates human neural data. Under this definition, our approach constitutes the first direct alignment of object recognition models using noninvasive human EEG signals. This novel approach opens new possibilities for enhancing brain-like representations in AI models. Furthermore, human EEG-optimized ReAlnets demonstrate improved alignment with human brain representations across different modalities (both human EEG and fMRI) and human behaviors.

Results

Aligning CORnet with human EEG representations

In this study, we developed a novel image-to-brain multi-layer encoding alignment framework that integrates human EEG data directly into the training of a deep convolutional vision model (Fig. 1). Given an input image, the model simultaneously performs object classification and generates

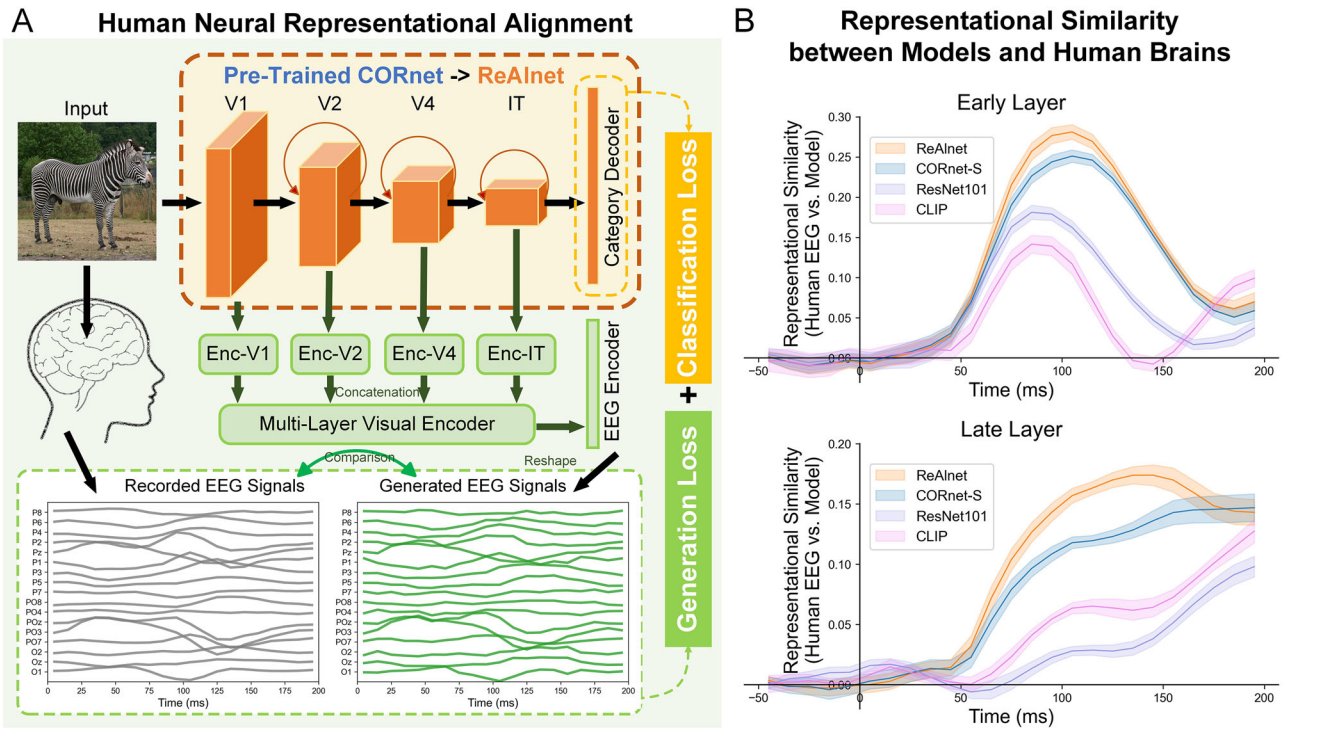

Fig. 1 | ReAlnets aligned with human EEG signals as more human brain-like vision models. A An overview of the ReAlnet alignment framework. Adding an additional multi-layer encoding module to an ImageNet pre-trained CORnet-S, the outputs contain the category classification results and the generated EEG signals. Using the THINGS EEG2 training dataset, we aim to minimize both classification loss and generation loss, enabling CORnet to not only stabilize the classification performance but also effectively learn human brain features and transform into ReAlnets. **B** Representational similarity between internal representations in models and human temporal EEG signals from the THINGS EEG2 test dataset. Models include ReAlnets and their primary comparison model CORnet-S, along with ResNet101 and CLIP (with a ResNet101 backbone) as additional baselines. These baselines are used in their off-the-shelf pretrained form without any EEG alignment; only ReAlnet is optimized with human EEG. Because the additional baselines have different number of layers, for all models we took the first layer as the Early Layer, and the layer before the classification layer (or last visual layer in CLIP) as the late layer for this analysis. The line labeled "ReAlnet" reflects the mean similarity across 10 individual ReAlnets, each trained on a different subject's EEG data ($N = 10$). For comparison models, each line reflects the mean similarity between the same 10 human EEG datasets and the single model instance. ReAlnets consistently show the highest similarity to the human brain. Lines and shading reflect mean ± SEM.

predicted EEG signals corresponding to that image. To instantiate this framework, we built ten individual models, termed ReAlnets, using the state-of-the-art CORnet-S model[9,34] as the foundational architecture. During training, ReAlnets were optimized jointly for object classification and EEG prediction using subject-specific EEG recordings collected while human subjects viewed a massive number of natural images from the THINGS EEG2[35] training set (Fig. 1A and see "Methods").

As an initial fidelity check, we verified that this joint training procedure preserved strong object recognition performance (average top-1 and top-5 accuracy on ImageNet: 69.31 and 88.88%, Supplementary Fig. 1A, B) while enabling the model to generate realistic EEG signals (average 0.7785 similarity, computed as Spearman correlation between generated and real EEG time series, treating the full images × channel × timepoint matrix as a vector, Supplementary Fig. 1C–E). To contextualize the high numerical correlation, we note that EEG signals exhibit highly consistent temporal patterns across images within our short 0–200 ms time-window, and EEG channels themselves show strong cross-channel covariance. Even under a more stringent metric removing the influence of these factors (an image-level channel-timepoint-wise correlation: for each of the 17 channels and 20 timepoints, we correlated the 200-image vector between real and generated EEG responses and then averaged the resulting 340 correlations), ReAlnet still exhibited a significant positive correlation (0.3081), whereas *Unpaired* and *Scrambled* controls were approximately zero (Supplementary Fig. 1F), confirming that ReAlnets captured image-specific EEG patterns rather than merely reproducing the shared ERP profile.

While these results demonstrate that the model successfully learned to reproduce realistic EEG activity, building the most accurate EEG generator is not our main objective. Instead, we seek to use EEG generation as a training mechanism to optimize the model's internal representations, with the goal of making them more similar to human brain representations. Our critical hypothesis is that modifying the model's internal representations in this way may carry benefits that generalize across image diet and modalities. Therefore, our core aim is to investigate whether aligning an artificial vision model with individual neural representations from actual human subjects can enhance the model's similarity to the human brain.

In the following sections, we evaluate whether the model's internal representations have indeed become more brain-like by conducting representational similarity analysis (RSA) to test the models' similarity to human EEG, human fMRI, and behavior. To preview, using an independent test dataset consisting of entirely novel (untrained) object categories, we calculated the temporal similarity between different models and the human brain EEG. As shown in Fig. 1B, for both early and late layers, comparing the ReAlnets to the original CORnet-S, along with ResNet101 and CLIP (with a ResNet101 backbone) as additional baselines, ReAlnets consistently show the highest similarity to the human brain patterns. Importantly, our assessment of the model's similarity to humans is not limited to its similarity with human EEG representations. We further evaluated the model's similarity to human brain fMRI representations (a completely different modality) from human subjects viewing a completely different dataset of novel image categories (based on Shen fMRI test set)[36]. Additionally, we measured the similarity between the model and human behavior in several object recognition tasks using the Brain-Score platform[37] based on two behavioral benchmarks. In the sections that follow, we describe these results in more detail, along with additional analyses and control experiments testing ReAlnets.

Improved similarity to human EEG

After training the ReAlnets, we employed an independent test dataset consisting of 200 images and associated EEG activity from the THINGS EEG2 dataset[35]. These test set images had not been presented at all during the training process, coming from entirely novel (untrained) object categories. We input these 200 test images to each model (the 10 subject-specific ReAlnets plus CORnet and other comparison models) and obtained the feature vectors corresponding to each image for each layer in the model. Then we obtained the actual human brain EEG patterns at each timepoint for each of the 10 human subjects viewing those same 200 test images from the THINGS EEG2 dataset[35], we calculated (1) the internal representational similarity between actual EEG data and CORnet (with the same structure as ReAlnets, but non aligned human neural data and non-individualized model), and (2) the internal representational similarity between actual EEG data and the subject-matched ReAlnet. As an additional control, we trained a *Scrambled*-model version of the ReAlnet framework by aligning on 10 subjects' *Scrambled* EEG signals (more details in the Control Experiments section and "Methods" section) and conducted the same RSA analysis.

ReAlnets exhibit significantly higher similarity to human EEG neural dynamics for all four visual layers (layer V1: 70–130 and 160–200 ms; layer V2: 60–200 ms; layer V4: 60–200 ms; layer IT: 70–160 ms) than the original CORnet without human neural alignment, or than *Scrambled* models trained on *Scrambled* EEG signals (Fig. 2A). The EEG similarity curves often show a two-peaked shape, potentially reflecting different stages of visual information processing. Based on prior studies on human neural dynamic representations[38,39], the earlier peak ($\tilde{1}$00 ms) likely corresponds to the processing of lower-level visual features, such as color and retinal size. And the later peak ($\tilde{1}$50–200 ms or even later) may reflect the processing of higher-level semantic features, such as real-world size and animacy information[40,41].

Further statistical analysis of each layer's similarity improvement (ReAlnet-CORnet) and improvement ratio ((ReAlnet- CORnet)/CORnet) also indicates that at the similarity peak timepoint, there is a maximum of a 6% similarity improvement and a 40% improvement ratio (Fig. 2B). Importantly, all individualized ReAlnets showed positive improvement over CORnet and *Scrambled* models, suggesting that our ReAlnet framework robustly generalizes across subjects and effectively enhances model-EEG similarity when trained on each individual's EEG data. It is worth noting that the test set doesn't overlap with the training set in terms of object categories (concepts). Therefore, these significant improvements reveal ReAlnets' generalization capability across different object categories. In addition to reporting layer-wise RSA curves, we also computed the maximum similarity across all layers for each model at each time point to provide a clearer summary of model-EEG alignment (Fig. 2C).

These results suggest three findings: (1) our multi-layer alignment framework indeed improves all layers' similarity to human EEG representations. (2) Every ReAlnet with individual neural alignment exhibits improved similarity to human EEG compared to the basic CORnet. (3) ReAlnets demonstrate the generalization of improvement in human brain-like similarity across object categories, as the image categories used for testing were entirely absent during the alignment training.

Additionally, unlike traditional models in computer vision, ReAlnet is a personalized model trained based on different individuals' neural data. Just as neuroscientists are interested in studying individual differences in the brain, we can similarly explore individual differences across the ten personalized ReAlnets—whether ReAlnets exhibit intra-model individual variabilities and how such variabilities change across different layers of the model. Note that CORnet serves as a single, publicly available reference model with fixed official weights, and thus does not have individualized versions analogous to ReAlnets. Therefore, inter-model variability can only be meaningfully assessed across individualized ReAlnets. Importantly, all ten ReAlnets were initialized with the same pretrained CORnet-S weights and trained using the same random seed, ensuring that any observed differences arise solely from the individual EEG data used for fine-tuning, not from variations in model initialization. We hypothesize that the observed individual differences in ReAlnets may provide insights into potential mechanisms underlying individual differences in human brains. To investigate this, we analyzed individual variability in human brain regions (V1, V2, V4, and LOC) using the Shen fMRI dataset based on a similar RDM-based correlation analysis. As a comparison, we similarly analyzed individual variability across different ReAlnet RDMs at each layer, using the 200 images in the THINGS EEG2 test set.

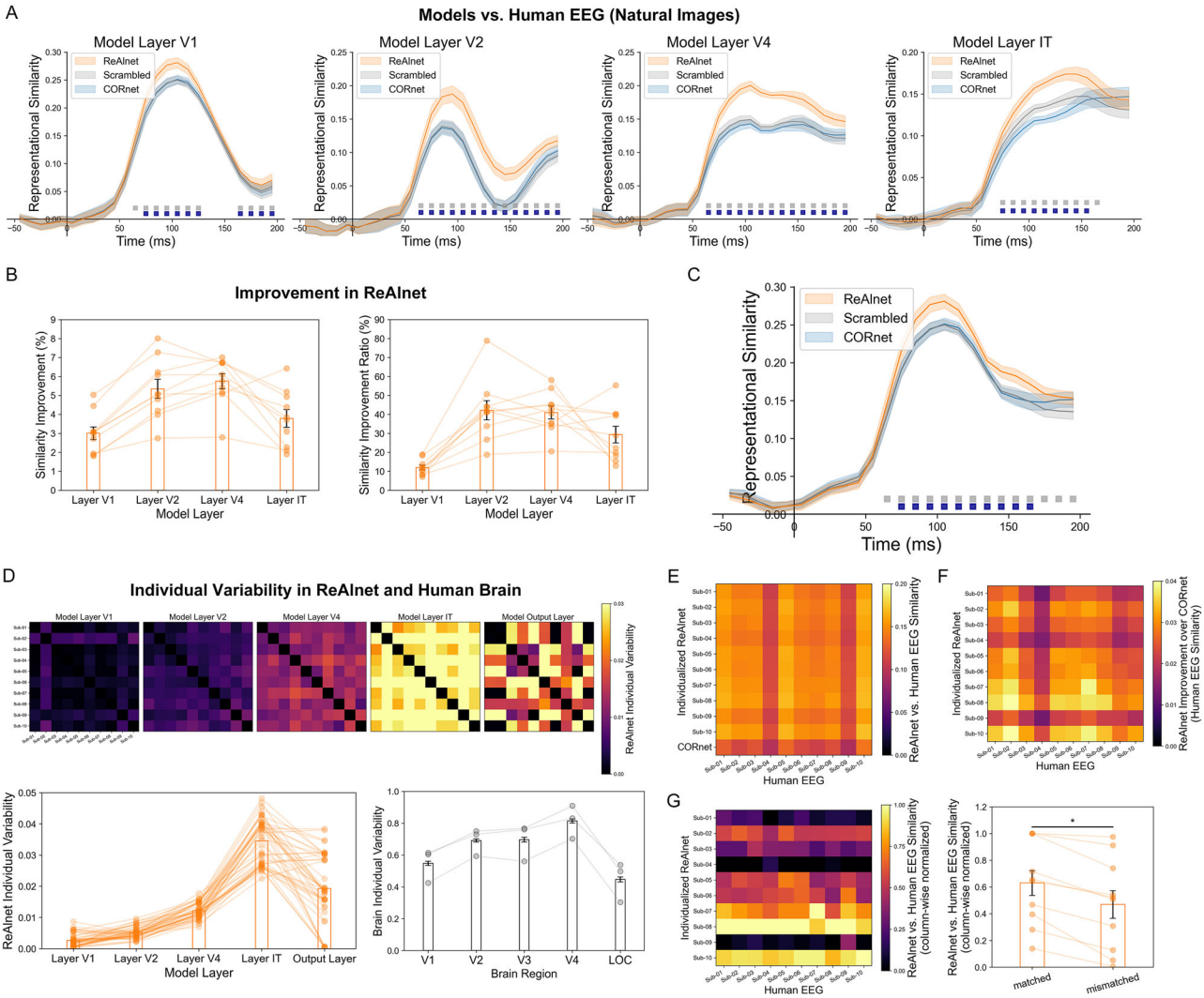

Fig. 2 | ReAlnets show higher similarity to human EEG and hierarchical individual variability. A Representational similarity time courses between human EEG and models (ReAlnets, *Scrambled* models, and CORnet) for different layers, respectively. Dark blue square dots at the bottom indicate the timepoints where ReAlnet vs. CORnet were significantly different ($p < 0.05$). Grey square dots at the bottom indicate the timepoints where ReAlnet vs. *Scrambled* were significantly different ($p < 0.05$). Lines and shading reflect mean ± SEM. **B** Similarity improvement and similarity improvement ratio of ReAlnets compared to CORnet at the similarity peak timepoint. Each circle dot indicates an individual ReAlnet. Error bar reflects ± SEM. **C** Time courses of the maximum representational similarity between human EEG and different models (ReAlnets, *Scrambled* models, and CORnet), computed by taking the highest similarity across all model layers at each timepoint. Dark blue square dots at the bottom indicate timepoints where ReAlnets significantly outperformed CORnet ($p < 0.05$). Grey square dots indicate significant differences between ReAlnets and *Scrambled* models ($p < 0.05$). Lines and shading reflect mean ± SEM. **D** Top: ReAlnet individual variability matrices of four visual layers. Bottom left: ReAlnet individual variability along layers. Bottom right: Human fMRI individual variability along the visual cortex. Each circle dot indicates a pair of two personalized ReAlnets or two human subjects. Error bar reflects ± SEM. **E** Cross-subject similarity matrix showing each individualized ReAlnet (rows) generalizes to EEG representations from all 10 subjects (columns). Each cell reflects the average representational similarity between human EEG and ReAlnets and CORnet across four model layers and the 50–200 ms time window. **F** Cross-subject generalization beyond baseline CORnet. Each cell reflects the ReAlnet-CORnet difference in EEG similarity, with positive values indicating that even mismatched ReAlnets outperform CORnet on other subjects' EEG data. **G** Left: column-wise normalized similarity matrix based on baseline-subtracted similarity matrix, where each column is scaled such that the highest similarity value is 1. Right: a statistical comparison between matched and mismatched pairs. Black asterisks indicate significantly higher similarity of matched pairs than mismatched pairs ($p < 0.05$). Error bar reflects ± SEM.

Our results of individual differences suggest: (1) personalized ReAlnets indeed exhibit individual variability (Fig. 2D). (2) This variability increases with the depth of the layers (from Layer V1 to Layer IT, Fig. 2D) and decreases in the output layer, which aligns with the trend observed in the human brain—where individual variability increases from V1 to V4 and then decreases in LOC. Importantly, we included the output layer of ReAlnets, which performs category classification, and found that its individual variability decreases, mirroring the pattern seen in the LOC region of the human brain (additional analysis in Supplementary Fig. 2A confirms that LOC is significantly more similar to the model's output layer than layer IT, suggesting that the model's output layer rather than layer IT is more comparable to brain LOC region).

While humans also show hierarchical increases in between-subject variability, the larger absolute magnitude observed in neural data likely reflects a combination of cross-participant representational heterogeneity and measurement noise inherent to neuroimaging. We further considered whether individual variability in ReAlnets could arise from factors unrelated to EEG alignment, such as random initialization or fine-tuning with Scrambled or unpaired EEG data. However, additional analyses showed that *Unpaired* and *Scrambled* models lack the systematic increase in variability from layer V1 to V4 observed in ReAlnets and exhibit overall weaker

variability (Supplementary Fig. 2B–D), indicating that ReAlnets optimized using true EEG alignment more faithfully capture structured individual variability. Although the overall magnitude of between-layer variability remains smaller in models than in neural data—likely due to the architectural regularity of the models—both humans and ReAlnets exhibit similar hierarchical trends.

To further evaluate the generalization ability of individualized ReAlnets within the EEG modality, we conducted a cross-subject analysis, testing each model against the EEG RDMs of all subjects. We computed a model-subject similarity matrix where each entry represents the average RSA similarity across four model layers and the 50–200 ms time window. This approach summarizes each ReAlnet's representational match with every subject and produces a compact confusion matrix (Fig. 2E). To more directly assess cross-subject generalization, we subtracted each subject's similarity with CORnet from all cells in that subject's column (Fig. 2F). Positive values in this normalized matrix indicate that even mismatched ReAlnets outperform the CORnet baseline, confirming that ReAlnets capture representational structures that generalize across individuals. Finally, to control for differences in signal quality across subjects, we applied column-wise normalization to the matrix (Fig. 2G). To more directly quantify subject-specificity, we compared similarity values for matched model-subject pairs (the diagonal: i.e., models evaluated on the same subject used for training) against unmatched pairs (the off-diagonal) and observed that matched model-subject pairs show significantly higher similarity than unmatched pairs ($t = 5.6068$, $p = 0.0003$). These results confirm that individualized ReAlnets capture subject-specific representational features and that these models generalize across individuals within the same neural modality.

Improved similarity in ReAlnets to human fMRI

Although ReAlnets demonstrate higher similarity to human EEG, a question arises: do ReAlnets learn representations specific to EEG, or more general neural representations of the human brain? To ensure that our alignment framework enables the model to learn representations beyond the single modality of EEG, we utilized additional human fMRI data of three human subjects viewing natural images to evaluate the model's cross-modality representational similarity to human fMRI.

Excitingly, we indeed observed a clear improvement in this cross-modal brain-like similarity. Based on human fMRI signals of three subjects viewing 50 natural images, the similarity results indicate that, overall, ReAlnets exhibit a stronger resemblance to human fMRI data than CORnet and *Scrambled* models (Fig. 3A), despite being aligned using a different type of neural signal with very different spatiotemporal properties (human EEG data), and that was collected from a different set of participants. Although a few conditions show CORnet may have higher similarity than ReAlnets, the general trend favors ReAlnets—highlighting their stronger alignment with human neural representations and showing that EEG-optimized ReAlnets generalize broadly to human representations.

We next asked whether the enhanced similarity between ReAlnets and human brain representations extends beyond natural image stimuli. We tested ReAlnets on additional stimulus sets included in the Shen fMRI dataset, including 40 artificial shape and 10 alphabetical letter images. Our results again demonstrate ReAlnets' improved similarity to human brain representations in comparisons to CORnet and *Scrambled* models (Fig. 3B, C). While some comparisons yield small absolute differences, we also examined effect sizes to better assess their statistical and practical significance (detailed statistical results including *t*-value, *p*-value, and Cohen's *d* are listed in Supplementary Table 1).

These findings further highlight three points: (1) across multiple regions-of-interest (ROIs), ReAlnets exhibits higher human fMRI similarity than CORnet. (2) Despite being trained with the EEG data of subjects not in the fMRI dataset, almost every ReAlnet shows higher fMRI similarity, suggesting that ReAlnets learn consistent brain information processing patterns across subjects. (3) Images from the fMRI datasets for evaluation were never presented during the alignment training, reaffirming the generalization of ReAlnets in improving brain-like similarity across object categories and images.

Moreover, the similarity improvements across modalities were significantly correlated (Supplementary Fig. 3): ReAlnets showing greater improvement in EEG similarity also showed greater similarity improvement in fMRI similarity, when correlating these improvement values across different ReAlnet instances ($r = 0.9204$, $p < 0.0001$). This cross-modal correspondence further indicates that the ReAlnet alignment framework is tapping into meaningful and robust human neural signals.

Improved similarity in ReAlnets to behavior

Does this neural-level alignment also translate into any behavioral alignment? To test whether ReAlnets show improved similarity to human behavior, we calculated the scores of CORnet and 10 personalized ReAlnets based on the human behavioral assessments, including two object recognition tasks, in the Brain-Score platform[37]. One task compared how well the ANN, compared to primates and humans, could recognize objects presented in the center of their visual field, even when the objects varied in position, size, viewing angle, and background[7]. The other paradigm tested the similarity of behavioral error between the errors made by humans and ANN on an image-by-image basis[42]. These scores serve as indicators of the models' similarity to human behavior. The average of two behavioral scores were used to compare model-behavior similarity. Excitingly, the result reveals that ReAlnets, aligned with human EEG data, exhibit representations significantly more akin to human behavior than CORnet and *Scrambled* models do (ReAlnets vs. CORnet: $t = 2.7702$, $p = 0.0217$, $d = 0.8762$; ReAlnets vs. *Scrambled* models: $t = 8.5582$, $p < 0.0001$, $d = 3.8273$) (Fig. 4A), further expanding and emphasizing ReAlnets' status as more human brain-like vision models. When looking separately at the two tasks, the results show that ReAlnets exhibit a greater improvement in human behavior similarity on the Geirhos2021 task compared to the Rajalingham2018 task. A possible explanation could be that because ReAlnets are trained with EEG data recorded from subjects viewing natural object images, it better captures human-like behavioral consistency when evaluated on naturalistic, colorful stimuli used in the Geirhos2021 task, compared to the grayscale images in Rajalingham2018. Differences in image manipulation between the two tasks (e.g., Rajalingham2018 but not Geirhos2021 manipulates objects by varying position, size, viewing angle, and background) may also contribute to this disparity.

In addition, we submitted our ReAlnets to the updated brain-score platform for evaluation and made the scores publicly available on the Brain-Score website (https://www.brain-score.org/vision/). Consistently, the results confirm that ReAlnets achieve significantly higher scores compared to CORnet (see Supplementary Fig. 4), further supporting their improved alignment with brain visual processing.

Refined object feature representations in ReAlnets

We found that ReAlnets exhibit improved similarity to the human visual system, suggesting that by incorporating human EEG data, they have learned brain representations that purely image-trained models could not capture. Thus, this raises an important question: how do their internal representations differ from those of CORnet? To further investigate what visual object feature representations have been enhanced in ReAlnets by learning human brain signals, we conducted internal representational analysis on purely image-trained CORnet and human EEG-aligned ReAlnets (for the detailed description of the methodology used to analyze these enhanced feature representations, refer to our Methods-Model internal representational analysis section). Here, we tracked feature representations of 49 object feature dimensions from THINGS ([43]) on models' layer IT to see which features could be better captured by ReAlnets than CORnet. These 49 object feature dimensions, derived from the THINGS dataset, describe various conceptual and perceptual properties of objects, providing a structured way to assess how different models encode object features beyond simple category labels. Figure 4B left shows that ReAlnets show significantly enhanced representations in food-related, artificial, and electronic

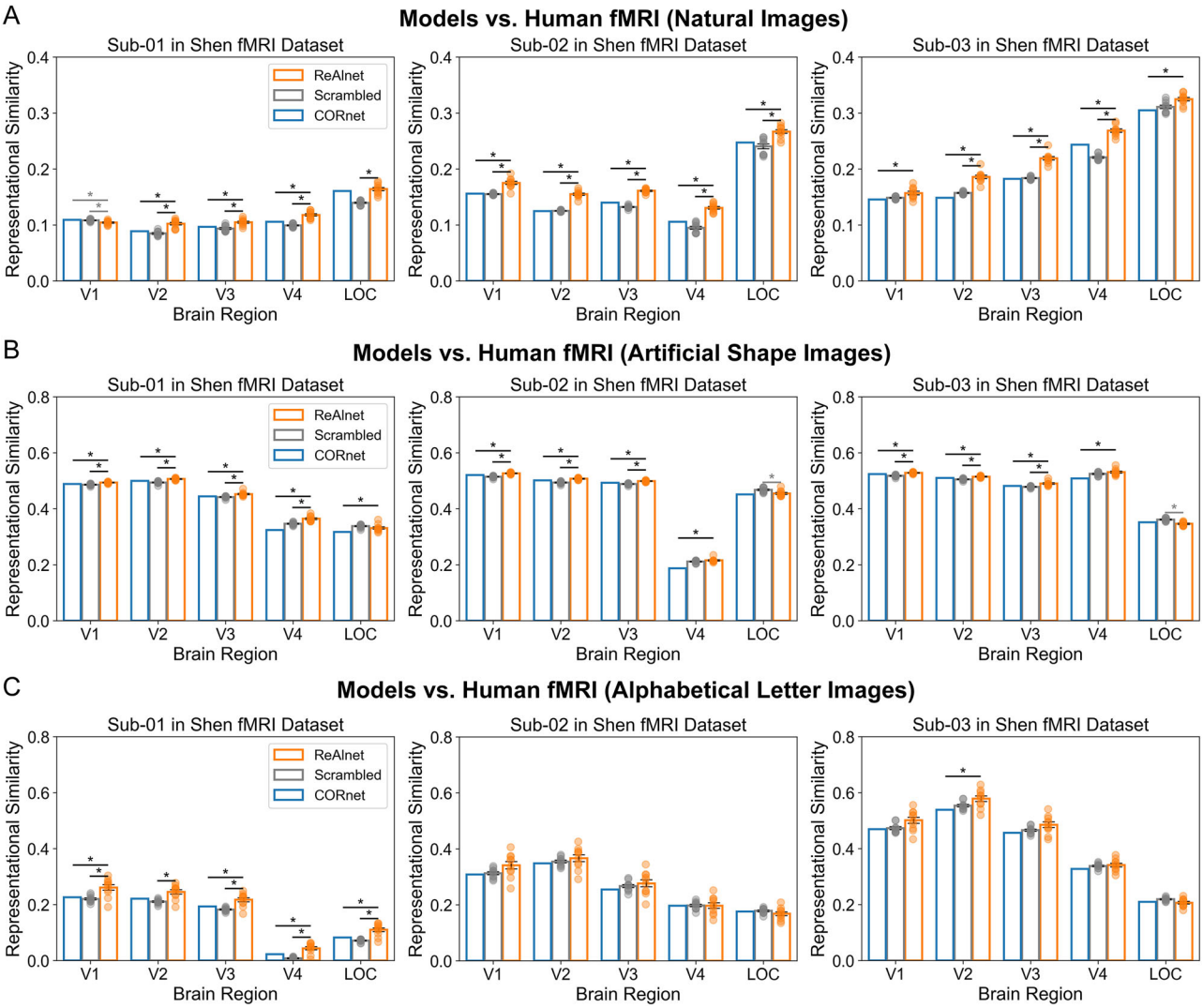

Fig. 3 | ReAlnets show higher similarity to human fMRI representations. Representational similarity between models and human fMRI of five different brain regions when three subjects in the Shen fMRI test dataset viewed **A** natural images, **B** artificial shape images, and **C** alphabetical letter images. Black asterisks indicate significantly higher similarity of ReAlnets than that of the *Scrambled* model or CORnet ($p < 0.05$), and grey asterisks indicate significantly lower similarity of ReAlnets than that of the *Scrambled* model or CORnet ($p < 0.05$). All pairwise comparisons were Bonferroni-corrected for multiple comparisons. Each circle dot indicates an individual ReAlnet or *Scrambled* model. Error bar reflects ± SEM.

information than CORnet (more detailed results of all 49 object feature dimensions are shown in Supplementary Fig. 5). Interestingly, *Scrambled* models also showed enhanced representations for some of these features, suggesting that even the statistical properties of EEG data can refine internal model representations—though to a lesser extent than real, temporally intact EEG signals.

To further dissociate neural-alignment effects from visual diet effects, we conducted the same analysis on ReAlnet $\beta = 0$, which is trained on the identical training images in THINGS EEG2 but without the EEG encoder. This model, therefore, isolates the influence of visual diet alone. Strikingly, ReAlnet $\beta = 0$ already shows a substantial increase in some feature dimensions, such as the food-related feature dimension—often even higher than the full ReAlnet—indicating that this representational change is primarily driven by the THINGS visual diet rather than EEG alignment (Supplementary Fig. 6).

To clarify which feature refinements genuinely reflect neural alignment, we compared both ReAlnets and *Scrambled* models relative to ReAlnet $\beta = 0$ (Fig. 4B right). Several feature dimensions, including electronic/technology-related, flat/patterned, and long-thin, exhibit additional enhancement in ReAlnets beyond ReAlnet $\beta = 0$. We note that flat/patterned and long-thin dimensions were enhanced in both ReAlnets and *Scrambled* models, suggesting that preserving the global statistical properties of the EEG might already bias these dimensions. In contrast, the electronic/technology-related dimension shows a greater increasing in ReAlnets, indicating that non-scrambled, paired EEG signals further strengthen this representation, above and beyond visual diet or scrambled EEG. Overall, these results clarify that while visual diet accounts for the main effects in representational changes (particularly food-related), EEG alignment contributes additional refinements in several object features.

Control experiments

To systematically evaluate how different training manipulations influence model-to-brain alignment, we conducted a set of control experiments by training four additional sets of ReAlnets. On the one hand, we would like to test the importance of the two loss components in generation loss—reconstruction loss and contrastive learning loss—in the model alignment. To evaluate the importance of the two loss components in the model, we tested model-to-brain alignment in (1) W/o ContLoss models (without the contrastive loss component), and (2) *W/o MSELoss* models (without the MSE loss component). Additionally, to evaluate the importance of different

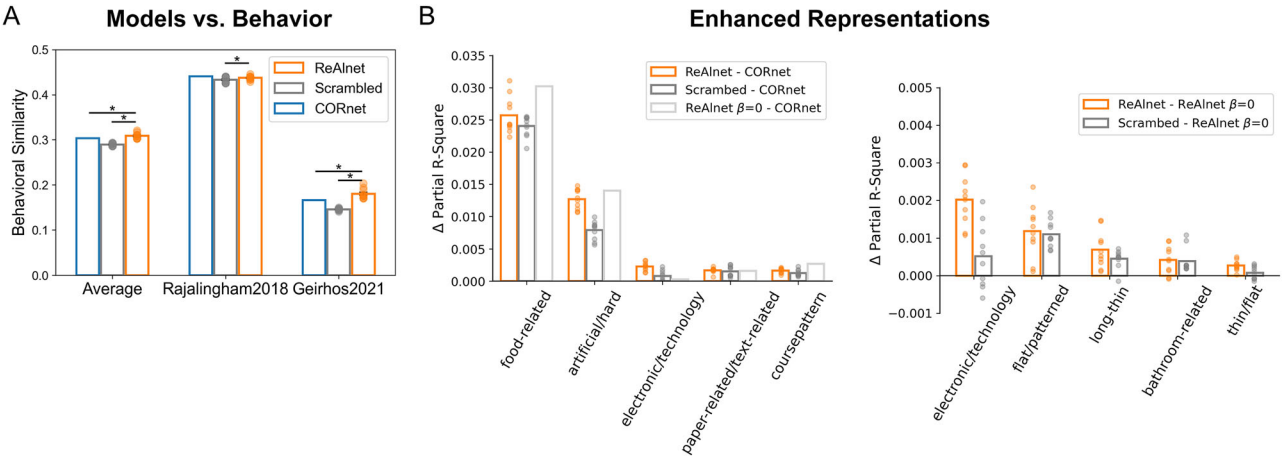

Fig. 4 | Enhanced behavioral similarity and feature representations in ReAlnets. **A** ReAlnets show higher similarity to human behavior based on the Brain-Score platform. Each orange circle dot indicates an individual ReAlnet. Each grey circle dot indicates an individual *Scrambled* model. Asterisks indicate significantly higher similarity of ReAlnets than that of CORnet or *Scrambled* models ($p < 0.05$). **B** Left: top-5 enhanced feature representations in ReAlnets, *Scrambled* models, and ReAlnets $\beta = 0$ compared to CORnet. Right: top-5 enhanced feature representations in ReAlnets and *Scrambled* models compared to ReAlnets $\beta = 0$. Each orange circle dot indicates an individual ReAlnet. Each grey circle dot indicates an individual *Scrambled* model. Error bar reflects ± SEM.

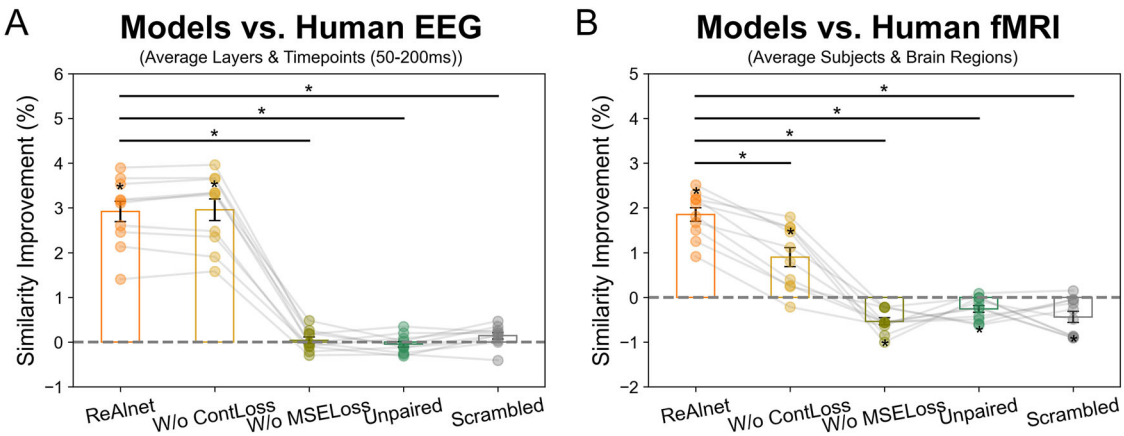

Fig. 5 | Results of control experiments. **A** Improvement in human EEG similarity of ReAlnets and control models compared to CORnet. **B** Improvement in human fMRI similarity of ReAlnets and control models compared to CORnet. Each circle dot indicates an individual model. Asterisks indicate the significance ($p < 0.05$). Error bar reflects ± SEM.

types of information in the EEG signal for the alignment[26,44], we tested (3) *Unpaired* models (where the pairing between images and EEG signals was disrupted), and (4) *Scrambled* models (where the EEG time-series were scrambled). Detailed definitions of *Unpaired* and *Scrambled* control models are provided in Methods, which target different information in EEG signals (associative vs. temporal). More detailed statistical results, including *t*-value, *p*-value, and Cohen's *d* are listed in Supplementary Tables 1 and 2.

We tested the control models on the THINGS EEG2 test dataset and the Shen fMRI dataset, and calculated the similarity improvement for each control model compared to ReAlnets and the CORnet baseline. Figure 5 plots the improvement in similarity (compared to the CORnet baseline) for ReAlnets and the four controls. Here, we averaged the EEG similarity improvement of all layers and timepoints between 50 and 200ms, and averaged the fMRI similarity improvement of three subjects and five brain regions (See more detailed EEG and fMRI similarity results in Supplementary Fig. 7). The *W/o MSELoss*, *Unpaired*, and *Scrambled* control models showed no significant similarity improvement over CORnet for human EEG (*W/o MSELoss*: $t = 0.5640$, $p = 0.5865$, $d = 0.1784$; *Unpaired*: $t = -0.6654$, $p = 0.5225$, $d = -2104$; *Scrambled*: $t = 1.9450$, $p = 0.0835$, $d = 0.6154$) and even significant similarity decreasing for human fMRI (*W/o MSELoss*: $t = -6.3837$, $p = 0.0001$, $d = -2.0187$; *Unpaired*: $t = -3.1304$, $p = 0.0121$, $d = -0.0900$; *Scrambled*: $t = -3.2647$, $p = 0.0098$, $d = -1.0324$). The *W/o ContLoss* control models were significantly improved over CORnet for both modalities (EEG: $t = 11.4987$, $p < 0.0001$, $d = 3.6362$; fMRI: $t = 4.1036$, $p = 0.0027$, $d = 1.2977$), but they didn't perform as well as ReAlnets in terms of fMRI similarity ($t = -8.0364$, $p < 0.0001$, $d = -1.5494$) but showed similar improvement for model-EEG similarity ($t = 0.8543$, $p = 0.4151$, $d = 0.0517$).

The results of the control experiments reveal: (1) *W/o ContLoss* models still exhibit an improvement in human brain similarity compared to CORnet. However, while the similarity to human EEG did not decrease compared to ReAlnets, the similarity to cross-modality human fMRI significantly decreased. This suggests that the contrastive loss in our alignment framework enables ReAlnets to learn broader and more generalized visual representation patterns across different neuroimaging modalities, such as EEG and fMRI data. (2) The *W/o MSELoss*, *Unpaired*, and *Scrambled* models failed to enhance brain similarity, which show no significant improvement in brain similarity compared to CORnet, indicating that the training process requires the model to effectively learn the specific neural visual features from the actual EEG signals corresponding to each image. Only in this way can the model become more human brain-like and then exhibit higher similarity to the human brain across different object images, categories, and human neuroimaging data modalities.

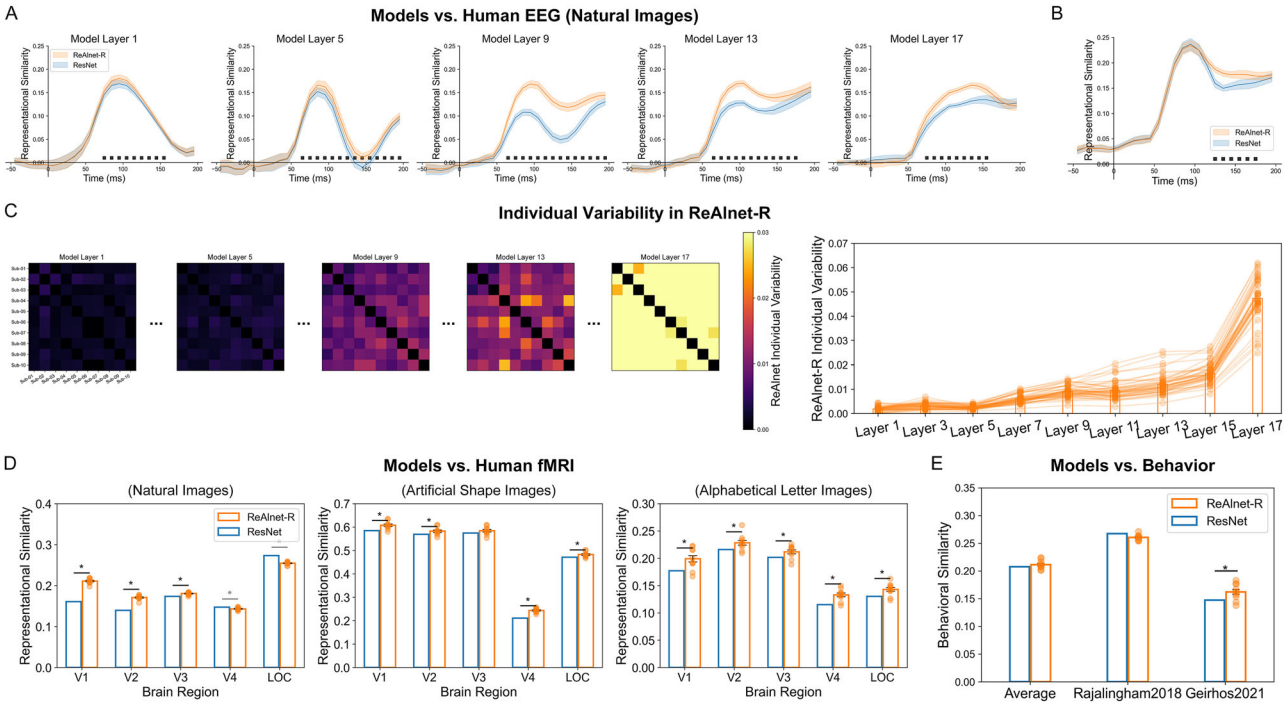

Fig. 6 | Similar improvements in ReAlnet-Rs. A Representational similarity time courses between human EEG and models (ReAlnet-Rs and ResNet) for layer 1, 5, 9, 13, and 17, respectively. Black square dots at the bottom indicate the timepoints where ReAlnet-Rs vs. ResNet were significantly different ($p < 0.05$). Lines and shading reflect mean ± SEM. **B** Time courses of the maximum representational similarity between human EEG and different models (ReAlnet-Rs and ResNet), computed by taking the highest similarity across all model layers at each timepoint. Black square dots at the bottom indicate timepoints where ReAlnet-Rs significantly outperformed ResNet ($p < 0.05$). Lines and shading reflect mean ± SEM. **C** ReAlnet-R individual variability matrices of layer 1, 5, 9, 13, and 17 and individual variability along layers. Each circle dot indicates a pair of two personalized ReAlnets. **D** Representational similarity between models and human fMRI of five different brain regions when subject 2 in the Shen fMRI test dataset viewed natural, artificial shape, and alphabetical letter images. Black asterisks indicate significantly higher similarity of ReAlnet-Rs than that of ResNet ($p < 0.05$). Grey asterisks indicate significantly lower similarity of ReAlnet-Rs than that of ResNet ($p < 0.05$). Each circle dot indicates an individual ReAlnet-R. Error bar reflects ± SEM. **E** Similarity between models and human behavior based on the Brain-Score platform. Each circle dot indicates an individual ReAlnet-R. Error bar reflects ± SEM.

Human EEG-aligned ResNet also becomes more brain-like

Although we trained ReAlnets based on CORnet and confirmed that they are more human brain-like, we also wondered whether our multi-layer encoding-based alignment framework could be extended to other models. Therefore, we chose ResNet18, a relatively larger model, and aligned it with the EEG representations of ten subjects from the THINGS EEG2 dataset using the same framework as above. We refer to the aligned model based on pretrained ResNet18 as ReAlnet-R. Subsequently, we tested ReAlnet-Rs for their similarity to human EEG, fMRI, and behavior, comparing the results with those of the purely image-trained ResNet18.

Firstly, ReAlnet-Rs show significantly higher similarity to human EEG neural dynamics compared to ResNet for nearly all visual layers (layer 1: 70–160 ms, layer 5: 60–200 ms, layer 9: 60–200 ms, layer 13: 60–180 ms, layer 17: 70–160 ms, Fig. 6A; maximum similarity across all layers: 120–180 ms, Fig. 6B; see all layers' EEG similarity results in Supplementary Fig. 8). Secondly, personalized ReAlnet-Rs, similar to ReAlnets, exhibit individual variability increasing with the depth of the layers (Fig. 6C; see all layers' individual variability matrices in Supplementary Fig. 9). Thirdly, ReAlnet-Rs also show higher similarity to human fMRI representations across multiple visual ROIs and different image categories (Fig. 6D; see fMRI similarity results on all three subjects in Shen fMRI dataset in Supplementary Fig. 10 and detailed statistical results including t-value, p-value, and Cohen's d are listed in Supplementary Table 4.). Fourthly, for human behavioral similarity, although it is not significant on the average score ($t = 1.6529$, $p = 0.1328$, $d = 0.5227$), seven out of ten ReAlnet-Rs show higher similarity than the original ResNet. And ReAlnet-Rs show significantly higher behavioral similarity on the Geirhos2021 task. These results collectively indicate that our alignment framework can be successfully extended to other visual models, such as ResNet, with ReAlnet-Rs still demonstrating improved similarity to human neural and behavioral representations.

ReAlnets trained across subjects

While the core focus of our study is on individualized ReAlnets that capture subject-specific neural representational patterns, we also evaluated the generalization capability of our alignment framework beyond individual tuning. To this end, we trained an additional model, ReAlnet-AcrossSub, using EEG data pooled across all ten participants in the THINGS EEG2 training set. Concretely, we pooled the EEG data across subjects by treating all trials as coming from a single "super-subject", without averaging across them, and the model was trained without subject identifiers. Thus, instead of 10 individualized ReAlnets, we have a single ReAlnet-AcrossSub. Analogous to the single CORnet, we then conducted RSA by comparing the across-subject model against individual subjects' EEG and fMRI RDMs, as in our main analyses.

ReAlnet-AcrossSub shows significantly higher similarity to human EEG and fMRI representations when human participants viewed natural images compared to CORnet (Supplementary Fig. 11A–C). These results demonstrate that this across-subject model learned unified neural representations across individuals and generalized across neural modalities. However, ReAlnet-AcrossSub did not show model-fMRI similarity improvement on artificial shape or alphabetical letter images. And they did not improve behavioral similarity, performing slightly worse than CORnet (Supplementary Fig. 11D). This dissociation implies that individual-specific neural tuning may be important for improving more brain-like representations of shape or letter visual inputs and capturing human behavioral patterns. When comparing ReAlnet-AcrossSub to the original

(individualized) ReAlnets, ReAlnet-AcrossSub performed better in some cases (model-EEG similarity, model-fMRI similarity of shape and letter images, and model-behavior similarity) but worse in others (model-fMRI similarity of natural images).

Discussion

Building upon previous research utilizing neural data for aligning object recognition models, we propose a novel framework for human neural representational alignment, along with the corresponding human brain-like model, ReAlnet. Unlike previous studies that focused on using animal neural signals to optimize models or were unable to use global neural activity for comprehensive model optimization[25–29], our approach leverages human EEG activity to simultaneously optimize multiple layers of the model, enabling it to learn the human brain's internal representational patterns for object visual processing. Notably, unlike prior research relying on behavioral or single modality neural recording data for model evaluation[25–29,33], we employed different modalities of human neuroimaging data and also human behaviors for model evaluation to ensure that ReAlnets learn broader, cross-modal brain representational patterns. Additionally, we have extended our alignment framework to another convolutional neural network model to obtain ReAlnet-Rs and observed a similar enhancement in the similarity to human brain representations.

Recent advances in brain-inspired AI have explored various approaches to enhancing biological alignment, including self-supervised learning, recurrent architectures, and direct optimization of similarity to neural data. Unlike prior studies that primarily focus on invasive recordings from nonhuman primates, ReAlnets directly incorporate noninvasive human EEG data, offering a more accessible alternative for modeling human visual representations. While previous work has attempted to improve model-to-brain alignment at a single layer or specific brain region, our multi-layer encoding-based alignment framework enables a more comprehensive alignment across the visual processing hierarchy. This framework is particularly useful in contexts where biological plausibility is a key objective, such as cognitive neuroscience and brain-inspired AI. Recent studies in cognitive neuroscience have begun to apply CNNs to obtain visual features from complex stimuli[45–47]. Building on this, more brain-like models can now be used to extract representations that are closer to human neural patterns. This approach offers a new way to probe how neural representations evolve across different modalities (e.g., EEG vs. fMRI), stimuli, tasks, and individuals, thereby providing insights into the mechanisms underlying visual perception.

Regarding ReAlnets themselves, they effectively learn not just the patterns of EEG data, but appear to capture something even broader about the brain's internal processing patterns of visual information. The fact that ReAlnets show higher similarity than the original CORnet not only to within-modality EEG but also to cross-modality fMRI and behavior suggests that the learned representations in ReAlnets capture shared neural patterns that are consistent across different neuroimaging modalities. One possibility is that these shared patterns may reflect the encoding of fundamental visual features, from lower- to higher-level, that are robustly represented in the brain irrespective of the neuroimaging method. In essence, the EEG data could be capturing a core, image-specific representation that generalizes across modalities, providing crucial cues for perceptual inference. And such cross-modal generalization highlights the potential of the ReAlnet alignment framework as a flexible framework for extracting the key neural representations beyond single modalities. Our control experiments also highlight that only the model trained on image-specific EEG signals and including both MSE and contrastive learning losses shows this generalization performance. If we remove the contrastive learning loss, we can still see significant model-fMRI similarity improvement, but not as great as ReAlnets. If we remove the MSE loss or we employ shuffled (unpaired or time-scrambled) EEG signals to train models, the model-fMRI similarity became even worse than that of the original CORnet. Future work could further explore the specific factors that drive this generalization, such as shared representational structures in EEG and fMRI neural features. And extending this alignment framework to directly incorporate fMRI data or combining fMRI and EEG for joint training could further enhance the model's ability to capture cross-modal brain representations. Additionally, the ability of ReAlnets to generalize from EEG to fMRI suggests a degree of representational consistency across neural modalities. This observation raises the possibility that different neuroimaging techniques, despite capturing distinct aspects of neural activity (e.g., temporal dynamics in EEG vs. spatial patterns in fMRI), may share a common representational basis[48–50]. Investigating this consistency further could provide insights into the shared neural mechanisms underlying human visual processing and inform the design of more versatile brain-aligned models.

Recent findings have highlighted the critical role of the training image diet in determining the model-to-brain fit[51]. In addition to the generalization to other neuroimaging modalities, our ReAlnets trained on the THINGS EEG2 training dataset also demonstrated robust generalization to held-out THINGS categories. This generalization suggests that the networks can capture shared representational structures that are meaningful both within and beyond the training distribution. Such robustness indicates that training on semantically rich and ecologically valid images and the corresponding brain signals, such as THINGS EEG2, may promote the emergence of brain-like visual representations that generalize across diverse visual domains. On the one hand, our findings demonstrate that our approach improves model-brain similarity. On the other hand, these results may also suggest that selecting appropriate training datasets might be important for letting artificial models learn generalized human brain representations.

EEG signals provide high temporal resolution data that capture rapid neural dynamics underlying visual processing. By leveraging EEG data, our alignment framework enables the model to align with temporal features that likely correspond to distinct stages of visual processing, such as early sensory features and later semantic attributes. Also, compared to fMRI or electrophysiology, EEG is significantly more cost-effective, making it more practical for widespread use. Recent studies in cognitive and computational neuroscience using the THINGS EEG2 dataset have traced evidence of human visual feature processing, including object categories, size, depth, image entropy[41,52], and even reconstructing visual information through EEG signals[53–55] and realizing inter-subject EEG conversions[6]. In our internal representational analyses, we initially observed that ReAlnets showed enhanced representations along several object dimensions, such as food-related, artificial/hard, electronic/technology-related, and others. Further analysis revealed that a substantial portion of these enhancements—most prominently the food-related dimension—are driven primarily by visual-diet differences between ImageNet and THINGS. This demonstrates that visual diet alone accounts for most of the feature shifts. Crucially, however, when we controlled for the visual diet effect, we still identified several dimensions—such as electronic/technology-related, flat/patterned, and long-thin—that showed additional enhancement only in the EEG-aligned models. Together, these findings highlight two complementary influences on model representations: (1) visual diet, which induces broad, dataset-level shifts in representational geometry; (2) neural alignment, which provides more targeted refinements in specific dimensions. It does warrant further exploration to ascertain what specific information has been learned from the alignment with human brains. More analyses of the neural network's internal representations may be needed to delve into this. Also, from a reverse-engineering perspective, attempting to understand the brain-like optimization process of the model could further aid in unraveling the mechanisms by which our brains process visual information[56–60].

Interestingly, when we applied this alignment framework to ResNet18, the resulting ReAlnet-Rs still demonstrate more human brain-like representations, akin to those exhibited by ReAlnets. This framework's generalizability may suggest that our alignment framework is potentially applicable to aligning other AI models with human EEG signals. Therefore, one potential direction for future research is to examine whether this alignment framework could generalize to other model architectures or neuroimaging modalities. First, while the current study only utilizes EEG signals for

alignment, the observed generalization to fMRI and behavior suggests potential cross-modal consistency in the learned representations. However, we acknowledge that our findings do not directly demonstrate generalization from fMRI to EEG. Future work incorporating direct fMRI-based alignment or joint EEG-fMRI training would be necessary to establish the full bidirectionality of cross-modal generalization. Future studies could test whether this alignment framework could be extended to other neural modalities, such as fMRI and MEG (dimensionality reduction might be necessary for extensive neural data features), suggesting potential applicability of this framework to variants aligned with fMRI or MEG data. Second, another ambition is to adapt this framework to a wider range of models and tasks in the future, including language and auditory processing and self-supervised or unsupervised models, which may be adapted for other modalities and learning paradigms, including auditory or language inputs and unsupervised learning settings. Third, one direction for refinement includes integrating loss functions designed to emphasize task-relevant neural features. Also, although our framework successfully demonstrates generalization to fMRI data, further work is needed to systematically analyze the factors contributing to this transfer. For instance, identifying representational dimensions that are consistent across EEG and fMRI, and relating them to specific brain regions, could provide deeper insights into the mechanisms of cross-modal generalization.

Additionally, as a coarse pooled-data control, we also trained an across-subject variant (ReAlnet-AcrossSub) that removed subject identifiers and treated trials as from a single "super-subject". This pooling simplifies the setup but likely injects noise by ignoring inter-individual variability. Consistent with this, ReAlnet-AcrossSub improved model-EEG and model-fMRI similarity on natural images compared to CORnet, yet did not improve behavioral alignment and underperformed individualized models on shapes/letters. These observations suggest that individualized tuning is beneficial for behavior alignment, and that richer across-subject designs (e.g., subject IDs or multi-head readouts) may better capture shared vs. idiosyncratic structure in future work.

While this study demonstrates encouraging progress, several limitations warrant consideration. First, the relatively small sample size of EEG data and the high level of signal noise may limit the precision of model-to-brain alignment. Second, the lack of shared category labels across datasets—such as the absence of ImageNet labels for THINGS stimuli—complicates consistent model evaluation. These issues may constrain the model's performance both in brain alignment and object recognition. In addition, although ReAlnets were fine-tuned on EEG signals from individual subjects, they still inherit architectural and representational biases from the pre-trained image recognition model, which may limit alignment fidelity. Beyond these data-related factors, certain limitations may also stem from the alignment methodology itself. For instance, our framework employs relatively simple alignment objectives—such as MSE and contrastive losses—that may not fully capture the complex, nonlinear, and distributed nature of brain representations. Furthermore, the shallow encoding modules assume a direct mapping between neural activity and model features, potentially overlooking intermediate transformations or multistage processing. These design choices could limit the model's capacity to learn more abstract or hierarchically structured neural patterns. Thus, the modest effect sizes observed may reflect not only data constraints, but also the current methodological limits of representational alignment strategies.

Despite these limitations, the framework consistently improves alignment with human EEG, fMRI, and behavior. To build on this work, future efforts could assess whether training the model jointly on EEG and fMRI improves cross-modal generalization and whether the framework remains robust when using smaller or noisier EEG datasets. Moreover, evaluating how alternative training losses or network architectures impact representational similarity may reveal which components are critical for capturing brain-like representations. Finally, although the THINGS EEG2 dataset is ecologically valid, further work is needed to test whether these findings generalize to datasets with different semantic or contextual structures.

Overall, this study underscores the importance of investigating the limitations and potentials of human EEG signals in shaping brain-like AI. We employ a novel alignment framework using human EEG data to achieve more human brain-like vision models—ReAlnets. Demonstrating significant advances in bio-inspired AI, ReAlnets not only align closely with human EEG and fMRI but also exhibit hierarchical individual variability and increased similarity to human behavior, mirroring human visual processing. We hope that our alignment framework stands as a testament to the potential synergy between computational neuroscience and machine learning and enables the enhancement of diverse AI models to be more human brain-like, opening up exciting possibilities for future research in brain-like AI systems.

Methods

Here, we describe the human neural data (EEG data for the alignment, and both EEG and fMRI data for testing the similarity between models and human brains) we used in this study, the alignment pipeline (including the structure, the loss functions, and training and test methods) for aligning the model representations with human neural representations, and the evaluation methods for measuring representational similarity between models and human brains and human behaviors.

Human EEG data for representational alignment

Human EEG data were obtained from an EEG open dataset, THINGS EEG2[35], including EEG data from 10 healthy human subjects in a rapid serial visual presentation paradigm. Stimuli were images sized 500 × 500 pixels from the THINGS dataset[61], which consists of images of objects on a natural background from 1854 different object concepts. Before imputing the images to the model, we reshaped image sizes to 224 × 224 pixels and normalized the pixel values of images to ImageNet statistics. Subjects viewed one image per trial (100 ms). Each participant completed 66160 training set trials (1654 object concepts × 10 images per concept × 4 trials per image) and 16000 test set trials (200 object concepts × 1 image per concept × 80 trials).

EEG data were collected using a 64-channel EASYCAP and a BrainVision actiCHamp amplifier. We use already pre-processed data from 17 channels (O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, and P2) overlying occipital and parietal cortex. We re-epoched EEG data ranging from stimulus onset to 200 ms after onset with a sample frequency of 100 Hz. Thus, the shape of our EEG data matrix for each trial is 17 channels × 20 time points. and we reshaped the EEG data as a vector including 340 values for each trial. Before the model training and test, we averaged all the repeated trials (4 trials per image in the training set and 80 trials per image in the test set) to obtain more stable EEG signals.

It is worth noting that the training and test sets do not overlap in terms of object categories (concepts), which means that the performance of ReAlnets trained on the training set, when evaluated on the test set, can effectively reveal the model's generalization capability across different object categories.

Human fMRI data for cross-modality testing

To demonstrate that our approach of aligning with human EEG not only enhances the model's similarity to human EEG but indicates that ReAlnets have effectively learned the human brain's representational patterns more broadly, we also performed cross-modal testing, testing ReAlnets on data from a different modality (fMRI), from a different set of subjects, viewing a different set of images. The fMRI data originate from ref. 36. This Shen fMRI dataset recorded human brain fMRI signals from three subjects while they focused on the center of the screen viewing images. We selected the test set from the Shen fMRI dataset, which comprises fMRI signals of each subject viewing 50 natural images of different categories from ImageNet, 40 artificial shape images, and 10 alphabetical letter images, with each image being viewed 24, 20, and 12 times, respectively. We averaged the fMRI signals across the repeated trials to obtain more stable brain activity for each image observation and extracted signals from five ROIs for subsequent

comparison of model and human fMRI similarity: V1, V2, V3, V4, and the lateral occipital complex (LOC).

Image-to-brain encoding-based alignment pipeline

Basic architecture of ReAlnets and ReAlnet-Rs. We have chosen the state-of-the-art CORnet-S model[9,34] as the foundational architecture for ReAlnets, incorporating recurrent connections akin to those in the biological visual system and proven to more closely emulate the brain's visual processing. Both CORnet and ReAlnets consist of four visual layers (V1, V2, V4, and IT) and a category decoder layer. Layer V1 performs a 7×7 convolution with a stride of 2, followed by a 3×3 max pooling with a stride of 2, and another 3×3 convolution. Layer V2, V4, and IT each perform two 1×1 convolutions, a bottleneck-style 3×3 convolution with a stride of 2, and a 1×1 convolution. Apart from the initial layer V1, the other three visual layers include recurrent connections, allowing outputs of a certain layer to be passed through the same layer several times (twice in layer V2 and IT, and four times in layer V4). We have also chosen another widely used model in image recognition, ResNet18, as the foundational architecture to obtain human EEG-aligned ReAlnet-Rs, which consists of 18 layers (the last layer is the final decoder to output the predicted category label).

EEG generation module. For ReAlnets, in addition to the original recurrent convolutional neural network structure, we have added an EEG generation module designed to construct an image-to-brain encoding model for generating realistic human EEG signals. Each visual layer is connected to a nonlinear $N \times 128$ layer-encoder (Enc-V1, Enc-V2, Enc-V4, and Enc-IT correspond to layer V1, V2, V4, and IT) that processes through a fully connected network with a ReLU activation. These four layer-encoders are then directly concatenated to form an $N \times 512$ multi-layer visual encoder, which is subsequently connected to an $N \times 340$ EEG encoder through a linear layer to generate the predicted EEG signals. Here, N is the batch size. For eAlnet-Rs, which is highly similar to ReAlnets, we extracted the features from layer 5, layer 9, layer 13, and layer 17 to connect to four nonlinear $N \times 128$ layer-encoders through fully connected networks with ReLU activations, and these four layer-encoders are then directly concatenated to form an $N \times 512$ multi-layer visual encoder, which is subsequently connected to an $N \times 340$ EEG encoder through a linear layer to generate the predicted EEG signals.

Therefore, we aim for the model to not only perform the object classification task but also to generate human EEG signals, which can be highly similar to the real EEG signals when a person views the certain image through the EEG generation module with a series of encoders. During this process of generating brain activity, ReAlnet(-R)s' visual layers are poised to effectively extract features more aligned with neural representations.

Alignment Loss. Accordingly, the training loss $\mathcal{L}^A$ of our alignment framework consists of two primary losses, a classification loss and a generation loss with a parameter β that determines the relative weighting:

$$\mathcal{L}^A = \mathcal{L}^C + \beta \cdot \mathcal{L}^G \tag{1}$$

$\mathcal{L}^C$ represents the standard categorical cross entropy loss for model predictions on ImageNet labels:

$$\mathcal{L}^C = -\sum_{i=1}^{N} y_i \log(p_i) \tag{2}$$

Here, y_i represents the i-th image, and p_i represents the probability that model predicts the i-th image belongs to class i out of 1000 categories. However, the correct ImageNet category labels for images in the THINGS dataset are not available. Therefore, we adopt the same strategy as in ref. 25, using the labels obtained from the ImageNet pre-trained CORnet without neural alignment as the true labels to stabilize the classification performance of ReAlnets.

$\mathcal{L}^G$ is the generation loss, which includes a mean squared error (MSE) loss $\mathcal{L}^{\mathrm{MSE}}$ and a contrastive loss $\mathcal{L}^{\mathrm{Cont}}$ between the generated and real EEG signals. To compute the contrastive loss, we originally aimed to use Spearman correlation to measure the similarity between predicted and target signals. However, as Spearman correlation involves a ranking operation that is non-differentiable, we replaced it with Pearson correlation, which is a differentiable similarity metric. Specifically, the dissimilarity index was calculated as 1 minus Pearson correlation. This substitution allowed us to compute gradients effectively during backpropagation, ensuring the compatibility of the contrastive loss with gradient-based optimization. The contrastive loss aims to bring the generated signals from the same image (positive pairs) closer to the corresponding real human EEG signals and make the generated signals from different images (negative pairs) more distinct. $\mathcal{L}^G$ is calculated as followed:

$$\mathcal{L}^G = \mathcal{L}^{\mathrm{MSE}} + \mathcal{L}^{\mathrm{Cont}} \tag{3}$$

$$\mathcal{L}^{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(S_i - \widehat{S}_i)^2 \tag{4}$$

$$\mathcal{L}^{\mathrm{Cont}} = 1 + \frac{1}{N}\sum_{i=1}^{N}[1 - r(S_i, \widehat{S}_i)] - \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}[1 - r(S_i, \widehat{S}_j)] \tag{5}$$

Here, S_i and $\widehat{S}_i$ represent the generated and real EEG signals corresponding to the i-th image.

Training procedures. Unlike CORnet and ResNet18, which trained on purely image-based ImageNet dataset, ReAlnets and ReAlnet-Rs additionally trained on individual EEG data. According to ten subjects in the THINGS EEG2 dataset, we obtained ten personalized ReAlnets. Each network was trained to minimize the alignment loss, including both classification and generation losses with a static loss weight β of 100 and a static training rate of 0.00002 for 30 epochs using the Adam optimizer. We used a batch size of 16, meaning the contrastive loss computed dissimilarities of 256 pairs for each gradient step.

Additionally, for ReAlnets, we applied other three different β weights ($\beta = 1$, 10, or 1000) separately to train the model to further explore the impact of this β value on the performance of ReAlnets. We observed that with an increase in β, ReAlnets show greater similarity to human EEG and fMRI and more pronounced individual variability within models. However, only ReAlnets with $\beta = 100$ show significantly higher similarity to human behaviors. Thus, we suggest that $\beta = 100$ could be the best parameter to conduct the human EEG alignment. Supplementary Figs. 12–15 show the performance and similarity results of ReAlnets with different β values.

We tested the classification accuracy of ReAlnets on ImageNet at different β values (Supplementary Fig. 1A). Importantly, to ascertain that the observed decrease in accuracy was not due to the additional generation task compromising classification performance, but rather the absence of correct ImageNet labels for images in the THINGS EEG2 dataset, we trained a ReAlnet with $\beta = 0$. This ReAlnet excluded the EEG signal generation module but underwent fine-tuning with images from the THINGS EEG2 dataset. The results indicated that the ReAlnet with $\beta = 0$ also experienced a similar level of decline.

Control experiments. To systematically evaluate how different losses and how shuffled EEG data influence our results, we conducted four control experiments. The corresponding four control conditions are defined as follows: (1) *W/o ContLoss* control: we removed the contrastive loss component; (2) *W/o MSELoss* control: we removed the MSE loss component; (3) unpaired EEG control: For each trial, the entire 17-channel $\times$ 20-timepoint trial-averaged EEG response matrix was

preserved but randomly reassigned to a different image. This broke the true image-EEG pairing while preserving the full temporal and channel structure of the EEG data; (4) *Scrambled* EEG control: after trial averaging, within each 17 × 20 EEG response matrix, the 20 timepoints in each of the 17 channels were independently permuted. This procedure destroyed all temporally structured information and cross-channel synchrony while preserving the image-channel pairing and overall signal statistics (e.g., mean and variance across time).

Model-human similarity measurement

Neural similarity via representational similarity analysis (RSA). RSA is used for representational comparisons between models and human brains[62] based on first computing representational dissimilarity matrices (RDMs) for models and human neural signals, and then calculating Spearman correlation coefficients between RDMs from two systems.

To evaluate the similarity between models and human EEG, we calculated EEG RDMs using classification-based decoding accuracy as the dissimilarity index. While fMRI RDMs are typically calculated using 1 minus the correlation coefficient (below)[2,62–64], decoding accuracy is more commonly used for EEG RDMs[65,66]. Since EEG has a low SNR and includes rapid transient artifacts, Pearson correlations computed over very short time windows yield unstable dissimilarity estimates[67,68] and may thus fail to reliably detect differences between images. In contrast, decoding accuracy—by training classifiers to focus on task-relevant features—better mitigates noise and highlights representational differences. For each image, we extracted the corresponding EEG response matrix (80 trials × 17 channels × 20 timepoints). For each subject and each timepoint, we referred 80 trials per image as training samples and 17 channels as features, resulting in a total of 160 samples for every pair of two images *i* and *j*. A linear SVM classifier was employed to classify the EEG responses between two images using a 5-fold cross-validation approach. In each fold, the classifier was trained on 4 folds of the data and tested on the remaining fold, ensuring that the decoding accuracy reflected generalizable patterns. The average decoding accuracy across the 5 folds was computed and used as the dissimilarity measure between images *i* and *j*. And this process was repeated for all possible pairs of the 200 images, resulting in a 200 × 200 RDM for each subject and each time point. For model RDMs, we input 200 images into each model and obtained latent features from each visual layer. Then, we constructed each layer's RDM by calculating the dissimilarity using 1 minus the Pearson correlation coefficient between flattened vectors of latent features corresponding to any two images. To compare the representations, we calculated the Spearman correlation coefficient as the similarity index between layer-by-layer model RDMs and timepoint-by-timepoint neural EEG RDMs.

To evaluate the similarity between models and human fMRI, we used the Shen fMRI dataset, which includes fMRI activation patterns for different image categories (natural images, artificial shape images, and alphabetical letter images). We calculated separate RDMs for each category, and the RDM dimensions correspond to the number of images in each category: 50 × 50 for natural images, 40 × 40 for artificial shape images, and 10 × 10 for alphabetical letter images. We used the GLM activation values (beta weights) provided in the open dataset for each image and each voxel, and calculated RDMs based on the common correlation-based method[2,62–64], as follows. For each subject and ROI, we defined the voxel-wise activation pattern for each image as the vector of activation values (e.g., if a given ROI had 250 voxels, this would be a 1 × 250 vector for each image). Then, for every pair of images, we computed the dissimilarity index as 1 minus the Pearson correlation coefficient between the two voxel-wise activation vectors corresponding to the two images. This was repeated for all image pairs, resulting in a symmetrical dissimilarity matrix (the RDM) for each ROI and each subject. For model RDMs, similar to the EEG comparisons above, we obtained the RDM for each layer from each model. Then, we calculated the Spearman correlation coefficient as the similarity index between layer-by-layer model RDMs and neural fMRI RDMs for different ROIs, assigning the final similarity for a certain brain region as the highest similarity result across model layers due to the lack of a clear correspondence between different model layers and brain regions. All RSA analyses were implemented based on the NeuroRA toolbox[69].

Behavioral similarity via Brain-Score. Brain-Score is a framework that evaluates how similar ANNs are to the primate visual system[37]. To measure the behavioral similarity between ReAlnets and humans (and monkeys) in visual recognition tasks, we used two behavioral benchmarks from the Brain-Score framework: (https://github.com/brain-score/vision). "Rajalingham2018public-i2n"[7] task uses grayscale images where objects are manipulated by varying position, size, viewing angle, and background, while "Geirhos2021-error_consistency"[42] task employs out-of-distribution colorful natural images. Both tasks calculate behavioral similarity between the model and human (and primates) observers using the error consistency method, which measure whether there is above-chance overlap in the specific images that humans and models classify incorrectly. The behavioral Brain-Score is calculated by taking the average of two behavioral benchmarks. We compared the results from ReAlnets and ReAlnet-Rs to the behavioral Brain-Score of CORnet and ResNet, respectively, using the same benchmarks. For more detailed information about the behavioral benchmarks used in this study, please refer to the original papers by refs. 7,37,42.

Individual variability in ReAlnets. To quantify inter-individual representational variability among the individualized ReAlnets, we computed a variability index at each model layer. Specifically, we extracted the RDM from each individualized ReAlnet (one per subject) for a given layer and then calculated all pairwise Spearman correlations between the RDMs of the ten models. The variability index was defined as the average of one minus Spearman correlation coefficient between two RDMs. A higher value indicates greater variability (i.e., less similarity) across models.

Model internal representational analysis

This section describes the methodology used for the analysis in results-refined object feature representations in ReAlnets section, where we examined which object feature dimensions were more strongly encoded in ReAlnets compared to CORnet. Specifically, we utilized 49 object feature dimensions from the THINGS dataset[43]. Each object concept is represented along these 49 feature dimensions. Our analysis focused on the 200 images in the test set of the THINGS EEG2 dataset, which were not part of the model's training data. We applied an RDM-based partial Spearman correlation method for the analysis. First, we computed the RDM for the IT layer of each model, which contains higher-level information, and 49 feature RDMs based on the 200 images by calculating the absolute differences in feature encoding strength between pairs of images as dissimilarity measures. Next, we computed the partial correlation between the model RDM and each feature RDM, while controlling for the other 48 feature RDMs. Finally, we calculated the square of the partial correlation coefficient to determine the variance explained by the model for each object feature dimension and got the top-3 improved feature dimensions of ReAlnets.

Statistics and reproducibility

All statistical analyses were conducted using custom scripts in Python. Statistical tests were two-sided, and exact *p*-values and effect sizes are reported in the main text, figure captions, or Supplementary Tables. Sample sizes were determined by the available datasets. Comparisons between models were performed using one-sample or paired-sample t-tests across different model instances. We considered $p < 0.05$ as the threshold for significance. To ensure that statistical significance reflects meaningful differences, we also report effect sizes in the Supplementary Materials. Random seeds were fixed across training runs to ensure reproducibility, and all analyses were conducted on held-out test datasets with no overlap with training data.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The EEG data (THINGS EEG2 dataset) used are available as open data via the Open Science Framework (OSF) repository: https://osf.io/3jk45/[35,70], and the fMRI data (Shen fMRI dataset) used are available as open data via the figshare repository: https://figshare.com/articles/Deep_Image_Reconstruction/7033577[36,71]. The numerical Source data for all graphs in this paper can be found in Supplementary Dataset 1 and Dataset 2.

Code availability

The models and the analysis code can be assessed at https://github.com/ZitongLu1996/ReAlnet.

Received: 22 October 2024; Accepted: 29 January 2026;
Published online: 20 February 2026

References

1. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 1–13 (2016).
3. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
4. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. USA* **116**, 21854–21863 (2019).
5. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619–8624 (2014).
6. Lu, Z. & Golomb, J. Generate your neural signals from mine: individual-to-individual EEG converters. In *Proc. Annual Meeting of the Cognitive Science Society* (CogSci, 2023).
7. Rajalingham, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
8. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
9. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2019).
10. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
11. Tang, H. et al. Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. USA* **115**, 8835–8840 (2018).
12. Bai, S., Li, Z. & Hou, J. Learning two-pathway convolutional neural networks for categorizing scene images. *Multimed. Tools Appl.* **76**, 16145–16162 (2017).
13. Choi, M., Han, K., Wang, X., Zhang, Y. & Liu, Z. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2023).
14. Han, Z. & Sereno, A. Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Comput.* **34**, 138–171 (2022).
15. Han, Z. & Sereno, A. Identifying and localizing multiple objects using artificial ventral and dorsal cortical visual pathways. *Neural Comput.* **35**, 249–275 (2023).
16. Sun, T., Wang, Y., Yang, J. & Hu, X. Convolution neural networks with two pathways for image style recognition. *IEEE Trans. Image Process.* **26**, 4102–4113 (2017).
17. Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K. & Grill-Spector, K. Topographic DCNNs trained on a single self-supervised task capture the functional organization of cortex into visual processing streams. In *Proc. NeurIPS 2022 Workshop SVRHM* (NeurIPS, 2022).
18. Lee, H. et al. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.09.185116 (2020).
19. Lu, Z. et al. End-to-end topographic networks as models of cortical map formation and human visual behaviour. *Nat. Hum. Behav.* **9**, 1975–1991 (2025).
20. Margalit, E. et al. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*. **112**, 2435–2451 (2024).
21. Konkle, T. & Alvarez, G. Cognitive steering in deep neural networks via long-range modulatory feedback connections. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2023).
22. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 1–12 (2022).
23. Prince, J. S., Alvarez, G. A. & Konkle, T. Contrastive learning explains the emergence and function of visual category-selective regions. *Sci. Adv.* **10**, eadl1776 (2024).
24. O'Connell, T. P. et al. Approximating Human-Level 3D visual inferences with deep neural networks. *Open Mind.* **9**, 305–324 (2025).
25. Dapello, J. et al. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *Proc. International Conference on Learning Representations* (ICLR, 2023).
26. Federer, C., Xu, H., Fyshe, A. & Zylberberg, J. Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Netw.* **131**, 103–114 (2020).
27. Li, Z. et al. Learning from brains how to regularize machines. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2019).
28. Pirlot, C., Gerum, R. C., Efird, C., Zylberberg, J. & Fyshe, A. Improving the accuracy and robustness of CNNs using a deep CCA neural data regularizer. Preprint at https://doi.org/10.48550/arXiv.2209.02582 (2022).
29. Safarani, S. et al. Towards robust vision by multi-task learning on monkey visual cortex. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2021).
30. Fong, R. C., Scheirer, W. J. & Cox, D. D. Using human brain activity to guide machine learning. *Sci. Rep.* **8**, 1–10 (2018).
31. Spampinato, C. et al. Deep learning human mind for automated visual classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6809–6817 (IEEE, 2017).
32. Palazzo, S. et al. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3833–3849 (2021).
33. Fu, K., Du, C., Wang, S. & He, H. Improved video emotion recognition with alignment of CNN and human brain representations. *IEEE Trans. Affect. Comput.* **14**, 1–15 (2023).
34. Kubilius, J. et al. CORnet: modeling the neural mechanisms of core object recognition.Preprint at *bioRxiv* https://doi.org/10.1101/408385 (2018).
35. Gifford, A. T., Dwivedi, K., Roig, G. & Cichy, R. M. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* **264**, 119754 (2022).

36. Shen, G., Horikawa, T., Majima, K. & Kamitani, Y. Deep image reconstruction from human brain activity. *PLoS Comput. Biol.* **15**, e1006633 (2019).
37. Schrimpf, M. et al. Brain-score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://doi.org/10.1101/407007 (2020).
38. Teichmann, L., Hebart, M. N. & Baker, C. I. Dynamic representation of multidimensional object properties in the human brain. *J. Neurosci.* e1057252026, https://doi.org/10.1523/JNEUROSCI.1057-25.2026 (2026).
39. Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D. & Oliva, A. Tracking the spatiotemporal neural dynamics of real-world object size and animacy in the human brain. *J. Cogn. Neurosci.* **30**, 1559–1576 (2018).
40. Wang, R., Janini, D. & Konkle, T. Mid-level feature differences support early animacy and object size distinctions: evidence from electroencephalography decoding. *J. Cogn. Neurosci.* **34**, 1670–1680 (2022).
41. Lu, Z. & Golomb, J. D. Human EEG and artificial neural networks reveal disentangled representations and processing timelines of object real-world size and depth in natural images. *eLife* **13**, RP98117 (2025).
42. Geirhos, R. et al. Partial success in closing the gap between human and machine vision. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)* Vol. 34 (Neural Information Processing Systems Foundation, Inc., 2021).
43. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
44. Shao, Z. et al. Probing Human Visual Robustness with Neurally-Guided Deep Neural Networks. Preprint at https://doi.org/10.48550/arXiv.2405.02564 (2025).
45. McMahon, E., Bonner, M. F. & Isik, L. Hierarchical organization of social action features along the lateral visual pathway. *Curr. Biol.* **33**, 5035–5047.e8 (2023).
46. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
47. Jagadeesh, A. V. & Gardner, J. L. Texture-like representation of objects in human visual cortex. *Proc. Natl. Acad. Sci. USA* **119**, e2115302119 (2022).
48. Cichy, R. M. & Oliva, A. A M/EEG-fMRI fusion primer: resolving human brain responses in space and time. *Neuron* **107**, 772–781 (2020).
49. Lee Masson, H. & Isik, L. Rapid processing of observed touch through social perceptual brain regions: an EEG-fMRI fusion study. *J. Neurosci.* **43**, 7700–7711 (2023).
50. Hu, Y. & Mohsenzadeh, Y. Neural processing of naturalistic audiovisual events in space and time. *Commun. Biol.* **8**, 1–16 (2025).
51. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 1–18 (2024).
52. Muukkonen, I. & Salmela, V. Entropy predicts early MEG, EEG and fMRI responses to natural images. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.21.545883 (2023).
53. Li, D., Wei, C., Li, S., Zou, J. & Liu, Q. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2024).
54. Du, C., Fu, K., Li, J. & He, H. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10760–10777 (2023).
55. Song, Y. et al. Decoding natural images from eeg for object recognition. In *Proc. International Conference on Learning Representations* (ICLR, 2024).
56. Ayzenberg, V., Blauch, N. & Behrmann, M. Using deep neural networks to address the how of object recognition. Preprint at https://doi.org/10.31234/osf.io/6gjvp (2023).
57. Cichy, R. M. & Kaiser, D. Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
58. Doerig, A. et al. The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
59. Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).
60. Lu, Z. & Ku, Y. Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *iScience* **26**, 108501 (2023).
61. Hebart, M. N. et al. THINGS: a database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE* **14**, 1–24 (2019).
62. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 249 (2008).
63. Nili, H. et al. A toolbox for representational similarity analysis. *PLOS Comput. Biol.* **10**, e1003553 (2014).
64. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
65. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J. Cogn. Neurosci.* **29**, 677–697 (2017).
66. Xie, S., Kaiser, D. & Cichy, R. M. Visual imagery and perception share neural representations in the alpha frequency Band. *Curr. Biol.* **30**, 2621–2627 (2020).
67. Kappenman, E. S. & Luck, S. J. The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology* **47**, 888–904 (2010).
68. Luck, S. J. *An Introduction to the Event-Related Potential Technique* 2nd edn (MIT Press, 2014).
69. Lu, Z. & Ku, Y. NeuroRA: a Python toolbox of representational analysis from multi-modal neural data. *Front. Neuroinformatics* **14**, 61 (2020).
70. Gifford, A. T., Dwivedi, K., Roig, G. & Cichy, R. M. A large and rich EEG dataset for modeling human visual object recognition https://osf.io/3jk45/ (2022).
71. Shen, G. et al. Deep image reconstruction dataset https://figshare.com/articles/Deep_Image_Reconstruction/7033577 (2019).

Acknowledgements

This work was supported by grants from the National Institutes of Health (R01-EY025648) and the National Science Foundation (NSF 1848939) to Julie D. Golomb. We thank the Ohio Supercomputer Center and Georgia Stuart for providing the essential computing resources and support. We thank Yuxuan Zeng for the "ReAlnet" name suggestion. We thank Tianyu Zhang, Shuai Chen, Jiaqi Li, and some other members in the Memory and Perception Reviews Reading Group (RRG) for helpful discussions about the methods and results. We thank Yuxin Wang for constructive feedback on the manuscript.

Author contributions

Conceptualization: Z.L. Formal analysis: Z.L., and Y.W. Funding acquisition: J.D.G. Investigation: Z.L., and Y.W. Methodology: Z.L., and Y.W. Resources: J.D.G. Project administration: Z.L. Visualization: Z.L. Writing—original draft preparation: Z.L. Writing—review and editing: Z.L., Y.W., and J.D.G.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-09685-w.

Correspondence and requests for materials should be addressed to Zitong Lu.

Peer review information *Communications Biology* thanks Ilya Kuzovkin, Bhavin Choksi and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Jasmine Pan. A peer review file is available.