Systematic comparison of color representations between humans and deep neural networks: towards predicting human color perception in a vast color space

Nipun Ravindu Wickramanayaka[1] and Masafumi Oizumi[1*]

1 Graduate School of Arts and Science, The University of Tokyo, Tokyo, Japan

* c-oizumi@g.ecc.u-tokyo.ac.jp

Abstract

The representational structure of human color perception, particularly within vast color spaces, remains incompletely understood. While representational structures for dozens of colors have been studied, exploring structures spanning hundreds or thousands of colors has been infeasible due to the time costs of psychophysical experiments. Given these constraints, deep neural networks (DNNs) have attracted attention as a promising tool for providing proxies or predictions of human perception beyond the scope of psychophysical experiments. However, it remains unclear which DNN models possess internal representations that geometrically align with human color perception. Furthermore, it is unclear which learning paradigm enables DNNs to acquire a color representation that is geometrically aligned with that of humans. Here, we systematically investigate which learning paradigm enables DNNs to produce a color representation that is structurally congruent with that of humans, with a focus on three types: self-supervised learning (SSL) that trains on images alone, supervised learning (SL) that trains on images with category labels, and contrastive language-image pre-training (CLIP) that trains on image-text pairs. We compared internal representations of DNNs with the human similarity judgments of 93 colors using a rigorous unsupervised method termed Gromov-Wasserstein Optimal Transport (GWOT), which reveals whether the representational structures of humans and models align at the fine-item level. Our results show that only the CLIP paradigm acquires color representations that strongly align with human data at the fine-item level. Furthermore, when we leveraged a key advantage of DNNs and investigated a substantial representational structure of 4096 colors, the human-aligned CLIP models consistently converged on a non-trivial distorted ring-like structure, which presents a plausible prediction for the large-scale human color representation. Our work demonstrates an approach for exploring unknown territories of human perception through the use of computational models validated in a limited empirical space, and provides predictions for future large-scale psychophysical experiments.

Author summary

How do we perceive the vast world of color? While scientists have mapped how humans organize a few dozen colors, we still do not know the structure of the massive "color map" that might underlie our perception of thousands of colors, as testing this directly is practically impossible. To explore this space, we turned to deep neural networks, a form of AI. Our first step was to identify models that "see" color in a way that matches

humans. We compared models trained under three different conditions: on images alone, on images with categorical labels, and on images paired with rich text descriptions. Using a powerful geometric comparison method, we found that only models trained jointly on images and language strongly matched the human color map. This allowed us to use these models as reliable computational proxies. We then used them to do what human experiments currently cannot: chart a vast map of 4,096 colors. The human-aligned models consistently converged on a unique, non-trivial ring-like structure with specific protrusions and indentations. Our work provides the first plausible, testable prediction for the large-scale structure of human color perception and demonstrates a new way to explore otherwise unreachable territories of our perceptual world.

Introduction

Understanding the detailed representational structure of human color perception is a fundamental challenge in vision and cognitive sciences. While the biological foundations of trichromatic color vision are well-established [1, 2], the precise geometric structure through which humans organize the vast spectrum of perceivable colors remains incompletely understood. Specifically, although previous studies have investigated the geometric structure of human color perception through a variety of psychophysical experiments, including behavioral judgments [3–9], electroencephalography (EEG) [10], magnetoencephalography (MEG) [11], and functional magnetic resonance imaging (fMRI) [12–14], the practical limitations of testing human subjects has typically restricted these investigations to limited stimuli sets numbering in the dozens. A recent study broadened the scale of investigation by successfully mapping the intricate similarity structure of a 93-color stimulus set that included different tones and achromatic colors [15]; nevertheless, the representational structure for even larger color spaces, such as those spanning thousands of colors, remains largely unexplored.

One powerful methodology for addressing these experimental limitations is the use of deep neural networks (DNNs). DNNs have emerged as powerful tools in computer vision [16–22]; they have achieved remarkable performance across a wide range of perceptual tasks [23–27] and are increasingly regarded as effective computational proxies for the human visual perception [28–30]. The use of DNNs enables experiments in ways that are not feasible with human subjects.

However, which DNN models exhibit internal representations that geometrically align with human color perception remain to be established. Previous researches on comparisons of DNN and human color representations have successfully captured correlations using stimulus sets of approximately 20 colors [31–33]. A recent study made progress by comparing a larger set of 93 colors more rigorously, showing that the output of the state-of-the-art multimodal large language model GPT-4 [34] is consistent with human color similarity judgments [35]. However, the study was unable to clarify whether the internal representations of GPT-4 are naturally consistent with human color perception.

Furthermore, learning paradigms that facilitate the acquisition of a color representation that is geometrically consistent with that of humans have not been identified. Even the study above [35], which provides the most direct insight to date, showed only that the output of the unimodal large language model GPT-3.5 had limited alignment with human color representation, whereas that of the multimodal large language model GPT-4, which integrates language and vision, achieved remarkable alignment. This finding suggests that language-only models are insufficient, but does not clarify the open question of whether visual information alone is sufficient, or if the integration of visual and linguistic information is necessary.

To address these gaps, our study investigates the contribution of visual information and linguistic grounding to forming a human-like color representation by systematically comparing the internal representations of DNNs trained under three distinct learning paradigms. Specifically, we analyze (1) self-supervised learning (SSL) models trained on images alone [36, 37], (2) supervised learning (SL) models trained on images with category labels [22], and (3) contrastive language–image pretraining (CLIP) models trained on images paired with rich textual descriptions [38]. We compare these models with human similarity judgments of 93 colors [15], representing the largest human-based color similarity dataset to date. This comparison allows us to assess which type of learning experience is effective for developing a representational structure that is consistent with that of human color perception, namely the use of visual information alone, visual information with categorical labels, or visual information grounded in rich language.

Of note, a rigorous comparison of these models requires a more sensitive method than standard correlation-based analyses [39–41]. Although such analyses can capture coarse-grained similarities, they are limited by an inherent ambiguity: a high correlation score cannot distinguish whether the alignment is due to a broad, group-to-group correspondence or to a fine-grained, one-to-one correspondence between individual elements [40, 42]. To overcome this limitation and directly assess the fine-grained geometric structure of the representations, we employed Gromov–Wasserstein Optimal Transport (GWOT) [43, 44]. This powerful unsupervised framework allows for sensitive comparison of representational geometry [15, 29, 35, 42, 45–47], making it particularly suitable for detecting subtle differences across models.

Our analysis reveals that strong alignment with human color similarity judgments is achievable only within the CLIP paradigm. In contrast, models trained under self-supervised or supervised learning paradigms failed to exhibit a color representation that is consistent with human data in unsupervised comparison, despite achieving higher correlations than the RGB baseline in supervised comparisons. Furthermore, by exploiting the unique advantage of DNNs to operate beyond human experimental constraints, we probed the geometry of a vast 4096-color representation and found that human-aligned CLIP models converged upon a unique ring-like structure with specific protrusions and indentations. This discovery offers a tangible computational prediction for the large-scale organization of human color perception, and paves the way for future, large-scale psychophysical validation. Our "select-and-predict" method thus demonstrates an approach for generating novel and nontrivial hypotheses about human color perception that are presently beyond the reach of direct empirical investigation.

Methods overview

Analysis framework

Our study employs a comprehensive analytical framework to systematically compare the structure of color representations between human behavioral data and various deep neural networks (DNNs) in order to probe which learning paradigms enable DNNs to develop internal representations that are structurally analogous to those of human color perception. As briefly illustrated in Figure 1, this process involves two primary stages.

The first stage is the preparation of representational dissimilarity matrices (RDMs) for both the human behavioral data and the various DNN models, which form the basis of our comparison (Fig. 1A). To establish the human-derived RDM, we first constructed a matrix of the rates from the discrete similarity judgments collected in the previous study [15] for 93 colors selected from the Practical Color Coordinate System (PCCS) [48]. We applied multidimensional scaling (MDS) to this matrix to produce a

set of continuous vector embeddings that are directly comparable to the embeddings of DNN, from which we then computed the RDM as the pairwise Euclidean distances between the 93 embeddings. For the DNN-derived RDMs, we first extracted the final-layer embeddings for the same 93 colors from models trained under three distinct learning paradigms: self-supervised learning (SSL) trained on images alone; supervised learning (SL) trained on images with category labels; and contrastive language-image pre-training (CLIP) trained on images paired with rich textual descriptions. The RDM for each model was then computed from the pairwise Euclidean distances between these 93 embeddings.

In the second stage, we compared the human-derived RDM with each of the DNN-derived RDMs using two complementary methods (Fig. 1B). The first method is a supervised comparison which is based on a fixed one-to-one correspondence between the colors. For this, we computed the Pearson correlation between the RDMs, known as conventional Representational Similarity Analysis (RSA). This supervised comparison is a standard and widely used approach in neuroscience [39–41], but is limited by an inherent ambiguity: a high correlation score cannot distinguish between a coarse, group-to-group correspondence (in our case, color category to color category alignment) and a fine-grained, one-to-one correspondence (in our case, color to color alignment) [42]. To capture these finer-grained structural properties, our second method is a more rigorous unsupervised comparison using Gromov-Wasserstein Optimal Transport (GWOT) [43, 44]. This method finds an optimal probabilistic mapping, or "transportation plan," that minimizes the difference between the internal distance structures of two point clouds (in our case, the two sets of color representations). This alignment process is entirely unsupervised, as it does not use the identity labels of the items being compared (in our case, the colors). This in turn allows for the direct comparison of their intrinsic geometric properties, which cannot be measured by correlation alone. After the alignment is complete, we then quantified the accuracy of the resulting alignment by calculating a "top-1 matching rate" score, which represents the percentage of colors for which the transportation plan correctly assigns the highest mapping probability to its corresponding color in the other structure. This dual-analysis framework provides a multi-faceted evaluation of which learning paradigms produce human-like color representations.

Selection of DNN models for comparing learning paradigms

To investigate the learning mechanism that enables DNNs to produce internal representations that are consistent with human color perception, this study systematically compares the color representations of DNN models trained under three distinct learning paradigms that represent different types of available information. We evaluated a total of 16 DNN models: four self-supervised (SSL) models trained on images alone, three supervised (SL) models trained on images with category labels, and nine CLIP models trained on images paired with rich textual descriptions. The models are detailed in the Materials and Methods section. The results of 93 colors for all 16 models are presented in the Supporting Information. A representative subset of this larger group are shown in the Results section to facilitate a clear and controlled comparison of the paradigms.

The specific models chosen for this representative subset were selected to ensure a fair comparison by controlling for the training dataset and varying only the model architecture. For the SSL and SL paradigms, all models featured in the main results were trained on the ImageNet dataset [49] using the Masked Autoencoder (MAE) objective and a standard supervised object classification, respectively. For the CLIP paradigm, which included models trained on three different datasets (LAION [50], MetaCLIP [51], and DFN [52]), we selected the DFN-trained models for the main

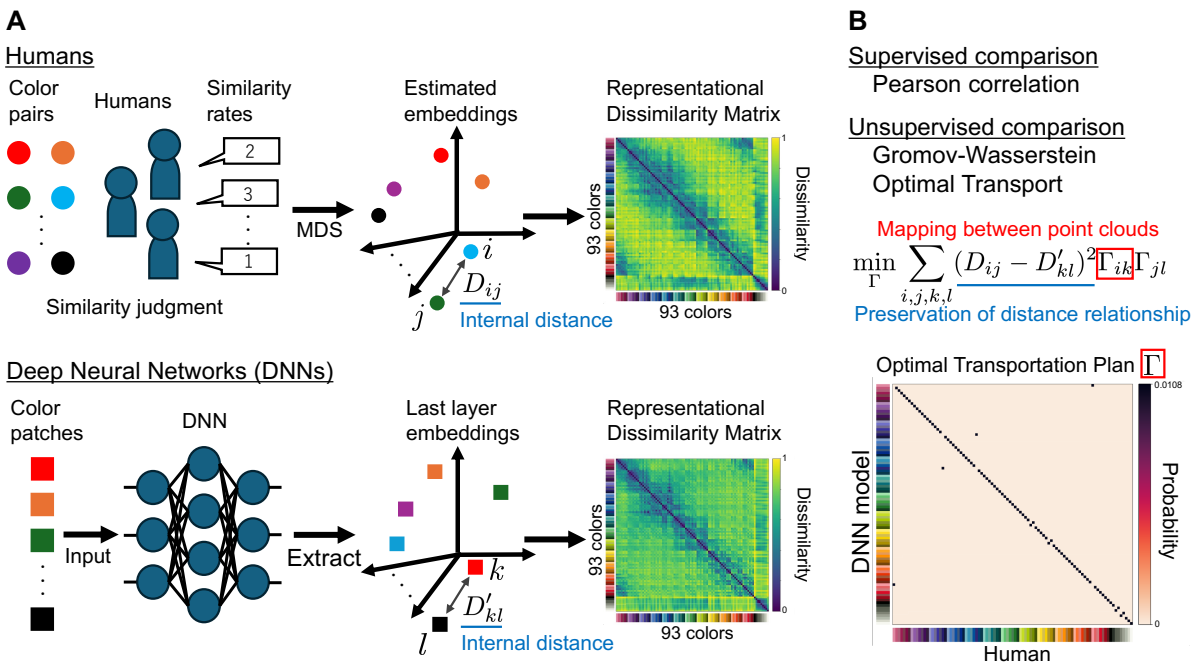

Fig 1. Overview of the analytical framework. (A) Preparation of the representational dissimilarity matrix (RDM). Human data were obtained as large-scale similarity judgments of 93 colors made by human participants in a previous study [15], who rated the similarity of color pairs on a scale from 0 ("very similar") to 7 ("very different"). Application of multidimensional scaling (MDS) to these ratings produced a set of continuous vector embeddings, which was used to make the human RDM used in this study. For the DNN models, embeddings for the same 93 colors were extracted from the final layer to make the DNN RDM. (B) Comparison of RDM. We compared the RDMs between human and DNNs by two methods, supervised comparison by Pearson correlation and unsupervised comparison by Gromov-Wasserstein Optimal Transport (GWOT).

results. This selection was guided by the results of our unsupervised alignment analysis, in which the DFN-trained models exhibited by far the strongest structural alignment with the human behavioral data. This presentation of the paradigm's most successful examples facilitates a clear comparison focused on the potential of each learning paradigm, rather than confounding the comparison with less successful instances of the same paradigm. Within each of these selected groups, the main results feature three models that differ only in their architecture size (Vision Transformer Base (ViT-B), Large (ViT-L), and Huge (ViT-H); listed in increasing order of size).

Results

Visualization of the color dissimilarity matrices and the color representations of the humans and DNN models

We first qualitatively compared the structure of color representations through a visual inspection of the representational dissimilarity matrices (RDMs) (Fig. 2A-E). In the human RDM, colors belonging to the same or similar hues are represented as being close to each other, resulting in a block-diagonal-like structure; conversely, colors with dissimilar hues are represented as being far apart (Fig. 2A). The RDMs from all three learning paradigms are visibly distinct not only from the simple structure of the baseline RGB model (Fig. 2B) but also from each other (Fig. 2C-E), indicating that

each learning process transforms the input color space to acquire its own unique representational structure. While models from all three paradigms exhibited a block-diagonal-like structure to some extent, visual inspection of the RDMs alone was insufficient to determine which model was most similar to the human data (Fig. 2A-E, S1).

The representational structures showed clearer qualitative similarities and dissimilarities when visualized using 2-dimensional Multidimensional Scaling (MDS), which revealed that the structure of the CLIP paradigm was visually similar to the human data (Fig. 2F-J). The human 93-color representation formed a wide, complex ring structure with intricate tones, and incorporated many achromatic colors (Fig. 2F). In contrast, although the chromatic colors sampled from PCCS form a distinct ring-like for each tone, the RGB baseline resulted in a filled-in structure when all 93 colors, including multiple tones and achromatic colors, were plotted together (Fig. 2G). Among the DNNs, the CLIP model exhibited a wide, complex ring-like structure similar to the human one, with intricate tones and achromatic colors (Fig. 2J). This tendency was consistent regardless of architectural size (Fig. S2I-K). The MAE model also formed a wide, complex ring-like structure with intricate chromatic colors (Fig. 2H), but this trend was dependent on architectural size (Fig. S2C-E). In contrast to these paradigms, the SL models failed to form a clear ring-like structure (Fig. 2I, S2F-H). To better understand these qualitative findings of varying degrees of similarity across the human and model representations, we conducted a quantitative analysis to objectively assess the degree of similarity between the representational structures.

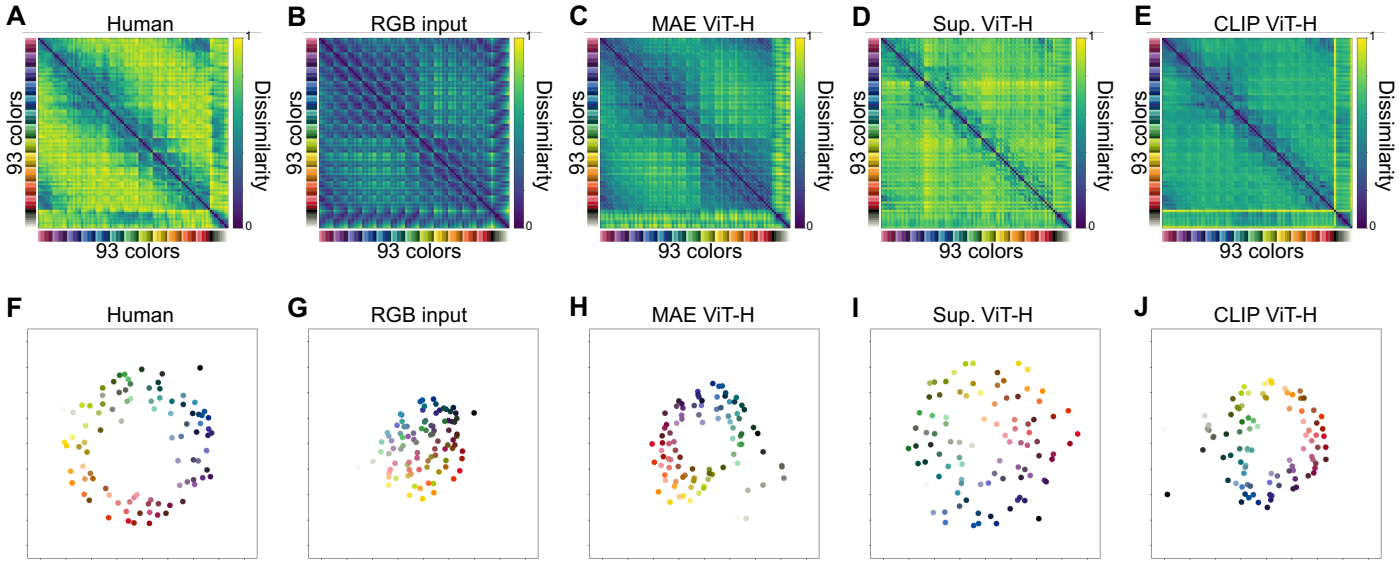

Fig 2. 93 colors representation of humans, RGB input, and DNN models. (A)-(E) Dissimilarity matrices of 93 colors from (A) the color-neurotypical participants group, (B) the RGB input, (C) a sample from the MAE models, (D) a sample from the SL models, and (E) a sample from the CLIP models. All matrices are normalized to have values between 0 and 1, where 0 means no difference between colors and 1 means the maximum difference for each dissimilarity matrix. (F)-(J) Representation of 93 colors from (F) the color-neurotypical participants group, (G) the RGB input, (H) a sample from the MAE models, (I) a sample from the SL models, and (J) a sample from the CLIP models. All representations are made by applying 2-dimensional MDS on each dissimilarity matrices.

Correlations between the color representations of humans and DNN models

A quantitative analysis using Pearson correlation revealed that most DNN models exhibited higher correlations with the human RDM than the RGB baseline (Fig. 3). Notably, even the baseline RGB model, which represents the input colors prior to any learning, showed a substantial correlation with the human RDM ($\rho = 0.64$). All featured DNN models, with the exception of one supervised learning (SL) model, surpassed this baseline score. A consistent trend was observed: on average, CLIP models achieved the highest correlations, followed by the MAE models and then the SL models. Among the different architectures within the same learning paradigms, only the SL models exhibited significant variations, with the ViT-B architecture showing a markedly lower value compared to the ViT-L and ViT-H architectures (Fig. 3, S3A).

Fig 3. Correlations of color similarity structures for 93 colors between humans and DNN models. Quantitative comparison based on Pearson correlations between the dissimilarity matrices of humans and DNN models.

A CLIP model aligns with humans in an unsupervised manner

While the preceding correlation-based analysis revealed high correlation scores for most DNN models, this finding is limited by an inherent ambiguity; it cannot distinguish whether the alignment is due to a coarse, group-to-group correspondence (in our case, alignment of colors within the same hue as represented in the block diagonal structure in RDMs) or a fine-grained, one-to-one correspondence (in our case, same color-to-same color alignment) [40, 42]. To resolve this ambiguity, we directly assess their geometric structure using the unsupervised alignment method, Gromov-Wasserstein Optimal Transport (GWOT) [43, 44]. Before presenting a systematic comparison across all models, we first illustrate the principles of this more sensitive analysis using a single, representative example: alignment of the human RDM (Fig. 4A) with that from a high-performing CLIP model (Fig. 4B).

The goal of GWOT is to find an optimal mapping, or the "optimal transportation plan", that minimizes the structural difference between two point clouds, quantified as the Gromov-Wasserstein Distance (GWD). Two RDMs (Fig. 4A and B) are used as inputs for finding the optimal transportation plan. As this optimization is non-convex, we ran the algorithm multiple times from different random initializations and selected

the solution that achieved the lowest GWD (Fig. 4C), which corresponds to the optimal transportation plan shown in Fig. 4D.

The GWOT analysis of this representative CLIP model revealed a high degree of structural congruence with the human data. The resulting optimal transportation (OT) plan, shown in Fig. 4D, visualizes the probabilistic mapping between the 93 colors of the human similarity structure and those of the CLIP model, exhibiting a remarkable concentration of probability mass along its main diagonal. This diagonal concentration indicates that for most colors, the highest mapping probability is assigned between the corresponding colors in the human and the CLIP model similarity structures. This correct one-to-one correspondence (top-1 matching rate) was achieved for 87.1% of the colors, even though the alignment was performed without any supervised information about color identities. Achievement of such a high matching rate via an unsupervised method strongly suggests that the intrinsic geometric structure of the CLIP model's color representation is highly consistent with that of human color perception. This example therefore demonstrates that GWOT can assess a detailed structural similarity that cannot be captured by simple correlation-based measures. We now leverage this powerful capacity to systematically compare all models.

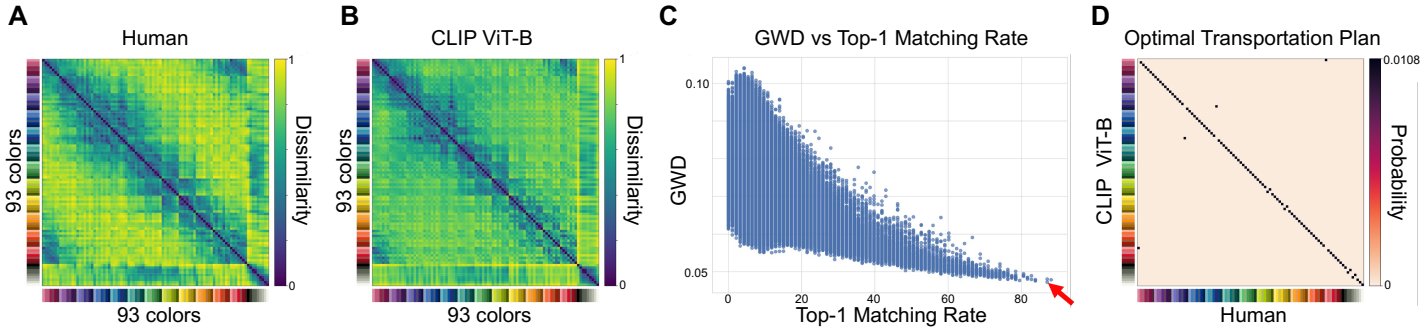

Fig 4. Unsupervised alignment of color similarity structures for 93 colors between humans and a CLIP model based on GWOT. (A)-(B) Dissimilarity matrices of 93 colors from (A) the color-neurotypical participants group, and (B) a sample from the CLIP model. All matrices are normalized to have values between 0 and 1, where 0 means no difference between colors and 1 means the maximum difference for each dissimilarity matrix. (C) Optimization results over 100,000 iterations with different initial values. The red arrow indicates the result with the lowest GW distance, whose transportation plan is shown in (D). (D) Optimal transportation plan between the dissimilarity matrices of human behavior and a CLIP model.

Unsupervised alignment between humans and DNN models

Having illustrated the principle of the GWOT analysis with a representative example, we now present the results of applying this method systematically to compare the human similarity structure with that of all DNN models and the RGB baseline (Fig. 5).

First, the RGB input achieved a top-1 matching rate of only 3.23% (Fig. 5), which is barely above the theoretical chance rate ($1/93 \approx 1.08\%$, shown as a dashed line in the figure). This result clearly exemplifies that a high correlation score does not guarantee fine-grained geometric alignment, as this model's extremely low matching rate was observed alongside its substantial correlation ($\rho = 0.64$) in the supervised analysis (Fig. 3).

In striking contrast, the CLIP models demonstrated a high degree of structural alignment with the human data. These models achieved high top-1 matching rates regardless of architecture size, with every score exceeding 70% and the top-performing model reaching 87.1% (Fig. 5). However, this result does not guarantee that models

trained on the CLIP paradigm will necessarily align with human data regardless of training dataset or architecture size. For CLIP models trained on the LAION and MetaCLIP datasets, alignment with human data showed a dependence on architecture size (Fig. S3B). Nevertheless, allowing this contingency of alignment on training dataset and architecture, the CLIP paradigm nevertheless demonstrated markedly higher matching rates compared to the other learning paradigms.

The MAE models and SL models all failed to align with the human similarity structure. For both paradigms, the matching accuracies did not substantially differ from the chance rate (Fig. 5). This poor performance in geometric alignment is particularly noteworthy given that these models exhibited high correlation scores in the supervised comparison (Fig. 3), and more so since the MAE models showed higher correlations than the LAION and MetaCLIP-trained CLIP models that did align with human data (Fig. S3A).

These results indicate that the GWOT analysis provides distinct information from the correlation-based analysis. Our findings demonstrate that only subsets of the CLIP models successfully achieve fine-grained, color-to-color alignment with the human data (Fig. 5). Notably, some CLIP models that aligned with human data exhibited lower correlation scores than the MAE models, which failed to align (Fig. S3A, B). Accordingly, unsupervised comparison is capable of measuring geometric alignments that cannot be discerned from the supervised comparison alone.

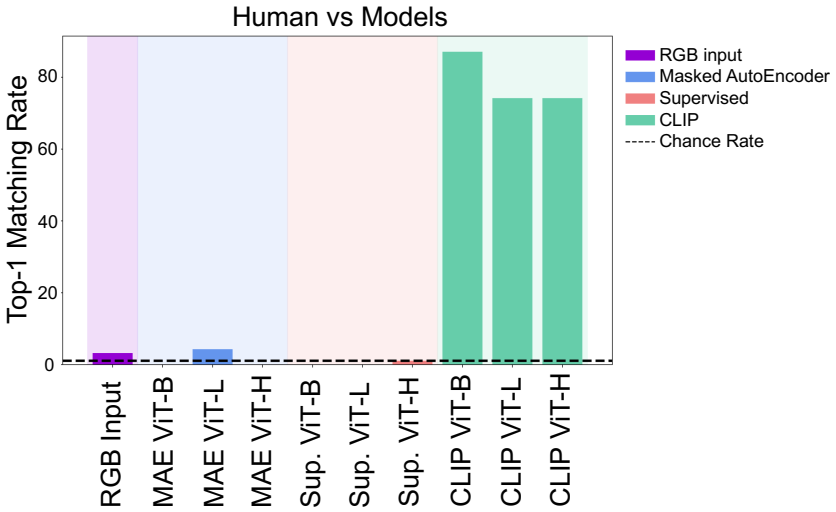

Fig 5. Unsupervised alignment of color similarity structures for 93 colors between humans and DNN models based on GWOT. Top-1 matching rates of unsupervised alignment based on GWOT between the dissimilarity matrices of humans and DNN models.

Analysis of consistency across DNN models

This subsection presents a direct model-to-model comparison to assess the consistency of the learned color representations both within and between the different learning paradigms. This analysis is driven by two primary motivations arising from our preceding human-vs-model results (Fig. 5). The first motivation is to investigate intra-paradigm consistency. Our finding that the CLIP models aligned well with human color perception leads to the hypothesis that they should also be highly consistent with one another. Conversely, since the MAE and SL models failed to align with the human data, we aim to determine whether they nonetheless converge to their own distinct, yet

internally consistent, representational structures. The second motivation is to assess inter-paradigm consistency. Having established each paradigm's relationship to the human benchmark, we now seek to quantify the structural similarity between the different learning paradigms themselves (e.g., CLIP vs. MAE), thereby assessing how (dis)similar their learned representations are. To address these questions, all sets of color representation were analyzed using both the supervised and the unsupervised comparison, but focusing primarily on the unsupervised GWOT alignment.

Our analysis of consistency within each learning paradigm revealed that only the CLIP models exhibited high consistency (Fig. 6B). Specifically, the models trained on the same dataset show high top-1 matching rates, with the matching accuracy for one pair reaching 100%. Notably, the human-aligned CLIP models also show high top-1 matching rates, regardless of different training dataset (Fig. S5). In sharp contrast, both the MAE and SL paradigms generally lacked such internal geometric consistency. Their top-1 matching rates remained low, with the exception of a single MAE model pair that also achieved a high matching rate. These model pairs with high top-1 matching rates also tended to show high inter-model correlations (Fig. 6A, S4).

The analysis of consistency between the different learning paradigms revealed that representations learned by one paradigm in most cases did not align with those learned by another (Fig. 6B). An exception was the MAE model with the largest architecture size, which exhibited moderate top-1 matching rates with the CLIP models trained on the DFN dataset (Fig. 6B). In contrast, other inter-paradigm combinations showed low alignment (Fig. S5).

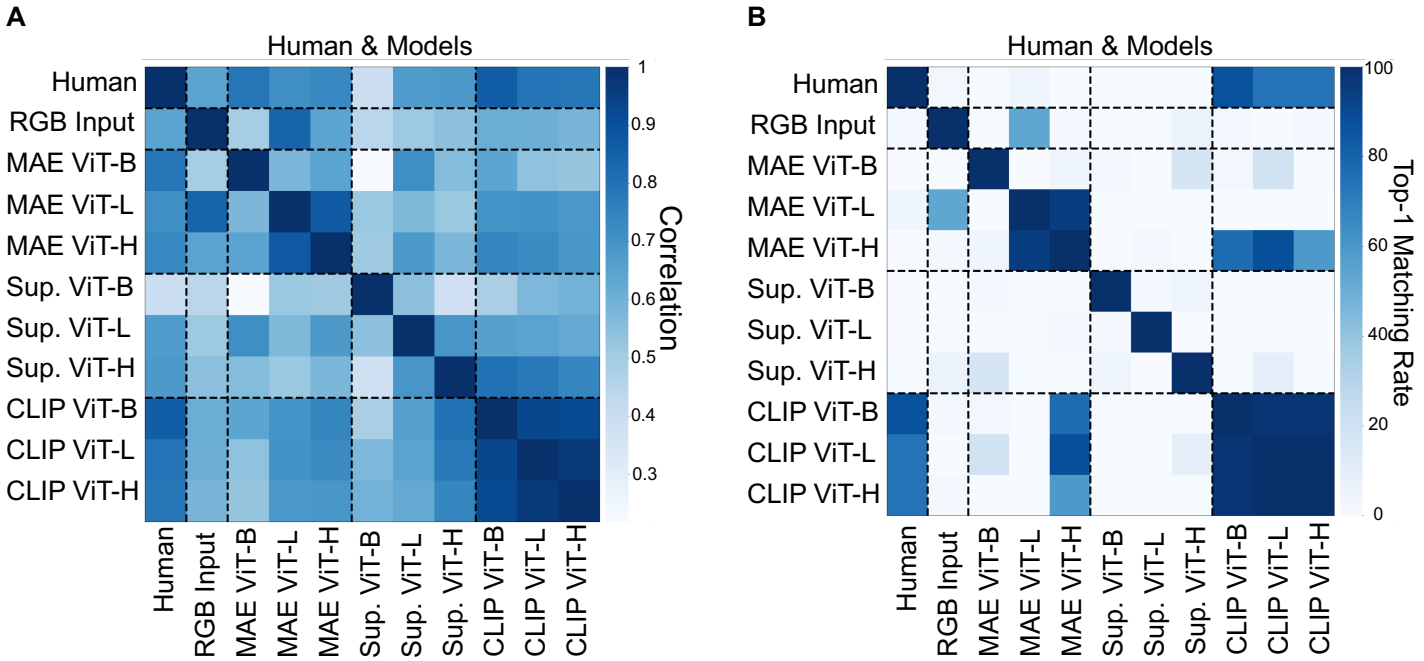

Fig 6. Supervised and unsupervised comparison of color similarity structures for 93 colors between each pair of humans, RGB input, and DNN models. (A) Supervised comparison based on Pearson correlations between each pair of humans, RGB input, and DNN models. (B) Unsupervised comparison based on GWOT between each pair of humans, RGB input, and DNN models.

Beyond human experimentation: probing the vast 4096-color representations in DNNs

Having identified the DNN models that best align with human color perception using the 93 PCCS colors, we now leverage a key advantage of computational models to explore a vastly larger color space that is inaccessible to human experimentation. For this analysis, we created a 4096-color set by uniformly sampling 16 values from each of the R, G, and B channels ($16^3 = 4096$), a sampling method that provides a more uniform and dense coverage of the entire digital RGB cube, in contrast to the PCCS's perceptually-structured sampling based on hue and tone. This allows us not only to test the robustness and generalizability of the learned representations by DNNs but also to use the best-performing models to make novel predictions about the large-scale structure of human color perception.

Visualization of the 4096-color representations

A qualitative analysis of the 4096-color representations, visualized using 2-dimensional MDS, reveals that the CLIP models appeared to converge on a mutually similar, non-trivial geometry (Fig. 7). As a baseline reference, the RGB model forms a filled-in, cube-like structure (Fig. 7A). This is a trivial outcome given that the colors were uniformly sampled from the raw three-dimensional input space. In clear distinction to this baseline, the CLIP models consistently produced a comparable geometric structure, namely a wide, hollow, distorted ring-like shape with specific protrusions and indentations to which the achromatic colors were generally located externally (Fig. 7H-J). In contrast to this observed similarity, neither the MAE nor SL models exhibited such consistency, instead yielding varied results; for example, some architectures formed ring-like structures while others resulted in more disorganized, filled-in point clouds (Fig. 7B-G).

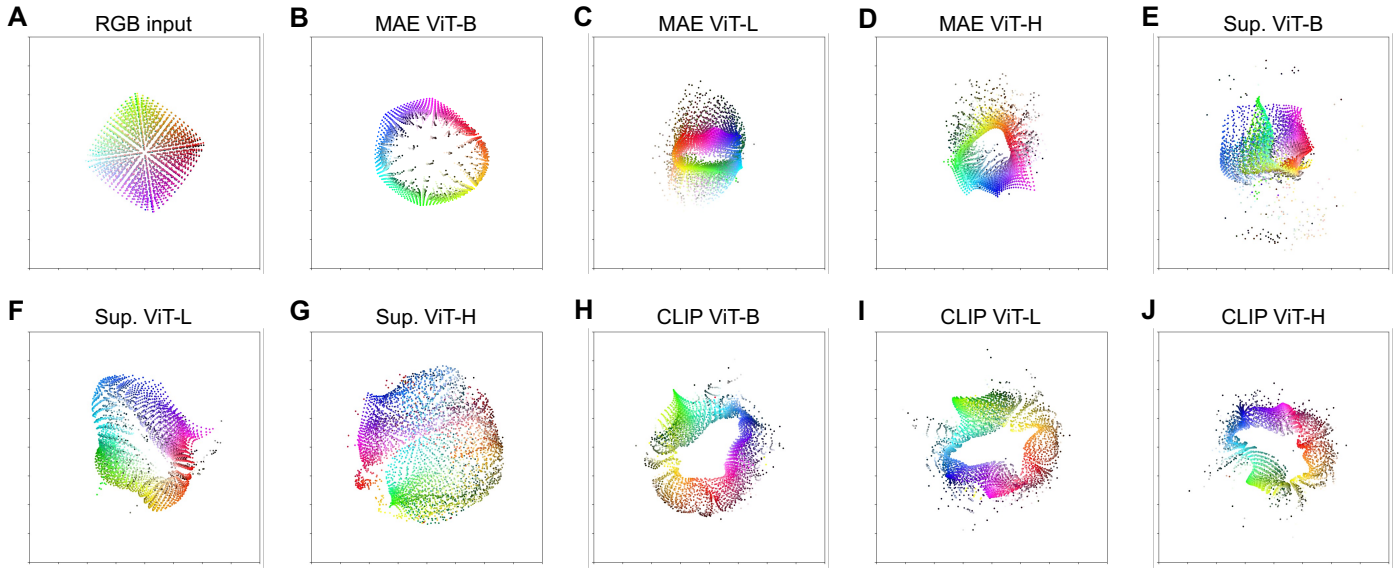

Fig 7. 4096 colors representation of RGB input and DNN models. Representation of the 4096 colors from (A) the RGB input, (B)-(D) the MAE models, (E)-(G) the SL models, and (H)-(J) the CLIP models. All representations are made by applying 2-dimensional MDS on the respective dissimilarity matrix.

Correlations and top-1 matching rates between 4096-color representations

Quantitative analysis of the measured correlations and top-1 matching rates between 4096-color representations provided objective evidence for the high consistency of the CLIP models (Fig. 8). Specifically, the CLIP model with the smallest architecture (ViT-B) exhibited moderate top-1 matching rates with the other two CLIP models (Fig. 8B). Although numerically smaller than the highest rates in the 93-color analysis, this result is arguably more significant given the combinatorial magnitude of the search space, comprising 4096! ($> 10^{10,000}$) possible permutations. Conversely, the MAE models and SL models showed significantly low intra-paradigm top-1 matching rates (Fig. 8B). Similarly, all inter-paradigm top-1 matching rates were significantly low, suggesting that the different paradigms did not converge on a shared geometric structure at this scale (Fig. 8B). As for their correlations, and consistent with the results for the 93-color analysis, pairs with high top-1 matching rates also showed high correlations (Fig. 8A). These quantitative results strengthen our finding that only the integration of visual input with rich, descriptive language produces a robust and consistent color representation. The remarkable consistency of this geometric structure across the human-aligned CLIP models allows us to speculate that the large-scale human color representation may share this specific, non-trivial geometry, which takes the form of a hollow, distorted ring with specific protrusions and indentations.

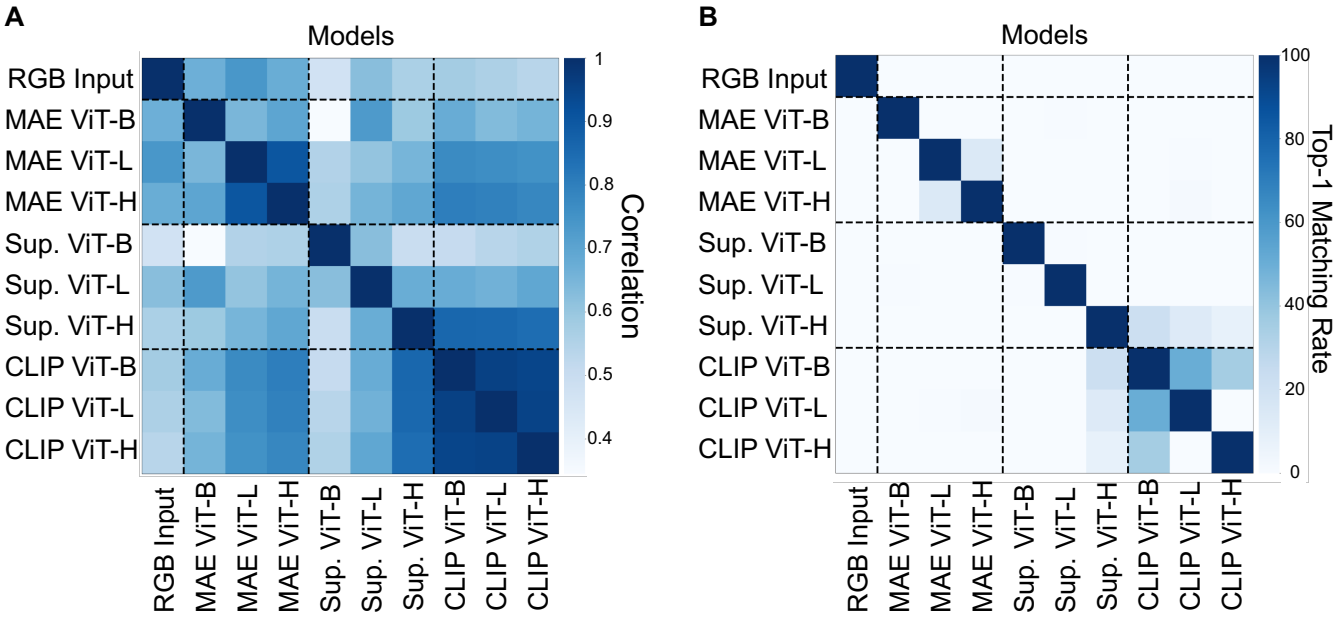

Fig 8. Supervised and unsupervised comparison of color similarity structures for 4096 colors between each pair of RGB input and DNN models. (A) Supervised comparison based of Pearson correlation between each pair of RGB input and DNN models. (B) Unsupervised comparison based on GWOT between each pair of RGB input and DNN models.

Discussion

In this study, we aimed to provide predictions of human color perception in a wide range of color space by systematically investigating the conditions under which DNNs acquire a color representation that mirrors human color perception. Our findings suggest that the integration of visual information with rich, descriptive language might

be a key factor in forming a human-like color similarity structure (Fig. 5). We found that the CLIP models that successfully aligned with human data also developed remarkably consistent and robust color representations in a vast color space (Fig. 7H-J, Fig 8), allowing us to make novel, testable predictions about the large-scale structure of human color perception. Here, we discuss the implications of these findings with a focus on two key areas: the role of language in forming these human-like representations, and future prospects for validating and extending these model-based predictions.

Effect of language on color representations of DNNs

The high degree of structural alignment with human color representations achieved exclusively by CLIP models (Fig. 5) suggests that the integration of visual information with rich, descriptive language might be critical for DNNs to acquire human-like color representations. Our finding that SSL models trained on vision alone and SL models trained on vision with simple categorical labels failed to align with human judgments indicates that visual information might be insufficient on its own. A previous study indicates that language alone also appears insufficient: the unimodal large language model (LLM) GPT-3.5 showed limited consistency with human color representation, whereas the multimodal-LLM GPT-4, which integrates language and vision, achieved remarkable consistency [35]. Taken together, these findings indicate that the acquisition of a highly human-like color similarity structure by DNNs appears to require not simply either language or vision alone, but the rich, descriptive linguistic information that is directly grounded in corresponding visual experience, as provided by the CLIP paradigm.

Nevertheless, the possibility that DNNs trained on vision-only or vision-with-simple-categories possess the potential to align with human color perception to some degree remains plausible. One potential reason why the MAE and SL models failed to align with human data in the unsupervised comparison (Fig. 5), despite exhibiting high correlations comparable to the CLIP models (Fig. 3, S3A), lies in the disparity of training data size. The ImageNet dataset [49] used for the MAE and SL models is significantly smaller than the massive datasets used for CLIP models [50–52]. Training these models on datasets of a comparable scale would allow for a more rigorous comparison between learning paradigms.

Interestingly, a finding that the MAE models still fail to align with the human data even when trained on datasets of a comparable size to the CLIP models may allow an intriguing hypothesis, namely that CLIP models correspond to the color representation of adults whereas MAE models might correspond to that of infants. If the results remain consistent with the current study (Fig. 5) after controlling for dataset scale, it would reinforce the necessity of language for acquiring the color representation of adults, as represented by the human data used in this study [15]. In contrast, the MAE models, which are characterized by high correlation with adult data (Fig. 3) but a lack of alignment in the unsupervised comparison (Fig. 5), might correspond to the color representation of pre-linguistic infants, potentially serving as a precursor to the adult color representation refined by language [53–57]. However, the detailed representational structure of infants or children in a large color space remains unexplored. Although a recent study reported that children as young as 3 years old already possess a color representation that resembles that of adults in a limited 9-color set [9], this coarse similarity does not guarantee alignment in a larger color space. Just as MAE models diverged from adult data in the fine-grained comparison despite high correlations, the color representation of children might also differ from that of adults when examined at a finer scale. Accordingly, attempts to determine whether language acquisition is indeed necessary for forming the adult-like color representation requires obtaining large-scale color structure data from children or infants and comparing them with MAE models.

This endeavor would yield significant insights into the developmental trajectory of human color perception.

Beyond the alignment with human data, the convergence of human-aligned CLIP models to a consistent structure in the 4096-color space (Fig. 7H-J) can be interpreted as evidence that language acts as a powerful constraint that guides representations toward a shared model. In this context, the convergence by CLIP models contributes to ongoing discussion about representational convergence in deep learning, where it is hypothesized that capable models may be converging toward a shared, underlying representation of the world [58]. This process can be understood through a constructive approach which views language not merely as a communication tool but also as a mechanism that facilitates the alignment of internal representations (or qualia structures) among agents [59]. In our study, the joint training on image-text pairs appears to function as such a powerful aligning force, constraining the vast possibility space of representations and enabling the different CLIP models to emerge with the same stable, shared color geometry.

Future prospects for predicting human color perception with CLIP models

The first future prospect involves empirically testing the plausibility of our prediction for the vast, 4096-color representation suggested by the human-aligned CLIP models (Fig. 7H-J). To validate this novel and non-trivial structure, we naturally need to compare our results with human data. This would require psychophysical experiments comparing the vast number of color pairs necessary to construct the 4096-color representation, which is infeasible for an individual subject to conduct. Therefore, one potential approach is to collect data from a group of people, as utilized in the previous research [15]. Although computationally and logistically demanding, empirical validation of this large-scale prediction remains a critical next step for future research.

A second important future direction is to investigate more naturalistic and nuanced color representations. A limitation of the current study, consistent with much foundational research in this area, is the use of idealized, single-color patches. This controlled approach is necessary for isolating variables, but does not capture the complexities of color perception in natural environments. Human color perception is heavily influenced by context, such as changing lighting conditions [60], the influence of surrounding backgrounds [61], and the complex interplay of multiple colors [62]. Future work should therefore extend this research to examine whether the human-like alignment of CLIP models holds under these more realistic conditions.

Materials and methods

Human Behavioral Data

Human behavioral data were drawn from a large-scale dataset collected from 426 color-neurotypical participants in a previous study [15]. In the experiment, participants performed a color similarity judgment task in which they were asked to rate the perceived similarity between pairs of 93 color stimuli selected from the Practical Color Coordinate System (PCCS) [48]. For each pair, ratings were provided on a scale from 0 ("very similar") to 7 ("very different"). The group-average human dissimilarity matrix was constructed by averaging the judgment responses for each color pair across all participants.

DNN Models Used in This Study

We used three types of deep neural network (DNN) learning method in this study: self-supervised learning, supervised learning, and CLIP (Contrastive Language–Image Pre-training). The details of each model, including their names, backbone architectures, and training data, are listed in the supporting information (Table. S1). A brief description is provided below.

Self-Supervised Learning Models: Self-supervised learning (SSL) is a framework where models learn representations from raw data by solving pretext tasks without relying on supervised signals. For our study investigating color representation, it was crucial to select models that had not been trained with color augmentation, as such techniques could artificially degrade or alter the intrinsic color information learned by the model. Our selection was therefore guided by this constraint, and lead us to focus on Masked Image Modeling (MIM) models [36, 37]. In this framework, the model learns by masking random patches of the input natural image and reconstructing the missing raw pixel values. As several publicly available checkpoints for these models were trained without color augmentation, we selected 4 Transformer-based MIM models. These models learned representations exclusively from the input images themselves, without any reliance on labels or linguistic information.

Supervised Learning Models: Supervised learning (SL) is a paradigm where models learn to classify or predict outputs from labeled data [22]. We used 3 Transformer-based architectures. These models were trained on labeled datasets of natural images. Although the labels were not primarily focused on color information, some inherently included color descriptors (e.g., "black swan," "red wine"), indirectly providing the models with color-related information during training.

CLIP Models: CLIP (Contrastive Language–Image Pretraining) models learn to associate natural images and text by jointly training on large-scale image-text pairs [38]. We used 9 transformer-based architectures. These models were trained with linguistic information that included color descriptions, enabling them to learn representations that link visual features with color-related linguistic information.

Deriving Representational Dissimilarity Matrices from DNNs

To create model-based dissimilarity matrices that were comparable to the human behavioral data, we performed the following procedure for each DNN model:

1. We generated 93 monochromatic images corresponding to the exact same PCCS color stimuli used in the human behavioral experiment [15].

2. Each image was individually fed into a model, from which we extracted the corresponding feature embedding from the final layer of its image encoder.

3. We then computed a 93×93 representational dissimilarity matrix (RDM) by calculating the pairwise Euclidean distances between the 93 extracted color embeddings.

The same procedure was also applied to derive RDMs for two larger color sets. The first was a richer 297-color set from the Practical Color Coordinate System (PCCS), comprising 288 chromatic colors (24 hues $\times$ 12 tones) and 9 achromatic colors. The second was a 4096-color set designed to exhaustively cover the entire digital RGB space, generated by uniformly sampling 16 values from each of the R, G, and B channels ($16^3 = 4096$). The findings from the 297-color set were highly consistent with the 93-color experiment. As this analysis confirmed our main results rather than uncovering novel large-scale structures as the 4096-color experiment did, we include it in the supporting information to support our primary conclusions (Fig. S6).

RGB Color Space Baseline

To establish a simple, non-learned baseline for comparison, we computed a representational dissimilarity matrix (RDM) for the RGB color space. For each color set (93 PCCS colors, extended 297 PCCS colors, and 4096 RGB colors), the RDM was computed by calculating the Euclidean distance between each color pair based on their coordinates in the 3-dimensional RGB space.

Supervised comparison based on Pearson correlation

To quantify the similarity between any two representational dissimilarity matrices (RDMs) in a supervised manner, we computed the Pearson correlation coefficient (ρ), known as conventional Representational Similarity Analysis (RSA) [39–41]. The correlation is calculated as:

$$\rho = \frac{\sum_{i<j}(D_{ij} - \bar{D})(D'_{ij} - \bar{D}')}{\sqrt{\sum_{i<j}(D_{ij} - \bar{D})^2}\sqrt{\sum_{i<j}(D'_{ij} - \bar{D}')^2}} \tag{1}$$

where D_{ij} and D'_{ij} are the dissimilarities between color i and color j in the first RDM (D) and the second RDM (D'), respectively. The terms $\bar{D}$ and $\bar{D}'$ represent the mean of all dissimilarity values in the respective RDMs. The summation is performed over the upper triangular elements of the matrices (where $i < j$) to include each color pair only once. The central assumption of this method is that a known one-to-one correspondence exists between the colors of the two RDMs, which makes a direct comparison between the dissimilarity of "color A vs. color B" in one RDM and "color A vs. color B" in the other a meaningful measure of similarity.

Unsupervised comparison based on Gromov–Wasserstein Optimal Transport (GWOT)

Gromov–Wasserstein Optimal Transport (GWOT)

We employed the Gromov–Wasserstein optimal transport (GWOT) algorithm [43, 44] for the unsupervised comparison of color similarity structures. This approach is used to assess structural similarity without relying on any pre-defined correspondence between the colors in the different structures. GWOT functions as an unsupervised alignment method that determines an optimal transportation plan between two distinct point clouds, and operates without any item-specific correspondence data. The core of this method involves the optimization (minimization) of the Gromov–Wasserstein distance (GWD), defined as:

$$\text{GWD} = \min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik}\Gamma_{jl} \tag{2}$$

GWD serves as a metric to quantify the level of correspondence between the two similarity structures (Fig. 9A). For this study, D_{ij} denotes the dissimilarity between colors i and j in the first similarity matrix D, and D'_{kl} denotes the dissimilarity between colors k and l in the second matrix D'. Prior to computation, every similarity matrix was normalized so that its values fell within a [0, 1] range. The solution to this GWD minimization problem is the optimal transportation plan, denoted as the matrix Γ. This matrix Γ provides an effective unsupervised alignment of the two color similarity structures. Each element Γ_{ik} within this matrix can be understood as the "probability" that the i-th color from the first domain maps to the k-th color in the second domain.

Optimizing the GWD (Eq. 2) presents a non-convex problem. Consequently, any solution found by the algorithm is a local optimum, and there is no assurance of reaching the global optimum. To manage this optimization challenge and identify good local minima, we employed the Python Optimal Transport (POT) library [63]. We executed the algorithm multiple times, each time beginning with a different random initialization of the Γ matrix. Specifically, 5×10^6 random initializations were performed for the 93-color dissimilarity matrices, and 1×10^5 initializations were performed for the 297-color matrices.

An alternative and more efficient method for optimizing GWD involves the inclusion of an entropy-regularization term, $H(\Gamma)$. This modification results in the following objective function:

$$\mathrm{GWD}_\epsilon = \min_{\Gamma} \sum_{i,j,k,l} (D_{ij} - D'_{kl})^2 \Gamma_{ik} \Gamma_{jl} + \epsilon H(\Gamma), \tag{3}$$

The introduction of this term is known to improve optimization efficiency [42, 64]. However, this regularized problem remains non-convex, and accordingly still requires multiple initializations to locate good local minima. We applied this entropy-regularized method to the 4096-color dissimilarity matrices, for which we performed 50 random initializations.

Quantifying Alignment with Matching Accuracy

To evaluate the quality of the unsupervised alignment produced by GWOT, we calculated a matching rate as the score of matching accuracy. This score measures the extent to which the optimal transportation plan Γ correctly maps corresponding colors between the two structures, using their ground-truth labels for verification(Fig. 9B). Specifically, for each color i in the first structure, we identify the color j in the second structure to which it is mapped with the highest probability. We consider this a "match" if the label of color i is identical to the label of color j.

This matching condition can be formalized as follows. Let the color labels for the items in the two dissimilarity matrices be c_1 and c_2. The match for the i-th color in the first matrix (c_{1i}) is determined by:

$$\mathrm{Match}(i) = \begin{cases} 1, & \text{if } \Gamma_{ij} = \max_{k \in \{1,\ldots,n\}}(\Gamma_{ik}) \text{ and } c_{1i} = c_{2j} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The overall matching accuracy is then the percentage of correctly matched colors, calculated as:

$$\text{Matching Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Match}(i) \times 100\% \tag{5}$$

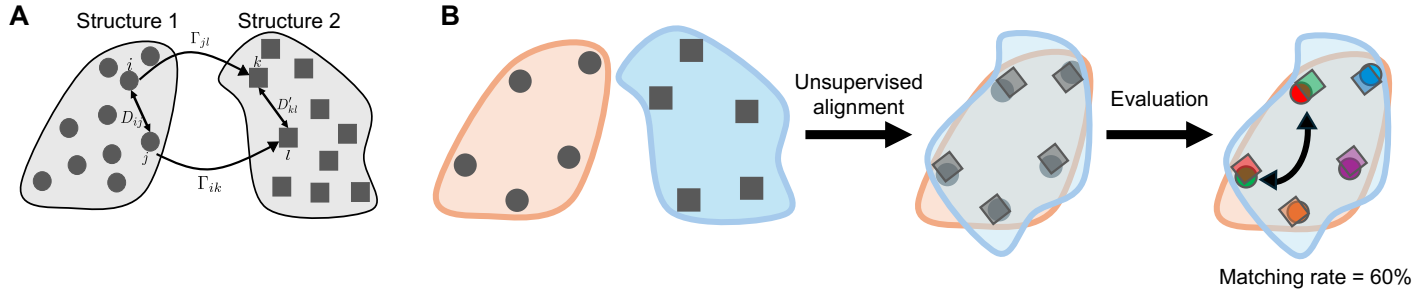

Fig 9. Schematic of the unsupervised comparison based on Gromov–Wasserstein Optimal Transport (GWOT). (A) Schematic of Gromov–Wasserstein optimal transport. The elements of matrices D and D' are the dissimilarities between the items. Γ is the transportation matrix, where each element indicates the probability of an item in one similarity structure corresponding to another item in the other similarity structure. (B) Evaluation of unsupervised alignment using external labels. Following unsupervised alignment based solely on the internal dissimilarities of each structure, the matching rate between them was calculated using ground-truth labels.

Supporting information

Contents

TableS1: List of DNN models used in this study and their correspondence with human data for 93 colors.

FigS1: Dissimilarity matrices of 93 colors from humans, RGB input, and DNN models.

FigS2: Embeddings of 93 colors from humans, RGB input, and DNN models.

FigS3: Supervised and unsupervised comparison of color similarity structure for 93 colors between humans and all DNN models.

FigS4: Supervised comparison of color similarity structure for 93 colors between each pair of humans, RGB input, and all DNN models.

FigS5: Unsupervised comparison of color similarity structure for 93 colors between each pair of humans, RGB input, and all DNN models.

FigS6: Qualitative and Quantitative analysis on 297-color similarity structure of RGB input and DNN models.

Table S1. List of DNN models used in this study and their correspondence with human data for 93 colors.

Category	Learning Method	Architecture	Training Dataset	Correlation	Matching rate/%
RGB input				0.64	3.23
SSL	MAE	ViT-B/16	ImageNet1k	0.79	0
		ViT-L/16	ImageNet1k	0.71	4.30
		ViT-H/14	ImageNet1k	0.74	0
	SimMIM	ViT-B/16	ImageNet1k	0.62	2.15
SL	Classification	ViT-B/16	ImageNet1k	0.40	0
		ViT-L/16	ImageNet1k	0.67	0
		ViT-H/14	ImageNet1k	0.68	1.08
CLIP	CLIP	ViT-B/16	laion2b_s34b_b88k	0.69	26.88
		ViT-L/14	laion2b_s32b_b82k	0.61	2.15
		ViT-H/14	laion2b_s32b_b79k	0.62	5.38
		ViT-B/16	metaclip_fullcc	0.71	53.76
		ViT-L/14	metaclip_fullcc	0.60	2.15
		ViT-H/14	metaclip_fullcc	0.63	4.30
		ViT-B/16	dfn2b	0.86	87.10
		ViT-L/14	dfn2b	0.79	74.19
		ViT-H/14	dfn5b	0.78	74.19

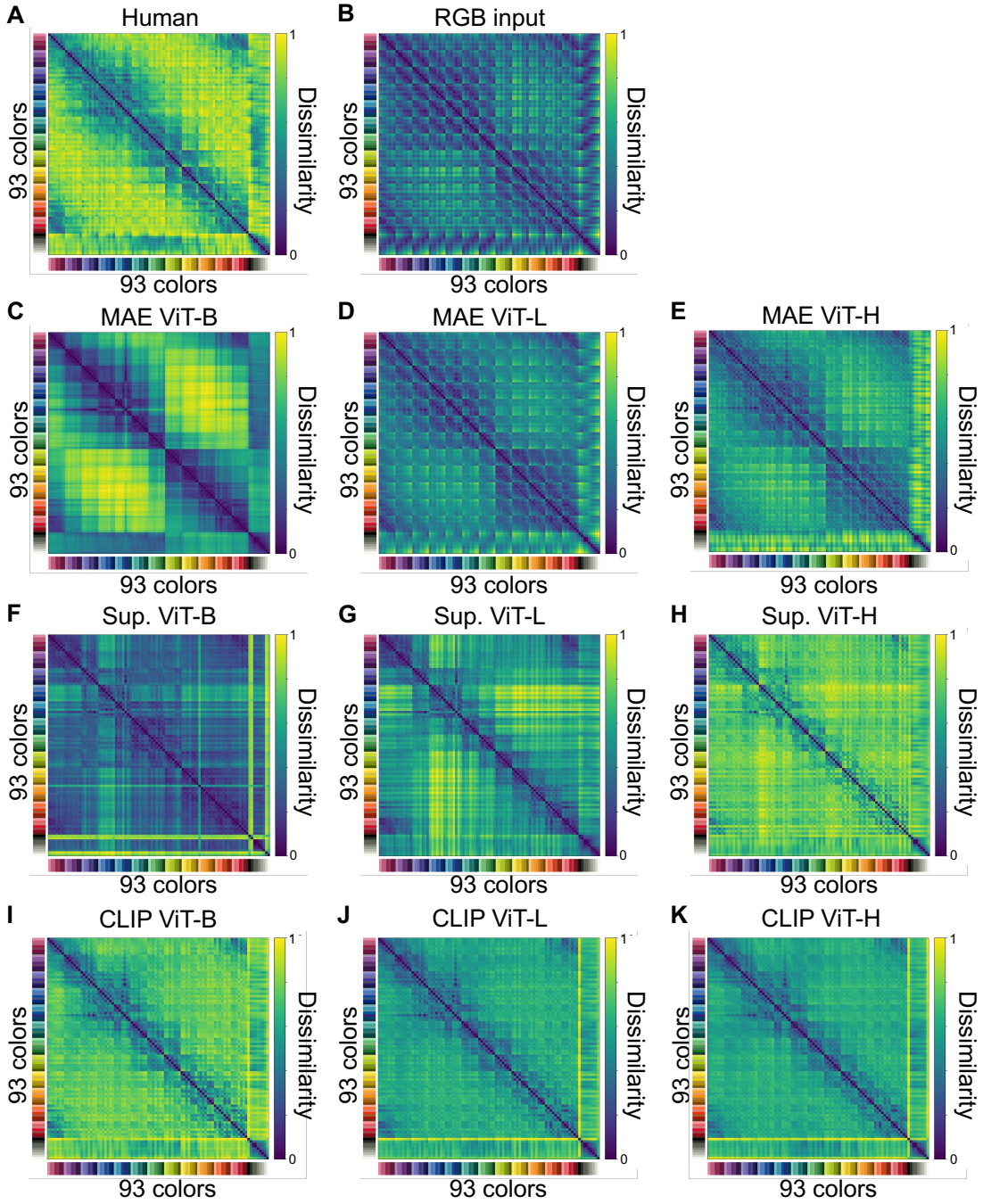

Fig S1. Dissimilarity matrices of 93 colors from humans, RGB input, and DNN models. (A)-(K) Dissimilarity matrices of 93 colors from (A) the color-neurotypical participants group, (B) the RGB model, (C)-(E) the SSL models, (F)-(H) the SL models, and (I)-(K) the CLIP models.

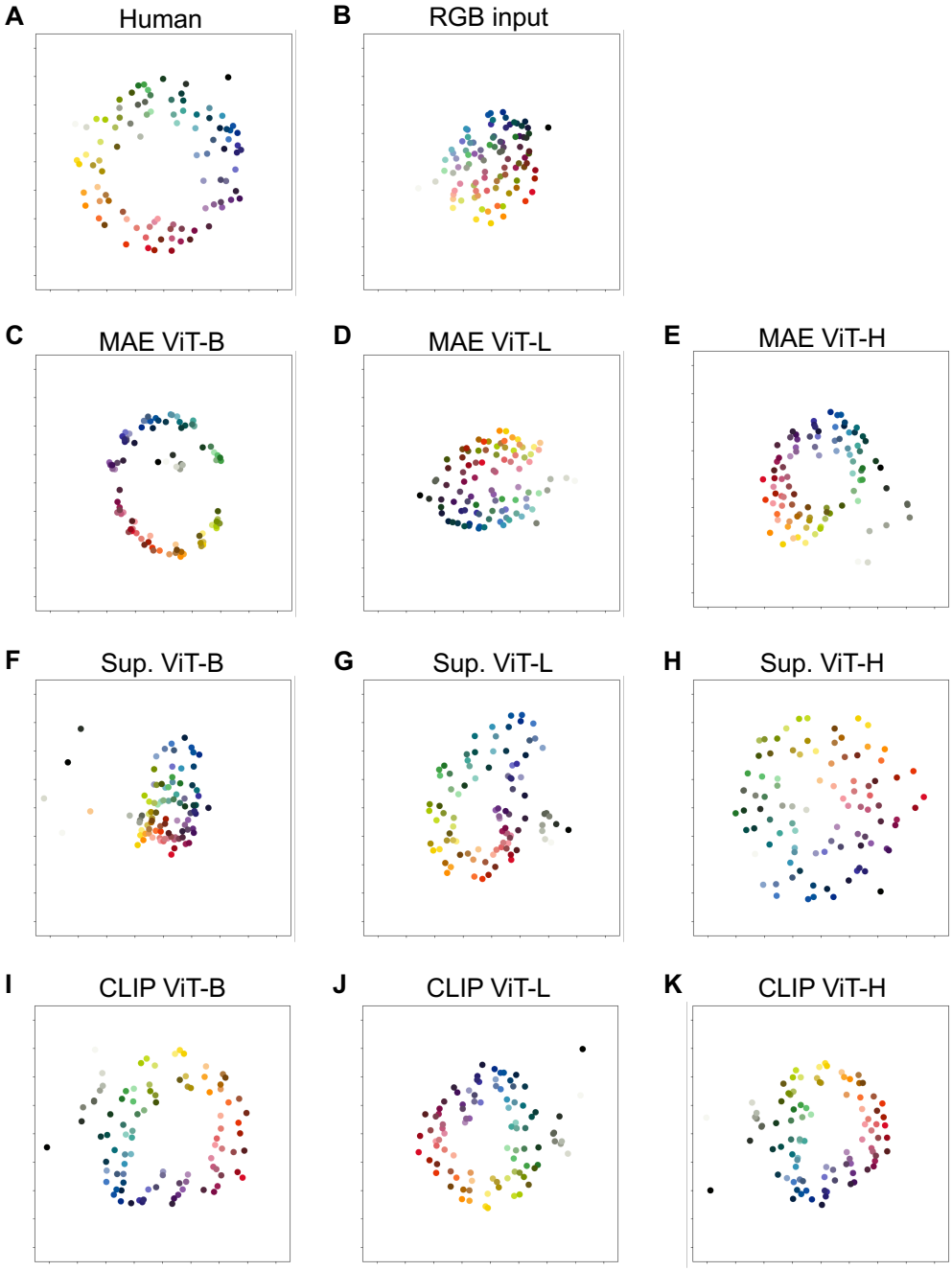

Fig S2. Embeddings of 93 colors from humans, RGB input, and DNN models. (A)-(K) Visualization of the embeddings of 93 colors by applying 2-dimensional MDS on the dissimilarity matrix from (A) the color-neurotypical participants group, (B) the RGB model, (C)-(E) the SSL models, (F)-(H) the SL models, and (I)-(K) the CLIP models.

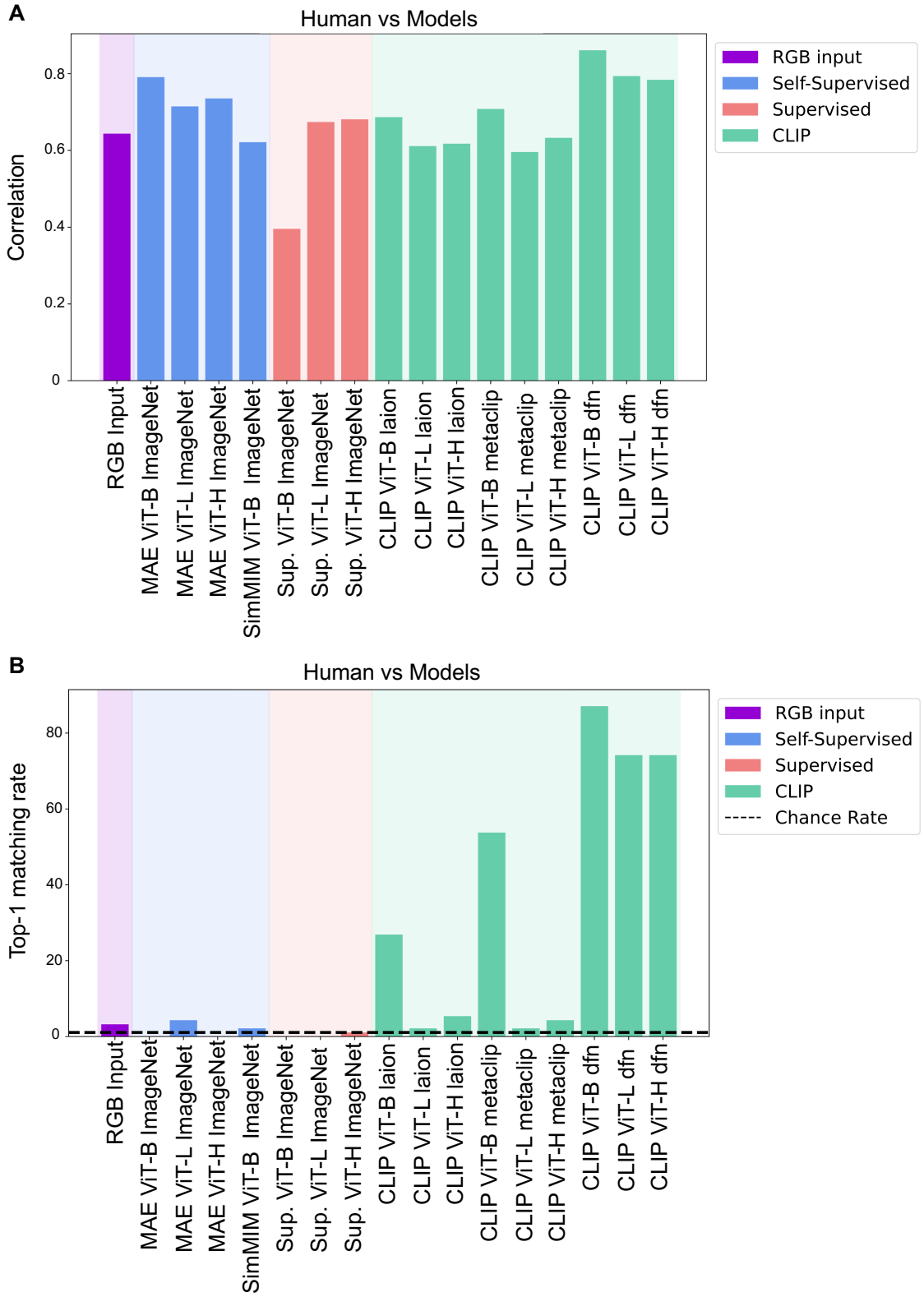

Fig S3. Supervised and unsupervised comparison of color similarity structure for 93 colors between humans and all DNN models. (A) Supervised comparison based on Pearson correlation between humans and all DNN models. (B) Unsupervised comparison based on GWOT between humans and all DNN models.

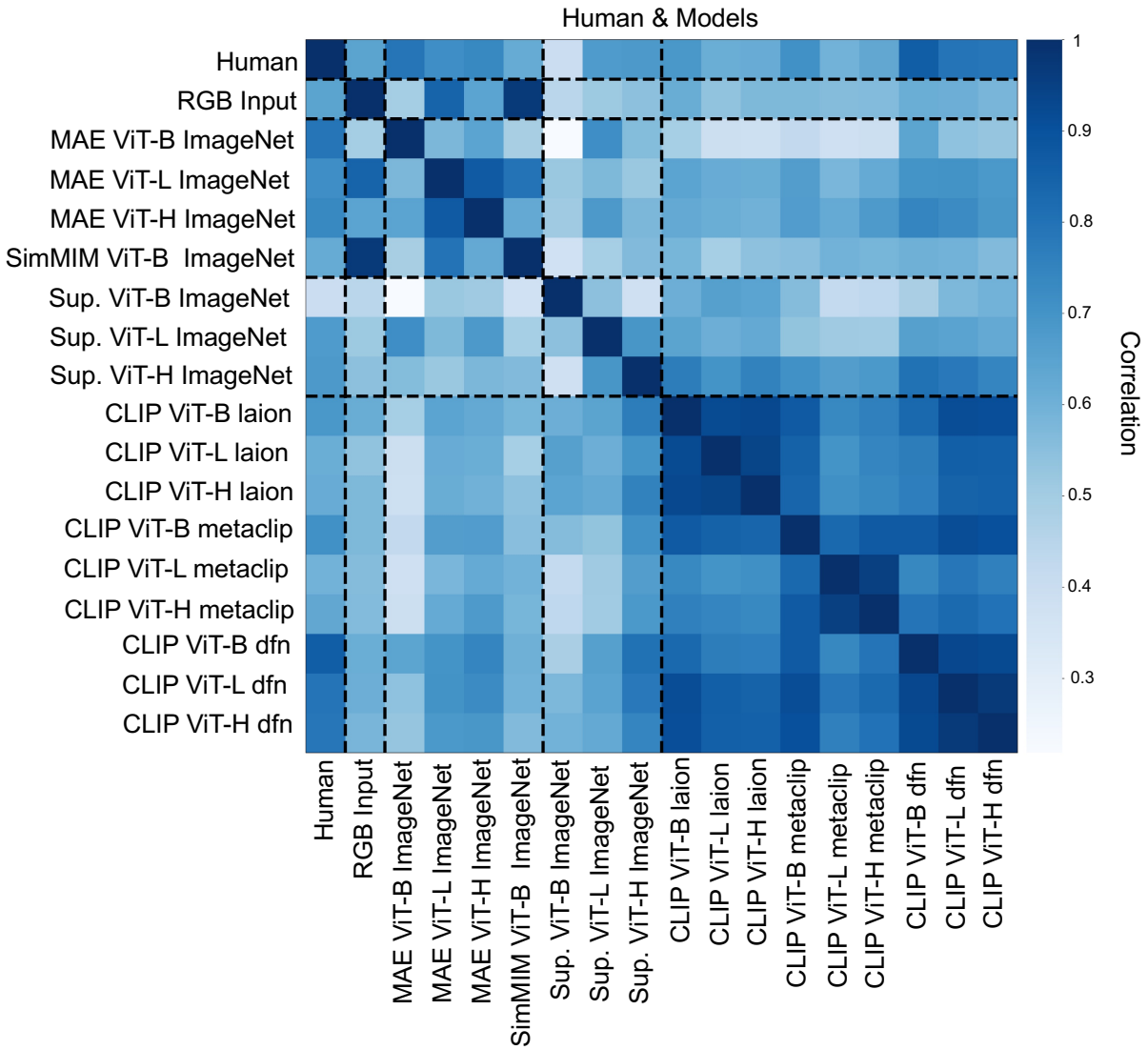

Fig S4. Supervised comparison of color similarity structure for 93 colors between each pair of humans, RGB input, and all DNN models. Supervised comparison based on Pearson correlation between each pair of humans, RGB input, and all DNN models.

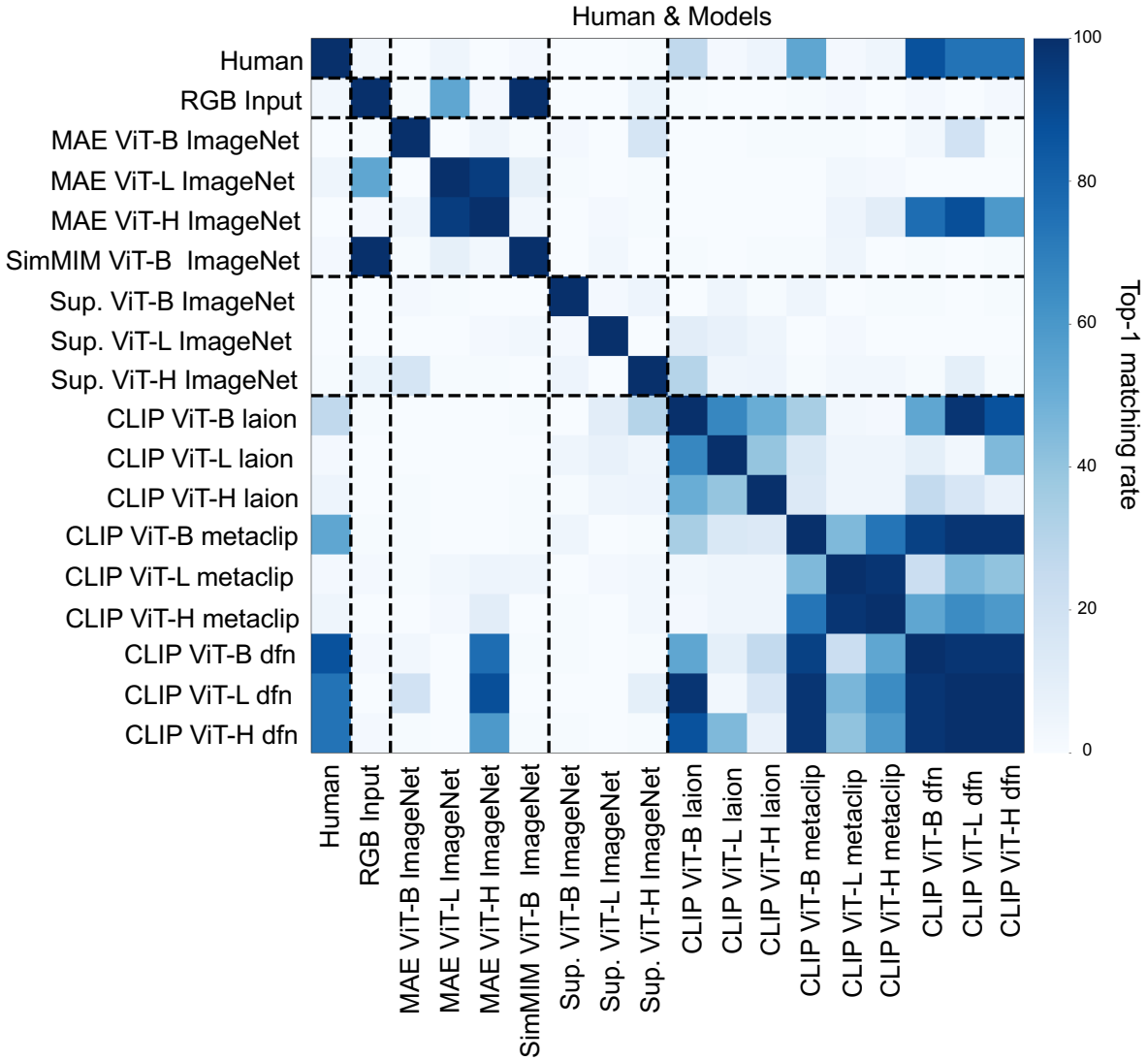

Fig S5. Unsupervised comparison of color similarity structure for 93 colors between each pair of humans, RGB input, and all DNN models. Unsupervised comparison based on GWOT between each pair of humans, RGB input, and all DNN models.

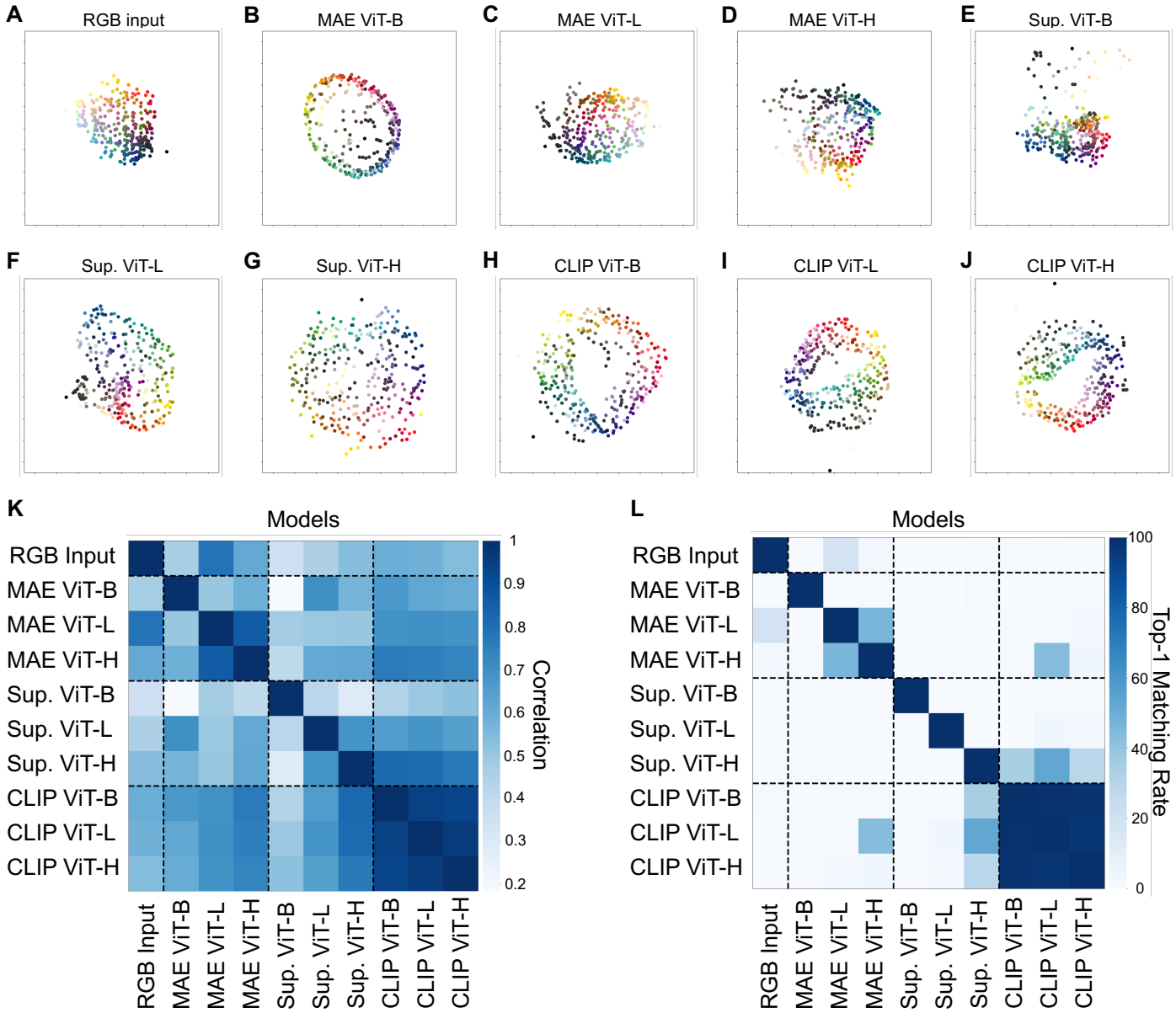

Fig S6. Qualitative and Quantitative analysis on 297-color similarity structure of RGB input and DNN models. (A)-(J) Visualization of the embeddings of 297 colors by applying 2-dimensional MDS on the dissimilarity matrix from (A) the RGB input, (B)-(D) the SSL models, (E)-(G) the SL models, and (H)-(J) the CLIP models. (K) Supervised comparison based on Pearson correlation between each pair of RGB input and DNN models. (L) Unsupervised comparison based on GWOT between each pair of RGB input and DNN models.

Acknowledgments

Masafumi Oizumi was supported by JST Moonshot R&D Grant Number JPMJMS2012 and Japan Society for the Promotion of Science, Grant-in-Aid for Transformative Research Areas Grant Number 23H04834.

Author contributions

Conceptualization: Nipun Ravindu Wickramanayaka, Masafumi Oizumi

Formal analysis: Nipun Ravindu Wickramanayaka

Funding acquisition: Masafumi Oizumi

Investigation: Nipun Ravindu Wickramanayaka

Methodology: Nipun Ravindu Wickramanayaka, Masafumi Oizumi

Project administration: Masafumi Oizumi

Resources: Masafumi Oizumi

Software: Nipun Ravindu Wickramanayaka

Supervision: Masafumi Oizumi

Validation: Nipun Ravindu Wickramanayaka, Masafumi Oizumi

Visualization: Nipun Ravindu Wickramanayaka

Writing - original draft: Nipun Ravindu Wickramanayaka, Masafumi Oizumi

Writing - review & editing: Nipun Ravindu Wickramanayaka, Masafumi Oizumi

References

1. Gegenfurtner KR. Cortical mechanisms of colour vision. Nat Rev Neurosci. 2003;4(7):563–572.
2. Conway BR. Color vision, cones, and color-coding in the cortex. Neuroscientist. 2009;15(3):274–290.
3. Helm CE. Multidimensional ratio scaling analysis of perceived color relations. J Opt Soc Am. 1964;54(2):256–262.
4. Indow T. Multidimensional studies of Munsell color solid. Psychol Rev. 1988;95(4):456–470.
5. Paramei GV, Cavonius CR. Color spaces of color-normal and color-abnormal observers reconstructed from response times and dissimilarity ratings. Percept Psychophys. 1999;61(8):1662–1674.

6. Paramei GV, Bimler DL, Mislavskaia NO. Colour perception in twins: individual variation beyond common genetic inheritance. Clin Exp Optom. 2004;87(4-5):305–312.

7. Boehm AE, MacLeod DIA, Bosten JM. Compensation for red-green contrast loss in anomalous trichromats. J Vis. 2014;14(13):19.

8. Zeleznikow-Johnston A, Aizawa Y, Yamada M, Tsuchiya N. Are color experiences the same across the visual field? J Cogn Neurosci. 2023;35(4):509–542.

9. Moriguchi Y, Watanabe R, Sakata C, Zeleznikow-Johnston A, Wang J, Saji N, et al. Comparing color qualia structures through a similarity task in young children versus adults. Proc Natl Acad Sci U S A. 2025;122(11):e2415346122.

10. Hajonides JE, Nobre AC, van Ede F, Stokes MG. Decoding visual colour from scalp electroencephalography measurements. Neuroimage. 2021;237(118030):118030.

11. Rosenthal IA, Singh SR, Hermann KL, Pantazis D, Conway BR. Color space geometry uncovered with magnetoencephalography. Curr Biol. 2021;31(3):515–526.e5.

12. Brouwer G, Heeger D. Decoding and reconstructing color from responses in human visual cortex. J Neurosci. 2009;29:13992–14003.

13. Kim I, Hong SW, Shevell SK, Shim WM. Neural representations of perceptual color experience in the human ventral visual pathway. Proc Natl Acad Sci U S A. 2020;117(23):13145–13150.

14. Hirao T, Miyamae M, Matsuyoshi D, Inoue R, Takado Y, Obata T, et al. A neuroimaging dataset during sequential color qualia similarity judgments with and without reports. Sci Data. 2025;12(1):389.

15. Kawakita G, Zeleznikow-Johnston A, Takeda K, Tsuchiya N, Oizumi M. Is my "red" your "red"?: Evaluating structural correspondences between color similarity judgments using unsupervised alignment. iScience. 2025;28(3):112029.

16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2012;60:84–90.

17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [csCV]. 2014;.

18. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2015. p. 1–9.

19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv [csCV]. 2015;.

20. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv [csCV]. 2015;.

21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017. p. 2261–2269.

22. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv [csCV]. 2020;.

23. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2014. p. 580–587.

24. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014; p. 3431–3440.

25. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014; p. 3156–3164.

26. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. arXiv [csCV]. 2014;.

27. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. arXiv [csCV]. 2017;.

28. Storrs KR, Anderson BL, Fleming RW. Unsupervised learning predicts human perception and misperception of gloss. Nat Hum Behav. 2021;5(10):1402–1417.

29. Liao C, Sawayama M, Xiao B. Probing the link between vision and language in material perception using psychophysics and unsupervised learning. PLoS Comput Biol. 2024;20(10):e1012481.

30. Du C, Fu K, Wen B, Sun Y, Peng J, Wei W, et al. Human-like object concept representations emerge naturally in multimodal large language models. Nat Mach Intell. 2025;7(6):860–875.

31. Marjieh R, Sucholutsky I, van Rijn P, Jacoby N, Griffiths T. Large language models predict human sensory judgments across six modalities. Sci Rep. 2023;14.

32. Fukushima M, Eshita S, Fukuhara H. Advancements and limitations of LLMs in replicating human color-word associations. Discov Artif Intell. 2025;5(1).

33. Xu J. Probing the representational geometry of color qualia: Dissociating pure perception from task demands in brains and AI models. arXiv [csNE]. 2025;.

34. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. arXiv [csCL]. 2023;.

35. Kawakita G, Zeleznikow-Johnston A, Tsuchiya N, Oizumi M. Gromov–Wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. Sci Rep. 2024;14(1):1–10.

36. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 16000–16009.

37. Xie Z, Zhang Z, Cao Y, Lin Y, Bao J, Yao Z, et al. SimMIM: A simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. p. 9643–9653.

38. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 8748–8763.

39. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2008;2:4.

40. Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci. 2013;17(8):401–412.

41. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014;10(4):e1003553.

42. Takeda K, Sasaki M, Abe K, Oizumi M. Unsupervised alignment in neuroscience: Introducing a toolbox for Gromov-Wasserstein optimal transport. J Neurosci Methods. 2025;419(110443):110443.

43. Mémoli F. Gromov–Wasserstein distances and the metric approach to object matching. Found Comut Math. 2011;11(4):417–487.

44. Peyré G, Cuturi M. Computational Optimal Transport. arXiv [statML]. 2018;.

45. Alvarez-Melis D, Jaakkola TS. Gromov-Wasserstein alignment of word embedding spaces. arXiv [csCL]. 2018;.

46. Asanuma H, Koide-Majima N, Nakamura K, Horii T, Nishimoto S, Oizumi M. Correspondence of high dimensional emotion structures elicited from video clips between humans and multimodal LLMs. Sci Rep. 2025;15(1):32175.

47. Takahashi S, Sasaki M, Takeda K, Oizumi M. Investigating fine- and coarse-grained structural correspondences between deep neural networks and human object image similarity judgments using unsupervised alignment. Neural Netw. 2025;195(108222):108222.

48. Nayatani Y. Adequateness of a newly modified opponent-colors theory. Color Res Appl. 2003;28(4):298–307.

49. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 248–255.

50. Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv [csCV]. 2022; p. 25278–25294.

51. Xu H, Xie S, Tan XE, Huang PY, Howes R, Sharma V, et al. Demystifying CLIP Data. arXiv [csCV]. 2023;.

52. Fang A, Jose AM, Jain A, Schmidt L, Toshev A, Shankar V. Data Filtering Networks. arXiv [csAI]. 2023;.

53. Bornstein MH. On the development of color naming in young children: data and theory. Brain Lang. 1985;26(1):72–93.

54. Sandhofer CM, Smith LB. Learning color words involves learning a system of mappings. Dev Psychol. 1999;35(3):668–679.

55. Kowalski K, Zimiles H. The relation between children's conceptual functioning with color and color term acquisition. J Exp Child Psychol. 2006;94(4):301–321.

56. Saji N, Imai M, Asano M. Acquisition of the Meaning of the Word Orange Requires Understanding of the Meanings of Red, Pink, and Purple: Constructing a Lexicon as a Connected System. Cogn Sci. 2020;44(1):e12813.

57. Scott ME, Kanero J, Saji N, Chen Y, Imai M, Golinkoff RM, et al. From green to turquoise: Exploring age and socioeconomic status in the acquisition of color terms. First Lang. 2023;43(1):3–21.

58. Huh M, Cheung B, Wang T, Isola P. The Platonic Representation Hypothesis. arXiv [csLG]. 2024;.

59. Taniguchi T, Oizumi M, Saji N, Horii T, Tsuchiya N. Constructive approach to bidirectional causation between qualia structure and language emergence. arXiv [csCL]. 2024;.

60. Foster DH. Color constancy. Vision Res. 2011;51(7):674–700.

61. Shevell SK, Kingdom FAA. Color in complex scenes. Annu Rev Psychol. 2008;59(1):143–166.

62. Brown RO, MacLeod DI. Color appearance depends on the variance of surround colors. Curr Biol. 1997;7(11):844–849.

63. Flamary R, Courty N, Gramfort A, Alaya MZ, Boisbunon A, Chambon S, et al. POT: Python Optimal Transport. Journal of Machine Learning Research. 2021;22(78):1–8.

64. Peyré G, Cuturi M, Solomon J. Gromov-Wasserstein averaging of kernel and distance matrices. ICML. 2016;48:2664–2672.