Comparing Representations in Static and Dynamic Vision Models to the Human Brain

Anonymous Author(s)
Affiliation
Address
email

Abstract

We compared neural responses to naturalistic videos and representations in deep network models trained with static and dynamic information. Models trained with dynamic information showed greater correspondence with neural representations in all brain regions, including those previously associated with the processing of static information. Among the models trained with dynamic information, those based on optic flow accounted for unique variance in neural responses that were not captured by Masked Autoencoders. This effect was strongest in ventral and dorsal brain regions, indicating that despite the Masked Autoencoders' effectiveness at a variety of tasks, their representations diverge from representations in the human brain in the early stages of visual processing.

1 Introduction

The human visual system is organized into distinct processing streams Ungerleider et al. [1982], Pitcher and Ungerleider [2021]: a ventral stream that extends from early visual regions into the inferior portions of temporal cortex, and a dorsal stream that extends into lateral occipital cortex and branches into a lateral stream (along the superior temporal sulcus) and a parietal stream (reaching the inferior parietal lobule). This organization likely results from the computational requirements of visual perception. Therefore, understanding the representations encoded by different visual streams could offer insights about the human brain and also about more general principles of vision.

The ventral stream has been proposed to encode static object identity Grill-Spector and Weiner [2014], while dynamic information has been associated with the dorsal, lateral and parietal streams Ganel and Goodale [2003], Culham et al. [2003]. Indeed, static images of objects are known to drive responses in ventral temporal regions in macaques Pasupathy and Connor [2002], Logothetis et al. [1995], Tanaka [1996], Hung et al. [2005] and in humans Edelman et al. [1998], Haxby et al. [2001]. Moving stimuli drive stronger responses in dorsal and lateral regions Zeki et al. [1991], Tootell et al. [1995], Saito et al. [1986]. In addition, disruption to lateral regions using TMS affects the processing of dynamic information Beckers and Hömberg [1992], Pitcher et al. [2014] as well as motion prediction Vetter et al. [2015].

However, other studies have challenged the hypothesis that visual streams in the human brain differ based on whether they encode static or dynamic visual features. These studies suggested that both static and dynamic features are represented in multiple visual streams Kourtzi et al. [2002], Freud et al. [2017], Cornette et al. [1998], Sunaert et al. [1999], Robert et al. [2023]. Here, we investigated the contribution of static and dynamic information to the representations encoded by different visual streams, by quantifying the convergence between neural representations and representations learned by deep network models.

Previous work compared neural responses to deep network models trained with static images Yamins et al. [2013], Khaligh-Razavi and Kriegeskorte [2014], Zhuang et al. [2021], Konkle and Alvarez

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

[2022]. The present work studies the additional contribution of dynamic information during the observation of quasi-naturalistic videos, by comparing neural responses to deep networks whose inputs are static images (e.g. convolutional ResNets He et al. [2016], image masked autoencoders He et al. [2022]) and to deep networks whose inputs are videos (e.g. hidden two-stream networks, video masked autoencoders Zhu et al. [2019], Tong et al. [2022], Feichtenhofer et al. [2022]). Critically, we include in our analyses a family of self-supervised models that are widely used in Computer Vision, but that are understudied in Cognitive Neuroscience: masked autoencoders (MAEs). We investigate the correspondence of representations in MAEs and Video MAEs to neural representations in the human brain.

2 Methods

2.1 Data

BOLD fMRI responses (3×3×3 mm) to eight movie segments of 'Forrest Gump' were obtained from the publicly available *studyforrest* audiovisual dataset (`http://studyforrest.org`). Fifteen right-handed participants took part in the study (6 females; age range 21-39 years, mean 29.4 years). The data was acquired with a T2*-weighted echo-planar imaging sequence, using a whole-body 3 Tesla Philips Achieva dStream MRI scanner equipped with a 32 channel head coil.

2.2 Preprocessing

Data were first preprocessed using fMRIPrep (`https://fmriprep.readthedocs.io/en/latest/index.html`): a robust pipeline for the preprocessing of diverse fMRI data. Anatomical images were skull-stripped with ANTs (`http://stnava.github.io/ANTs/`), and FSL FAST was used for tissue segmentation. Functional images were corrected for head movement with FSL MCFLIRT (`https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT`), and were subsequently coregistered to their anatomical scan with FSL FLIRT. Finally, the skull-stripped anatomical images were normalized to the MNI template using SPM. We denoised the data with CompCor Behzadi et al. [2007] using 5 principal components extracted from the union of cerebrospinal fluid and white matter.

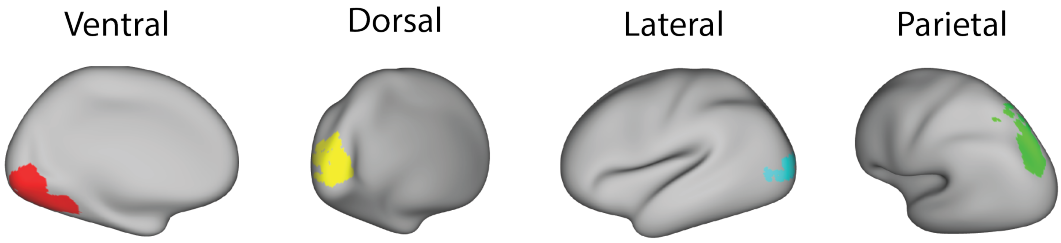

Figure 1: Masks of the visual streams in the human brain projected on an inflated cortical surface in MNI space.

2.3 Regions of Interest (ROI)

To identify the regions of interest (ROI), we used an atlas of probabilistic maps of visual topography in the human cortex from a previous study Wang et al. [2015]. The atlas contains twenty-five cortical regions and spans multiple visual streams: ventral, dorsal, parietal, and lateral (Figure 1).

A list of probabilities is associated with each voxel to reflect the likelihood of that voxel being part of each of the twenty-five brain regions ($R_i, i = 1, \ldots, 25$). We calculated the transformation from MNI space to each participant's native space and co-registered the probability maps with each participant's anatomy. To prevent overlap between the regions of interest in the participants' native space, we followed a procedure analogous to Wang et al. [2015]. Specifically, we calculated the maximum probability map for each participant, using which we exclusively classified each voxel as either belonging to a specific ROI or as being outside of all the ROIs.

The inclusion probability was computed as the probability of a voxel of being in any of the defined regions ($P(\cup_{i=1}^{25} v \in R_i)$), and The exclusion probability is the probability of a voxel not belonging to

Table 1: Models of visual cortex

Model	Input	Output	Training dataset	#Selected layers
Supervised static	image	object identity	Image-net	11
Supervised static	image	action identity	HAA-500	11
Self-supervised dynamic	video	optic flows	HAA-500	11
Supervised dynamic	optic flow	action identity	HAA-500	11
pre-trained Masked Autoencoder	(masked) image	(unmasked) image	Image-net	12
fine-tuned Masked Autoencoder	image	object identity	Image-net	12
pre-trained Masked Video Autoencoder	(masked) video	(unmasked) video	Kinetics-400	12
fine-tuned Masked Video Autoencoder	video	action identity	Kinetics-400	12
pre-trained Masked Video Distillation	(masked) video	MAE & VideoMAE high-level features	Kinetics-400	12

any of the ROIs (P($\cap_{i=1}^{25} v \notin R_i$)). If the exclusion exceeded the inclusion probability, we discarded the voxel. Otherwise, we classified the voxel as belonging to the region with the highest probability. The resulting ROIs were grouped into four sets associated with distinct visual streams. The ventral stream contains V1v, V2v, V3v, hV4, VO1, VO2, PHC1, PHC2; the dorsal stream V1d, V2d, V3d, V3a, V3b; the lateral stream LO1, LO2, hMT, and finally the parietal stream IPS0, IPS1, IPS2, IPS3, IPS4, IPS5, SPL1, and FEF. While often the term "dorsal stream" is used to refer to the combination of the regions we labeled as "dorsal stream" and the regions we labeled as "parietal stream", here we sought to distinguish between the initial branch of the dorsal stream and its parietal and lateral temporal continuations, without implying that the initial segment is disproportionately associated with one or the other.

2.4 Models of human visual cortex

To study representations of quasi-naturalistic visual stimuli, we used a variety of vision models, including feed-forward convolutional neural networks, as well as state-of-the-art foundation vision models. The models vary in architecture, learning objective, and training data (Table 1). Here, we propose an overview of the models. Training details for the HAA-trained CNNs are presented in supplementary materials. The trained versions of all other models are adopted from their official implementation repository. For model details, refer to the original papers.

Supervised (sup) static net is the spatial stream of the hidden two-stream convolutional neural network model Zhu et al. [2019]. The sup static net has a resnet18 architecture and encodes static features of visual stimulus. Two versions of the model were included in the models' pool: one is trained on Image-Net Deng et al. [2009] and predicts object identity, and the other is trained on HAA-500 action dataset Chung et al. [2021] and predicts action label. Both versions take a single frame as input.

Self-supervised (s-sup) dynamic net is the first part of the temporal stream (i.e., motion net) in the hidden two-stream convolutional neural network model Zhu et al. [2019]. The self-supervised dynamic net takes 11 consecutive frames as input and infers the optic flow between each pair of

consecutive frames. The network is trained to minimize an self-supervised learning objective obtained by combining three loss functions:1) a pixel-wise reconstruction error, 2) a smoothness loss addressing the ambiguity problem of optic flow estimation (also known as the aperture problem), and 3) a structural dissimilarity between the original and the reconstructed image patches (see Zhu et al. [2019] for details of loss functions). The models' pool contains one version of the self-supervised dynamic net, trained on the HAA-500 action dataset Chung et al. [2021].

Supervised (sup) dynamic net is the second part of the temporal stream in the hidden two-stream convolutional neural network model Zhu et al. [2019]. The model has resnet18 architecture and takes optic flows from the self-supervised dynamic net as input. We used the HAA-500 dataset Chung et al. [2021] and trained the supervised dynamic net to predict action labels using optic flows.

Masked Autoencoder (MAE) learn image representations, required to reconstruct original uncorrupted images from corrupted (masked) input through a series of transformer blocks He et al. [2022]. The models' pool contains two versions of the MAE model: 1) a pre-trained version, where the model is trained to reconstruct pixel values, and 2) a fine-tuned version, where the pre-trained model is further fine-tuned to predict object identities. Both versions were trained on Image-net Deng et al. [2009].

Video Masked Autoencoder (VMAE) learns a spatiotemporal representation of videos, required to reconstruct original uncorrupted videos, from corrupted (tube masked) input through a series of transformer blocks Tong et al. [2022]. We added two versions of the VMAE to our models' pool. The first is a pre-trained version, where the model is trained to reconstruct missing pixels of the input set of frames. The second version is the fine-tuned version obtained by fine-tuning the pre-trained version to predict action labels of input videos. Both models take a consecutive set of frames as input, and were trained on the Kinetics-400 action dataset Kay et al. [2017].

Masked Video Distillation (MVD) learns a higher-level spatial and spatiotemporal representation of the input video, required to reconstruct the representation of teacher MAE and VMAE while taking corrupted (tube-masked) videos as input Wang et al. [2023a]. Unlike VMAE and MAE, the MVD model does not learn pixel-level features. Rather, it learns high-level features of the input video using pre-trained MAE and VMAE models' features as masked prediction targets. Using the Kinetics-400 action dataset Kay et al. [2017], a pre-trained version was obtained and added to the models' pool.

2.5 Models' Representational Dissimilarity Matrices (RDM)

In order to compare the models and the fMRI data, we computed representational dissimilarity matrices (RDMs) for the models' layers with a multi-step procedure. First, since the temporal resolution of the models' representations (25Hz) is much higher than the temporal resolution of fMRI data, we down-sampled each layer's activation timecourses over time by selecting one data point every five time points(down to 5 Hz). Then, we convolved the layer's activations with a standard Hemodynamic Response Function (HRF). Given that the fMRI data's repetition time (TR) is 2 seconds, we took a layer's activation every $25 \times 2 = 50$ time points.

Finally, for each layer we computed the dissimilarities between all pairs of timepoints, obtaining RDMs in which the entry at column j and row i contains correlation dissimilarity (1-Pearson's r) between the layer activations at time i and time j. We repeated this procedure for BOLD responses to all eight movie segments, resulting in eight RDMs.

2.6 Brain Representational Dissimilarity Matrices (RDM)

RDMs were constructed separately for each brain stream in the subject's native space. The voxels for each brain stream were obtained as the union of the region voxels for individual regions within that stream. For each brain stream, we calculated the correlation dissimilarity ($1 - r$ where r is Pearson's correlation) of fMRI response patterns for all pairs of TRs. This yielded eight RDMs, corresponding to BOLD responses in eight video segments.

2.7 Measuring models similarity with brain data

To evaluate how well each model accounts for the activity in the brain streams, we used a cross-validated linear regression to predict the left-out movie segment brain stream RDM and computed the correlation between the predicted and the true RDM in each brain stream. The correlation captures how well a model's layers can predict a brain stream's responses to the visual stimuli. First, we

used each model's layers' RDMs corresponding to seven (out of eight) video segments to train a linear regression model that predicts the corresponding seven RDMs in each brain stream. Then, we averaged the linear regression model's coefficients along the seven segments and used the averaged coefficients to predict the brain stream RDM of the left-out segment, using the model layers' RDMs of the corresponding segment. Finally, we calculated the Pearson's correlation between the predicted and the true RDMs. We repeated the leave-one-out cross-validation process for all the segments and averaged over the obtained correlations.

2.8 Measuring combined models similarity with brain data

We sought to study whether a combination of features from two models can improve similarity with brain data. We followed the procedure in 2.7 and used RDMs of all the layers in a pair of combined models to estimate the coefficients of a linear regression model that best predicts the RDM of a brain stream in seven (out of eight) of the video segments. Using leave-one-out cross-validation, we predicted the brain stream RDM of the left-out video segment using the average of the coefficients obtained from the seven video segments during training. Finally, we measured the correlation between the predicted RDM and the actual brain stream RDM to measure the correspondence between the combined models' features and the brain activity.

2.9 Measuring unique and shared similarity of a pair of models with brain data

To evaluate how well unique and shared features among a pair of computational models correspond to the brain data, we used Pearson's r to measure the accuracy of a "target" model's layers prediction of a brain stream RDM while controlling for the variation of a "control" model layers. Using leave-one-out cross-validation, first, we estimated the coefficients of a linear regression model that predicts a brain stream's RDM from the control model's layers in training video segments (seven out of eight). Second, we subtracted the predicted from the actual brain stream RDM in the training and the left-out video segments to obtain training and left-out residuals. Third, we estimated the coefficients of a linear regression model that predicts training residuals of each video segment using the target model layers. Finally, we measured Pearson's correlation between the target model's prediction of the left-out video segment residuals and the residuals obtained from the prediction of the control model.

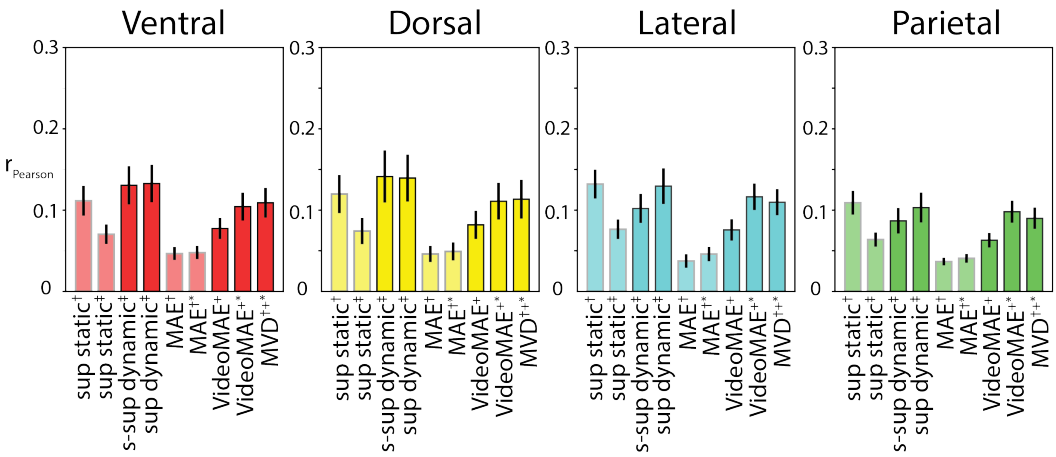

Figure 2: Pearson's correlation between actual and predicted brain stream RDMs, averaged over participants. Predicted RDMs were obtained by training and test a leave-one-out cross-validation linear regression model using each model's layers. Error bars show standard deviation over participants. Lighter bars correspond to models containing static, and darker ones to models containing dynamic visual information (sup: supervised, s-sup: self-supervised, †: Image-net-trained, ‡: HAA-500-trained, +: Kinetics-400-trained, *: fine-tuned; MVD was trained on pre-trained MAE (Image-net) and VideoMAE (Kinetics-400))

3 Results

The human visual system does not consist of a single processing stream. Instead, it is organized into distinct neural pathways. To the extent that the structure of the human visual system is shaped by computational optimality, understanding the visual representations encoded in these pathways can offer insights into more general principles of vision. The contribution of this work is to quantify the similarity between the representations in the different visual pathways in the human brain and representations in models of vision that are widely used in Computer Science but understudied in Cognitive Neuroscience (e.g. masked autoencoders, two-stream networks, masked video distillation). In a first set of analyses 3, we leverage differences between models to reveal differences between the information encoded in different visual pathways in the brain. In a second set of analyses, we quantify the unique contribution of different deep network models to account for neural responses 3.2.

3.1 DNN models similarity with human brain streams

While numerous research studies have been conducted on model-to-brain correspondence using static images Rose et al. [2021], Doshi and Konkle [2023], Tsao et al. [2006], the impact of dynamic information on neural responses to naturalistic videos is understudied. To fill this gap, we tested the correspondence between neural representations in different visual pathways (ventral, dorsal, lateral, and parietal) and deep network models that can process dynamic information (two-stream networks, video masked autoencoders, and masked video distillation). Comparing the correspondence of neural responses with these models and their correspondence with models that only process static information (standard convolutional ResNets, masked autoencoders) made it possible to study the contribution of dynamic information independently of whether the learning objective is supervised (as in two-stream networks) or unsupervised (as in masked autoencoders). In addition, the study of the correspondence between neural representations and representations in masked autoencoders is of interest in its own right: masked autoencoders are effective and widely used, but little is known about their similarity to neural representations.

3.1.1 Static and dynamic information in brains and feed-forward convolutional neural networks (CNN)

Functional MRI responses recorded during the observation of naturalistic videos in the ventral, dorsal, lateral and parietal visual pathways were compared to the representations in feed-forward convolutional neural networks. The same dataset (HAA-500) was used to train the different branches of a hidden-two-stream network: the "supervised static" branch (a ResNet that takes as input individual frames of a video and computes as output the action category), the "unsupervised dynamic" branch (a convolutional network trained to compute optic flow by minimizing a self-supervised loss), and the "supervised dynamic" branch (a ResNet that takes as input optic flow and computes as output the action category). In addition, to facilitate parallels with prior work, we compared neural responses to a widely studied feed-forward model: a ResNet trained with Image-net.

Comparing deep network models trained with the same dataset (HAA-500) showed that models including dynamic information correlated with neural responses more than the Spatial model, that does not use dynamic information (Figure 2). This effect was observed for all visual pathways. In addition, representations in the lateral and parietal pathways correlated more with the supervised dynamic model than with the unsupervised dynamic model (fisher-transformed t-values with Bonferroni-corrected threshold). Lateral and parietal regions are located downstream compared to the dorsal regions, thus this result is complementary to earlier work that reported a correspondence between subsequent stages of processing in deep neural networks and in neural pathways in the case of static visual stimuli [Khaligh-Razavi and Kriegeskorte, 2014] and in the case of auditory stimuli [Kell et al., 2018].

Supervised CNNs trained with Image-Net performed well, achieving correspondence with neural responses that was close to that of HAA-trained models that included dynamic information. This could indicate that some of the variance in neural responses that correlates with dynamic models might also be accounted for by models trained exclusively with static information, as long as a suitable training dataset is used (in this case, Image-Net). However, an alternative possibility is that the supervised static model trained with Image-Net and the dynamic models trained with HAA might

account for different portions of the variance in neural responses. We investigate these alternative possibilities in section 3.2.

3.1.2 Static and dynamic information in brains and masked autoencoders

Masked Autoencoders (MAE, He et al. [2022]) and Video Masked Autoencoders (VideoMAE, Tong et al. [2022], Feichtenhofer et al. [2022]) models are trained to reconstruct masked pixels of input (image or video) during pre-training and are further fine-tuned to predict object/action labels. MAE and VideoMAE models are very effective in learning visual representations and have been shown to outperform competing models in several visual tasks He et al. [2022], Tong et al. [2022], Feichtenhofer et al. [2022], Wang et al. [2023b], Venkatesh et al.. However, it is still unknown whether the representations learned by models based on masked autoencoding are similar to visual representations in the human brain. Here we investigated this question, quantifying the correlation between neural responses measured with fMRI while participants watched naturalistic videos, and representations learned by models trained with masked autoencoding.

We compared the correspondence between neural responses and MAEs trained with images (which learn spatial relationships between component of an image, Wang et al. [2023a]) as well as VideoMAEs (which learn temporal relationships in videos, Wang et al. [2023a]). Finally, we also compared neural responses to masked video distillation (MVD, Wang et al. [2023a]), which combines image MAEs and videoMAEs to better capture both spatial and temporal relationships. Unlike MAE and VideoMAE, the MVD model does not aim to reconstruct missing patches at the level of pixel values. Instead, MVD adopts a knowledge-distillation approach, reconstructing missing information at the level of features extracted from pre-trained MAE and VideoMAE teachers.

As in the case of supervised models trained with the HAA dataset, models that included dynamic information (VideoMAEs) outperformed models using only static information (Image MAEs). This pattern was observed across all visual pathways. Image MAEs did not correlate well with neural responses, even compared to supervised models trained with static inputs. Overall, the representations learned by Image MAEs were very different from neural representations. By contrast, VideoMAEs showed greater correspondence with neural responses. In particular, fine-tuning with an action recognition task (Figure 2, VideoMAE fine-tuned) improved the correspondence between VideoMAE representations and neural representations across all streams (fisher-transformed t-values with Bonferroni-corrected threshold). Across all the pre-trained models, pre-trained MVD showed the highest similarity to neural representations in all brain streams. Further, MVD showed comparable similarity with brain streams to that of fine-tuned VideoMAE.

3.2 Vision models capture shared and unique neural activity variation in human brain streams

The results described in 3 show that representations from models trained with dynamic information are more correlated with neural representations compared to representations from models trained with static information. This overall pattern is broken by the exception of ResNets trained with ImageNet, which performed on par with models trained with dynamic information. This raises the question of whether ResNets trained with ImageNet and dynamic models explain overlapping variance in neural responses or whether, instead, they are complementary, capturing non-overlapping portions of the variance. This question can be posed more generally for any pair of models studied in section 3. We investigated this first by combining layers from two models and measuring whether a combination of models can better predict the pattern of neural activity in visual pathways. Second, we measured the correspondence between a "target" model's representations and the representations in each brain stream while controlling for the representations encoded in a "control" model. To this end, we predicted neural representations using the representations of the control model and obtained the residuals. Then, we predicted the residuals using the representations in the target model (see Methods for details).

Each matrix in Figure 3.a shows how well a combination of models' layers can predict the pattern of neural activity in a brain stream. Each column of each row demonstrates the correlation between the neural response pattern of a brain stream and the combined models' layers' prediction of that brain stream's neural activity pattern. Model-to-brain-stream similarity increased in all brain streams when combining features from static models with features from dynamic models. Notably, the correspondence between combined models' features with both dorsal and ventral streams improved

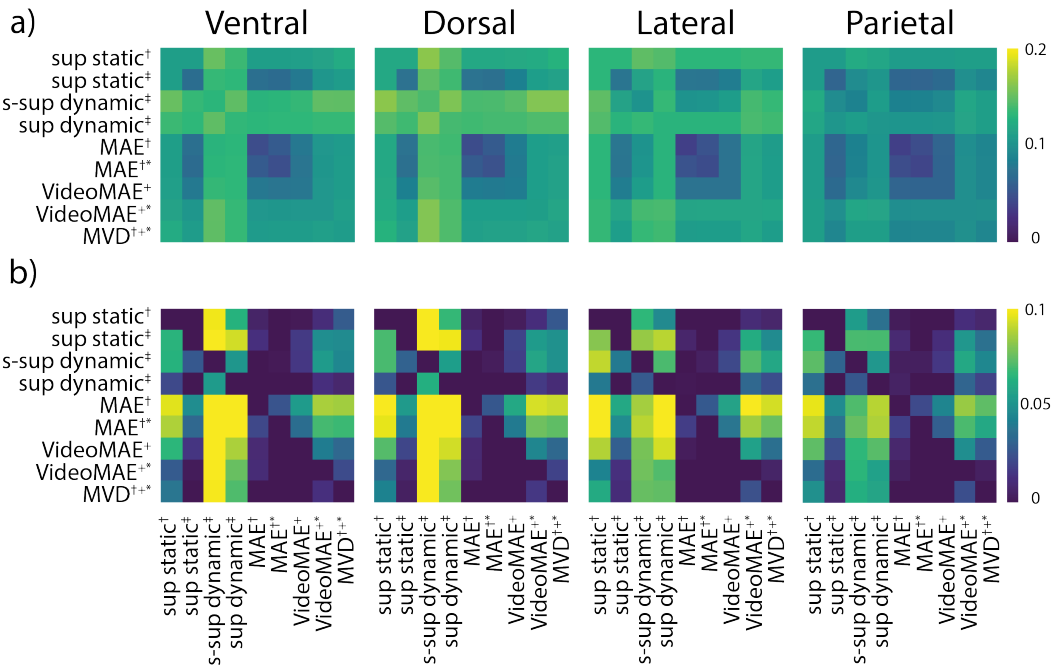

Figure 3: **a)** Model combination similarity with brain streams. The similarity was calculated using Pearson's correlation between a brain stream's actual and predicted RDMs. These predictions were obtained by combining layers from two models (corresponding to the row and column names), and averaged across participants. A linear regression model was trained and tested using leave-one-out cross-validation to generate the predictions.. **b)** Models unique similarity with brain streams. The similarity was calculated using Pearson's correlation between the actual RDM of a brain stream and the RDM predicted by a target model while controlling for the variation explained by a control model in the brain stream. These correlations were averaged across participants. Each row corresponds to a different control model, and each column corresponds to a different target model used for prediction. (sup: supervised, s-sup: self-supervised, †: Image-net-trained, ‡: HAA-500-trained, +: Kinetics-400-trained, *: fine-tuned; MVD was trained on pre-trained MAE (Image-net) and VideoMAE (Kinetics-400))

in two cases: 1) combined features from Image-net-trained static supervised models with dynamic features from self-supervised model and 2) features from the combination of the self-supervised dynamic model with either VideoMAE or MVD. These cases shows that, first, ventral and dorsal brain streams both represent static and dynamic visual features, and second, different types of dynamic information are represented in both of these human streams.

Figure 3.b demonstrates the correspondence between a target model's features and each brain stream when we controlled for the features of a control model in the brain stream's neural responses. The results are visualized as a matrix in which each row corresponds to a control model and each column to a target model. The first row displays the correlations between models and neural responses after controlling for the Image-net-trained static model. The high values for the columns corresponding to the self-supervised dynamic and the supervised dynamic models indicate that these models and the Image-net-trained static model capture non-overlapping variance in neural responses. Representations learned by these models also capture non-overlapping variance with those learned by the unsupervised dynamic models: the VideoMAEs. This finding shows that despite VideoMAEs exhibit relatively high correlations with neural responses (outperforming Image MAEs), they nonetheless fail to capture some variance in human visual representations that is accounted for by self-supervised and supervised dynamic models.

VideoMAEs and MVD accounted for additional variance in neural responses compared to MAEs (as expected given the results in Figure 2) but also compared to the HAA-trained static and self-supervised dynamic models. However, they accounted for a minimal amount (if any) of additional

variance compared to the supervised dynamic model, suggesting some degree of convergence on common representations across models trained with different learning objectives.

4 Limitations

This study focused on a set of models selected to enable comparing the contribution of static and dynamic information and the impact of supervised and unsupervised learning objectives. The selection of models in this study includes only a subset of the existing models, future work will be needed to expand the set of models tested. In addition, the present work centered on the comparison between models and entire visual streams. A finer-grained analysis comparing models to individual regions within each stream will require further work.

5 Discussion

Three main findings emerged. First, models including dynamic information outperformed models using exclusively static information, not only in the dorsal, lateral and parietal streams, but also in the ventral stream. This is in line with recent evidence of responses to dynamic features in ventral brain regions [Robert et al., 2023]. Patients with deficits for motion perception typically present with lesions affecting dorsal regions (such as area V5, [McLeod, 1996, Vaina et al., 1990, Zihl et al., 1983]) or parietal regions [Battelli et al., 2003]. By contrast, patients with damage to ventral regions typically do not present with deficits for motion perception [Gilaie-Dotan et al., 2015]. This raises the question of what might be the use of dynamic information represented in the ventral stream. We hypothesize that this information might be used to support object segmentation, as proposed in recent computational models [Chen et al., 2022] inspired by classic work in Developmental Psychology [Spelke, 1990].

Second, Image MAEs showed little correspondence with neural representations, even compared to other models trained exclusively with static information. These results indicate that despite the effectiveness of Image MAEs for learning visual representations that can transfer to a variety of visual tasks [He et al., 2022], these models do not converge on representations that are similar to those observed in the human brain, suggesting that human vision and image MAEs rely on different computational mechanisms.

Third, models based on optic flow representations accounted for unique variance in all streams, even compared to video masked autoencoders that can make use of dynamic information. Fine-tuning video MAEs with an action classification task increased their correspondence with neural representations, but did not fully bridge the gap with neural responses compared to optic flow models, which still explained additional unique variance compared to the fine-tuned video MAEs. The additional contribution of optic flow models was particularly strong in ventral and dorsal regions, suggesting that representations based on optic flow exhibit greater correspondence with representations in early stages of visual processing in the human brain compared to both image and video MAEs.

References

Leslie G Ungerleider, Mortimer Mishkin, et al. Two cortical visual systems. analysis of visual behavior. *Ingle DJ, Goodale MA, Mansfield RJW*, 1982.

David Pitcher and Leslie G Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25(2):100–110, 2021.

Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

Tzvi Ganel and Melvyn A Goodale. Visual control of action but not perception requires analytical processing of object shape. *Nature*, 426(6967):664–667, 2003.

Jody C Culham, Stacey L Danckert, Joseph FX De Souza, Joseph S Gati, Ravi S Menon, and Melvyn A Goodale. Visually guided grasping produces fmri activation in dorsal but not ventral stream brain areas. *Experimental brain research*, 153:180–189, 2003.

Anitha Pasupathy and Charles E Connor. Population coding of shape in area v4. *Nature neuroscience*, 5(12):1332–1338, 2002.

Nikos K Logothetis, Jon Pauls, and Tomaso Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current biology*, 5(5):552–563, 1995.

Keiji Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

Shimon Edelman, Kalanit Grill-Spector, Tammar Kushnir, and Rafael Malach. Toward direct visualization of the internal shape representation space by fmri. *Psychobiology*, 26:309–321, 1998.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Semir Zeki, JD Watson, CJ Lueck, Karl J Friston, C Kennard, and RS Frackowiak. A direct demonstration of functional specialization in human visual cortex. *Journal of neuroscience*, 11(3): 641–649, 1991.

Roger B Tootell, John B Reppas, Kenneth K Kwong, Rafael Malach, Richard T Born, Thomas J Brady, Bruce R Rosen, and John W Belliveau. Functional analysis of human mt and related visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience*, 15(4):3215–3230, 1995.

Hide-aki Saito, Masao Yukie, Keiji Tanaka, Kazuo Hikosaka, Yoshiro Fukada, and Eiichi Iwai. Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *Journal of Neuroscience*, 6(1):145–157, 1986.

G Beckers and V Hömberg. Cerebral visual motion blindness: transitory akinetopsia induced by transcranial magnetic stimulation of human area v5. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 249(1325):173–178, 1992.

David Pitcher, Bradley Duchaine, and Vincent Walsh. Combined tms and fmri reveal dissociable cortical pathways for dynamic and static face perception. *Current Biology*, 24(17):2066–2070, 2014.

Petra Vetter, Marie-Helene Grosbras, and Lars Muckli. Tms over v5 disrupts motion prediction. *Cerebral cortex*, 25(4):1052–1059, 2015.

Zoe Kourtzi, Heinrich H Bülthoff, Michael Erb, and Wolfgang Grodd. Object-selective responses in the human motion area mt/mst. *Nature neuroscience*, 5(1):17–18, 2002.

Erez Freud, Jody C Culham, David C Plaut, and Marlene Behrmann. The large-scale organization of shape processing in the ventral and dorsal pathways. *elife*, 6:e27576, 2017.

L Cornette, Patrick Dupont, A Rosier, Stefan Sunaert, P Van Hecke, J Michiels, Luc Mortelmans, and GA Orban. Human brain regions involved in direction discrimination. *Journal of Neurophysiology*, 79(5):2749–2765, 1998.

Stefan Sunaert, Paul Van Hecke, Guy Marchal, and Guy A Orban. Motion-responsive regions of the human brain. *Experimental brain research*, 127:355–370, 1999.

Sophia Robert, Leslie G Ungerleider, and Maryam Vaziri-Pashkam. Disentangling object category representations driven by dynamic and static visual input. *Journal of Neuroscience*, 43(4):621–634, 2023.

Daniel L Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. *Advances in neural information processing systems*, 26, 2013.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 363–378. Springer, 2019.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (compcor) for bold and perfusion based fmri. *Neuroimage*, 37(1):90–101, 2007.

Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, 25(10):3911–3931, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13465–13474, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6312–6322, 2023a.

Olivia Rose, James Johnson, Binxu Wang, and Carlos R Ponce. Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. *Nature communications*, 12(1):6723, 2021.

Fenil R Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023b.

Rahul Venkatesh, Honglin Chen, Klemen Kotar, Kevin Feigelis Wanhee Lee Daniel Bear, and Daniel Yamins. Climbing the ladder of causation with counterfactual world modeling.

P McLeod. Preserved and impaired detection of structure from motion by a'motion-blind" patient. *Visual Cognition*, 3(4):363–392, 1996.

Lucia M Vaina, Marjorie Lemay, Don C Bienfang, Albert Y Choi, and Ken Nakayama. Intact "biological motion" and "structure from motion" perception in a patient with impaired motion mechanisms: A case study. *Visual neuroscience*, 5(4):353–369, 1990.

Josef Zihl, D Von Cramon, and Norbert Mai. Selective disturbance of movement vision after bilateral brain damage. *Brain*, 106(2):313–340, 1983.

Lorella Battelli, Patrick Cavanagh, and Ian M Thornton. Perception of biological motion in parietal patients. *Neuropsychologia*, 41(13):1808–1816, 2003.

Sharon Gilaie-Dotan, Ayse Pinar Saygin, Lauren J Lorenzi, Geraint Rees, and Marlene Behrmann. Ventral aspect of the visual form pathway is not critical for the perception of biological motion. *Proceedings of the National Academy of Sciences*, 112(4):E361–E370, 2015.

Honglin Chen, Rahul Venkatesh, Yoni Friedman, Jiajun Wu, Joshua B Tenenbaum, Daniel LK Yamins, and Daniel M Bear. Unsupervised segmentation in real-world images via spelke object inference. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Michael Hanke, Nico Adelhöfer, Daniel Kottke, Vittorio Iacovella, Ayan Sengupta, Falko R Kaule, Roland Nigbur, Alexander Q Waite, Florian Baumgartner, and Jörg Stadler. A studyforrest extension, simultaneous fmri and eye gaze recordings during prolonged natural stimulation. *Scientific data*, 3(1):1–15, 2016.

5.1 Supplementary materials

5.1.1 Training and testing the Two-stream CNN for action recognition

We adopted the models in Zhu et al. [2019] and trained on the HAA500 dataset Chung et al. [2021]. The dataset contains over 591k labeled frames with 500 action classes. 85% of the data points were used for training, 5% for validation, and 10% for testing 5.1.1. The training dataset was converted to the Webdataset format, i.e., shards of tar files. We used 4 V100 GPUs and 8 workers to load the dataset and train the models. All the analyses were performed on the same version of the movie that was used to acquire fMRI responses in the StudyForrest dataset Hanke et al. [2016].
The *supervised static model* have a ResNet18 architecture He et al. [2016], and were trained for 47 epochs with a batch size of 128. The training was done with the stochastic gradient descent algorithm with a 0.001 initial learning rate and a 0.0001 weight decay. During training, the gradients were accumulated and backpropagated for every two batches. Each frame in an input batch is a 224×224 frame and was randomly flipped horizontally.
The *unsupervised dynamic model* was trained for 12 epochs with a batch size of 32 and an initial learning rate of 0.01. No weight decay was used during training. Input to this model consists of a set of 11 frames each with dimensions of 224×224.
The *supervised dynamic model* was trained for 50 epochs with a batch size of 128 and an initial learning rate of 0.001. A weight decay of 0.0005 was used to train the models, and the gradients were accumulated and backpropagated every 5 batches.

Table 2: Test performance of models on the HAA500 dataset

Model	epochs	Performance Top-1	Top-3
sup static	47	30.80%	49.38%
unsup dynamic + sup dynamic	12, 50	22.72%	37.90%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**The abstract contains a summary of conclusions, and the introduction clearly states the motivation and the contributions made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**A separate part is dedicated to describing the limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**All the datasets used in the paper, including image and video datasets as well as the fMRI dataset, are publicly available, and the proper citation is included in the manuscript. All the steps required to reproduce the results are clearly explained in the Methods section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: **[TODO]**All the datasets used in the paper, including image and video datasets as well as the fMRI dataset, are publicly available, and the proper citation is included in the manuscript. The code will be released shortly after the submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**full details of the model that were trained from scratch are explained in supplementary materials. Other models were adopted from the official repository, without any changes to the original setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: **[TODO]**error bars were demonstrated in result figures. significant statistical tests were reported throughout the manuscript

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: **[TODO]**Yes. The details are explained in the supplementary materials

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**Proper citation and URL are included in the manuscript.

    Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes] The human data is adopted from a publicly available dataset. For provided essential instruction for referring to the study and the accompanying dataset.

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: **[TODO]**We did not collect any data and only used publicly available repositories.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.