Multi-modal brain encoding models for multi-modal stimuli

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite participants engaging in single modality stimuli, such as watching images or silent videos, recent work has demonstrated that multi-modal Transformer models can predict visual brain activity impressively well, even with incongruent modality representations. This raises the question of how accurately these multi-modal models can predict brain activity when participants are engaged in multi-modal stimuli. As these models grow increasingly popular, their use in studying neural activity provides insights into how our brains respond to such multi-modal naturalistic stimuli, i.e., where it separates and integrates information from different sensory modalities. We investigate this question by using multiple unimodal and two types of multi-modal models—cross-modal and jointly pretrained—to determine which type of models is more relevant to fMRI brain activity when participants were engaged in watching movies (videos with audio). We observe that both types of multi-modal models show improved alignment in several language and visual regions. This study also helps in identifying which brain regions process unimodal versus multi-modal information. We further investigate the impact of removal of unimodal features from multi-modal representations and find that there is additional information beyond the unimodal embeddings that is processed in the visual and language regions. Based on this investigation, we find that while for cross-modal models, their brain alignment is partially attributed to the video modality; for jointly pretrained models, it is partially attributed to both the video and audio modalities. The inability of individual modalities in explaining the brain alignment effectiveness of multi-modal models suggests that multi-modal models capture additional information processed by all brain regions. This serves as a strong motivation for the neuro-science community to investigate the interpretability of these models for deepening our understanding of multi-modal information processing in brain.

1 Introduction

The study of brain encoding aims at predicting the neural brain activity recordings from an input stimulus representation. Recent brain encoding studies use neural models as a powerful approach to better understand the information processing in the brain in response to naturalistic stimuli (Oota et al., 2023a). Current encoding models are trained and tested on brain responses captured from participants who are engaged in a *single stimulus modality*, using stimulus representations extracted from AI systems that are pretrained on single modality, such as language (Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2020; Schrimpf et al., 2021; Toneva et al., 2022; Aw & Toneva, 2023), vision (Yamins et al., 2014; Eickenberg et al., 2017; Schrimpf et al., 2018; Wang et al., 2019) or speech (Millet et al., 2022; Vaidya et al., 2022; Tuckute et al., 2023). In this paper, we build encoding models where participants are engaged with *multi-modal stimuli* (e.g., watching movies that also include audio). We explore multi-modal stimulus representations extracted

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

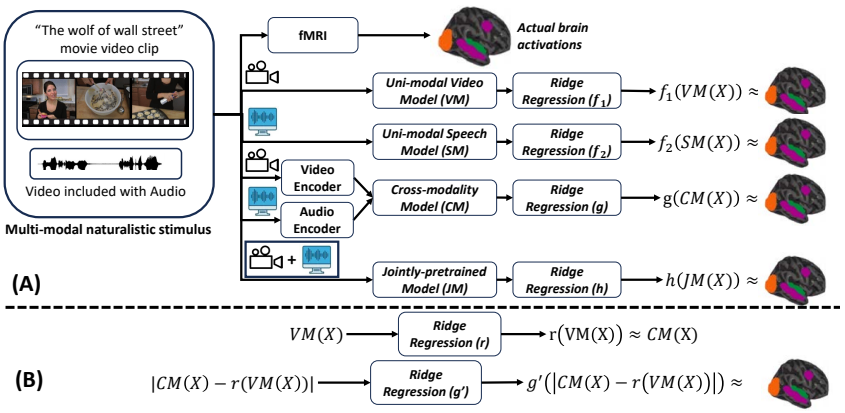

Figure 1: (A) Overview of our proposed Multi-modal Brain Encoding Pipeline. (B) Residual Analysis.

using Transformer (Vaswani et al., 2017) based multi-modal models. Our analysis focuses on their alignment with both uni- and multi-modal brain regions.

There is a growing evidence that the human brain's ability for multi-modal processing is underpinned by synchronized cortical representations of identical concepts across various sensory modalities (Gauthier et al., 2003; Bracci & Op de Beeck, 2023). Reflecting similar principles, the recent advances in AI systems have led to the development of multi-modal models (like CLIP (Radford et al., 2021), ImageBind (Girdhar et al., 2023), and TVLT (Tang et al., 2022)) using massive interleaved image-text data, speech-text data or video-audio-text data to represent multi-modal input. This recent progress in AI has stimulated advancements in brain encoding models (Doerig et al., 2022; Oota et al., 2022; Popham et al., 2021; Wang et al., 2022; Tang et al., 2024; Nakagi et al., 2024) that learn effectively from multiple input modalities, despite participants being engaged with single stimulus modality during experiments, e.g., watching natural scene images, or silent movie clips. However, these studies have experimented with subjects engaged with single-modality stimulus, leaving the full potential of these models in true multi-modal scenarios still unclear.

Using brain recordings of participants watching several popular movies included with audio (St-Laurent et al., 2023), we investigate several research questions. First, we investigate the effectiveness of multi-modal stimulus representations obtained using multi-modal models versus unimodal models for brain encoding. Multi-modal models are of two broad types: (i) cross-modal pretrained models, where first individual modality encoders are trained and then cross-modal alignment is performed, and (ii) jointly pretrained models, which involve combining data from multiple modalities and training a single joint encoder. Hence, we also investigate which of the two types (cross-modal versus joint) are better for encoding. In this work, we focus on one cross-modal (ImageBind), one jointly pretrained (TVLT), three video and two speech models. Additionally, we explore which modality representations are more brain relevant, and identify which brain regions process uni- and multi-modal information. Overall, this research utilizes various modality representations to develop encoding models based on fMRI responses within a multi-modal model framework (see Fig. 1 for workflow).

Using our multi-modal brain encoding approach, we examine several insights. First, we use previous neuroscience findings that have identified brain regions involved in visual, language and auditory processing, and investigate how well our model aligns with these regions when both the model and a human participant watch the same multi-modal video stimuli. Second, we expect that multi-modal models which can learn cross-modal and joint embeddings across modalities in a brain-relevant way would significantly align with these regions. However, alignment with these brain regions doesn't necessarily mean that the model is effectively learning from multiple modalities, as unimodal models for vision or language or audio have also been shown to significantly align with these brain regions (Wehbe et al., 2014; Toneva et al., 2022; Schrimpf et al., 2021; Millet et al., 2022; Vaidya et al., 2022). To check the second aspect, we investigate this question via a direct approach, closely related to previous studies (Toneva et al., 2022; Oota et al., 2023b,c). For each modality, we analyze how the alignment between brain recordings and multi-modal model representations is affected by the elimination of information related to that particular modality from the model representation.

Our analysis of multi-modal brain alignment leads to several key conclusions: (1) Both cross-modal and jointly pretrained models demonstrate significantly improved brain alignment with language

regions (AG, PCC, PTL, and IFG) and visual regions (EVC and MT) when analyzed against unimodal video data. In contrast, compared to unimodal speech-based models, all multi-modal embeddings show significantly better brain alignment, except in the OV (object visual processing) region. This highlights the ability of multi-modal models to capture additional information—either through knowledge transfer or integration between modalities—which is crucial for multi-modal brain alignment. (2) Using our residual approach, we find that the improved brain alignment in cross-modal models can be partially attributed to the removal of video features alone, rather than auditory features. On the other hand, the improved brain alignment in jointly pretrained models can be partially attributed to the removal of both video and auditory features.

Overall, we make the following contributions in this paper. (1) To the best of our knowledge, this study is the first to leverage both cross-modal and jointly pretrained multi-modal models to perform brain alignment while subjects are engaged with multi-modal naturalistic stimuli. (2) We evaluate the performance of several unimodal Transformer models (three video and two audio) and measure their brain alignment. (3) Additionally, we remove unimodal features from multi-modal representations to explore the impact on brain alignment before and after their removal. We will release code upon publication of this paper.

2 Related Work

Multi-modal models. Pretrained Transformer-based models have been found to be very effective in various tasks related to language (Devlin et al., 2019; Radford et al., 2019), speech (Baevski et al., 2020), and images (Dosovitskiy et al., 2020). To learn associations between pairs of modalities, Transformer models have been pretrained on multiple modalities, showing excellent results in multi-modal tasks like visual question answering and visual common-sense reasoning. These multi-modal models are pretrained in two different ways: (i) cross-modal models that integrate information from multiple modalities and learn a joint encoder, such as VisualBERT (Li et al., 2019) and ImageBind (Girdhar et al., 2023), and (ii) jointly pretrained models like LXMERT (Tan & Bansal, 2019), CLIP (Radford et al., 2021), ViLBERT (Lu et al., 2019), and TVLT (Tang et al., 2022) which fuse individual modality encoders at different stages, transferring knowledge from one modality to another. In this work, we investigate how the representations extracted from *cross-modal and jointly-pretrained Transformer models* align with human brain recordings when participants engage with multi-modal stimuli.

Brain Encoding using Multi-modal Models. Since human brain perceives the environment using information from multiple modalities (Gauthier et al., 2003), examining the alignment between language and visual representations in the brain by training encoding models on fMRI responses, while extracting joint representations from multi-modal models, can offer insights into the relationship between the two modalities. For instance, it has been shown that multi-modal models like CLIP (Radford et al., 2021) better predict neural responses in the high-level visual cortex as compared to previous vision-only models (Doerig et al., 2022; Wang et al., 2022). Additionally, Tang et al. (2024) demonstrate the use of multi-modal models in a cross-modal experiment to assess how well the language encoding models can predict movie-fMRI responses and how well the vision encoding models can predict narrative story-fMRI. Nakagi et al. (2024) analyzed fMRI related to video content viewing and found distinct brain regions associated with different semantic levels, highlighting the significance of modeling various levels of semantic content simultaneously. However, these studies have experimented with subjects engaged with single-modality stimulus, leaving the full potential of these models in true multi-modal scenarios still unclear. Recently, Dong & Toneva (2023) interpreted the effectiveness of pretrained versus finetuned multi-modal video transformer using video+text stimuli-based brain activity. However, they did not perform any cross-modal vs jointly-pretrained model analysis or analysis of multi-modal versus unimodal models, leaving it unclear which type of multi-modal models perform best for brain activity prediction. Further, unlike them, we study video+audio stimuli, and perform comprehensive residual analysis.

3 Dataset Curation

Brain Imaging Dataset. We experiment with a multi-modal naturalistic fMRI dataset, Movie10 (St-Laurent et al., 2023) obtained from the Courtois NeuroMod databank. This dataset was collected while six human subjects passively watched four different movies: *The Bourne supremacy (∼100*

mins), *The wolf of wall street (∼170 mins)*, *Hidden figures (∼120 mins)* and *Life (∼50 mins)*. Among these, *Hidden figures* and *Life* are repeated twice, with the repeats used for testing and the remaining movies for training. In this work, we use *Life* movie for testing where we average the two repetitions to reduce noise in brain data. This dataset is one of the largest publicly available multi-modal fMRI dataset in terms of number of samples per participant. It includes 4024 TRs (Time Repetitions) for *The Bourne supremacy*, 6898 TRs for *The wolf of wall street* used in train and 2028 TRs for *Life* in test. The fMRI data is collected every 1.49 seconds (= 1TR).

The dataset is already preprocessed and projected onto the surface space ("fsaverage6"). We use the multi-modal parcellation of the human cerebral cortex based on the Glasser Atlas (which consists of 180 regions of interest in each hemisphere) to report the ROI (region of interest) analysis for the brain maps (Glasser et al., 2016). This includes four visual processing regions (early visual (EV), object-related areas (LO), face-related areas (OFA) and scene-related areas (PPA)), one early auditory area (AC), and eight language-relevant regions, encompassing broader language regions: angular gyrus (AG), anterior temporal lobe (ATL), posterior temporal lobe (PTL), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), middle frontal gyrus (MFG), posterior cingulate cortex (PCC) and dorsal medium prefrontal cortex (dmPFC), based on the Fedorenko lab's language parcels (Milton et al., 2021; Desai et al., 2022). We list the detailed sub-ROIs of these ROIs in Appendix B.

Estimating dataset cross-subject prediction accuracy. To account for the intrinsic noise in biological measurements, we adapt Schrimpf et al. (2021)'s method to estimate the cross-subject prediction accuracy for a model's performance for the Movie10 fMRI datasets. By subsampling fMRI datasets from 6 participants, we generate all possible combinations of s participants ($s \in$ [2,6]) for watching movies, and use a voxel-wise encoding model (see Sec. 5) to predict one participant's response from others. Note that the estimated cross-subject prediction accuracy is based on the assumption of a perfect model, which might differ from real-world scenarios, yet offers valuable insights into model's performance. We estimate cross-subject prediction accuracy in three settings: (i) training with *The Bourne supremacy* and testing with *Life* data, (ii) training with *The wolf of wall street* and testing with *Life* data, and (iii) training with both *The Bourne supremacy* and *The wolf of wall street* and testing with Life data. We present the average cross-subject prediction accuracy across voxels for the *Movie10 fMRI* dataset and across the three settings in Appendix A.

4 Methodology

4.1 Multi-modal models

To analyse how human brain process information while engaged in multi-modal stimuli, we use recent popular deep learning models to explore multiple modalities information and build the encoding models in two different ways: "cross-modality pretraining" and "joint pretraining".

Cross-modality Pretrained Multi-modal Models. Cross-modality representations involve transferring information or learning from one modality to another. For example, in a cross-modal learning scenario, text descriptions can be used to improve the accuracy of image/video recognition tasks. This approach is often used in scenarios where one modality might have limited data or less direct relevance but can be informed by another modality.

Recently, a cross-modal model called ImageBind (IB) (Girdhar et al., 2023) has shown immense promise in binding data from six modalities at once, without the need for explicit supervision. ImageBind model uses separate encoders for each individual modality and learns a single shared representation space by leveraging multiple types of image-paired data. ImageBind consists of 12 layers and outputs a 1024 dimensional representation for each modality.

Jointly Pretrained Multi-modal Models. Jointly pretrained multi-modal model representations, on the other hand, involve combining data from multiple modalities to build a more comprehensive joint understanding to improve decision-making processes. The system processes these diverse inputs concurrently to make more informed and robust decisions.

TVLT (Zellers et al., 2022) is an end-to-end Text-less Vision-Language multi-modal Transformer model for learning joint representations of video and speech from YouTube videos. This joint encoder model consists of a 12-layer encoder (hidden size 768) and uses masked autoencoding objective for both videos and speech. Given the video-speech pairs, the TVLT model provides 768 dimensional representations for each modality across 12 layers.

Extraction of multi-modal features. To extract video and audio embedding representations from multi-modal models for the brain encoding task, we input video and audio pairs at each TR and obtain the aligned embeddings for the two modalities. Here, we first segment the input video and audio into clips corresponding to 1.49 seconds, which matches the fMRI image rate. For both the models, ImageBind and TVLT, we use the pretrained Transformer weights. ImageBind generates an embedding for each modality (IB video and IB audio) in an aligned space. We concatenate these embeddings to create what we refer to as IB concat embeddings. On the other hand, TVLT provides a joint embedding across all modalities at each layer. Only for the last layer, TVLT provides an embedding for each modality.

4.2 Unimodal Models

To investigate the effectiveness of multi-modal representations in comparison to representations for individual modalities, we use the following methods to obtain embeddings for individual modalities.

Video-based models. To extract representations of the video stimulus, we use three popular pretrained Transformer video-based models from Huggingface (Wolf et al., 2020): (1) Vision Transformer Base (ViT-B) (Dosovitskiy et al., 2020), (2) Video Masked Autoencoders (VideoMAE) (Tong et al., 2022) and (3) Video Vision Transformer (ViViT) (Arnab et al., 2021). Details of each model are reported in Table 1 in Appendix.

Speech-based models. Similar to video-based models, we use two popular pretrained Transformer speech-based models from Huggingface: (1) Wav2Vec2.0 (Baevski et al., 2020) and (2) AST (Baade et al., 2022). Details of each model are reported in Table 1 in Appendix.

Extraction of video features. ViT-B (Dosovitskiy et al., 2020), the underlying video encoder model for ImageBind is used for extracting representations for all frames in each TR for every video. To extract embedding at each TR, we average all frame embeddings and obtain the corresponding video representation. For VideoMAE and ViViT, we directly obtain the video embeddings for each TR. All 3 models provide 768 dimensional representations and all of them are 12-layer Transformer encoders.

Extraction of speech features. To explore whether speech models incorporate linguistic information, we extract representations beyond 1.49 secs, i.e., we considered context window of 16 secs with stride of 100 msecs and considered the last token as the representative for each context window. The pretrained speech-based models output token representations at different layers. Both Wav2Vec2.0 and AST models provide 768 dimensional representations and all of them are 12-layer Transformer encoders. Finally, we align these representations with the fMRI data acquisition rate by downsampling the stimulus features with a 3-lobed Lanczos filter, thus producing chunk-embeddings for each TR.

5 Experimental Setup

Encoding Model. We train bootstrap ridge regression based voxel-wise encoding models (Deniz et al., 2019) to predict the fMRI brain activity associated with the stimulus representations obtained from the individual modalities (speech and video) and multi-modal embeddings from cross-modal and jointly pretrained multi-modal models. For each subject, we account for the delay in the hemodynamic response by modeling hemodynamic response function using a finite response filter (FIR) per voxel with 5 temporal delays (TRs) corresponding to $\sim$7.5 seconds (Huth et al., 2022). Formally, at each time step t, we encode the stimuli as $X_t \in \mathbb{R}^D$ and brain region voxels $Y_t \in \mathbb{R}^V$, where D denotes the dimension of the concatenation of delayed 5 TRs, and V denotes the number of voxels. Overall, with N such TRs, we obtain N training examples.

Train-test Setup. We build encoding models in three settings: (i) We used all data samples from 10 training sessions of the *The Bourne supremacy* movie for training and tested generalization on samples from the test sessions (5 sessions) of the *Life* movie. (ii) We used data from 17 training sessions of the *The wolf of wall street* movie for training, with the *Life* movie used for testing. (iii) We combined data from the *The Bourne supremacy* and *The wolf of wall street* movies for training, and tested on the *Life* movie.

Removal of a single modality features from multi-modal representations. To remove features for a particular modality m from multi-modal model representations, we rely on a simple method proposed previously by Toneva et al. (2022) and Oota et al. (2023b), in which the linear contribution

of the features to the multi-modal model activations is removed via ridge regression. Specifically, for this ridge regression the feature vector corresponding to modality m is considered as input and the multi-modal representations are the target. We compute the residuals by subtracting the predicted multi-modal feature representations from the actual multi-modal features resulting in the (linear) removal of feature vector for modality m from the pretrained multi-modal embeddings. Because the brain prediction method is also a linear function, this linear removal limits the contribution of features for modality m to the eventual brain alignment. See Fig. 1(B).

Evaluation Metrics. We evaluate our models using Pearson Correlation (PC) which is a standard metric for evaluating brain alignment (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022). Let TR be the number of time repetitions in the test set. Let $Y = \{Y_i\}_{i=1}^{TR}$ and $\hat{Y} = \{\hat{Y}_i\}_{i=1}^{TR}$ denote the actual and predicted value vectors for a single voxel. Thus, $Y \; and \; \hat{Y} \; \in \mathbb{R}^{TR}$. We use Pearson Correlation (PC) which is computed as $corr(Y, \hat{Y})$ where corr is the correlation function.

The final measure of a model's performance is obtained by calculating Pearson's correlation between the model's predictions and neural recordings. This correlation is then divided by the estimated cross-subject prediction accuracy and averaged across voxels, regions, and participants, resulting in a standardized measure of performance referred to as normalized brain alignment. For calculating normalized alignment, we select the voxels whose cross-subject prediction accuracy is ≥ 0.05.

Implementation Details for Reproducibility. All experiments were conducted on a machine with 1 NVIDIA GeForce-GTX GPU with 16GB GPU RAM. We used bootstrap ridge-regression with the following parameters: MSE loss function; L2-decay (λ) varied from 10^1 to 10^3; the best λ was chosen by tuning on validation data that comprised a randomly chosen 10% subset from the train set used only for hyper-parameter tuning.

Statistical Significance. To determine if normalized predictivity scores significantly higher than chance, we run a permutation test using blocks of 10 contiguous fMRI TRs (considering the slowness of hemodynamic response) rather than individual TRs. By permuting predictions 5000 times, we create an empirical distribution for chance performance, from which we estimate the p-value of the actual performance. To estimate the statistical significance of performance differences, such as between the model's predictions and chance or residual predictions and chance, we utilized the Wilcoxon signed-rank test (Conover, 1999), applying it to the mean normalized predictivity for the participants. In all cases, we denote significant differences (p≤ 0.05) with a $*$ or $\wedge$.

6 Results

6.1 How effective are multi-modal representations obtained from multi-modal models?

In Fig. 2, we present the average normalized brain alignment scores for both multi-modal and individual modality features. Specifically, we show the normalized brain alignment for cross-modality (ImageBind), jointly pretrained multi-modal (TVLT), and the average from individual video and speech models. The results are shown for whole brain, and also for average across language and visual ROIs. Results for individual ROIs are in Fig. 3.

Baseline comparison. To compare the brain predictivity of multi-modal and unimodal models against baseline performance, we employ randomly generated vector embeddings to predict brain activity as baseline. We observe that the brain alignment from a random vector is significantly lower than that of both multi-modal and unimodal models across the whole brain and language-visual processing regions. This shows that the representations from these multi-modal models are significant enough for learning non-trivial alignment with the fMRI recordings of multi-modal stimuli.

Cross-modal vs. Jointly pretrained multi-modal models vs. Unimodal Models. Fig. 2(left) displays results for whole brain analysis, where the IB Concat bar plot corresponds to results for representations from a cross-modal model, while TVLT Joint bar plot corresponds to results for representations from a jointly pretrained multi-modal model. From Fig. 2(left), we make the following observations: (i) At the whole brain level, the Wilcoxon signed-rank test shows that the differences in embeddings from the IB Concat and TVLT models are not statistically significant. (ii) The multi-modal embeddings show improved brain alignment compared to unimodal models. Specifically, cross-modal embeddings are significantly better than both unimodal video and speech models, while jointly pretrained embeddings are significantly better than speech models. This implies that cross-

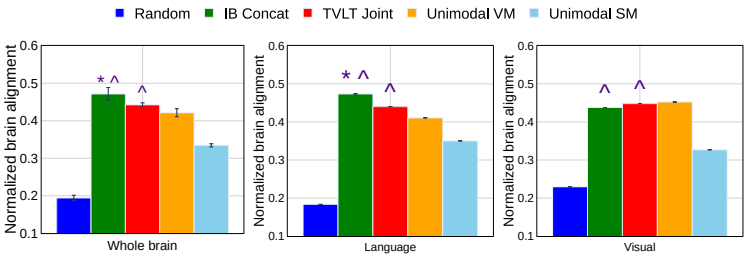

Figure 2: Average normalized brain alignment for both multi-modal and individual modality features across whole brain, language, and visual regions. Error bars indicate the standard error of the mean across participants. $*$ indicates cases where multi-modal embeddings are significantly better than unimodal video models (VM), i.e., p≤ 0.05. $\wedge$, indicates cases where multi-modal embeddings are significantly better than unimodal speech models (SM), i.e., p≤ 0.05.

modal embeddings contain additional information beyond the two modalities, while embeddings from a jointly pretrained model do not provide extra information beyond unimodal visual information but do contain additional information beyond unimodal speech.

We also present average results across language and visual regions in Figs. 2 (middle), and 2(right), respectively. The Wilcoxon signed-rank test shows that the differences in embeddings from the IB Concat and TVLT models are not statistically significant when averaged across language and visual regions. Similar to whole brain performance, in the language regions, cross-modal embeddings are significantly better than both unimodal video and speech models, while jointly pretrained embeddings are significantly better than unimodal speech models. In contrast, for the visual regions, the normalized brain alignment of cross-modal and jointly pretrained embeddings is similar to the performance of unimodal video models. This implies that when we average across visual regions, there is no additional information beyond unimodal video features. However, when compared to unimodal speech features, both multi-modal embeddings show significant improvement.

Since we didn't observe any significant difference at the whole brain level and when averaged across language and visual regions, between cross-modal and jointly pretrained multi-modal models, we attempt to seek if there any any differences when we pay a closer look at the individual ROIs. We present results for language and visual regions such as Angular gyrus (AG), the posterior temporal lobe (PTL), and the inferior frontal gyrus (IFG) in Fig. 3. Additionally, we cover visual regions like early visual cortex (EVC), scene visual areas (SV) and middle temporal gyrus (MT), as well as early auditory cortex (AC). In this figure, we also report the average normalized brain alignment of each modality obtained from multi-modal models. Unlike the whole brain analysis, we observe some differences between cross-modal and jointly pretrained models in several language and visual ROIs. Results for other ROIs are in Fig. 7 in Appendix. Our observations are as follows: (i) Cross-modal IB Concat embeddings are significantly better than TVLT Joint embeddings in semantic regions such as AG and PCC, as well as the multi-modal processing region MT. (ii) Conversely, TVLT Joint embeddings are significantly better than IB Concat embeddings in dmPFC regions. While considering both joint and each modality embeddings from multi-modal models, we make the following observations from Fig. 3: (1) Cross-modal IB video embeddings exhibit improved brain alignment compared to unimodal video in the AG and MT regions with the exceptions of the PTL and AC regions. But this is not the case for IB audio vs unimodal audio. This suggests that video modality information is more relevant and beneficial in the brain for IB Concat embeddings from cross-modality models. (2) TVLT video embeddings show improved brain alignment in the AG, PTL, PCC, dmPFC and EVC regions, with other regions displaying similar normalized brain alignment unimodal video embeddings. (3) Consistent with the cross-modality models, in jointly pretrained TVLT models, TVLT video embeddings significantly outperform TVLT audio embeddings, except in PTL region. These observations indicate that video information is advantageous for both cross-modal and jointly pretrained models, whereas audio embeddings mainly benefit the PTL region.

6.2 Which brain regions process uni- and multi-modal information?

From Fig. 3, we observe that multi-modal video embeddings exhibit improved brain alignment not only in the whole brain but also in various language, visual and multi-modal regions. For instance, the cross-modal IB Concat embeddings demonstrate superior brain alignment compared to unimodal video-based models in areas such as the AG, PTL, IFG, and PCC. Moreover, TVLT-joint embeddings

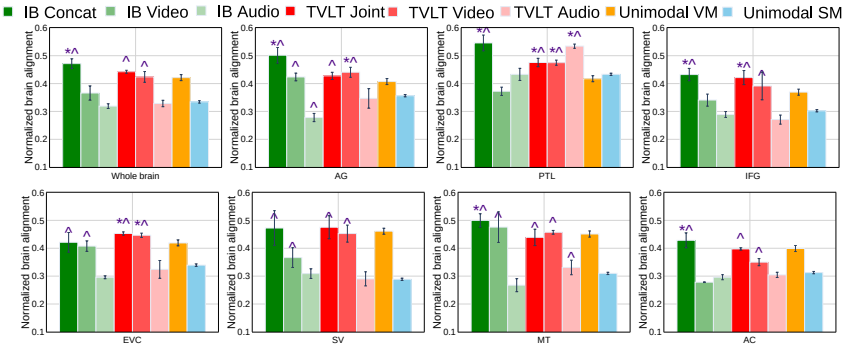

Figure 3: Average normalized brain alignment for video and audio modalities from multi-modal and individual modality features across whole brain and several ROIs of language (AG, PTL and IFG), visual (EVC, SV and MT) and auditory cortex (AC). Error bars indicate the standard error of the mean across participants. $*$ indicates cases where multi-modal embeddings are significantly better than unimodal video models (VM), i.e., p≤ 0.05. $\wedge$ indicates cases where multi-modal embeddings are significantly better than unimodal speech models (SM), i.e., p≤ 0.05.3

show notable enhancements in the AG, PTL, IFG, PCC, dmPFC and EVC regions. In contrast, compared to unimodal speech-based models, all multi-modal embeddings display significantly better brain alignment, except the OV (object visual processing) region. Overall, this observation suggests that integrating multiple modalities leads to transferring information from one modality to another, resulting in improved brain predictability. Based on these, it can be inferred that these multi-modal models can indeed learn multi-modal linkages that are relevant to the brain.

When subjects engage with multi-modality stimuli, we observe that multi-modal embeddings show improvements in semantic regions such as the AG, PCC and dmPFC, and syntactic regions such as the PTL and IFG. Overall, we find that multi-modal information is processed in only a few regions. Furthermore, several regions, including the SV (scene visual area), EVC (early visual cortex), ATL (anterior temporal lobe), IFGOrb, MFG, and dmPFC, exhibit similar brain alignment with both unimodal and multi-modal embeddings.

6.3 How is the brain alignment of multi-modal features affected by the elimination of a particular modality?

To understand the contribution of each modality to the multi-modal brain alignment for multi-modal naturalistic stimulus, we perform residual analyses by removing the unimodality features from multi-modal joint representations as well as multi-modal video or audio representations from joint representations and measure the differences in brain alignment before and after removal modality-specific features. Fig. 4 displays the normalized brain alignment for language (AG) and visual regions (MT). We note a decrease in brain alignment for both the AG and MT regions following the removal of video embeddings from cross-modality models, whereas the removal of audio embeddings does not affect the brain alignment. On the other hand, for jointly pretrained models, removal of both video and audio embeddings partially impacts the brain alignment. We observe similar findings for language ROIs such as PTL, MFG, ATL, PCC and visual regions EVC, OV and FV, as shown in Figs. 9 and 10 in Appendix. These results suggest that there is additional information beyond the unimodal embeddings considered in this study that is processed in the visual and language regions.

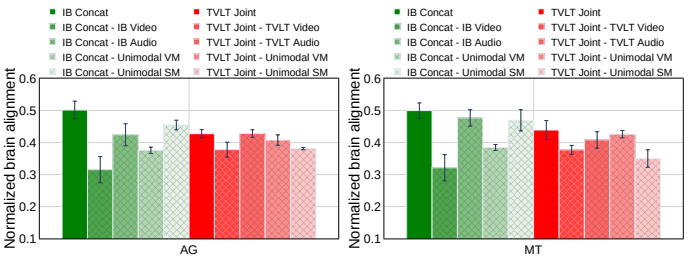

Figure 4: Residual analysis: Average normalized brain alignment was computed across participants before and after removal of video and audio embeddings from both jointly pretrained and cross-modality models. Error bars indicate the standard error of the mean across participants.

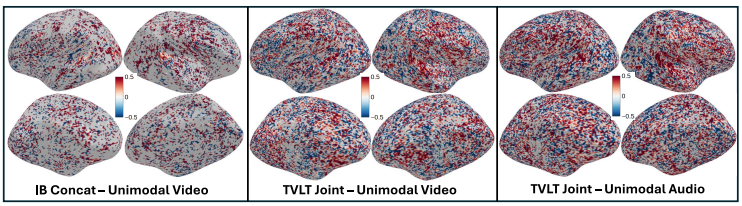

Figure 5: Percentage decrease of brain alignment after removal of (left) Unimodal VM embeddings from IB-Concat (middle) Unimodal VM embeddings from jointly pretrained TVLT, and (right) Unimodal SM embeddings from TVLT Joint. Colorbar indicates the percentage of decrease where red denotes higher and white denotes zero.

Qualitative analysis. We compute the percentage decrease in alignment for each voxel following the removal of unimodal video embeddings from the IB Concat (cross-modality) and the TVLT Joint (jointly pretrained model), with projections onto the brain surface averaged across participants, as depicted in Fig. 5. The colorbar shows the percentage decrease in brain alignment, where red voxels indicate a higher percentage decrease and white voxels indicate areas where unimodal video features do not contribute any shared information within the multi-modal context. We observe that removal of unimodal video features leads to a significant drop (40-50%) in performance in the visual regions for IB Concat, and in language regions (PTL & MFG) for TVLT Joint.

7 Discussion

Using multi-modal model representations, including both cross-modal and jointly pretrained types, we evaluated how these representations can predict fMRI brain activity when participants are engaged in multi-modal naturalistic stimuli. Further, we compared both multi-modal and unimodal representations and observed their alignment with both unimodal and multi-modal brain regions. This is achieved by removing information related to unimodal stimulus features (audio and video) and observing how this perturbation affects the alignment with fMRI brain recordings acquired while participants are engaged in watching multi-modal naturalistic movies.

Our analysis of multi-modal brain alignment yields several important conclusions: (1) The improved brain alignment of the multi-modal models over unimodal models, across several language, visual, and auditory regions is only partially attributable to the video and audio stimulus features presented to the model. A deeper understanding of these models is required to shed light on the underlying information processing of both unimodal and multi-modal information. (2) Cross-modal representations have significantly improved brain alignment in language regions such as AG, PCC and PTL. This variance can be partially attributed to the removal of video features alone, rather than auditory features. (3) Video embeddings from multi-modal models exhibit higher brain alignment than audio embeddings, except in the PTL and AC regions. This suggests that audio-based models may encode weaker brain-relevant semantics. (4) Both cross-modal and jointly pretrained models demonstrate significantly improved brain alignment with language regions (AG, PCC, PTL and IFG) compared to visual regions when analyzed against unimodal video data. In contrast, when compared to unimodal audio-based models, all multi-modal embeddings display significantly better brain alignment, with the exception of the OV region. This underscores the capability of multi-modal models to capture additional information—either through knowledge transfer or integration between modalities—crucial for multi-modal brain alignment.

Limitations. The low alignment scores clearly show that despite the increasing popularity of multi-modal models in tackling complex tasks such as visual question answering, we are still far from developing a model that fully encapsulates the complete information processing steps involved in handling multi-modal naturalistic information in the brain. In the future, by fine-tuning these multi-modal models on specific tasks such as generating captions for videos, we can better leverage their alignment strengths. This approach will allow us to explore task-level brain alignment of three modalities—video, audio, and text—more effectively. Further, multi-modal large language models (MLLMs) (Zhang et al., 2023; Ataallah et al., 2024; Wu et al., 2023) that align visual features from video frames into the LLM embedding space via a trainable linear projection layer, offer promise for enhanced multi-modal capabilities. We would further extend this work by comparing the region-wise brain alignment performance of these multi-modal LLM models with existing approaches.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *Interspeech 2022*, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.

Cordell M Baker, Joshua D Burks, Robert G Briggs, Andrew K Conner, Chad A Glenn, Kathleen N Taylor, Goksel Sali, Tressie M McCoy, James D Battiste, Daniel L O'Donoghue, et al. A connectomic atlas of the human cerebrum—chapter 7: the lateral parietal lobe. *Operative Neurosurgery*, 15(suppl_1):S295–S349, 2018.

Stefania Bracci and Hans P Op de Beeck. Understanding human object vision: a picture is worth a thousand representations. *Annual review of psychology*, 74:113–135, 2023.

Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.

William Jay Conover. *Practical nonparametric statistics*, volume 350. john wiley & sons, 1999.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 2019.

Rutvik Desai, Usha Tadimeti, and Nicholas Riccardi. Proper and common names in the semantic system, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 2022.

Dota Tianai Dong and Mariya Toneva. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

Isabel Gauthier, Thomas W James, Kim M Curby, and Michael J Tarr. The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, 20(3-6):507–523, 2003.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Alexander G Huth, Shinji Nishimoto, An T Vu, and T Dupre La Tour. Gallant lab natural short clips 3t fmri data. *50 GiB*, 2022.

Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pp. 6629–6638, 2018.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13–23, 2019.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.

Camille K Milton, Vukshitha Dhanaraj, Isabella M Young, Hugh M Taylor, Peter J Nicholas, Robert G Briggs, Michael Y Bai, Rannulu D Fonseka, Jorge Hormovas, Yueh-Hsin Lin, et al. Parcellation-based anatomic model of the semantic network. *Brain and behavior*, 11(4):e02065, 2021.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. The brain tells a story: Unveiling distinct representations of semantic content in speech, objects, and stories in the human brain with large language models. *bioRxiv*, pp. 2024–02, 2024.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. In *COLING*, pp. 116–133, 2022.

Subba Reddy Oota, Manish Gupta, Raju S Bapi, Gael Jobard, Frédéric Alexandre, and Xavier Hinaut. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *arXiv preprint arXiv:2307.10246*, 2023a.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *NeurIPS*, 2023b.

Subba Reddy Oota, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *24th INTERSPEECH Conference*, 2023c.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 2021.

Marie St-Laurent, Basile Pinsard, Oliver Contier, Katja Seeliger, Valentina Borghesani, Julie Boyle, Pierre Bellec, and Martin Hebart. cneuromod-things: a large-scale fmri dataset for task-and data-driven assessment of object representation and visual memory recognition in the human brain. *Journal of Vision*, 23(9):5424–5424, 2023.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36, 2024.

Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. *Advances in Neural Information Processing Systems*, 35:9617–9632, 2022.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022.

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.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.

Aditya Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pp. 21927–21944. PMLR, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *NeurIPS*, 32:15501–15511, 2019.

Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Natural language supervision with a large and diverse dataset builds better models of human high-level visual cortex. *BioRxiv*, pp. 2022–09, 2022.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 11, 2014.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624, 2014.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

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: We have ensured that the main claims made in the abstract and introduction are directly correlating to the research findings and the methods we have employed.

   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: The paper discusses the main limitations of the work performed by the authors in the discussion section.

   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: Our paper does not require any explicit theorems and proofs.

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: The paper has delineated all the information related to the experimental setup in the experimental setup 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: [NA]

Justification: We will release the code upon acceptance. The dataset is publicly available through a licence.

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: We provide all the training and test details in the experimental setup.

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: We conducted our experiments multiple times across 6 participants and took the average results. We also include error bars in the plots.

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: We have included the specifications of the hardware and software environments to ensure the reproducibility of our results.

   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: The research conducted in this paper fully conforms with the NeurIPS Code of Ethics in every respect.

   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: [Yes]

    Justification: The paper explores how the advancements and applications of our findings could benefit society in terms of computational neuroscience research by specifically investigating the effectiveness of the current state of the art multimodal models in encoding brain activity.

    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: Our research does not pose any risks for misuse.

    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: We have explicitly cited the datasets, code and models used.

    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: We will try to opensource the code and provide complete documentation for our assets upon acceptance.

    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: [NA]

    Justification: We use publicly available fMRI dataset and do not collect any new data.

    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: We use publicly available fMRI dataset and do not collect any new data.

    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.

A Cross-subject prediction accuracy

We estimate cross-subject prediction accuracy in three settings: (i) training with *The Bourne supremacy* and testing with *Life* data, (ii) training with *The wolf of wall street* and testing with *Life* data, and (iii) training with both *The Bourne supremacy* and *The wolf of wall street* and testing with Life data. We present the average cross-subject prediction accuracy across voxels for the *Movie10 fMRI* dataset and across the three settings in Fig. 6.

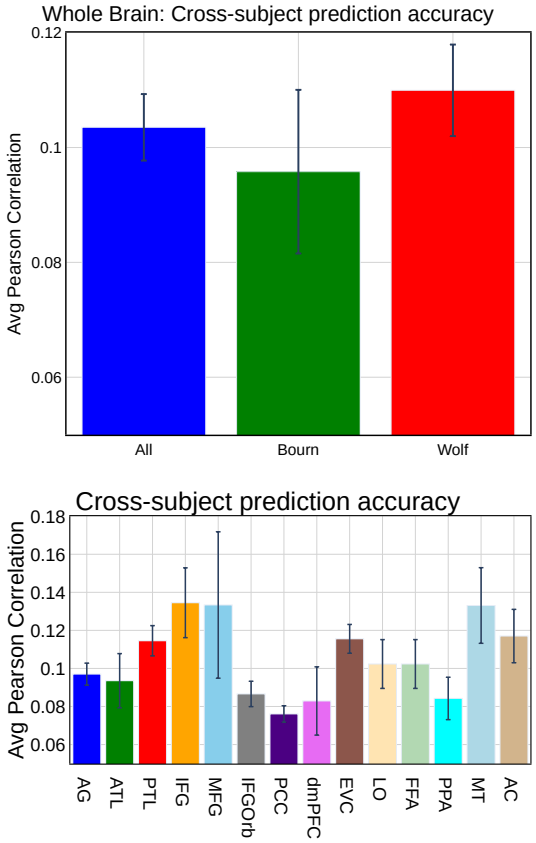

Figure 6: Cross-subject prediction accuracy: (top) across whole brain, (bottom) across language, visual and auditory regions.

B Detailed sub-ROIs of language, visual and auditory regions

The data covers seven brain regions of interest (ROIs) in the human brain with the following sub-divisions: (i) early visual (EV: V1, V2, V3, V3B, and V4); (ii) object-related areas (LO1 and LO2); (iii) face-related areas (OFA), (iv) scene-related areas (PPA), (v) middle temporal (MT: MT, MST, LO3, FST and V3CD), (vi) late language regions, encompassing broader language regions: angular gyrus (AG: PFm, PGs, PGi, TPOJ2, TPOJ3), lateral temporal cortex (LTC: STSda, STSva, STGa, TE1a, TE2a, TGv, TGd, A5, STSdp, STSvp, PSL, STV, TPOJ1), inferior frontal gyrus (IFG: 44, 45, IFJa, IFSp) and middle frontal gyrus (MFG: 55b) (Baker et al., 2018; Milton et al., 2021; Desai et al., 2022).

C Details of pretrained Transformer models

Table 1: Pretrained Transformer-based Encoder Models. All models have 12 layers.

Model Name	Pretraining
Cross-modal Pretrained (ImageBind)	Video & Audio
Jointly Pretrained (TVLT)	Video & Audio
ViT-B	Image
VideoMAE	Video
ViViT	Video
Wav2Vec2.0-base	Speech
AST	Speech

D Effectiveness of multi-modal vs unimodal representations for various brain regions

We now present the results for per unimodal video model and per speech model in Fig. 8. Similar to the average results of unimodal video and speech models, we observe that multi-modal models exhibit better normalized brain alignment than individual unimodal video and speech models across language and visual regions. Among unimodal speech models, the AST model shows better normalized brain alignment than the Wav2vec2.0 model. Among unimodal video models, each unimodal video model displays notably consistent performance across regions.

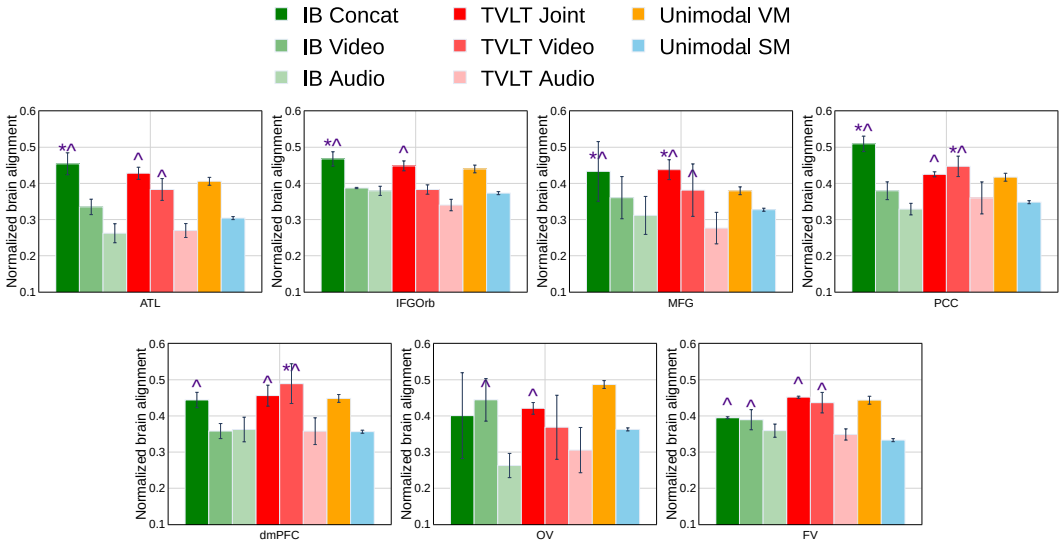

Figure 7: Average normalized brain alignment for per video and audio modalities from multi-modal and individual modality features across whole brain and several ROIs of language (ATL, IFGOrb, MFG, PCC, dmPFC) and visual (OV, FV). Error bars indicate the standard error of the mean across participants.

E How is the brain alignment of multi-modal features affected by the elimination of a particular modality?

To understand the contribution of each modality to the multi-modal brain alignment for multi-modal naturalistic stimulus, we perform residual analyses by removing the unimodality features from multi-modal joint representations as well as multi-modal video or audio representations from joint representations and measure the differences in brain alignment before and after removal modality-specific features. Figs. 9 and 10 display the normalized brain alignment for language ROIs such as PTL, MFG, ATL, PCC and visual regions EVC, OV and FV. We note a decrease in brain alignment for these regions following the removal of video embeddings from cross-modality models, whereas the removal of audio embeddings does not affect the brain alignment. On the other hand, for jointly pretrained models, removal of both video and audio embeddings partially impacts the brain alignment.

F Layerwise brain alignment

We now plot the layer-wise normalized brain alignment for the Unimodal models and TVLT joint model, as shown in Fig. 11. Observation from Fig. 11 indicates a consistent drop in performance from early to lower layers, specifically for both TVLT joint and unimodal video models. The key finding here is that our results that TVLT joint embeddings showcase improved brain alignment across all the layers compared to unimodal video and speech embeddings.

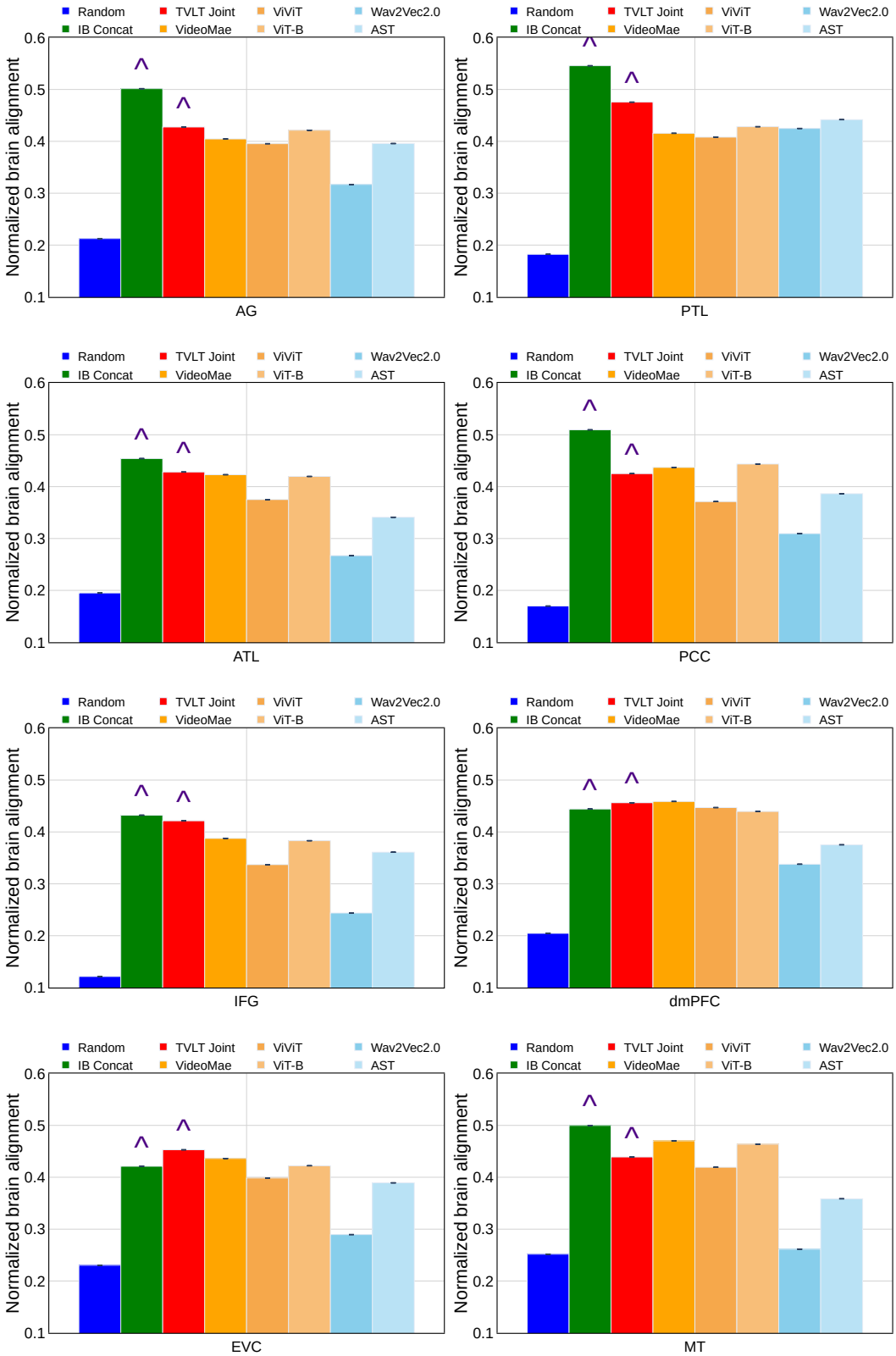

Figure 8: Average normalized brain alignment for video and audio modalities from multi-modal and individual modality features across whole brain and several ROIs of language (ATL, ATL, PTL, IFG, PCC, dmPFC) and visual (EVC, MT). Error bars indicate the standard error of the mean across participants.

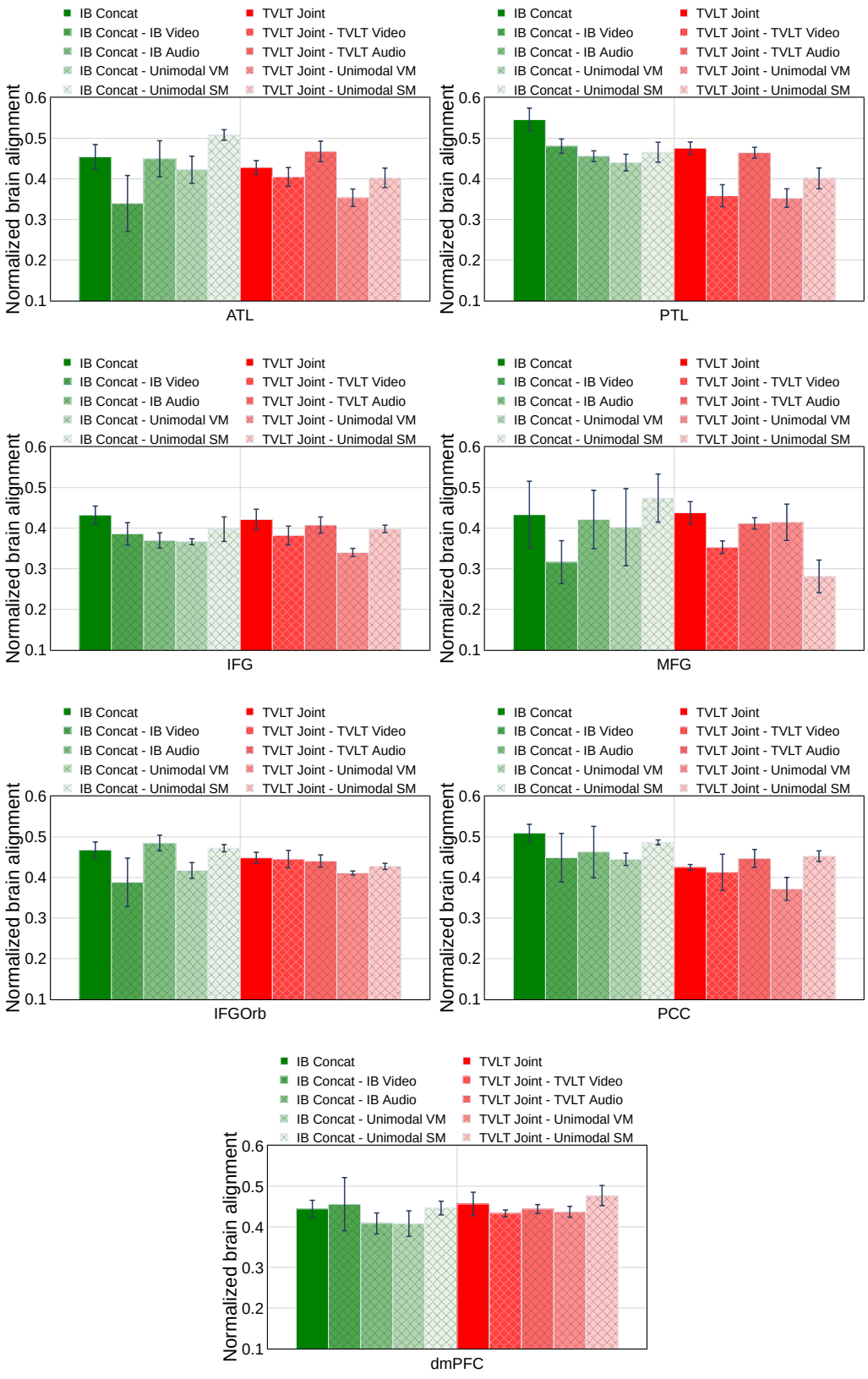

Figure 9: Residual analysis for ATL, PTL, IFG, MFG, IFGOrb, PCC and dmPFC regions: Average normalized brain alignment was computed across participants before and after removal of video and audio embeddings from both jointly pretrained and cross-modality models. Error bars indicate the standard error of the mean across participants.

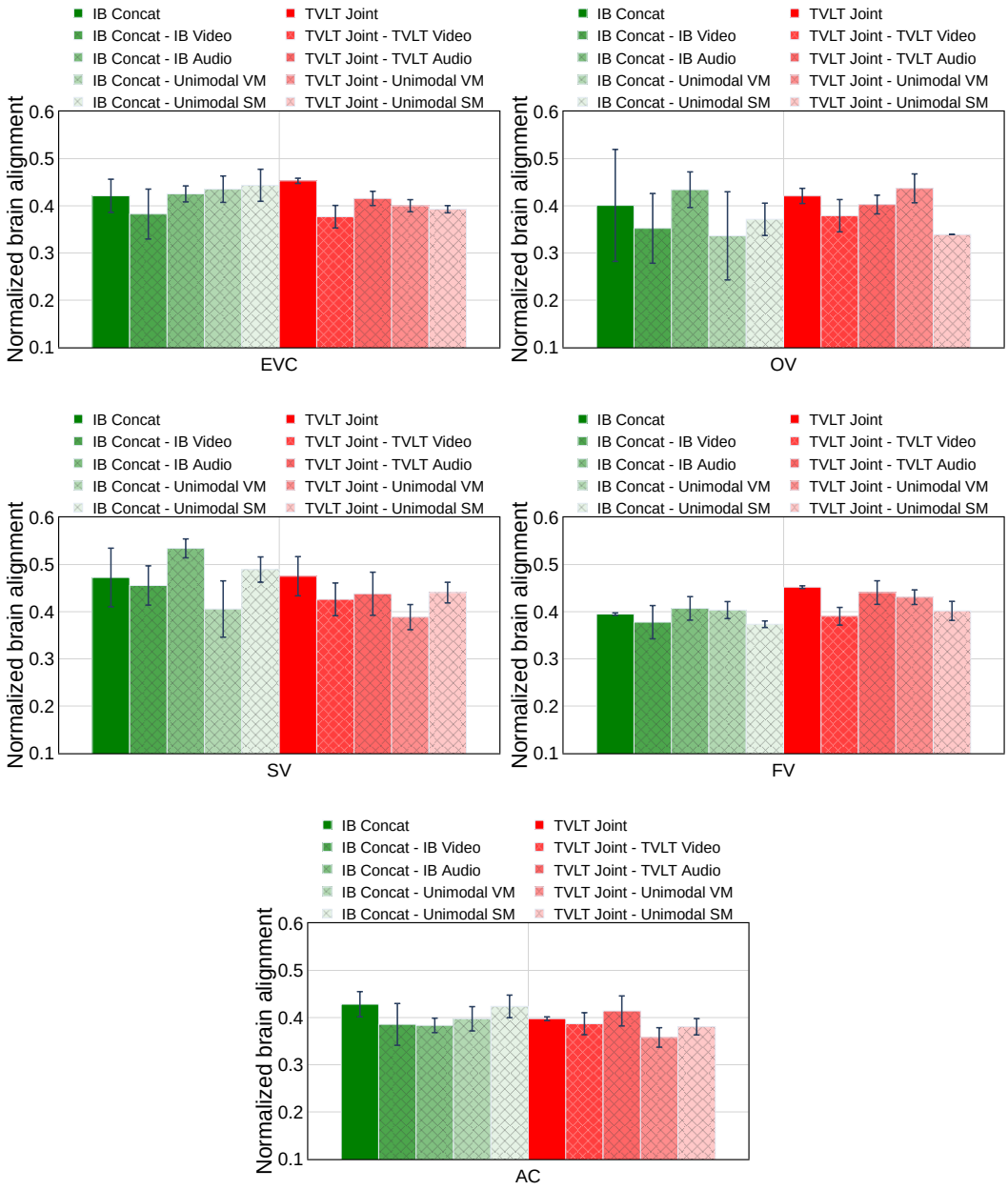

Figure 10: Residual analysis for EVC, OV, SV, FV and AC regions: Average normalized brain alignment was computed across participants before and after removal of video and audio embeddings from both jointly pretrained and cross-modality models. Error bars indicate the standard error of the mean across participants.

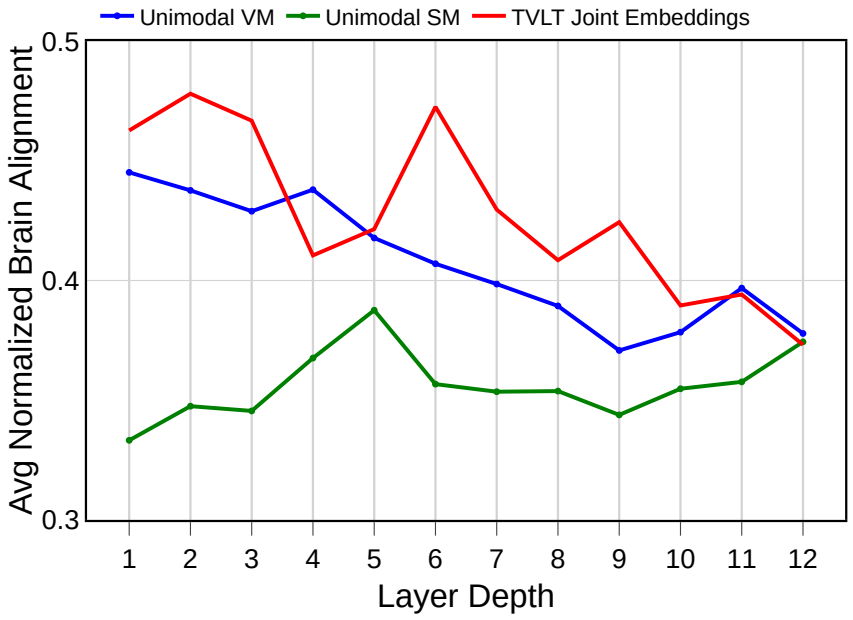

Figure 11: Normalized brain alignment across layers for multi-modal model (TVLT joint embeddings) and unimodal video and speech models.