Two cortical mechanisms for natural audiovisual processing

Subha Nawer Pushpita[1,2]
Leila Wehbe[1,2*]

[1]Machine Learning Department, Carnegie Mellon University
[2]Neuroscience Institute, Carnegie Mellon University

Abstract

Understanding how the human brain processes natural audiovisual information remains a central challenge in cognitive neuroscience. Progress has been limited by the difficulty of modeling complex audiovisual stimuli - most prior work has therefore relied on short, controlled stimuli, or on stimuli from one modality at a time, leaving cortical mechanisms that support real-world comprehension poorly characterized. Further, while recent advances in artificial intelligence now enable the extraction of high-dimensional, time-resolved features from naturalistic stimuli, how cortical regions dynamically process and prioritize auditory and visual information as time unfolds remains largely unexplored. Using large-scale fMRI data collected while participants watched movies, we developed two complementary computational approaches relying on prediction performance to map the moment-by-moment dynamics of sensory processing across cortical regions: one detects sustained periods when one modality predicts a region substantially better than the other, identifying regions that switch the modality they encode for meaningful stretches of time; while the other identifies periods when both modalities predict the region well, revealing regions maintaining balanced representation of both auditory and visual information. Together, these analyses reveal two types of audiovisual processing across the cortex: a pair of "bows" that switch modalities (one posterior bow encircling category-selective visual cortex and another anterior bow spanning dorso-lateral frontal areas) and an arrow-like axis of regions that jointly represents both modalities (extending from lateral occipital cortex into the temporal cortex). The coexistence of these systems points to a cortical architecture that adaptively reweights sensory inputs while maintaining balanced multimodal representations, supporting robust comprehension of complex natural events.

*Corresponding author: `lwehbe@cmu.edu`

Main

Humans effortlessly navigate a rich multimodal world, comprehend complex continuous stream of audiovisual stimuli like movies and TV series, and seamlessly integrate rich information conveyed across multiple sensory modalities. However, the vast majority of cognitive neuroscience experiments, even those using naturalistic stimuli, focus on one modality at a time, such as vision [1, 2, 3, 4] or audition/language processing [5, 6, 7, 8], not studying how these modalities are used together in everyday life. There are a few prevailing theories of how the brain supports multimodal comprehension. The hub-and-spoke view holds that modality-specific "spokes" in sensory cortices converge on an amodal hub in the anterior temporal lobe (ATL), consistent with findings that ATL degeneration yields semantic dementia while syntactical, numerical abilities, and executive functions are fairly intact [9, 10, 11, 12]. In contrast, the convergence-zone view proposes that modality-specific and amodal codes integrate across multiple cortical sites beyond ATL[13]. Extending beyond these models, Driver and Noesselt [14] emphasized that multisensory integration itself can occur at multiple hierarchical levels - from direct interactions between unisensory cortices, to specialized multisensory patches, to higher-order hubs that flexibly coordinate and reshape the balance between unisensory and multisensory representations. Crucially, border regions such as angular gyrus, precuneus, and middle temporal gyrus encode the same semantic categories across visual and linguistic input [15, 16]. There has also been evidence for the semantic alignment hypothesis, suggesting that each location on the anterior edge of the visual cortex that encodes a particular visual category is followed directly anteriorly by a region that encodes the same category in linguistic form [17]. In parallel, distinct patches in superior temporal and parietal cortex were shown to represent social interactions—an intrinsically multimodal construct—highlighting that the brain's capacity for multimodal comprehension extends beyond objects and semantics to the dynamic understanding of how people relate to one another [18]. Beauchamp's work demonstrated that the superior temporal sulcus integrates auditory and visual inputs, with responses exceeding those to either modality alone, establishing it as a key cortical locus for the comprehension of audiovisual speech and action [19].

However, prior work has primarily examined the anatomical loci of multisensory integration and comprehension using primarily short, carefully controlled audiovisual stimuli, and often using non-simultaneous auditory and visual stimuli [19, 1, 2, 3, 4]. Far less is known about the neural mechanisms that govern how content demands at different moments within long, naturalistic audiovisual streams drive cortical regions to adopt diverse modes of audiovisual processing. Further, studies that do use naturalistic stimuli and analysis methods such as encoding models [20] or representation similarity analysis [21], focus

on identifying fixed feature sensitivity of a given region, and do not evaluate whether the sensitivity of these regions changes in time - leading to implications that a region's modality preference or integrative role is static. The assumption of representations being static is problematic for natural narratives, where the informational burden can shift dramatically between modalities over time.

To address this gap, we introduce a new hypothesis space (Figure 1) that incorporates the temporal dynamics of sensory processing: some regions show sustained unisensory dominance - favoring either visual (Figure 1a) or auditory (Figure 1b) inputs throughout (here, auditory refers broadly to all perceptual information received through the ears, including speech, language, prosody, and non-speech environmental sounds) - whereas others act as jointly predicted regions (Figure 1c), reflecting information from both modalities, potentially supporting integrative processing or amodal concept representation, and others emerge as "switching regions" that flexibly shift their preference (Figure 1d), significantly favoring auditory input at some moments and visual input at others.

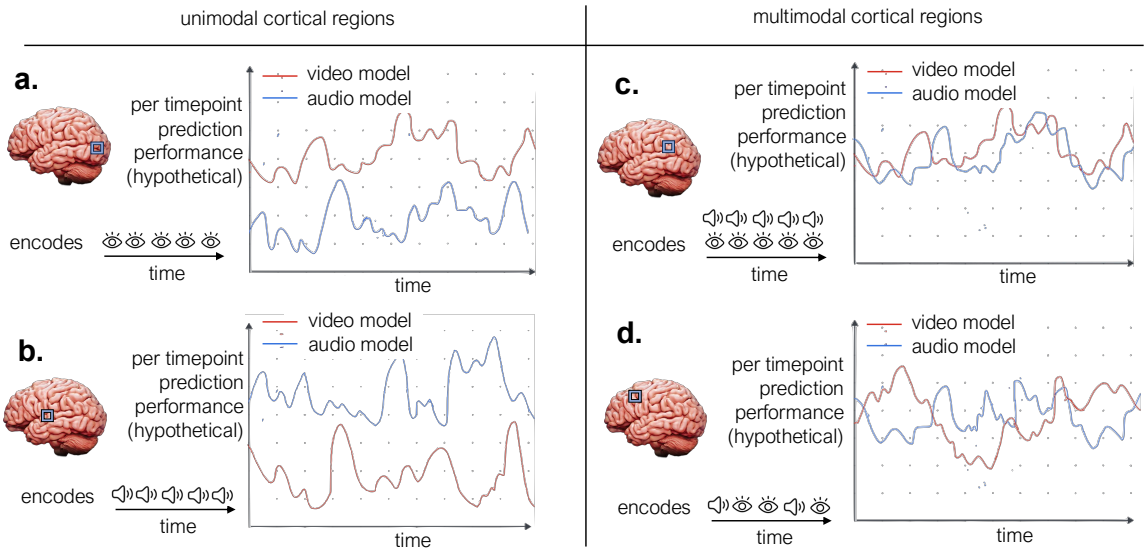


Figure 1: Hypothesis space for unimodal and multimodal regions. We illustrate four canonical patterns of audiovisual encoding that can explain regional differences in prediction performance. **(a)** In unimodal visual regions, BOLD responses will be predicted substantially better by models fitted on visual features (red) than by those fitted on auditory features (blue), reflecting stable visual selectivity. **(b)** Conversely, unimodal auditory/language regions are consistently better predicted by auditory feature–based models, reflecting stable auditory/language selectivity. **(c)** In some multimodal regions, BOLD responses are predicted well by both feature types throughout time, indicating concurrent encoding of auditory and visual information. **(d)** Other multimodal regions alternate between periods dominated by auditory versus visual feature prediction, suggesting dynamic switching of modality preference. Together, these four patterns capture the range of possible temporal dynamics by which cortical regions may support comprehension of naturalistic audiovisual stimuli.

To empirically test this hypothesis space, we adopted a data-driven approach using the Courtois NeuroMod dataset[22, 23], which consists of fMRI BOLD responses from four hu-

man subjects to more than 36 hours of TV series and movies. The data for each participant has been averaged into 1000 cortical parcels from the Schaefer atlas [24], and we further average the cross-participant activity for each parcel. We fit encoding models that predict each parcel activity as a function of distinct auditory and visual feature sets (Figure 2a, and details in Methods). We extracted auditory features from an audio LLM, Qwen2 Audio-7B instruct [25]. This model is trained on large audio–text corpora for speech recognition, emotion, speaker and event identification, and higher-level instruction-following tasks such as transcription, translation, and spoken question answering. Therefore, the collected features capture a hierarchy of paralinguistic and semantic information-from low-level acoustic structure to higher-order features such as sentence meaning, speaker identity, affect, and communicative intent-providing a rich representation of the auditory and semantic content beyond the speech signal itself. We extracted visual features from Qwen 2.5 VL[26] and VideoMAE[27]. Qwen 2.5 VL is trained on large-scale image–text and video–text datasets to optimize object- and region-level grounding, temporal reasoning, and action recognition across short and long timescales, and multimodal instruction-following tasks including captioning, question answering, and event localization whereas Videomae is trained for fine-grained short term temporal understanding. Hence, the collected features encode a broad spectrum of visual information-from static object and scene semantics to dynamic motions and event structure-capturing both momentary visual details and temporally extended context relevant for interpreting naturalistic audiovisual stimuli. Note that the audio features are derived solely from the audio stream, whereas the visual features are computed solely from the video frames. Together, the audio and visual features can predict a large swathe of cortex with robust prediction performance (see Figure 2c and Extended Figure E2 for the components of the explained variance).

We introduce two complementary frameworks to track the temporal dynamics of modality engagement across the cortex: the Dominance Indication Paradigm (DIP) and the Performance Indication Paradigm (PIP). DIP marks timepoints where prediction accuracy on held-out data is substantially better accounted for by the audio-based model than by the video-based model (or vice versa), evaluated independently for each region, thereby identifying regions that alternate their dominant sensory modality over time (Figure 2b). PIP, in contrast, identifies time windows where either model robustly predicts the neural response—and particularly where both do so concurrently—computed separately for each region to reveal those regions that encode information from both modalities in parallel (Figure 2b). We further developed a specifically modified transformer model that predicts whole-brain fMRI responses while dynamically weighting audio and visual inputs for each cortical region, providing an automated means to analyze cortical regions by their modality-specific tuning profiles over time.

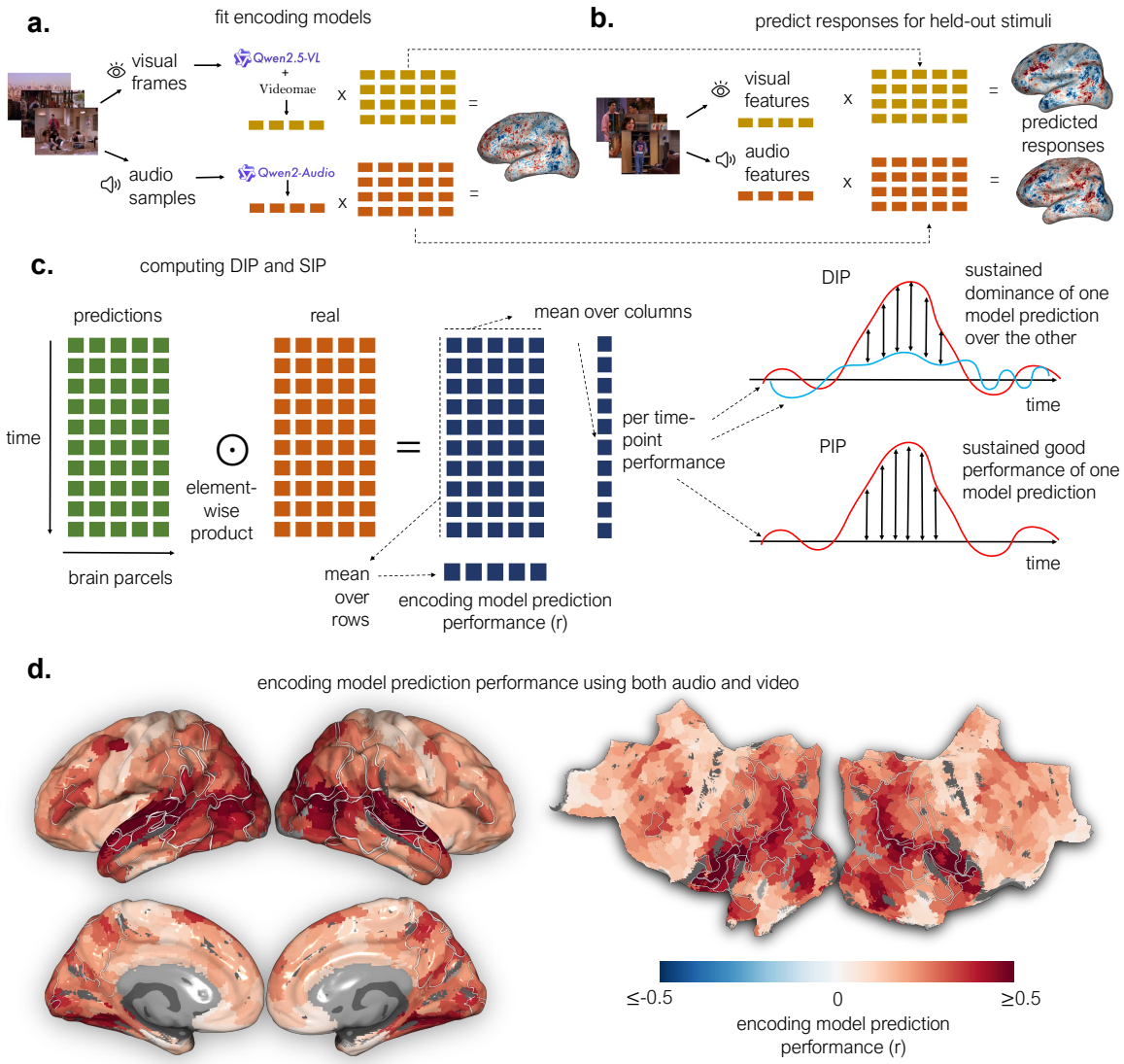


Figure 2: Overview of encoding model framework and computational paradigms. (a) We fit separate encoding models using features derived from the auditory stream (Qwen2-Audio) and the visual stream (VideoMAE and Qwen2.5-VL). The auditory model captures both acoustic and paralinguistic/language-related structure rather than only low-level sound features, while the visual model encodes spatiotemporal information and actions from video frames. Ridge regression was used to learn parcel-wise linear mappings between each feature space and measured BOLD responses (averaged across participants). (b) The trained models were then applied to predict BOLD activity for held-out movie segments, yielding time-resolved predictions from the visual and auditory models that could be compared across cortical regions. (c) We introduce two complementary data-driven approaches —the Dominance Indication Paradigm (DIP) and Performance Indication Paradigm (PIP)—to characterize temporal dynamics of modality-specific and shared encoding. Instead of computing model performance across all timepoints for every parcel/voxel as in traditional encoding analyses, we quantified prediction performance at each timepoint within each region by taking the element-wise product between model predictions and observed responses, then averaging across parcels in that region. DIP captures periods of sustained dominance by one modality's model over the other, whereas PIP captures periods of jointly strong performance by both models. (d) Parcel-wise prediction accuracy (Pearson correlation, r) of a concatenated audiovisual model on held-out stimuli. Both inflated (left) and flattened (left) cortical surfaces reveal strong prediction performance along posterior visual areas, extending anteriorly into lateral, ventral, dorsal and medial regions, spanning the temporal cortex and parts of the frontal cortex (see region names in Extended Figure E1). These widespread high-performing zones reflect the robust encoding of visual and auditory–linguistic features across large portions of cortex.

Results

Two Circular Bands of Regions Around Category-Selective Visual Cortex and Lateral Frontal Cortex Show Remarkable Switching

To operationalize DIP, we evaluated separately trained audio- and video-based encoding models on held-out movie data, quantifying differences in their prediction accuracy at each region of interest (ROI, defined using the Harvard–Oxford cortical atlas [28]) and time point (Figure 2a-b). Reliable differences between the time point performance of models should be large enough and sustained for a long enough period of time. Instead of picking these parameters arbitrarily, we adapted the cluster free cluster enhancement method (TFCE) [29]. We identified TRs (fMRI time unit) for which the observed TFCE-transformed difference in model performance exceeded the maximum TFCE-transformed difference in the null distribution in at least 80% of the block-wise permutations (details in Methods). By conducting this analysis in both directions - testing when the audio-based model outperformed the video-based model and vice versa - we isolated timepoints where one modality's model had a reliable advantage in predictive performance over the other, rather than merely a marginal difference in fit. This yielded, for each ROI, temporal segments of audio- or video-predictive dominance, providing a fine-grained measure of when and where cortical responses were better accounted for by one modality over the other.

Aggregating, for each region, the total duration of audio-dominant segments (D_{audio}) and the total duration of video-dominant segments (D_{video}) across all sessions revealed a striking organization of "switching" regions forming two distinct bow-like bands (Figure 3a): a posterior bow encircling visual cortex—including the medial precuneus, superior parietal lobule, posterior angular gyrus, anterior lateral occipital cortex, and temporo-occipital divisions of the middle and inferior temporal gyri—and a smaller anterior bow spanning dorso-lateral and dorso-medial frontal association areas, including parts of the superior, middle, and inferior frontal gyri. To determine which regions significantly alternated between modalities, we defined our switching index as $S = \text{minimum}(D_{\text{audio}}, D_{\text{video}})$, which allows us to identify regions that both show switching between the two modalities and switch for a large number of time points. We conducted exhaustive pairwise bootstrap tests across ROIs with reliable data (with average explainable variance > 0.1, with more details in Methods). For every ordered region pair, we evaluated the hypothesis that one region's switching index (S_i) exceeded that of the other (S_j) (null: $S_i \leq S_j$; alternative: $S_i > S_j$) using bootstrap sampling across fMRI runs. Benjamini-Hochberg false discovery rate correction[30] was applied to all resulting p-values. Regions that were significantly higher than others and were never significantly outperformed by any other (corrected $p < 0.05$) were defined as switching leaders (Figure 3b and Figure 3a). Thirteen such leaders emerged, distributed across the

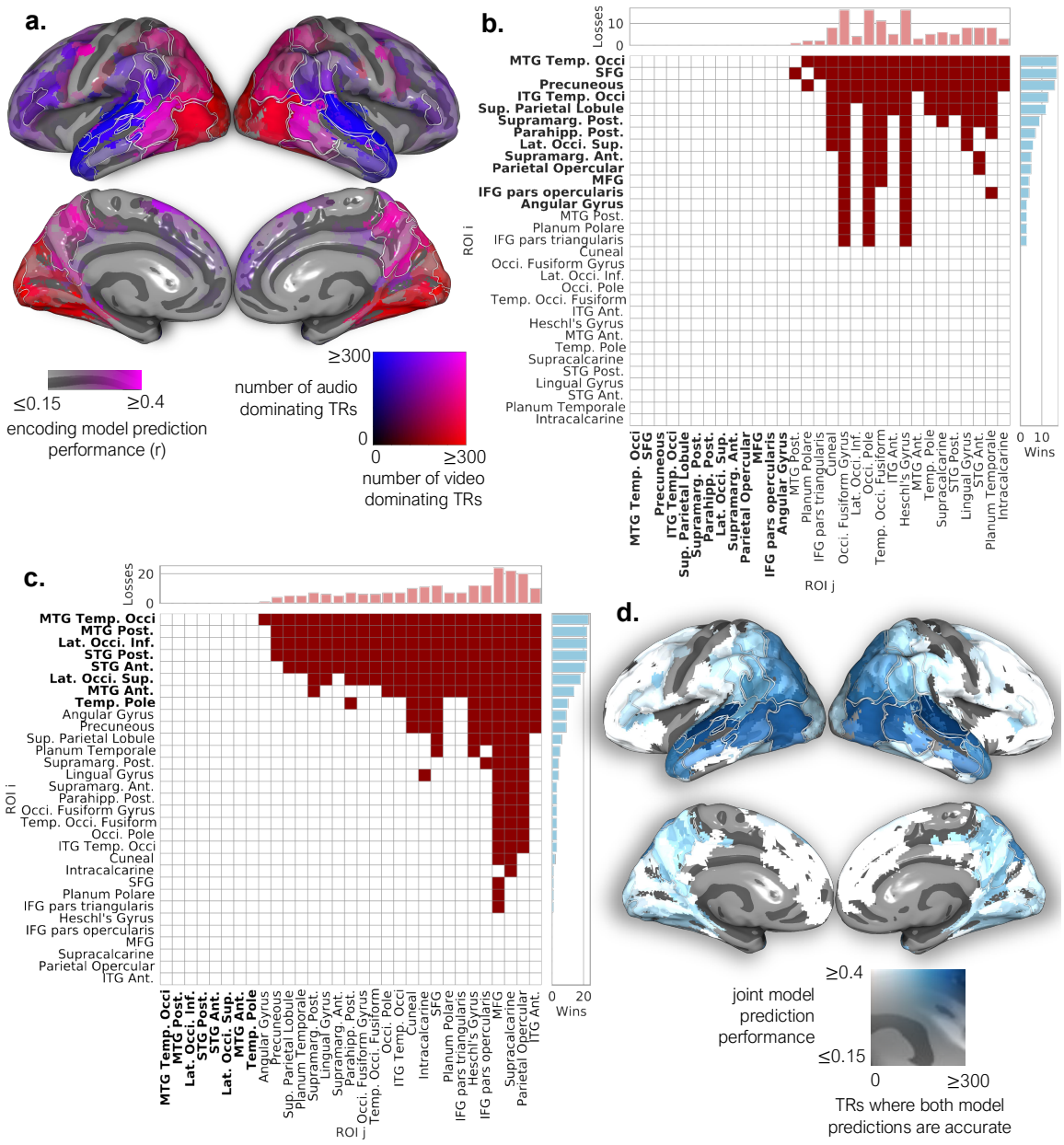


Figure 3: Cortical organization of modality switching and jointly-predicted dynamics. (a) Spatial distribution of the total duration of audio- and video-dominant time segments from the DIP analysis. For each cortical parcel, red intensity reflects the total duration of video-dominant segments, blue intensity reflects audio-dominant segments, and pink/purple intensity reflects parcels with substantial durations for both modalities. The encoding model prediction accuracy (Pearson r) from the concatenated audiovisual embedding model is used as a transparency layer to focus on well predicted regions. Two bands of regions, one surrounding the visual cortex and one in the frontal cortex, appear to show notable switching. (b) Results of pairwise bootstrap comparisons of the switching index. Each filled cell represents a win for region i (rows) over region j (columns), if region i exhibits a significantly higher switching index than region j ($p < 0.05$ corrected). Regions that never lose to any other region are bolded as switching leaders.(c) Pairwise bootstrap comparison of the jointly-predicted index, computed analogously to (b). Each filled cell marks a win for region i over region j, meaning region i exhibits a significantly greater proportion of jointly-predicted timepoints after Benjamini–Hochberg correction. Bolded regions denote jointly-predicted leaders. (d) Spatial distribution of the total duration of jointly-predicted time segments from PIP. Higher blue intensity indicates parcels with longer periods where both audio and video-based model predictions are high, while the concatenated-model prediction accuracy lightens regions with low overall fit. Regions extending from the lateral occipital to the superior and middle temporal gyri are the most jointly-predicted.

posterior and anterior bows, jointly delineating two hubs of dynamic modality reweighting across the cortex. Notably, several of these regions - including the precuneus, posterior middle and inferior temporal gyri, angular gyrus, and lateral prefrontal cortex - overlap with areas previously implicated in amodal conceptual representation[16]. Whereas prior work demonstrated that these regions encode object meaning invariant to what modality it was presented in, our findings suggest that such supramodal representations may arise from a dynamic process of contextual weighting, where the same regions transiently prioritize one modality to maintain stable understanding across sensory formats.

An Axis of Regions spanning the occipito-temporal cortex shows remarkable joint prediction

To operationalize the performance indication framework(PIP), we again evaluated audio- and video-based encoding models on held-out movie data, this time assessing when each model independently achieved robust prediction of brain responses (Figure 2). For each ROI, we identified TRs where the TFCE-transformed performance of each model exceeded the maximum of the null distribution in at least 80% of block-wise permutations (details in Methods) and then determined their temporal overlap. These overlapping segments marked periods when ROI responses were robustly explained by both auditory and visual features. For every ROI, we used as a metric for joint predictivity the total duration B of such jointly predicted segments across all sessions, yielding a measure of each ROI's overall tendency to jointly encode auditory and visual information over time. This procedure revealed a continuous axis of jointly predicted regions extending from the lateral occipital cortex to the temporo-occipital, posterior, and anterior middle temporal gyrus, encompassing the superior temporal gyrus (anterior and posterior) and reaching the temporal pole (Figure 3d). We conducted an exhaustive pairwise bootstrap tests across ROIs with average explainable variance > 0.01, following the same resampling procedure used for modality switching. For each region pair, we estimated the probability that one region's joint-predictivity index (B_i) exceeded that of the other (B_j), and corrected all p-values using the Benjamini–Hochberg procedure. Regions that outperformed others and were never significantly outperformed by any other (corrected $p < 0.05$) were defined as jointly-predicted leaders (Figure 3c). Eight such leaders emerged, forming a continuous axis that links lateral occipital cortex to temporal cortex — a configuration that closely resembles an "arrow" of audiovisual engagement that intersects the posterior switching bow at temporo-occipital MTG (Figure 4a). Prior work has repeatedly identified the superior temporal sulcus (STS)[31, 32, 19] as a locus of audiovisual integration (e.g., supradditive AV responses). Although our atlas does not have a separate ROI for the STS, we observe the adjacent superior temporal gyrus parcels to be robustly predicted by both auditory and visual features, consistent with

integration-related processing in nearby STS-adjacent cortex.

Variance decomposition distinguishes switching and jointly predicted ROIs

We compared the variance components of switching and jointly predicted leader regions during audio- and video-dominated segments identified via DIP. For each ROI, we tested whether the dominant-modality unique variance (audio-unique during audio-dominated TRs; video-unique during video-dominated TRs) exceeded both the non-dominant unique component and the shared component. Four paired Wilcoxon tests were within each ROI (dominant-unique > non-dominant-unique; dominant-unique > shared, for both audio-dom and video-dom segments), and all p-values were FDR-corrected across all tests and ROIs ($4 \times$ N tests total).

Switching leaders exhibited significantly greater unique variance for the dominant modality (audio-unique during audio-dominated clips and video-unique during video-dominated clips) compared with both non-dominant unique and shared components ($p < 0.05$ corrected). This pattern extended across almost all switching regions, up to the lateral occipital cortex superior, whereas most jointly predicted leaders showed no statistically significant differences between dominant-modality unique and other variance components (Figure 4b). These results indicate that during modality-dominated contexts, switching regions rely more heavily on features of the dominant modality, whereas jointly predicted regions maintain a more balanced reliance across modalities. When extending this analysis to include all canonical visual and language regions alongside switching and jointly predicted ROIs, we found that virtually all regions showing significant dominant-unique variance were switching regions, with only two high-level visual areas also exhibiting this pattern (Extended Figure E3).

Variance decomposition reveals sensitivity to modality specific and shared structure in jointly-predicted ROIs

Variance decomposition in jointly-predicted ROI leaders during jointly-predicted TRs revealed that shared variance was significantly higher than the audio- or video-unique components for some of the regions. During the rest of the TRs, the shared variance was not greater than the other components. One potential explanation for this result is that jointly predicted time-points correspond to moments when auditory and visual feature spaces are possibly more correlated (Figure 4c). However, this explanation is not sufficient, as it predicts that all highly predictable language and visual regions—including Heschl's gyrus, occipital fusiform gyrus, occipital pole, etc would emerge as jointly-predicted leaders, which they do not. Instead, we observe that many jointly predicted regions also have a robust

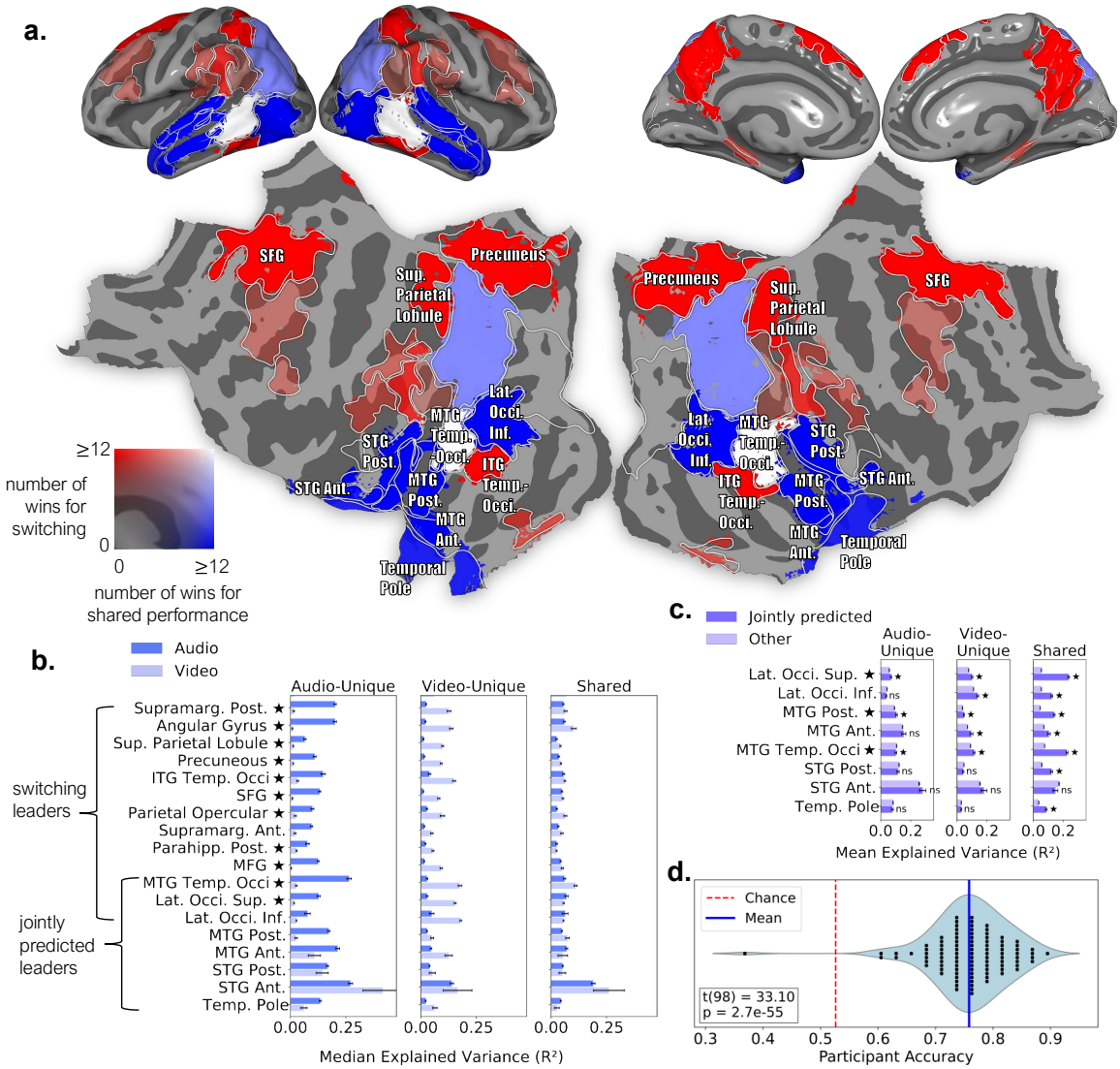


Figure 4: Spatial motifs and variance structure of switching and jointly-predicted cortical regions. (a) Cortical distribution of *switching* and *jointly-predicted leaders*, shown by the number of significant wins each region achieved in pairwise bootstrap tests of the switching index (red) and joint-predictivity index (blue). Color intensity indicates the number of wins (see scale); regions high on both metrics appear white. Red regions form two bow-shaped structures - a posterior bow encircling visual cortex, and an anterior bow spanning lateral frontal cortex. Blue regions form an arrow-like axis extending from lateral occipital into the temporal cortex. The MTG temporo-occipital region, appearing white, is both a switching and jointly-predicted leader and lies at the intersection of the bow and arrow structures. **(b)** Variance-decomposition results for switching and jointly-predicted leaders during audio-dominating and video-dominating segments (DIP). Bars show cross-TR mean unique variance (R^2) explained by audio and video features and the shared variance. Asterisks mark regions where the dominant-modality-unique variance significantly exceeds other components (Wilcoxon paired t-test, corrected at $p < 0.05$ across all tests). Regions up to and including lateral occipital superior are switching leaders, whereas the remaining ones are jointly predicted (including MTG temp. occi. and lateral occipital superior). Most switching regions show significant modality-unique dominance, indicating reliance on the prevailing sensory modality, whereas jointly-predicted regions maintain more balanced encoding even during unimodal-dominant segments. **(c)** Variance components for jointly-predicted leaders during jointly-predicted TRs (labeled both) versus the remaining TRs (labeled other). Asterisks next to region names indicate that shared variance significantly exceeds both unique components (Wilcoxon paired t-test, corrected at $p < 0.05$ across all tests), and asterisks on horizontal bars denote that the corresponding variance component during jointly-predicted TRs is significantly greater than that during the rest. All divisions of MTG and Lat. Occi. retain significant modality-unique components, suggesting that even within periods of high audiovisual joint-prediction, these regions encode complementary modality-specific information. **(d)** Participants viewed audio and video dominant segments identified via DIP and indicated whether the audio or the video was contextually more important for understanding the scene's meaning. Human responses matched the DIP-derived dominant modality labels significantly above chance (one-sample t test, $t(98) = 33.10, p = 2.7e^{-55}$).

amount of unique variance explained by audio and video features during jointly predicted TRs, suggesting that these regions are engaged in encoding both shared and unique information from audio and video features, which would correspond to cross-modality integration or consolidation.

Behavioral evidence links contextual modality importance to DIP-derived dominance

To test whether modality-dominated segments identified from switching regions reflect meaningful contextual structure in the stimuli, we conducted a behavioral experiment probing human judgments of modality importance. We first applied DIP to all switching regions treated as a single unified ROI, yielding a global set of audio-dominated and video-dominated segments. Participants ($N = 100$) viewed these same clips and indicated whether the audio or video was more important for understanding the central meaning of each scene. Human choices aligned closely with the DIP-derived labels: participants selected the DIP-identified dominant modality on average for 75.83% of trials, significantly above chance ($t(98) = 33.10, p = 2.7e^{-55}$) (Figure 4d). These results suggest that the modality-dominant segments detected via DIP correspond to scenes in which one modality is contextually prioritized by human observers.

Importantly, this behavioral alignment does not imply that switching arises simply because the non-dominant modality carries no informative content. Early sensory regions such as Heschl's gyrus and occipital pole showed comparable unique and shared variance components across audio- and video-dominated segments derived from switching regions, indicating that both modalities are processed robustly at low levels. The sharp preference of one modality over the other emerged specifically in switching regions, which are located in higher-order association cortices, consistent with a mechanism in which contextual demands, not the absence of information, drive the prioritization of one modality over the other (Extended Figure E4).

A multimodal transformer reveals dynamics of modality weighting

To investigate how cortical regions dynamically reweight sensory information, we developed a multimodal transformer that learns to assign content-dependent weights to auditory and visual features for every region. Transformers have been recently proposed to build encoding models for datasets such as the one used here in the aim of improving prediction performance [33, 34]. Here, we define a targeted transformer architecture that can directly model per-ROI modality weighting by focusing on making the components of the learned model interpretable. Unlike conventional linear models that simply concatenate audio and

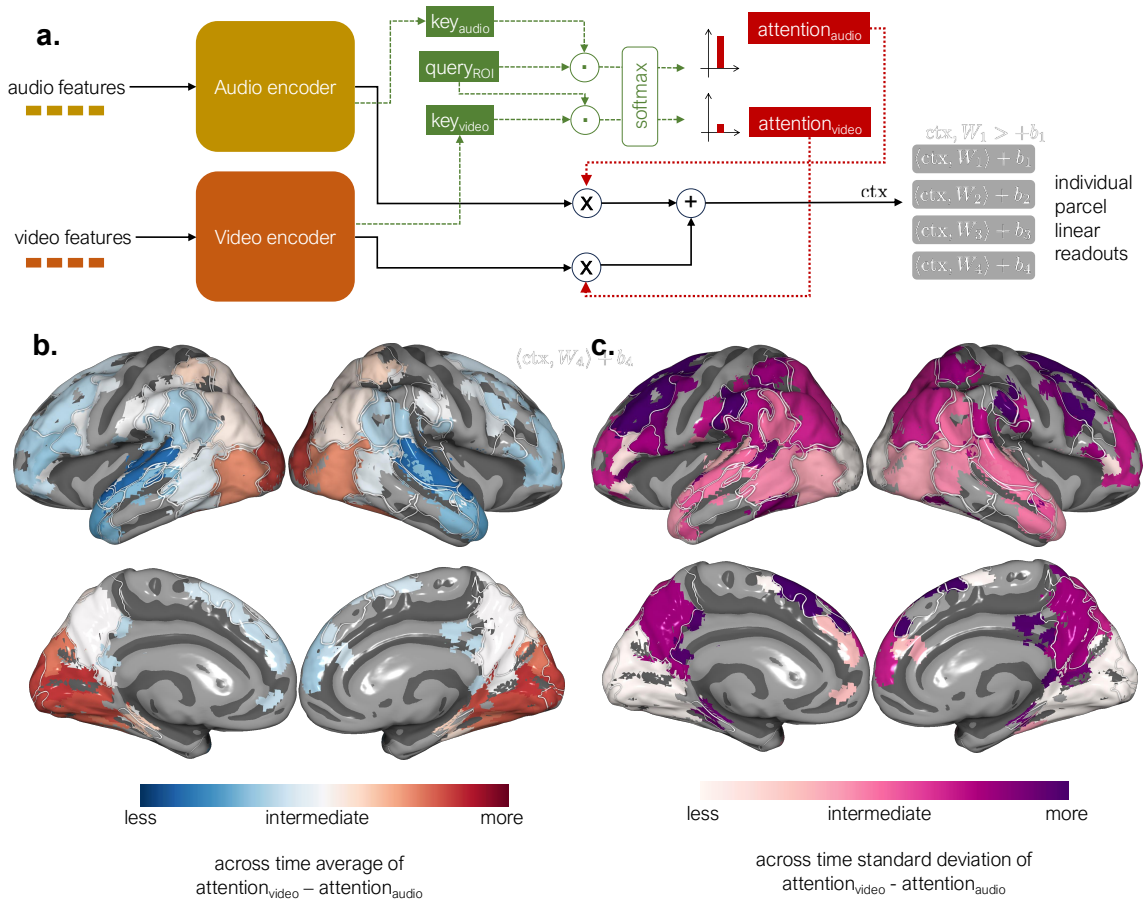


Figure 5: Modeling switching and joint prediction with a transformer. (a) Schematic of the ROI-specific audiovisual transformer used to predict parcel-wise BOLD responses. Audio and video embeddings at each timepoint are passed through modality-specific encoder layers. A query matrix with ROI-specific learnable parameters determines the relative attention assigned to each modality as a function of the current embeddings for a specific region. The resulting context vector—a weighted combination of the audio and video streams—is passed through linear readouts to generate the predicted parcel responses. Thus, the model learns how much weight to assign to audio versus video features independently for each ROI and timepoint. **(b)** Mean difference across time between attention weights for video and audio features ($\text{attention}_{\text{video}} - \text{attention}_{\text{audio}}$) for every parcel. Positive values (red) indicate regions where the transformer consistently assigns a higher weight to video inputs, primarily around occipital lobe, both laterally and medially. Negative values (blue) indicate higher weights for audio inputs, concentrated in superior and anterior temporal areas. Regions with intermediate mean differences can reflect two distinct cases: (1) the model alternates between assigning high weight to one modality and then the other across time, yielding an average near zero but reflecting temporal switching, or (2) the model maintains a balanced weighting of both modalities throughout, indicating sustained joint attention. **(c)** Temporal standard deviation of the attention-weight difference ($\text{attention}_{\text{video}} - \text{attention}_{\text{audio}}$). High variance (magenta) indicates strong temporal fluctuations in relative attention, often co-localized with switching regions in frontal cortex, precuneus, and parietal areas. Low standard deviation (light pink) indicates stable weighting of the two modalities across time, characteristic of unimodal regions focusing on one modality and jointly-attentive regions that maintain a consistent balance between auditory and visual information. Our transformer independently recovers that most switching regions show high variability and near-zero mean differences, whereas jointly-predicted regions show low variability and stable, balanced attention profiles.

video embeddings—treating each modality's contribution as static across time—our architecture uses a transformer attention mechanism (distinct from the neuroscience definition of attention) that adaptively weights the two modalities for every region at every time point when predicting cortical activity (Figure 5a). Specifically, for each ROI, embeddings from the audio and video streams are passed through a weighting mechanism that computes an attention weight for audio and an attention weight for video at each time point as a function of the modality-specific embeddings at that time point (note this makes the combination of information from the two feature streams a convex combination, where both attention weights are between 0 and 1 and sum to 1). Thus, the attention mass varies dynamically over time but is common to all parcels in a given ROI, while each parcel retains its own linear readout for response prediction. This design enables the model to capture temporally varying modality weighting that mirrors cortical dynamics during naturalistic perception. This model performs similarly to the model with combined audiovisual embedding (Extended Figure E5). To characterize these weighting patterns, we computed for each ROI the difference between the attention weight for video and audio features for each timepoints,

$$\Delta\text{attention}_t = \text{attention}_{\text{video}}(t) - \text{attention}_{\text{audio}}(t),$$

where t indexes timepoints. We examined the mean and standard deviation of $\Delta\text{attention}_t$ over time for every ROI. Strikingly, the transformer independently recovered our previous results. Despite no explicit priors or anatomical constraints guiding the model, the transformer spontaneously discovered the brain's canonical modality structure—assigning higher video attention weight to occipital regions and higher audio attention weight to temporal-language areas. Most regions previously identified as switching or jointly predicted showed a mean $\Delta\text{attention}_t$ close to the middle of the ROI distribution (Figure 5b), indicating relatively equal average attention weights on audio and video across time. Switching leaders showed a large standard deviation in $\Delta\text{attention}_t$, indicating fluctuations between times with high attention weights on audio features and times with high attention weights on video features. Jointly predicted regions showed a small standard deviation in $\Delta\text{attention}_t$, consistent with them jointly focusing on features from both modalities (Figure 5c).

Discussion

Our analyses uncovered two complementary forms of cortical multimodality that dynamically balance sensory evidence during naturalistic perception. Using the Dominance Indication Paradigm (DIP) and Performance Indication Paradigm (PIP), we found two circular belts of regions that alternated their dominant sensory modality over time, and a continuous occipito-temporal axis that jointly tracked both modalities, together forming a bow-and-

arrow-like pattern. These findings reveal that multimodal processing is not monolithic but composed of two interacting systems that are also spatially separated: one that flexibly switches between modalities depending on context, and another that integrates shared audiovisual structure to maintain perceptual coherence.

The posterior switching bow encompasses regions including the precuneus, angular gyrus, and posterior middle and inferior temporal gyri—areas traditionally linked to amodal or conceptual representation [15, 16]. Their fluctuating dominance suggests that such territories are not permanently supramodal but dynamically prioritize whichever modality carries more contextually relevant or reliable information. A smaller anterior bow within the lateral prefrontal cortex exhibits the same pattern, implying that dynamic modality reweighting also extends into higher-order control regions. By contrast, jointly predicted regions form a continuous occipito-temporal arrow spanning the lateral occipital gyri and the superior and middle temporal gyri and reaching the temporal pole, consistent with the long-recognized role of superior temporal sulcus–adjacent cortex in audiovisual integration [19, 31, 32].

The spatial arrangement of the posterior bow closely parallels the organization reported in Popham *et al.* [17], which found that each category-selective patch along the anterior border of visual cortex has an immediately anterior region selective for the same semantic category in language. The authors proposed that this border acts as a convergence zone where visual semantic information enters an amodal system through parallel, category-aligned pathways. Our results may represent a dynamic counterpart of this motif, where each switching region might encode specific semantic categories, but the origin of what is being encoded might alternate between the audio and the video streams.

What allows the switching regions to determine which modality to encode? Researchers interested in the role of attention in audiovisual integration have studied the role of bottom-up effects, such as stimulus saliency, or top-down effects, such as task demands and expectations [35]. This line of inquiry usually investigates how the brain allocates gain to sensory sources (e.g. the sight and sound of the same object, or the speech of a person speaking and their mouth movements) on short time scales, to increase fusion between modalities or reduce interference. In contrast, our approach reveals modality-dominant scenes on the order of tens of seconds, and the observed switching pattern is likely driven by high-level processing such as understanding the ongoing narrative. Indeed, work on event segmentation during narrative comprehension [36] shows that high-level regions such as precuneus and angular gyrus shift their latent event model only every tens of seconds, suggesting that these regions maintain long-duration situation models rather than rapidly changing stimulus features, though this hypothesis should be further tested in future work. While some existing work indicates that frontal regions are modulated by overt attention tasks

during narratives [37], other work shows that our more posterior switching ROIs also switch their tuning towards an explicitly attended category, such as the precuneus or the posterior parahippocampal gyrus [38]. Our switching ROIs also aligns considerably with the default mode network (DMN) [39] which is classically considered to be a region that is deactivated during tasks, but has recently been shown to be involved in complex tasks that require sustained representations over multiple seconds such as narratives [7, 40], whether these tasks are internally or externally driven, as argued in De Soares *et al.* [37]. Based on this hypothesis, in our behavioral experiment, we asked participants to label the audio and video dominating scenes based on the importance of audio or video information to understand the global meaning of the scenes (see Methods). As we mentioned above, the participant's answers matched our labels 75.83% of the time. Interestingly, the responses of the participants had a lot of agreement (see Extended Figure E6), with many scenes having a matching close to 100% between the behavioral ratings and our labels. A few scenes had a strikingly low matching rate, with most participants indicating the opposite modality as the one chosen by our approach, hinting that for these scenes, the dominance might not have been driven by relevance to high-level meaning. For these scenes, future work can evaluate if the reason for the dominance of a modality may be due instead to bottom-up effects, such as the saliency of information in that modality.

We also argue that switching is not optional or trivial; it is computationally and pragmatically necessary for understanding natural audiovisual narratives where information content is asymmetrically distributed across modalities over extended timescales. If cortex contained only (i) strictly unimodal sensory regions and (ii) strictly "always-joint" multimodal regions, then during narrative moments when one modality carries the more contextually relevant information, the system would be forced either to (a) always combine both modalities equally (suboptimal when one modality is misleading/irrelevant), or (b) rely solely on upstream unimodal pathways without a mechanism for higher-order regions to reweight which modality they privilege. Our contribution is to isolate and map that mechanism empirically, revealing for the first time two contiguous "bows" of such dynamic switching alongside a distinct axis of jointly predicted regions which can be understood under the more classical notion of multi-modal integration.

What semantic information do jointly predicted regions encode? Recent work using visual videos (with no audio) has found evidence that social features from the video stream predict the STS [41]. Social interaction information might thus be a part of the video information that predicts unique activity in our jointly predictive superior temporal regions. A more recent experiment using audiovisual stimuli has shown additional evidence that social information in video features is processed in the STS even in the presence of an audio stream [42]. Future work can focus on identifying what other types of information

are provided by each modality in different regions.

Several methodological and interpretive limitations warrant consideration. First, although the audio and video models used to extract embeddings have demonstrated strong performance on diverse benchmarks, we cannot fully characterize what features they encode. Their internal representations likely conflate low-level sensory features with higher-order semantics, making it difficult to pinpoint the precise information driving each region's predictivity. Nonetheless, the models' robust performance across speech, prosody, and visual-scene understanding tasks indicates that their embeddings capture ecologically relevant structure sufficient for brain-level mapping. Second, spatial localization is constrained by the Harvard–Oxford cortical atlas [28], which is both probabilistic and relatively coarse: each parcel's regional assignment reflects the probability of belonging to a region across subjects, rather than a definitive boundary, and parcel units are big and hence could sometimes exist within two or three regions sharing a boundary. As a result, ROI borders can overlap or blur fine-scale functional distinctions, particularly near the superior temporal sulcus and temporal pole. Notably, when we repeat our analyses at the parcel level, we obtain a similar general pattern of switching regions (Extended Figure E7), indicating that our results are not an artifact of the chosen atlas. Higher-resolution, subject-specific surface parcellations would further mitigate this limitation and better separate interdigitated subregions.

Our multimodal transformer, by allowing each ROI to share a time-varying attention weight across all its parcels, captured region-specific dynamics of modality weighting: switching regions showed high temporal variance and near-zero mean in attention differences, while unimodal cortices exhibited stable positive or negative mean in attention differences. This convergence between model-inferred weighting and empirical dominance patterns adds more evidence in favor of the switching and jointly predicted regions. Together, our findings suggest that the human brain sustains comprehension through two intertwined strategies: one that flexibly shifts between sensory channels and another that maintains persistent multimodal engagement. The result is a brain architecture finely tuned to extract meaning from complex, ever-changing natural audiovisual experience.

Methods

Feature Collection

For both the Qwen2-Audio and Qwen2.5-VL models (open-sourced on HuggingFace[43]), we extracted contextualized hidden-state representations for short audiovisual segments. Each stimulus corresponded to a $6 \times$ TR window (audio waveform for the audio model; silent video clip for the visual model, 1TR = 1.49s). The hidden-state vectors from the final layer of the language model backbone were averaged across all tokens within each window

to yield a single feature vector per segment. For the video model, we passed the stimuli at 15 frames/second rate.

For the video model, we used the following instruction prompt: *"You are a helpful, smart assistant. You have watched many famous movies and sitcoms throughout your training process that feature the day-to-day actions and events of the main characters. You are now watching a clip from such a movie. Carefully comprehend all the events in the video and how the scenes progress, including the important actions, emotional reactions, and changes in setting."*

For the audio model, the prompt was: `<|im_start|>system\nYou are a helpful assistant.<|im_end|>\n<|im_start|>user\nAudio 1:<|audio_bos|><|AUDIO|><|audio_eos|>\nYou have listened to many famous movies and sitcoms throughout your training process that feature the day-to-day conversations and events of the main characters. You are now listening to such a clip from a movie. Carefully listen to what events are happening in the audio, what the characters are talking about, and their intentions.<|im_end|>\n<|im_start|>assistant\n`

Because each feature segment covers the preceding $6 \times$ TR interval, audiovisual features are aligned to the fMRI data beginning at the 6th TR. Accordingly, we used only fMRI volumes from that point onward for model fitting and testing.

For VideoMAE feature extraction, each $4 \times$ TR segment was uniformly sampled at 16 frames, which were passed through the pretrained model. We collected the hidden states from the final transformer layer and averaged them across the sampled frames. The first two feature segments were discarded to maintain temporal alignment with the other feature sets.

Explainable Variance Computation

To estimate the reliability of fMRI responses across subjects and thereby quantify the explainable variance of each parcel, we computed a noise-ceiling measure based on repeated presentations of the same stimuli [44]. For each fMRI session, data from four participants viewing identical segments were treated as repeated measurements. Let $X \in \mathbb{R}^{R \times T \times V}$ denote the data matrix, where R is the number of repeats (subjects), T the number of time points, and V the number of parcels or voxels. Each time course was z-scored along the temporal dimension to remove baseline differences across repeats. For each parcel v, the residual variance across repeats was computed as

$$\text{residvar}_v = \frac{1}{R} \sum_{r=1}^{R} \text{Var}_t(X_{r,t,v} - \bar{X}_{\cdot,t,v}),$$

and the explainable variance was then defined as

$$\mathrm{EV}_v = 1 - \mathrm{residvar}_v,$$

corresponding to the proportion of response variance that is reproducible across subjects. A small-sample bias correction was applied based on the number of repeats,

$$\mathrm{EV}_v = \mathrm{EV}_v - \frac{1 - \mathrm{EV}_v}{R - 1},$$

following the same formulation used in prior noise-ceiling analyses. This measure served to identify parcels with a sufficiently reliable signal for downstream analyses.

Dataset and Encoding model fitting

The Algonauts Project-2025 challenge[23] aims to predict human brain responses to complex, naturalistic audiovisual stimuli. Conducted in collaboration with the Courtois Project on Neuronal Modeling (CNeuroMod)[22], it provides the largest publicly available fMRI dataset for modeling neural responses to multimodal movies. We used a subset of the released *FRIENDS* data, which includes the first three seasons, and the four available full-length films—*The Bourne Supremacy*, *The Wolf of Wall Street*, *Hidden Figures*, and the documentary *Life*. The total duration of fMRI data we used was more than 36 hours. While *FRIENDS* is relatively dialogue-heavy and therefore biased toward auditory content, the movies offer a more balanced mix of visual and auditory information: *Bourne Supremacy* and *The Wolf of Wall Street* feature rich visual dynamics, whereas *Hidden Figures* and *Life* emphasize speech, social interaction, and narrative structure.

The fMRI data were acquired from English-speaking participants recorded in a 3T Siemens PrismaFit scanner. Four participants' data was made available in the Algonauts competition. Each Friends episode and movie was divided into fMRI runs of about 12 minutes. Participants watched segments sequentially, though not necessarily in a single session. The data was preprocessed by the Courtois Neuromod group using `fMRIPrep` [45], which included motion correction, slice-timing correction, and co-registration of the functional images of each participant to their T1-weighted anatomical scan. The fMRI time series were resampled to the MNI152NLin2009cAsym and then averaged into 1000 parcels from the Schaeffer atlas [24]. For more information on data preprocessing, please see the Courtois Neuromod documentation website [46]. We further averaged the parcel data across participants.

To ensure a fair evaluation of modality-specific and multimodal effects, we adopted a two-split testing scheme. For each movie, half of the fMRI sessions were included in the

training set together with the *FRIENDS* data, and the remaining sessions were held out for testing. All reported results are based on these held-out movie sessions.

To fit the encoding models, we used the extracted audiovisual features to predict parcel-wise fMRI responses. To account for the hemodynamic lag, we concatenated features from the current TR and the preceding five TRs when predicting the BOLD response at the current time point. The concatenated feature vectors were reduced to 1000 dimensions via principal component analysis (PCA). For audio features, all components were taken from a single model; for video features, we combined two models by reducing each to 500 dimensions and concatenating them to match the 1000-dimensional audio representation. Separate ridge regression models were then trained for the audio and visual features on the training set, and model performance was evaluated on the held-out movie sessions as described above, by computing Pearson correlation across time.

We define our ROIs, we use the Harvard–Oxford cortical atlas [28]. We further compute another performance measure for evaluating the time dynamics of encoding model predictions. We take the dot product of the data and the predicted data across the parcels in an ROI for each time point (see Figure 2b) and divide the product by the number of parcels in that ROI. This yields a measure of model fit for each ROI and each time point.

Dominance Indication Paradigm (DIP)

For each ROI and each fMRI run (around 12 minutes), to quantify modality dominance, we computed the per time point difference in model fit between the video-based and audio-based encoding models. This yielded two directional performance-difference series: one capturing where the video model outperformed the audio model, and the other capturing the reverse pattern. Then for each timepoint and for a range of thresholds $h \in [0.01, 0.3]$ (in increments of 0.01), we adapted the threshold free cluster enhancement (TFCE) method [29] and computed

$$\mathrm{TFCE}(t) = \sum_{k=1}^{K} e(h_k)\, h_k\, \Delta h$$

where $e(h_k)$ denotes the temporal extent—that is, the number of consecutive time points for which the performance difference exceeds threshold h_k, and $\Delta h = 0.01$ in our case. This integral-like operation enhances contiguous segments of strong differences while penalizing isolated fluctuations.

To estimate a null distribution, we computed performance-difference and TFCE scores in the same manner from temporally permuted data. Specifically, the order of stimulus feature segments was randomly shuffled in contiguous blocks of seven TRs before prediction, and the encoding models were then applied to these permuted feature sequences. This blockwise

permutation preserved the temporal autocorrelation while disrupting its true alignment with the brain data. The resulting maximum TFCE scores, one from each permutation, constituted the empirical null distribution against which the observed scores were evaluated. For each TR in the original unpermuted comparison, we compared its TFCE score to the empirical null: if the observed score exceeded the 80th percentile of the maximum score distribution across permutations, that TR was labeled as **audio-dominant** (for the audio − video direction) or **video-dominant** (for the reverse). Formally speaking,

$$F_{\text{perm}}\big(\text{TFCE}(p)\big) = \Pr\Big(\text{TFCE}_{\text{max}}^{(\text{perm})} \leq \text{TFCE}(p)\Big) > 0.8$$

would quality TR as position p to be audio or video dominating, depending on the direction. Here $\text{TFCE}(p)$ is the observed threshold-free cluster enhancement (TFCE) value at position p; $\text{TFCE}_{\text{max}}^{(\text{perm})}$ is a random variable representing the maximum TFCE value obtained under permutation of the data; and $F_{\text{perm}}(\cdot)$ denotes its cumulative distribution function (CDF).

These dominant TRs thus represent moments when one modality uniquely accounts for more variance in the neural signal than the other, beyond chance fluctuations estimated by permutation baselines. We collect dominating TRs from both modalities for each ROI from every fMRI run. For each ROI, we compute the total duration of audio-dominating segments (D_{audio}) and the total duration of video-dominant segments (D_{video}). We define our switching index for that ROI as $S = \text{minimum}(D_{\text{audio}}, D_{\text{video}})$. This allows us to identify TR that switch multiple times and that switch to both modalities.

Performance Indication Paradigm (PIP)

The PIP analysis followed the same TFCE-based framework as the DIP procedure, but instead of performance differences, it operated directly on the time-resolved prediction performance of each modality. For each ROI and run, TFCE scores were computed separately for the audio and video model performance vectors using the same threshold range and scoring parameters described in the previous section. Null distributions were obtained using the identical blockwise permutation procedure applied to the feature space, with models and fMRI responses held fixed.

TRs whose TFCE scores exceeded the 80th percentile of the permutation-derived maximum distribution were labeled as **highly predictable** for that modality. The intersection of highly predictable TRs from the audio and video analyses within each run defined the set of **jointly-predictable TRs**, corresponding to timepoints where both modalities robustly explained variance in the neural response. for each ROI, we compute B, the total length jointly-predictable TRs, which serves as our joint-prediction index.

Pairwise bootstrap testing

Pairwise, one-sided bootstrap tests were used to assess whether the switching index (S) or joint-predictivity index (B) of one region exceeded that of another. For each ordered region pair (i, j), the null hypothesis ($H_0 : M_i \leq M_j$) was evaluated separately for $M \in \{S, B\}$. In each test, the indices M from individual fMRI runs were resampled with replacement 20,000 times. For each bootstrap sample, the sum of the resampled indices was computed for each ROI and compared between pairs of ROIs. This procedure yielded an empirical sampling distribution of pairwise differences ($M_i - M_j$), from which the one-sided p-value was estimated as the proportion of iterations where $M_i - M_j \leq 0$ (alternative: $M_i > M_j$). Exhaustive pairwise comparisons were conducted among all regions with mean explainable variance above 0.1, and all resulting p-values were corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate[30] procedure ($p_{\text{corr}} < 0.05$). Regions that were never significantly outperformed by any other were designated as "leaders" for the corresponding measure (S for switching, B for joint predictivity).

Behavioral Experiment

We ran a behavioral experiment approved by the Carnegie Mellon University Institutional Review Board. We recruited 100 participants from Prolific (US native English speaker pool; 6USD compensation). One participant failed two attention checks and was excluded, leaving 99 participants for analysis. DIP was applied to the switching network considered as a single unified ROI to derive a global set of audio- and video-dominated segments. On the welcome screen, the following instruction was provided - `In this experiment, you'll watch short clips (average length 20s) from four Hollywood movies. Your task is to decide whether the audio (speech and non-speech sounds) or the video (visual content) is more important for understanding the core meaning of the scene. A helpful way to think about this is to ask yourself: Which modality is providing information that feels especially relevant for what happens next in the movie? We will show a few examples before the actual trials to make the idea clearer. This study usually takes around 25 minutes.`

Before the main task, participants completed four instructional examples demonstrating how contextual relevance can make one modality more informative than the other to help them understand the task better. On each trial, participants viewed one of the DIP-identified segments and judged which modality—audio or video—was more important for understanding the core meaning of the scene. Responses were collected via a two-alternative multiple-choice question presented immediately after the clip ended.

Transformer training and evaluation

To capture how each cortical region dynamically attends to auditory and visual information over time, we designed a transformer-based model that learns modality-specific attention weights for every region that vary across time. At each time point, audio and video embedding vectors ($x_a, x_v \in \mathbb{R}^{B\times 1\times 1000}$) were independently processed by audio and video transformer encoders, each comprising two layers ($N_a = N_v = 2$) of multi-head self-attention with four attention heads ($n_{\text{heads}} = 4$) and feedforward sublayers of dimension 256 ($d_{\text{ff}} = 256$), each followed by pre-layer normalization and dropout ($p_{\text{dropout}} = 0.5$). Learned positional embeddings ($\text{pos}_a, \text{pos}_v$) were added to preserve order, after which the encoded representations were concatenated along the modality dimension to yield $h = [h_a; h_v] \in \mathbb{R}^{B\times 2\times 1000}$. Region-of-interest (ROI) level query matrices $Q_{\text{roi}} \in \mathbb{R}^{N_{\text{roi}}\times 1000}$ were used to compute attention weights over the two modalities via

$$\text{attn}_{\text{roi}} = \text{softmax}\left(\frac{1}{\sqrt{1000}} Q_{\text{roi}} h^\top\right),$$

producing **ROI-specific attention distributions that were broadcast to all parcels within each ROI** according to parcel-ROI assignments. For each parcel, a modality-weighted context vector was then computed as

$$\text{ctx} = \sum_{t=1}^{2} \text{attn}[:,:,t] \cdot h[:,t,:],$$

followed by dropout ($p_{\text{context_dropout}} = 0.3$) to prevent overfitting. Parcelwise BOLD predictions were generated through independently learned linear readouts with parameters $W_{\text{read}} \in \mathbb{R}^{N_{\text{parcel}}\times 1000}$ and $b \in \mathbb{R}^{N_{\text{parcel}}}$, optimized jointly with all other model parameters during training. The model was trained to minimize mean squared error between predicted and observed parcel responses using Adam (learning rate 10^{-4}, batch size 400, weight decay 10^{-2}), with early stopping if validation loss failed to improve for 15 consecutive epochs. For *inference* on held-out stimuli, we passed the audio/video embeddings through the trained encoders, computed ROI-level attention weights from the learned query matrix, broadcast them to parcels, and formed a parcel-level representation $z \in \mathbb{R}^{B\times N_{\text{parcel}}\times 1000}$ by the attention-weighted sum over the two modality embeddings. Predictions were then obtained as

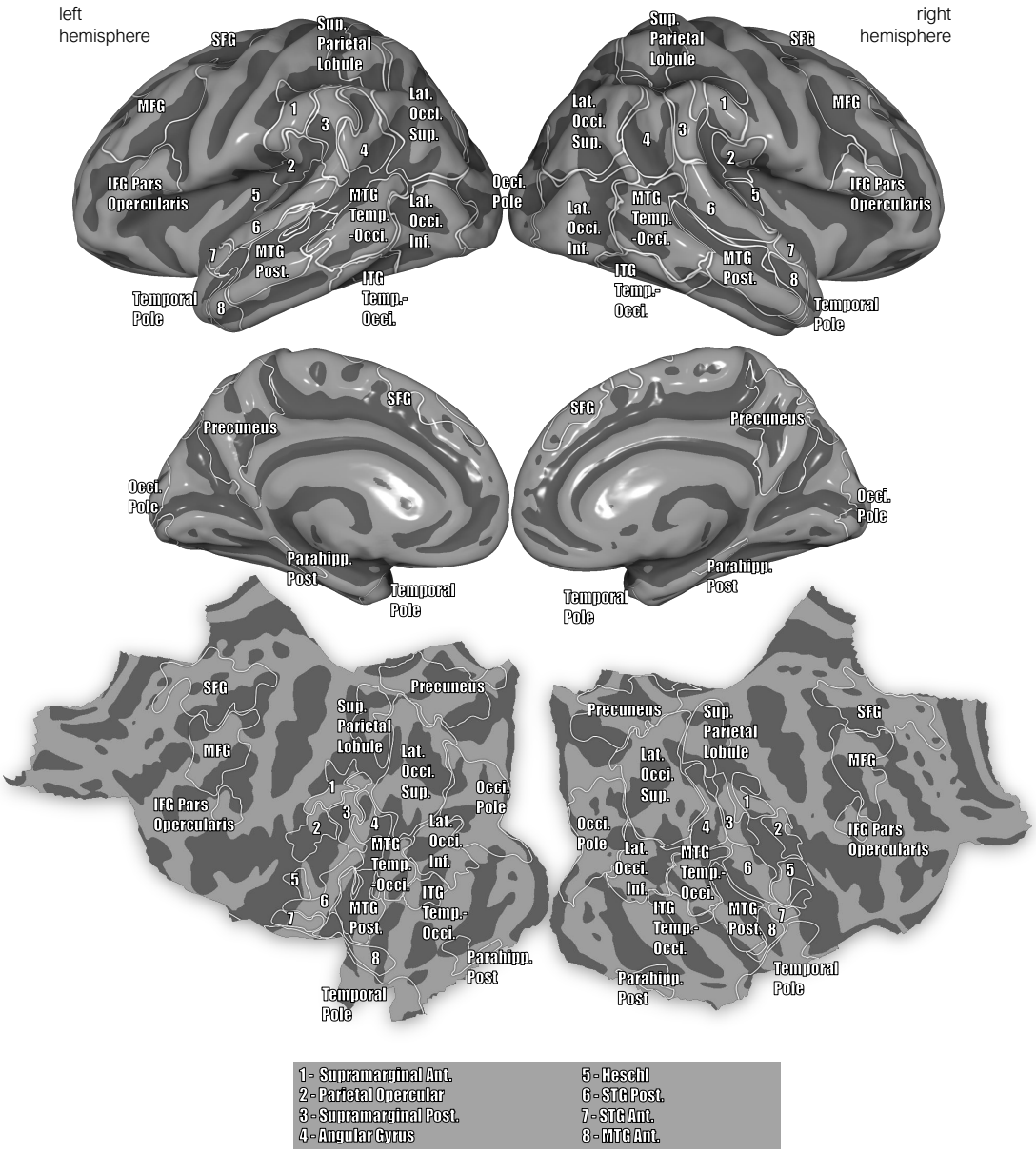
$$\hat{r} = \langle z, W_{\text{read}} \rangle + b,$$

using the trained W_{read} and b. The audio and video attention weights used to compute the predicted responses were taken from this same held-out forward pass and used for the analysis reported in Figure 5.

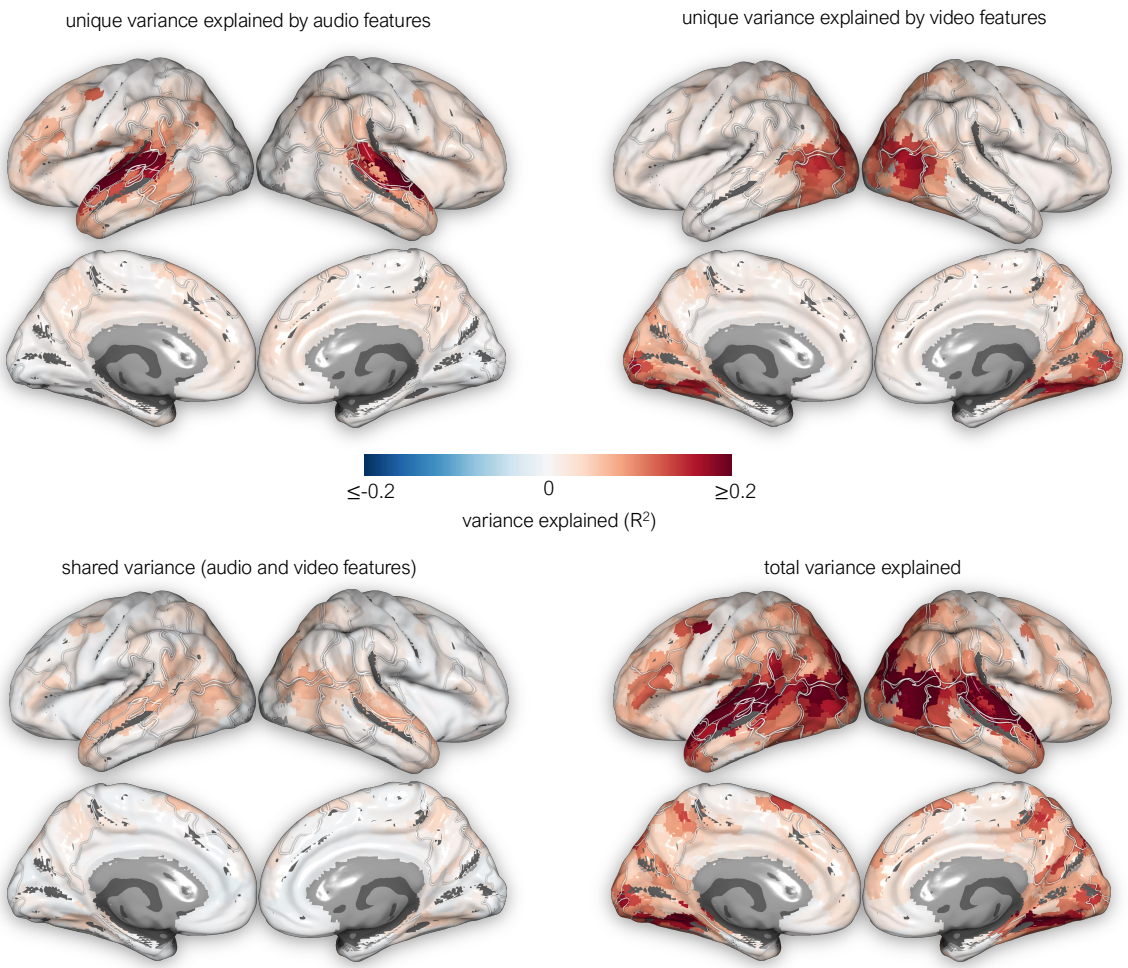
Acknowledgements

This research was supported by NSF CAREER grant 2237064. The authors thank the Courtois Neuromod team for collecting the large amount of fMRI data and the Algonauts team for creating a platform and a competition to make it available. The authors also thank Michael Tarr and Bradford Mahon for scientific discussions of the work, Aaditya Ramdas for insightful discussions regarding the statistical analyses, and Isil Bilgin for assistance with the dataset.

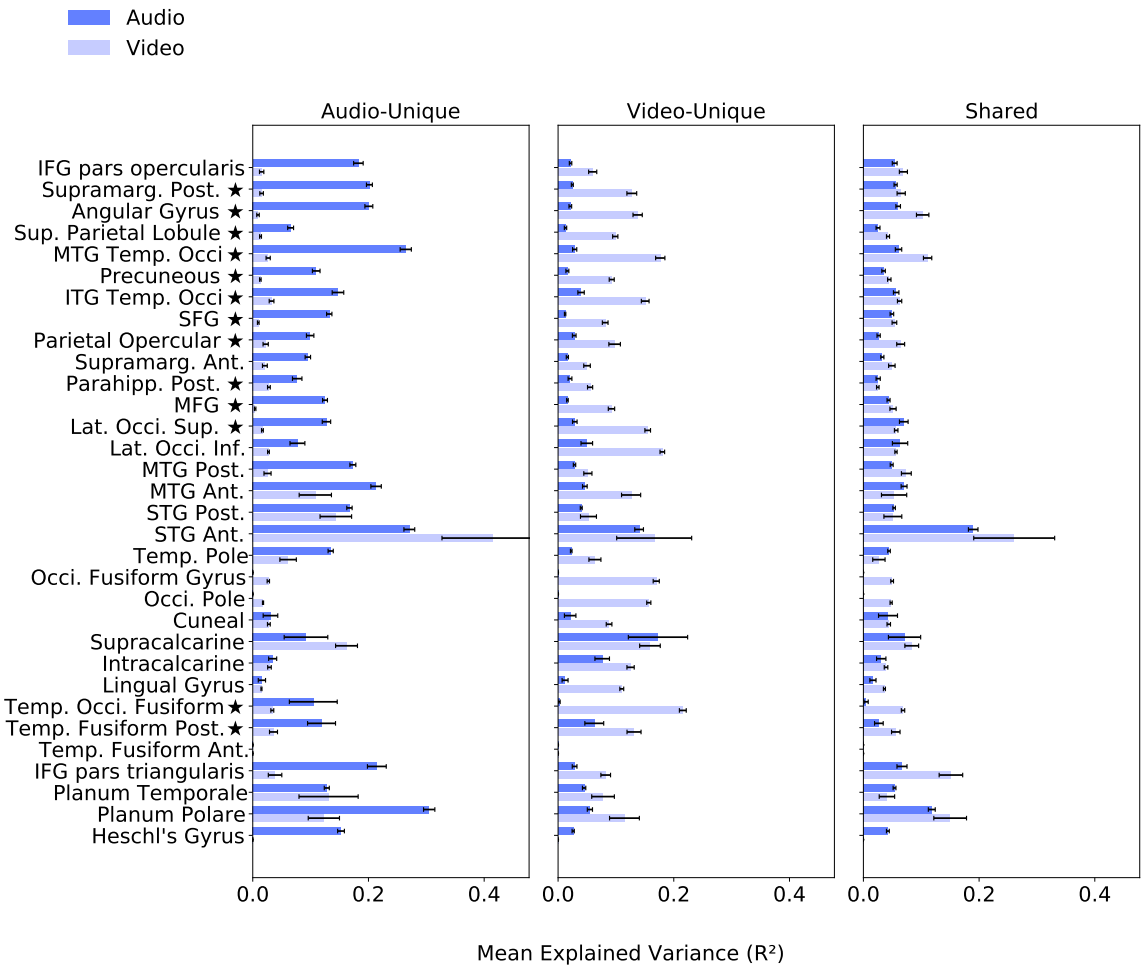
Extended figures



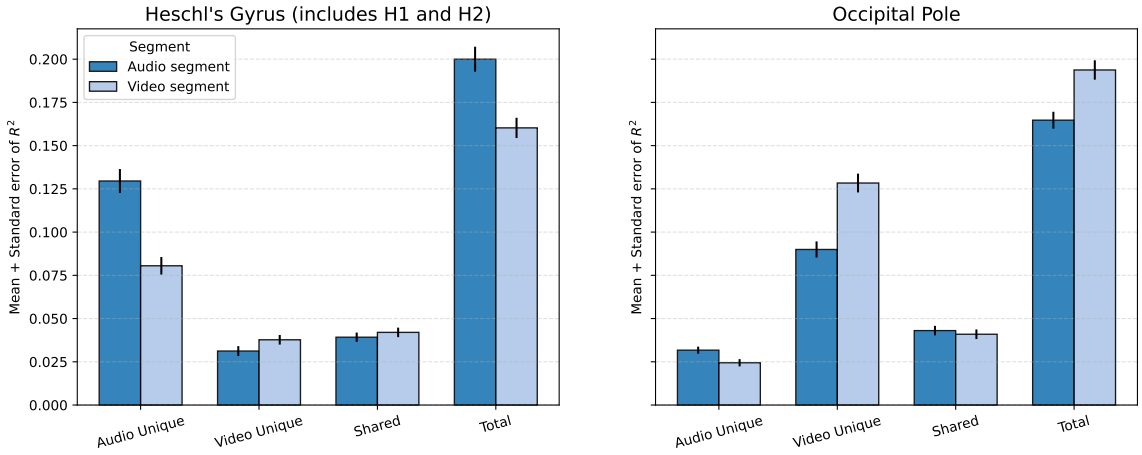
Extended Figure E1: Labeled ROIs with notable prediction performance on the inflated and flattened surface. Abbreviations as follows. MFG: middle frontal gyrus, SFG: superior frontal gyrus, IFG: inferior frontal gyrus, MTG: middle temporal gyrus, STG: superior temporal gyrus, ITG: inferior temporal gyrus, Sup.: superior, Inf.: inferior, Lat.: lateral, Ant.: anterior, Post.:posterior, Occi.: occipital, Temp.: temporal, Parahipp.: parahippocampal.



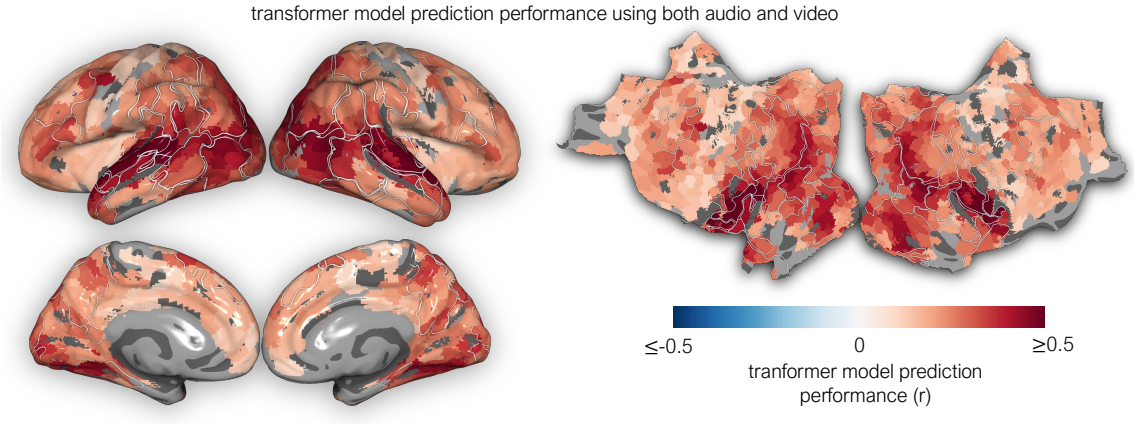
Extended Figure E2: Components of explained variance, showing variance uniquely explained by the audio (top left), uniquely explained by the video (top right), jointly explained by both modalities (bottom left), and the total variance explained (bottom right).



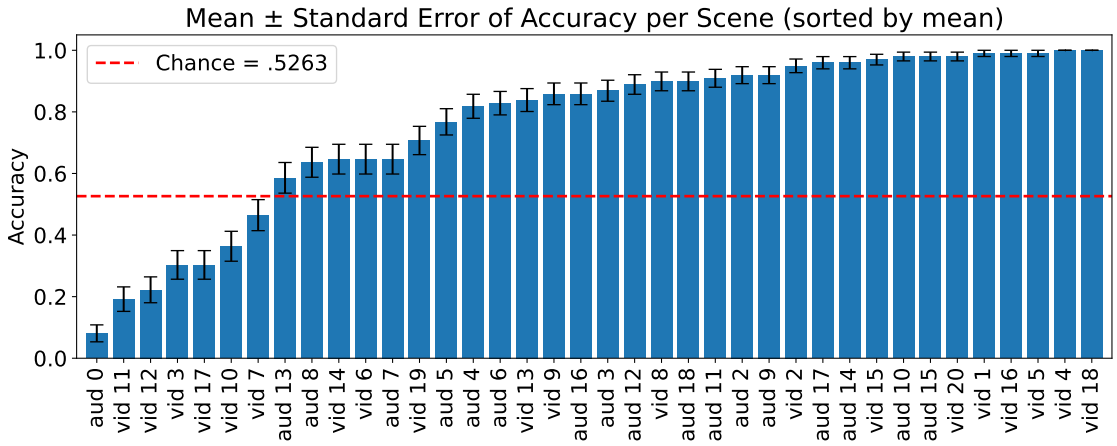
Extended Figure E3: Variance-decomposition results for all ROIs during their respective audio-dominating and video-dominating segments (DIP). Bars show the cross-TR mean unique variance (R^2) explained by audio and video features and the shared variance explained. Asterisks mark regions where the dominant-modality-unique variance significantly exceeds both other components (Wilcoxon paired t-test, corrected at $p < 0.05$ across all tests).



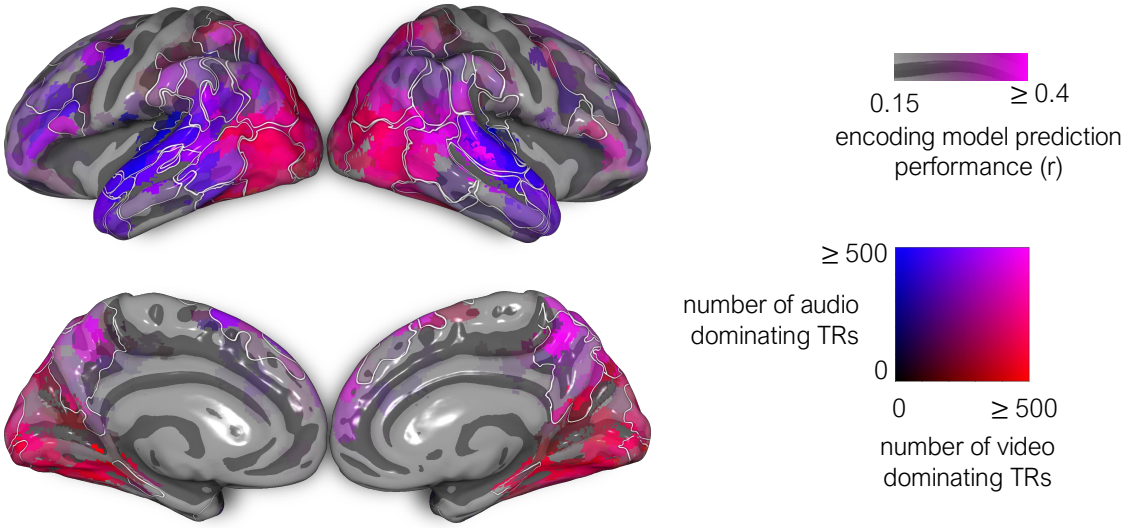
Extended Figure E4: Prediction performance in Heschl's gyrus (primary auditory cortex) and occipital pole (primary visual cortex) during audio dominating and video dominating segments estimated jointly from switching regions. Activity in Heschl's gyrus is predicted by audio features during video dominating segments, and activity in the occipital pole is predicted by video features during audio dominating segments.



Extended Figure E5: Transformer encoding model prediction performance.



Extended Figure E6: Average and standard error of the matching rate between the rating of behavioral subjects for each scene and our labels, ranked in increasing order.



Extended Figure E7: Each parcel is colored according to the total duration of time segments across all episodes where one modality's prediction performance exceeded the other's by ≥ 0.05 for at least 10 consecutive TRs. Redness of a parcel indicates total duration of segments where video-based predictions surpassed audio-based predictions that way; blueness of a parcel indicates the reverse. Thus parcels appearing pink or purple mean they exhibit both trends—reflecting frequent reversals in predicted dominance. Regions identified as "switching" via formal DIP analysis tend to include a mixture of such pink and purple parcels, along with pockets of predominantly audio- or video-biased parcels.

References

[1] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011. Publisher: Elsevier.

[2] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. Publisher: Elsevier.

[3] Emily J Allen, Ghislain St-Yves, ..., Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

[4] Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, pages 1–12, 2023. Publisher: Nature Publishing Group UK London.

[5] Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173, February 2012. Publisher: Elsevier.

[6] Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of neurophysiology*, 112(5):1105–18, September 2014.

[7] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, Jack L Gallant, Wendy a De Heer, Thomas L Griffiths, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. Publisher: Nature Publishing Group.

[8] Sam V Norman-Haignere and Josh H McDermott. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS biology*, 16(12):e2005127–e2005127, December 2018. Publisher: Public Library of Science.

[9] Julie Snowden, Peter J Goulding, and David Neary. Semantic dementia: a form of circumscribed cerebral atrophy. *Behavioural neurology*, 2(3):167–182, 1989.

[10] Elizabeth K Warrington. The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, 27(4):635–657, 1975.

[11] Arnold Wilkins and Morris Moscovitch. Selective impairment of semantic memory after temporal lobectomy. *Neuropsychologia*, 16(1):73–79, 1978.

[12] Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1):42–55, 2017.

[13] Antonio R Damasio. The brain binds entities and events by multiregional activation from convergence zones. *Neural computation*, 1(1):123–132, 1989.

[14] Jon Driver and Toemme Noesselt. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron*, 57(1):11–23, 2008.

[15] Barry J Devereux, Alex Clarke, Andreas Marouchos, and Lorraine K Tyler. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, 33(48):18906–18916, 2013.

[16] Scott L Fairhall and Alfonso Caramazza. Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, 33(25):10552–10558, 2013.

[17] 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.

[18] Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152, 2017.

[19] Michael S Beauchamp, Brenna D Argall, Jerzy Bodurka, Jeff H Duyn, and Alex Martin. Unraveling multisensory integration: patchy organization within human sts multisensory cortex. *Nature neuroscience*, 7(11):1190–1192, 2004.

[20] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56:400–410, 2011. Publisher: Elsevier.

[21] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008. Publisher: Frontiers Research Foundation.

[22] Julie A Boyle, Basile Pinsard, Amal Boukhdhir, Sylvie Belleville, Simona Brambatti, Jeni Chen, Julien Cohen-Adad, André Cyr, P Fuente Rainville, and Pierre Bellec. The courtois project on neuronal modelling-first data release. In *26th annual meeting of the organization for human brain mapping*, 2020.

[23] Alessandro T Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2024.

[24] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex*, 28(9):3095–3114, 2018. Publisher: Oxford University Press.

[25] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[26] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[27] 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.

[28] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[29] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

[30] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

[31] Michael S Beauchamp. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current opinion in neurobiology*, 15(2):145–153, 2005.

[32] Michael S Beauchamp, Kathryn E Lee, Brenna D Argall, and Alex Martin. Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5):809–823, 2004.

[33] Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Predicting brain activity using Transformers. *bioRxiv*, pages 2023–08, 2023. Publisher: Cold Spring Harbor Laboratory.

[34] Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction. *arXiv preprint arXiv:2507.22229*, 2025.

[35] Durk Talsma, Daniel Senkowski, Salvador Soto-Faraco, and Marty G Woldorff. The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences*, 14(9):400–410, 2010. Publisher: Elsevier.

[36] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017. Publisher: Elsevier.

[37] Alexandra De Soares, Tony Kim, Franck Mugisho, Elen Zhu, Allison Lin, Chen Zheng, and Christopher Baldassano. Top-down attention shifts behavioral and neural event boundaries in narratives with overlapping event scripts. *Current Biology*, 34(20):4729–4742, 2024. Publisher: Elsevier.

[38] Tolga Çukur, Shinji Nishimoto, Alexander G. Huth, and Jack L. Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, 2013.

[39] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38(1):433–447, 2015. Publisher: Annual Reviews.

[40] Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1):12141, July 2016.

[41] Emalie McMahon, Michael F Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047, 2023. Publisher: Elsevier.

[42] H. Small, H. Lee Masson, E. Wodka, S. H. Mostofsky, and L. Isik. Ubiquitous cortical sensitivity to visual information during naturalistic, audiovisual movie viewing, 2025. Published: OSF Preprint.

[43] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

[44] Mark D. Lescroart, Dustin E. Stansbury, and Jack L. Gallant. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, Volume 9 - 2015, 2015.

[45] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, and others. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1):111–116, 2019. Publisher: Nature Publishing Group.

[46] Courtois NeuroMod Team. Courtois NeuroMod Documentation, 2022.