Frontal cortex organization supporting audiovisual processing during naturalistic viewing

Faxin Zhou[1*], Amirhossein Khalilian-Gourtani[2], Patricia Dugan[2], Andrew Michalak[2], Orrin Devinsky[2], Peter Rozman[3], Werner Doyle[3], Daniel Friedman[2], Adeen Flinker[1,2*]

[1*]Department of Biomedical Engineering, Tandon School of Engineering, New York University, New York, 11201, NY, USA.
[2]Department of Neurology, School of Medicine, New York University, Street, New York, 10016, NY, USA.
[3]Department of Neurosurgery, School of Medicine, New York University, Street, New York, 10016, NY, USA.

*Corresponding author(s). E-mail(s): fz2185@nyu.edu; adeen@nyu.edu;

Abstract

Our brains dynamically adapt to a multisensory world by orchestrating diverse inputs across sensory streams. This process engages multiple brain regions, but it remains unclear how audiovisual stimuli are represented and evolve over time, especially in naturalistic scenarios. Here, we employed movie-watching to explore this question. We recorded intracranial electrocorticography (iEEG) to measure brain activity in 19 participants watching a short multilingual movie. Using unsupervised clustering and supervised encoding models, we identified a robust modality-specific gradient in the frontal cortex, wherein the ventral division primarily processes auditory information and the dorsal division processes visual inputs. Further, we found that this cortical organization dynamically changed, adapting to different movie contexts. This result potentially reflects flexible audiovisual-resource assignment to construct a coherent percept of the movie. Leveraging behavioral ratings, we found that the frontal cortex is the primary site in this modality assignment process. Together, our findings shed new light on the functional architecture of the frontal cortex underlying flexible multisensory representation and integration in natural contexts.

Keywords: audiovisual processing, intracranial EEG, encoding model, frontal cortex

Introduction

The brain has a remarkable capacity to process multiple sensory signals simultaneously for a coherent percept [1]. This multisensory processing leverages complementary sensory information to enhance environmental awareness and guide behavior [2]. Neuroimaging and electrophysiological studies have implicated the frontal cortex as a key region in this process [3–5]. Increased activation in the frontal regions occurs when congruent multisensory information is presented [6–8] and when sensory inputs are successfully integrated [9]. Additionally, the frontal lobe also selectively represents relevant information originating from perceptual cortical regions [7, 10].

These findings are primarily derived from highly controlled, task-based experiments that contrast multi- and uni-sensory conditions. There remains a paucity of research addressing how the brain dynamically organizes multiple sensory inputs in natural settings. To this end, movies can serve as ideal stimuli to study multisensory processing. Movies contain abundant audiovisual information, and observers need to dynamically allocate resources for different modalities to optimize perceptual understanding. Further, naturalistic materials are ecologically valid [11] and potentially generalizable outside the laboratory [12].

Only two functional Magnetic Resonance Imaging (fMRI) studies have directly investigated audiovisual processing during naturalistic movie viewing. One employed deep learning (DL)-based encoding models to quantify cortical representations of auditory and visual features and found that audiovisual models provided better fits for frontal cortex versus unimodal models [13]. A related study implicated frontal regions in watching audiovisual movies, as well as in audio-only and visual-only controls, indicating that the frontal cortex represents cross-modal information [14]. These studies, however, did not explicitly differentiate audio and visual representations in the frontal cortex. Moreover, due to the relatively low temporal resolution of fMRI, these studies could not assess moment-to-moment temporal changes in audiovisual processing. Additionally, these studies primarily identified static neural representations, without focusing on how audiovisual representations change throughout the movie. As a result, it remains unclear how the frontal cortex processes auditory and visual information (i.e., in a modality- specific or general manner) and how this functional organization evolves temporally.

To bridge these gaps, we examined neural responses in a cohort of neurosurgical patients using intracranial electrocorticography (iEEG) while they watched a multilingual naturalistic movie. Intracranial EEG directly records from cortex with excellent spatiotemporal resolution (millimeters and milliseconds) and high signal-to-noise ratio (SNR), enabling investigation of neural dynamics in audiovisual processing that cannot be assessed on fMRI or scalp EEG [15]. Applying functional clustering and encoding models to these recordings, we identified a ventral-to-dorsal gradient in the lateral frontal cortex, associated with auditory (ventral) and visual (dorsal) representations. Further, this modality-specific frontal pattern varied with changes in language context, indicating flexible modality assignment of neural resources. Additional analyses showed that the neural responses related to the audiovisual assignment were predominantly located in the frontal cortex. Overall, these results highlight a

dynamic functional organization in the lateral frontal cortex, advancing our understanding of natural audiovisual processing. Our work contributes to the development of neurobiologically plausible models for flexible multisensory processing in real-world scenarios.

Results

We measured neural activity during free viewing of a movie across 19 patients, including 2688 electrode contacts with wide coverage across the brain (Fig. S1A). The movie contained four different stories with a common theme, alternating between English and other languages (Greek, German, and French; see Methods: Task and procedure). All participants were native English speakers with no prior knowledge of any of the foreign languages in the movie. In this study, we focus on the high-gamma (70 to 150 Hz) broadband field potentials, given recent evidence showing that high-gamma activity reflects local neuronal firing [16, 17] and is highly correlated with the fMRI blood-oxygenation level-dependent (BOLD) response [18, 19].

Unsupervised clustering reveals modality-specific neural responses

We were first interested in understanding the neural responses to diverse auditory and visual scenes. To this end, we divided the movie into four conditions, each representing a distinct audiovisual scenario (Fig. 1A): (1) English (EN) condition, in which dialogue is spoken in English; (2) Foreign Language (FL) condition, in which characters communicate in non-English languages such as French or German (English subtitles are provided for movie comprehension); (3) Other Sound (OS) condition, in which non-speech sounds co-occur with the visual scene; and (4) Silent (SI) condition, in which visual scenes are presented without audio.

By analyzing neural signals within a one-second post-onset epoch for each condition (onsets were defined as the beginning of the sound in EN, FL, OS or the visual scene in SI, see Methods: Active electrodes selection), we observed various neural characteristics related to audiovisual processing in different regions. For instance, contexts including auditory information (i.e., the EN, FL, and OS conditions) elicited early responses (around 200 ms) in an example electrode from superior temporal gyrus (STG) (Fig. 1C, leftmost panel). Similarly, an electrode in primary visual cortex (V1) exhibited strong activation in the Foreign Language condition in an early epoch (Fig. 1C, the middle left panel). Additionally, frontal electrodes demonstrated delayed responses exclusively for language conditions (after 400 ms), while signals in non-language conditions remained stable (Fig. 1C, the right two panels). The electrodes with significant responses were predominantly distributed in the superior temporal gyrus (STG), occipital regions, frontal cortex, and inferior parietal lobe (IPL) (Fig. 1B, Fig. S1C; permutation test, $p < 0.05$, FDR corrected; see Methods: Active electrodes selection).

To characterize the prototypical patterns of these diverse neural signals, we applied a non-negative matrix factorization (NMF) analysis to the significant electrodes across the four conditions (Fig. S3A, see Methods: Functional clustering analysis) [20–22].

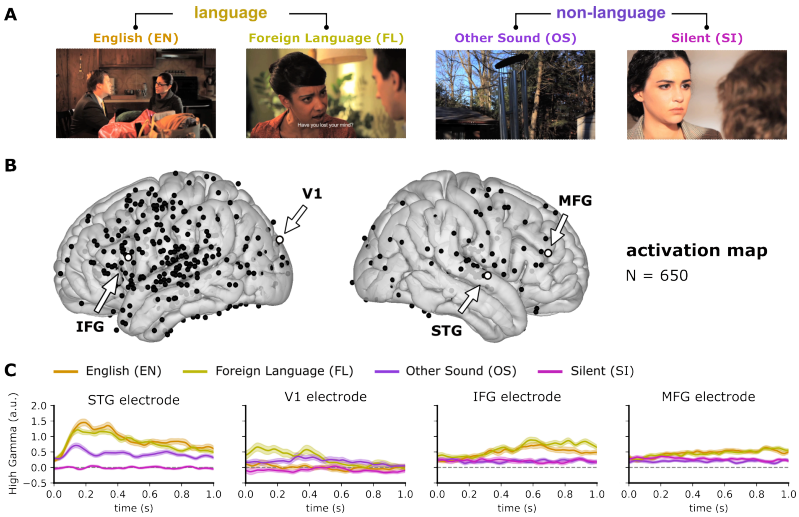

Fig. 1 Neural responses during movie viewing. (A) Four conditions identified from the movie: (1) English (EN) condition: dialogue is in English; (2) Foreign Language (FL) condition: communications in other languages such as French or German (English subtitles are provided for movie comprehension); (3) Other Sound (OS) condition: non-speech sounds co-occur with the visual scenes; (4) Silent (SI) condition: visual scenes are presented without audio. (B) All active electrodes survived the permutation test (Fig. S2) across the four conditions (Fig. S1C). (C) Averaged signals across the four conditions for the representative electrodes marked in (B). The shade areas around the curves represent the standard error of the mean (SEM) across trials.

This clustering approach, combined with a silhouette test, identified two distinct clusters (Fig. S3B). The first cluster exhibited higher weights for electrodes in the STG and ventrolateral prefrontal cortex (vlPFC), while the second cluster contained the occipital areas, dorsolateral prefrontal cortex (dlPFC), and pericentral gyrus (Fig. 2A). Based on anatomical structures and their functional responses, we assigned the two clusters as the auditory cluster (red) and the visual cluster (blue), respectively. Furthermore, we calculated the weighted average of neural responses for both clusters (Fig. 2B, left panel) and observed that the language conditions generally showed higher amplitudes than non-language conditions. By computing the weighted activity according to distinct anatomical regions (Fig. 2B middle and right panels; Fig. S1B), we found that the auditory and visual clusters showed earlier peaks in the STG and occipital regions respectively (STG electrodes: 365 ± 242 ms; occipital electrodes: 389 ± 221 ms; computed based on the language conditions). In addition, the two clusters showed later responses in the frontal areas (auditory cluster $= 657 \pm 237$ ms; visual cluster peak timing: 608 ± 278 ms; computed based on the language conditions). Notably, the frontal electrodes appeared to be spatially organized according to their respective modalities, wherein the vlPFC electrodes were primarily assigned to the auditory cluster, and the dlPFC electrodes were predominantly assigned to the visual cluster (Fig. 2A). These results indicated a modality-specific cortical activation map during movie viewing.

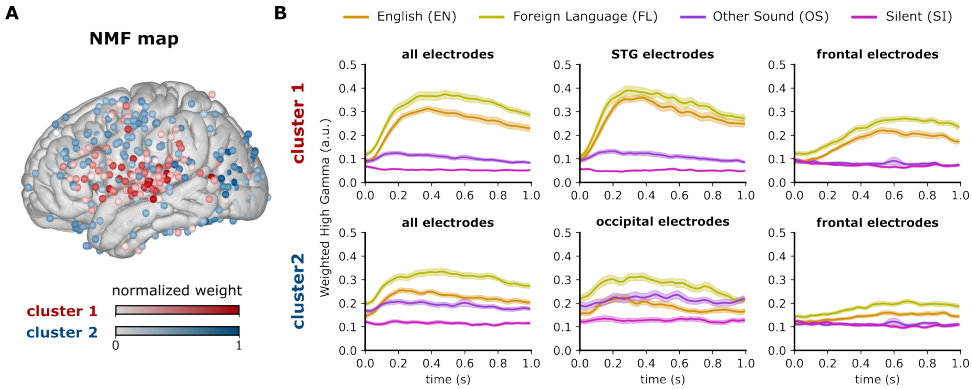

Fig. 2 Functional clustering of the active responses. (A) Spatial map showing two identified clusters with the non-negative matrix factorization (NMF) approach. For both clusters, the weights were normalized to the range 0-1, with higher values indicating greater contributions to each cluster. (B) Weighted average signals of the two clusters across all electrodes (left panel), superior temporal gyrus (STG) and occipital electrodes (middle panel), as well as frontal electrodes (right panel). The shade areas around the curves represent the SEM.

Audiovisual encoding exhibits modality-specific representations

We aimed to determine whether the spatial distributions identified by the functional clusters reflect how distinct auditory and visual information is processed during free viewing. To this end, we employed the encoding models to examine how audiovisual information is represented in the brain [23]. First, we utilized a series of computational models to extract low- and high-level features from auditory and visual streams (Fig. 3A, see Methods: Audiovisual feature extraction). At the low level, we were interested in capturing fine-grained acoustic representations from audio (i.e., the spectrogram) and spatiotemporal frequencies from visual scenes (i.e., motion energy Gabor filters [24]). At the high level, by tapping into the significant improvements of transformer architecture in complicated downstream tasks [25, 26], we extracted the vectorial embeddings from two representative models for auditory (the wav2vec 2.0 model [27]) and visual modalities (the vision transformer (ViT) model [28]). Then, we concatenated the low- and high-level embeddings to ensure a comprehensive feature space.

We then used the multivariate temporal response function (mTRF) model to quantify neural encoding [29]. Briefly, the mTRF is a linear model that incorporates time lags (τ) between features (X) and neural activity (y) (Fig. 3B), which has been widely used in neural representational analysis of acoustic, phonological, and linguistic features in connected speech [22, 30, 31] as well as semantic novelty in naturalistic movie segments [32]. To prevent model overfitting and facilitate computational efficiency, we used principal component analysis (PCA) to reduce the dimensions of the audio and visual features (Fig. S4A). We conducted the encoding analysis on all electrodes and evaluated model performance based on the Pearson correlation (r) between predicted and actual neural signals in the withheld test set under a 4-fold cross-validation

procedure (see Methods: Encoding modeling procedure). Auditory and visual mTRF models were trained separately.

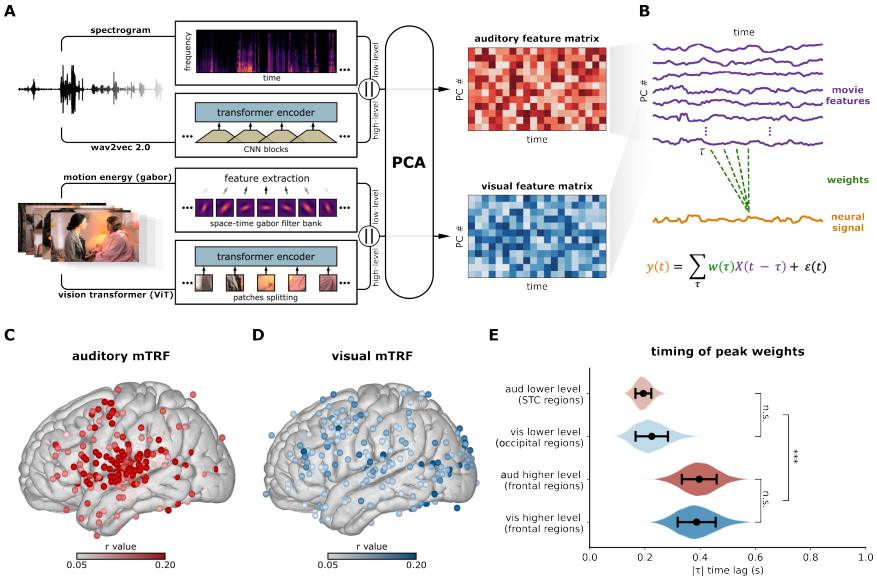

Fig. 3 Neural representations of the auditory and visual features in the movie. (A) Pipeline for extracting the audiovisual features, including: 1) low-level auditory (spectrogram) and visual features (motion energy filters); 2) high-level auditory (transformer embeddings from the wav2vec 2.0) and visual features (transformer embeddings from the vision transformer (ViT); see Methods: Audiovisual feature extraction). Low- and high-level features were concatenated (denoted by "||") before undergoing dimensionality reduction using principal component analysis (PCA). (B) Overview of the multivariate temporal response function (mTRF). The processed movie features (X, purple) were used to predict the neural responses (y, oranges) with a set of weights (w, green) across various time lags (τ). (C, D) The correlation brain map of the auditory and visual mTRF models, showing only significant electrodes. (E) The peak timing for all the electrodes in four anatomical region-of-interests (ROIs). The distributions are plotted based on a bootstrap procedure, and the significance test between different regions was conducted with a permutation test (see Methods: Timing analysis of encoding models). The error bars represent the standard deviation (SD). Significance levels are set as $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and $p \geq 0.05$ (n.s.).

Spatially, significant auditory electrodes were primarily located in STG, vlPFC, and middle precentral gyrus (midPrCG) (Fig. 3C and S5A), while significant visual electrodes were largely distributed in occipital regions, dlPFC, STG, and IPL (Fig. 3D and S5B). The anatomical distributions identified by auditory mTRF models replicated our previous NMF analysis of the auditory cluster (Chi-square test, $\chi^2(9) = 8.921$, $p = 0.445$; Fig. S6A-C). Similarly, the spatial patterns identified by the visual mTRF models were consistent with those of the visual cluster, showing no significant difference (Chi-square test, $\chi^2(9) = 14.337$, $p = 0.111$).

Temporally, based on our functional clustering results showing that frontal regions are recruited later than perceptual cortices (e.g., STG and occipital region), we observed similar temporal patterns by examining the lags (τ) of the peak weights in the

mTRF filters (Fig. S4B): the STG (auditory: 194 ± 28 ms) and occipital regions (visual: 224 ± 58 ms) peaked at relatively earlier stages, while the frontal electrodes exhibited later peak latencies (auditory: 396 ± 63 ms; visual: 387 ± 68 ms; Fig. 3E). Moreover, we found no significant differences between modalities in either the lower-level regions (permutation test, $p = 0.512$, FDR corrected) or the frontal regions (permutation test, $p = 0.745$, FDR corrected), but observed a significant difference between them (permutation test, $p < 0.001$, FDR corrected; Fig. 3E). These results collectively suggest distinct encoding of auditory and visual information across multiple timescales.

A ventral-dorsal gradient in lateral frontal cortex shows audiovisual selectivity

Both the NMF and mTRF results suggest that the ventral and dorsal areas of the frontal cortex are selective for auditory and visual information, respectively. To quantify this relationship, we constructed an index to capture the audiovisual (AV) tuning strength of the significant electrodes in the frontal region. This AV index is computed for each significant electrode by taking the r-value difference between the visual and auditory models, divided by their sum (see Methods: Frontal ventral-dorsal gradient analysis). Therefore, an AV index of 1 indicates that the electrode is completely visually tuned, while an index of -1 indicates complete auditory tuning. To further delineate the ventral-dorsal gradient, we constructed a frontal polar coordinate system with the intersection of the precentral sulcus and the Sylvian fissure as the origin (MNI coordinate: [-55, 15, -8]). The smaller values of the radius (d, in mm) indicate more ventral positions, while the larger values correspond to more dorsal areas (Fig. 4A). We found a significant linear relationship between the radius and the AV index ($r(63) = 0.447$, $p < 0.001$), quantitatively validating an audiovisual gradient in the frontal cortex (Fig. 4B). In addition, we conducted a control analysis to ensure that the frontal gradient was not biased by the semantic or linguistic components of the subtitles in the Foreign Language condition. By extracting the BERT embeddings of the subtitles and applying a model partitioning procedure (see Methods: Audiovisual feature extraction and Encoding model partitioning), we found that adding the subtitle embeddings neither significantly increased r-values in the frontal cortex (Mann-Whitney U-Test, $U(72) = 763$, $p = 0.399$, $r = -0.115$; FDR corrected; Fig. S7A-B) nor affected the frontal audiovisual gradient ($r(47) = 0.594$, $p < 0.001$; Fig. S7C-D), indicating that the modality-specific frontal pattern was not confounded by subtitle-related semantic processing.

We further examined what type of audiovisual information drives this frontal gradient (Fig. 4A-B). Our feature construction procedure allows us to dissociate low-level perceptual effects (e.g., spectrogram and Gabor features) from high-level representations (e.g., wav2vec 2.0 and ViT features). To assess the contribution of each feature space to neural encoding performance, we computed the reduction in r-values when either low- or high-level features were removed from the model (see Methods: Encoding model partitioning). We only observed pronounced effects in the frontal cortex for the features derived from transformer-based models (one-sample t-test, auditory: $t(21) = 9.097$, $p < 0.001$, $d = 1.940$; visual: $t(18) = 4.789$, $p < 0.001$,

$d = 1.099$; FDR corrected), but not for the low-level features (one-sample t-test; auditory: $t(21) = 1.267$, $p = 0.219$, $d = 0.270$; visual: $t(18) = 2.074$, $p = 0.063$, $d = 0.476$; FDR corrected; Fig. 4C). These findings indicated that the frontal gradient is primarily driven by the high-level information embedded in transformer-based models (see Discussion).

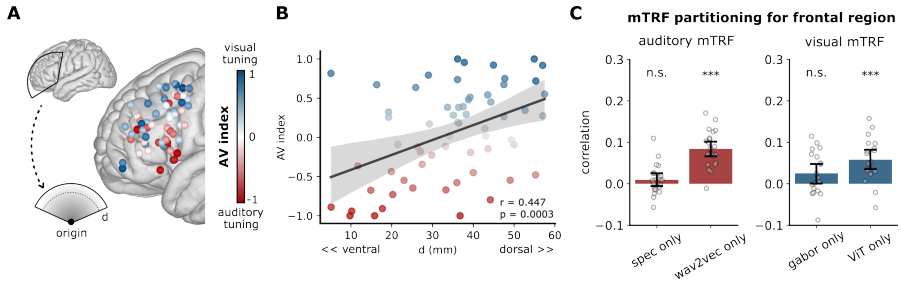

Fig. 4 Frontal gradient for audiovisual modalities. (A) The significant electrodes within the frontal sector ROI are represented by audiovisual (AV) index. The sector ROI was defined using a polar coordinate system, with the origin at the intersection of the precentral sulcus and the Sylvian fissure, where the radius (d, mm) quantifies the ventral-to-dorsal gradient. The AV index is computed for each significant electrode by taking the r-value difference between the visual mTRF and auditory mTRF, normalized by their sum (see Methods: Frontal ventral-dorsal gradient analysis). (B) A significant correlation was found between radius and AV index. The gray area around the regression line represents the 95% confidence interval (CI). (C) Results of the encoding model partitioning procedure. We obtained the unique low-level effects (auditory: spectrogram; visual: motion energy Gabor filter) by subtracting the high-level effects (auditory: wav2vec 2.0; visual: vision transformer) from the full model (including both the low- and high-level features). Similarly, the unique high-level effects can be obtained by reversing the procedure (see Methods: Encoding model partitioning). The error bars represent the 95% CI. Significance levels are set as $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and $p \geq 0.05$ (n.s.).

A flexible audiovisual-resource assignment supported by behavioral and neural evidence

Next, we were interested in how bimodal information is represented in the brain throughout movie watching and whether neural patterns vary across different movie context. To this end, we examined the mTRF results for multiple conditions. Although we did not detect frontal representations in the non-language conditions (i.e. Other Sound and Silent; Fig. S5), we observed significant representations changing with language conditions. Specifically, the auditory electrodes were more pronounced in the frontal cortex during the English condition while the visual electrodes were more prominent in the Foreign Language condition (Fig. 5A). Moreover, there was a significant interaction effect between language and modality conditions (two-way ANOVA, language-by-modality interaction effect: $F(1, 196) = 55.178$, $p < 0.001$, partial $\eta^2 = 0.220$; Fig. 5B), suggesting that audiovisual information is reallocated in the frontal cortex depending on the context.

Why did distinct frontal patterns emerge under different language conditions? Past studies suggest that combining multisensory information operates as a weighting process [33–35], indicating that audiovisual processing during movie watching involves weight-based assignment of different modalities. From this perspective, we would expect the auditory domain to receive greater weights in the English condition and visual cues to be prioritized in the Foreign Language condition, consistent with our findings (Fig. 5A-B). To explicitly test this idea, we conducted a series of movie rating tasks using the Amazon Mechanical Turk (AMT) platform (see Methods: Task and procedure; Fig. 5E). In the task, the entire movie was segmented into 2-3 seconds clips, and participants were asked to evaluate (1) how important the clip is for understanding the entire movie (global context; Fig. 5C); (2) what type of information is more important for understanding the current movie clip (local modality; Fig. 5D). As a result, the point-wise multiplication of the global context and local modality ratings (i.e., the interaction term) could serve as an indicator of modality assignment, quantifying which modality contributes more to overall movie comprehension (global modality; Fig. 5F). Additionally, we performed condition analyses and found that the global context ratings in the language conditions were significantly higher than that in the non-language conditions (two sample t-test, $t(196) = 3.055$, $p = 0.003$, $d = 0.425$; Fig. 5G). This result suggests that the language conditions (i.e., English and Foreign Language) demanded greater cognitive involvement — consistent with our NMF and mTRF findings showing stronger responses for these conditions (Fig. 2B and Fig. S5). Moreover, our findings for the language conditions also matched our prediction, where the English condition showed significant tuning to auditory information (one sample t-test, $t(43) = -2.562$, $p = 0.014$, $d = 0.386$, FDR corrected), while the Foreign Language condition showed greater selectivity for the visual stream (one sample t-test, $t(45) = 5.027$, $p < 0.001$, $d = 0.741$, FDR corrected; Fig. 5H).

Furthermore, we predicted that specific neural resources should be responsible for modality assignment in support of movie comprehension. To this end, we fitted the mTRF models to the audiovisual assignment variable (i.e., global modality; see Methods: Encoding modeling procedure). Significant electrodes were predominantly located in the lateral frontal cortex and were also sparsely scattered in the anterior temporal lobe (ATL), temporal-parietal junction (TPJ), and other regions (Fig. 6A and Fig. S6D). To rule out potential confounding influences of audiovisual features in the frontal cortex, we performed an additional partitioning procedure (see Methods: Encoding model partitioning) and detected significant effects only for the audiovisual assignment (one sample t-test, $t(33) = 7.968$, $p < 0.001$, $d = 1.366$, FDR corrected, Fig. S8A).

Additionally, these electrodes were distinct from those representing audiovisual features. The r-values from the assignment models did not correlate with either auditory (Pearson correlation, $r = -0.122$, $p = 0.145$) or visual r-values (Pearson correlation, $r = -0.086$, $p = 0.306$; Fig. 6B). In contrast to the ventral-to-dorsal frontal gradient observed for the audiovisual features, the electrodes related to audiovisual assignment were more diffusely distributed in the frontal lobe (Fig. 6C). Notably, 73.856% of the electrodes (113 out of 153) were associated with only one specific function, and only four electrodes showed overlap across all three models (2.614%; Fig. 6D). An

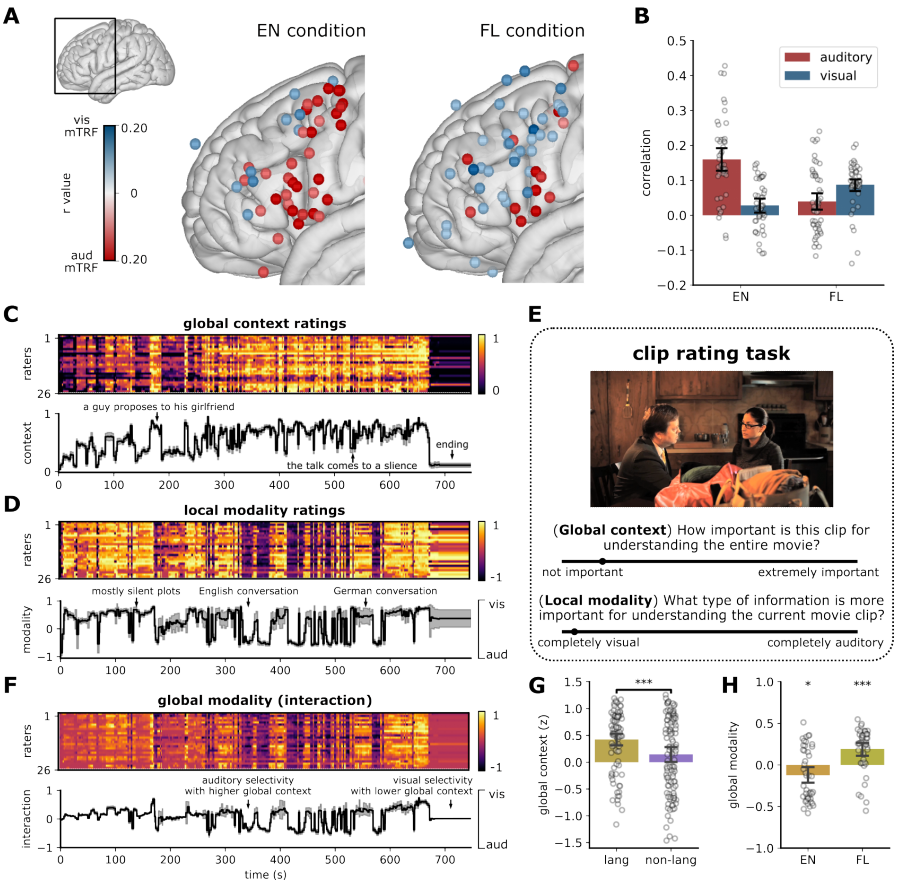

Fig. 5 Neural dynamics and the behavior experiment. (A) The mTRF correlation map in the frontal region for English (EN) and Foreign Language (FL) conditions. (B) An interaction effect between the auditory and visual mTRF models in the frontal cortex for the EN and FL conditions. (C, D) The global context and local modality ratings. Upper panel: ratings across all participants over time. Lower panel: the average time series across participants. (E) The pipeline of the clip rating task conducted on the Amazon Mechanical Turk platform. (F) Pointwise multiplication of global context and local modality (the interaction effect). This interaction effect reflects dynamic changes in modality assignment for movie comprehension. (G) The bar plot of the global context ratings for the language (lang) condition, which includes the EN and FL conditions, and the non-language (non-lang) condition, which includes the OS and SI conditions. The dots represent trials in each condition. (H) The bar plot of the global modality for the EN and FL conditions. The shaded areas in figure C, D, and F represent SE; the error bars in figure B, G, and H represent 95% CI. Significance levels are set as $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and $p \geq 0.05$ (n.s.).

analysis of the timing (peak lag of modality assignment) showed that the audiovisual assignment in the frontal cortex (297 ± 47 ms) occurred significantly later than the audiovisual representation in the perceptual regions (permutation test, $p = 0.008$, FDR corrected), but earlier than the frontal encoding of audiovisual features (permutation test, $p = 0.001$, FDR corrected; Fig. S8B).

Taken together, our neural and behavioral findings suggest that frontal cortex representations flexibly adapt to shifts in movie plots and linguistic scenarios. Furthermore, audiovisual assignment likely relies on distinct frontal neural substrates that complement the representation of audiovisual features.

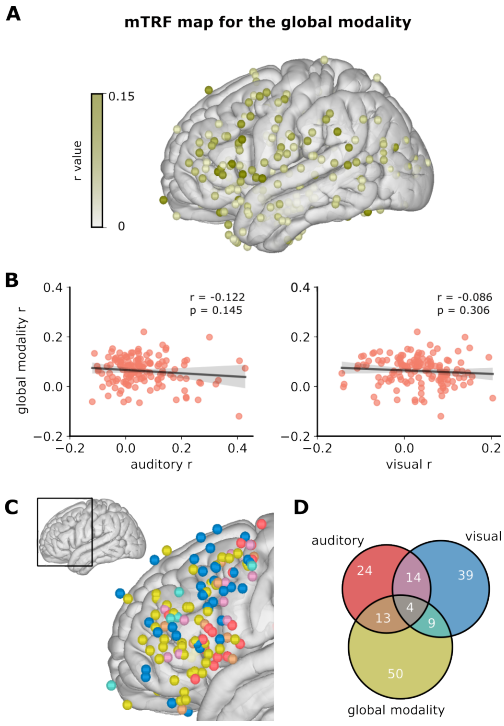

Fig. 6 Neural substrates for the audiovisual assignment. (A) The mTRF results for the global modality (i.e., the audiovisual assignment). (B) The correlation analysis between r-values from the assignment mTRF models and r-values from the auditory/visual mTRF models for all significant electrodes in the frontal cortex. The shaded areas around the lines represent the 95% CI. (C) The brain map of significant frontal electrodes across the three mTRF models, rendered with the color code in (D). (D) The Venn diagram of the significant frontal electrodes from the three mTRF models.

Discussion

The human brain dynamically reconfigures auditory and visual information during free movie viewing. Here, we identified auditory and visual neural responses and representations spanning from the early perceptual areas to the frontal cortex. Notably, the frontal cortex exhibited a modality-specific topography, which shifted depending on the language context. Together with converging behavioral data, we provide evidence that observers dynamically allocate audiovisual resources in response to contextual demands, revealing a flexible and adaptive mechanism for real-world audiovisual processing.

Our results demonstrate a frontal gradient in audiovisual representation during free viewing. Although previous studies have shown that the frontal cortex is broadly involved in audiovisual processing during natural movie watching [13, 14], they primarily focused on specific ROIs and lacked a comprehensive view of representational organization across the frontal lobe. In contrast, we systematically examined the lateral frontal cortex using both unsupervised functional clustering and supervised encoding models, revealing a robust segregation between auditory and visual modalities. Further, this gradient was solely driven by high-level features extracted from transformer-based models and exhibited relatively late peak latencies (around 400 ms). These findings align with recent unimodal auditory and visual studies reporting dissociable semantic representations in the frontal cortex [20, 36–40]. Importantly, rather than restricting our analysis to predefined anatomical areas, we constructed a frontal sector ROI and identified a modality-specific gradient, transitioning from the ventral region associated with auditory information to the dorsal region associated with visual information. This suggests a spatially continuous coordinate system along the ventral–dorsal axis of the frontal cortex [41, 42], reflecting a functional topographic organization that supports audiovisual processing under naturalistic experiences.

Prior perceptual research has shown that observers integrate multimodal cues by assigning weights proportional to their reliability across different task contexts [34, 35, 43–45]. Under this view, more reliable modalities tend to exert greater influence on unified perception. Our findings suggest a similar "modality assignment strategy" during movie viewing, in which frontal neural resources are dynamically allocated between auditory and visual modalities based on contextual demands. Unlike prior studies that manipulated stimulus reliability by adding noise, movies are carefully crafted and contain minimal artificial distortion [11, 46], with both auditory and visual streams maintaining consistently high fidelity. Therefore, the modality assignment effect we observed is unlikely to stem from sensory reliability *per se*, but instead reflects how different modalities contribute to forming a coherent perceptual object. This aligns with a framework of goal-driven, top-down control over multisensory processing [47]. Moreover, we found that the modality assignment effect peaked around 300 ms, consistent with the temporal window associated with multisensory integration [48–50]. Collectively, our results provide an expansion from previous well-controlled laboratory experiments to naturalistic settings, revealing a goal-oriented modality assignment strategy for audiovisual integration during free viewing of movies.

In summary, our study demonstrates that the lateral frontal cortex serves as a key region for processing audiovisual information under naturalistic experiences. We reveal

modality-specific representations in the frontal cortex, with auditory information encoded toward ventral regions and visual information toward dorsal areas. We also show that distinct frontal substrates are largely involved in the modality assignment process. Together, these findings provide evidence for a distinct spatial organization in the frontal cortex supporting flexible multisensory processing of real-world scenarios.

Methods

Participants

The first cohort consisted of 19 patients undergoing neurosurgical evaluation for refractory epilepsy (12 females; all right-handed; 32.053 ± 11.297 years old; 12 grid and 7 stereotactic coverages; see Supplementary Table S1). These patients participated in a multilingual movie watching task during their stay in the hospital. Electrode implantation was determined exclusively based on clinical requirements. All participants were fluent in English and had no knowledge of the foreign languages used in the movie. Written informed consent was obtained from all participants either prior to the neurosurgical procedure or upon admission to the hospital unit for evaluation.

A second cohort of 26 crowdsourced raters was included in an online movie rating task (11 females; 3 left-handed; 42.154 ± 8.694 years old). Participants self-identified as either native English speakers (23 raters) or proficient English speakers (3 raters), none reported knowledge of the foreign languages used in the movie.

The current study protocol was approved by the NYU Langone Medical Center Committee on Human Research. All participants were reimbursed for their participation.

Task and procedure

In the intracranial experiment (**Movie watching task**), patients were asked to silently watch a 12-minute long multilingual movie called Foreign Language Movie (FLM; https://vimeo.com/61040183). This short movie contains four distinct storylines: 1) a Greek mother reveals her terminal disease to her daughter (Greek and English); 2) two German girls discuss saving an unborn life (German); 3) a married couple's relationship comes to an end (English); 4) a young couple conversing in French and English about getting married. The stories are intricately interwoven, forming a cohesive and unified story. English subtitles are provided in the foreign language conversations to aid in movie comprehension.

The movie was presented on a 15" laptop placed 0.5–1.0 meters in front of the participants. Audio was delivered through a nearby speaker beside the laptop, with the volume adjusted to a clear and comfortable level for each participant. To synchronize neural activity with the movie, trigger pulses were transmitted to the EEG recording system at the movie onset and at one-minute intervals thereafter.

The behavioral experiment (**Online movie-rating task**) was conducted using the Amazon Mechanical Turk (AMT) platform and required participants to watch the movie twice. Before the experiment, raters were instructed to "find a quiet place, make sure that your internet has a stable connection and will not be disturbed during the experiment". During the first viewing, participants watched the entire movie and were asked to provide continuous engagement ratings by adjusting a scale bar from 1 (not engaging) to 99 (extremely engaging) [51]. The scale bar remained visible below the video window throughout the session. If participants did not rate for longer than 5 seconds (i.e., did not move the cursor), the movie would automatically pause, and a flashing reminder would prompt them to rate. Immediately after the first viewing,

participants completed a free-recall task to assess comprehension. The instruction for this section was "Please describe what you remember from the movie. Please try to recount events in their original order, if possible, and describe events in writing in as much detail as possible (spend approximately 5-10 minutes). During your description, completeness and detail are more important than the order of events, typos, and grammar." [52]. Raters whose responses lacked sufficient detail or omitted one or more storylines were excluded from further analysis.

During a second viewing, the movie was presented as a sequence of 2-3 second clips. The segmentation was performed using the FFmpeg toolbox (https://www.ffmpeg.org/). For each clip, the raters were asked to evaluate two aspects using scale bars: 1) how important is this clip for understanding the entire movie from 1 (not important) to 99 (extremely important); 2) what type of information is more important for understanding the current movie clip from 1 (completely visual) to 99 (completely auditory) (Fig. 5E). If unsure of their ratings, the raters could replay the clips before proceeding to the next clip. The resulting clip-specific ratings were concatenated (based on frame counts) to reconstruct continuous movie rating signals (Fig. 5C-D).

Statistical testing

For both behavioral and neural data, the D'Agostino test was used to determine the normality of the distribution prior to statistical analysis. If data followed a normal distribution, a parametric test was used (e.g., paired t-test); otherwise, a non-parametric test was used (e.g., Wilcoxon signed-rank test). In addition, we applied a permutation test to identify active electrodes (e.g., Fig. 1B) and assess significance in the encoding models (e.g., Fig. 3C-D). Detailed procedures are described in the corresponding Methods sections (see Methods: Active electrodes selection and Encoding modeling procedure).

All statistical tests were two-tailed with a significance threshold of $p < 0.05$, unless stated otherwise. The false discovery rate (FDR) correction was applied when multiple comparisons were conducted.

Data acquisition and preprocessing

During the movie watching task, iEEG signals were recorded using one of two amplifier types (dictated by clinical location during acquisition): (1) NicoletOne amplifier (Natus Neurologics, Middleton, WI), with signals bandpass filtered from 0.16 to 250 Hz and digitized at 512 Hz; (2) Neuroworks Quantum Amplifier (Natus Biomedical, Appleton, WI), with signals recorded at a sampling rate of 2048 Hz, bandpass filtered between 0.01 to 682.67 Hz and then decimated to 512 Hz. For grid electrodes, a two-contact subdural strip facing the skull near the craniotomy site served as the reference electrode, and a similar strip screwed to the skull was used as the instrument ground. For sEEG electodes, a five-lead subgaleal strip was placed facing the skull and used as both the ground and reference. Electrodes showing epileptiform activity, line noise artifacts, or large amplitude shifts were excluded. The data were then re-referenced using the common average reference (CAR) approach, in which the averaged signal

was subtracted from all electrodes for each subject. To extract high-gamma broadband activity (70–150 Hz), we applied a multi-band averaging method. Specifically, we computed the z-scored analytic amplitude across eight logarithmically spaced frequency bands within this range using the Hilbert transform, and then averaged them [20, 53]. This method enhances signal robustness as it takes the mean across multiple frequency bands and mitigates the bias of lower frequencies given the z-score normalization. Data were then corrected relative to the baseline (with silent, black scenes) for each channel.

Electrode localization in individual space was performed by co-registering post-operative brain MRI or CT scans to preoperative MRI scans using a rigid-body transformation. Electrodes were then projected into the MNI space using a nonlinear DARTEL algorithm [54]. Further, anatomical locations of electrodes were determined using automated FreeSurfer segmentation of the preoperative MRI.

Active electrodes selection

To identify the active electrodes involved in audiovisual processing during free viewing, we extracted the neural responses in four different audiovisual conditions (Fig. 1A): 1) English (EN) condition, in which dialogue is in English; 2) foreign language (FL) condition, in which characters communicate in other languages like French or German (English subtitles are provided to aid movie comprehension); 3) other sound (OS) condition, in which non-speech sounds are presented with visual scenes; and 4) silent (SI) condition, in which only visual scenes are presented without sound. Specifically, for the conditions with sound (i.e., EN, FL, and OS), we extracted trials from the onset to the offset of each continuous sound segment. Trials shorter than one second were included to ensure data stability and robustness. As a result, 44 EN trials, 46 FL trials, 58 OS trials, and 51 SI trials were included for analysis.

For each electrode, we aligned the onset of all trials and analyzed the potential changes within a one-second epoch in each condition (Fig. S2). An electrode was considered active if its activity across trials was significantly greater than zero (one sample t-test) for a duration exceeding a threshold determined by the permutation procedure (Fig. S1B-C). Specifically, in each permutation, we distorted the temporal alignment of the signals using a phase randomization approach [55], then identifying the maximum number of consecutive significant time points (i.e., significantly above zero) using the same analysis. Repeating this procedure 1000 times yielded a null distribution of significant duration. Then, the p-value can be derived by calculating the proportion in the null distribution that exceeded the real significant duration. An electrode was considered significant if $p < 0.05$ and survived the false discovery rate (FDR) correction of multiple comparisons (Fig. S1D). The union of electrodes identified as active across the four conditions was used for subsequent analyses.

Functional clustering analysis

We performed unsupervised non-negative matrix factorization (NMF) to summarize neural underpinnings associated with audiovisual processing [20, 22]. By introducing

the non-negativity constraint, NMF is able to effectively derive the parts-based representation [56]. Mathematically, a non-negative data matrix ($\mathbf{D} \in \mathbb{R}^{m \times n}$) can be factorized into two components $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$ by minimizing the error function iteratively (using the Frobenius norm):

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & ||\mathbf{D} - \mathbf{W}\mathbf{H}||_F \\ \text{s.t.} \quad & \mathbf{W} \geq 0, \mathbf{H} \geq 0 \end{aligned} \tag{1}$$

where the term k is a hyperparameter representing the number of factorized clusters.

In practice, the matrix $\mathbf{D}$ represents the concatenated neural responses of active electrodes across four conditions (Fig. S3A). We imposed non-negativity in $\mathbf{D}$ by setting all negative elements to zero, as no significant negative activations were identified with the permutation test (see Methods: Active electrodes selection). $\mathbf{W}$ is the weighting matrix, representing the extent to which electrodes were assigned to each cluster. $\mathbf{H}$ is the temporal prototype matrix, demonstrating the typical temporal dynamics for each cluster. To identify the best number of clusters (k), we applied the silhouette method to measure the similarity of each sample to its own cluster relative to others [57]. The highest silhouette score, which indicates the optimal clustering assignment, was found when $k = 2$ (Fig. S3B). Finally, electrodes were clustered according to their maximum contributions in the weighting matrix $\mathbf{W}$, and were normalized (0-1) for visualization purposes (Fig. 2A).

Audiovisual feature extraction

We employed several computational models to derive generalizable feature spaces for both auditory and visual modalities (Fig. 3A). Specifically, we extracted features at two levels. At the low levels, we captured the spectrogram and motion energy features to represent auditory and visual information, respectively. For the spectrogram, we applied the librosa Python package [58] to extract spectral frequencies ranging from approximately 20 Hz to 8 kHz, which serves as an effective acoustic representation of auditory perception. For the motion energy features, we utilized a pyramid of nonlinear spatiotemporal Gabor filters, which are designed to capture information across multiple positions, orientations, spatiotemporal frequencies, and motion directions [24]. The extraction procedure was implemented with the pymoten package (https://github.com/gallantlab/pymoten).

At the high levels, we leveraged transformer-based deep-learning models, which have demonstrated significant improvements in complicated downstream tasks [25, 26]. The core mechanism of the transformer architecture is multihead self-attention, an algorithm that calculates context-dependent weights to integrate information across input vectors [59]. In practice, we extracted features from two representative models: wav2vec 2.0 for auditory information [27] and Vision Transformer (ViT) for visual information [28]. To avoid task-specific influences and capture representative features of audiovisual information, we selected vector embeddings from the middle layers, as prior research indicated that these layers produce the most comprehensive features for both the wav2vec 2.0 [37, 60] and the ViT models [61]. In detail, for the auditory

domain, we extracted 7th layer embeddings of transformer blocks in wav2vec 2.0-Base model (number of convolution layers = 7, number of transformer layers = 12, hidden size = 768, number of self-attention heads = 8, total parameters = 95 M) [27]. We kept the same sampling rate (16 kHz) and batch size (15.6 seconds) with the pretrain models during feature extraction to minimize potential input format biases. For the visual domain, we extracted the [CLS] embeddings from the 16th layer of the ViT-Huge model (number of transformer layers = 32, hidden size = 1280, number of self-attention heads = 16, total parameters = 632 M) [28]. [CLS] is an extra learnable "token" introduced to the transformer encoder, representing image representations for the classification task.

Furthermore, to control the potential effect of visual subtitles in the Foreign Language condition (Fig. S7), we extracted semantic embeddings of these sentences utilizing the robustly optimized BERT approach (RoBERTa) model [62]. RoBERTa is a generalized BERT representational model, featuring a multi-layer bidirectional transformer encoder conditioned on both left and right contexts [63]. Compared with the original BERT model, RoBERTa model has a larger architecture, with a bigger batch size and more training data. Here, we applied the RoBERTa-Large model (the number of layers = 24, the hidden size = 1024, the number of self-attention heads = 16, total parameters = 355 M) [62], and selected the features from the penultimate hidden layers, as the final layers tend to be biased by the model training objectives (i.e., masked language model and next sentence prediction). Practically, we represented sentence embeddings by averaging all word vectors within each sentence [64–66].

All feature extraction procedures associated with deep-learning models were performed using the PyTorch (v2.1.0) framework [67] in Python (v3.10).

Encoding modeling procedure

We utilized encoding models to probe the neural substrates responsive to various features by leveraging the multivariate temporal response function (mTRF) [29]. Specifically, the mTRF is a linear model that incorporates time lags (τ) between features (X) and neural activities (y) (Fig. 3B). It can be formulated in a convolutional form:

$$y(t) = \sum_{\tau} w(\tau) X(t - \tau) + \epsilon(t) \tag{2}$$

where w represents the weights (also known as the mTRF filter) to be estimated, and ϵ is the additive noise.

Both neural signals and features were z-scored prior to model training. Given the relatively large parameter spaces (number of lags $\times$ feature dimensions), we applied L2-norm regularization during parametric estimation to mitigate the potential overfitting problem. The encoding model was implemented using four-fold cross-validation (CV), where the data were split into four groups - three groups used for training and the withheld one for testing. Model performance was assessed using Pearson's correlation (r-value) between predicted ($\hat{y}$) and real signal (y). The ridge parameter was optimized using a grid search approach to maximize the r-value during training. Final model performance was calculated as the average r-value across all four folds. We included neural responses with delays up to one second (i.e., $\tau_{max} = 1$ s) considering the longer

processing time associated with natural movie watching. We conducted the encoding analysis on all electrodes rather than the active electrodes that survived from the permutation test, avoiding the potential double dipping problem [68]. Specifically, three types of mTRF models were primarily employed:

1) Auditory and visual mTRF models. The auditory and visual mTRF models were trained independently. For each modality, low- and high-level features were first concatenated, followed by a principal component analysis (PCA) procedure to reduce dimensionality to enhance computational efficiency (Fig. 3A). To determine the optimal number of PCs, we trained mTRF models using 5 to 50 principal components (PCs) in increments of 5 PCs (incremental PCs test). Then, we compared the average r-values to identify the elbow point. As a result, this procedure revealed that 15 PCs were optimal (Fig. S4B).

2) Audiovisual assignment mTRF model. The averaged global modality feature obtained from the behavior experiment (Fig. 5F) was fed into the encoding model.

3) Subtitle mTRF model. The dimension of the subtitle BERT features in the foreign language (FL) condition was also relatively large (i.e., 1024 dimensions). Therefore, PCA was also employed here to reduce the feature dimensions. In the end, 50 PCs were selected for the subtitles, explaining 95.01% variance of the feature space (Fig. S7).

Furthermore, we evaluated the statistical significance level of the model performances using a permutation test, wherein the neural data was phase-randomized for 1000 iterations. We repeated the above encoding model training procedure for each permutation to generate a null distribution of r-values. Then, the p-value can be derived by calculating the proportion of r-values in the null distribution that exceeded the real r-value. An electrode was considered significant if $p < 0.05$ and survived the false discovery rate (FDR) correction of multiple comparisons.

Timing analysis of encoding models

To estimate the latencies of the neural responses, we identified the time delays corresponding to the peak weights in the mTRF filters for each electrode. The peak timings were first grouped by region-of-interest (ROI) (Fig. 3E and Fig. S3C). Then, we applied a bootstrap procedure to assess the response lags for each ROI. Specifically, we estimated the 95% confidence intervals (CI) by sampling the data with replacement and calculated the mean for 10000 iterations.

A permutation test was conducted to determine statistical significance across different ROIs (Fig. 3E). For each comparison, the group labels of each element were randomly shuffled for 1000 iterations, yielding a null distribution of mean differences. Then, the p-value was calculated by determining the position of the real difference in the null distribution. The resulting p-value was corrected for multiple comparisons using the FDR method.

Frontal ventral-dorsal gradient analysis

We constructed the AV index and frontal coordinate to quantify the representational distribution of audiovisual information in the prefrontal cortex. Specifically, the AV

index (I_{AV}) was applied to capture the electrode selectivity for auditory or visual modalities, based on r-values obtained from the auditory and visual mTRF models (r_A and r_V):

$$I_{AV} = \frac{r_V - r_A}{r_V + r_A} \tag{3}$$

By definition, the range of I_{AV} is from -1 to 1, where 1 represents the electrode is complete visual tuning and -1 represents complete auditory tuning.

In addition, a coordinate was created for quantifying the ventral-to-dorsal gradient of frontal cortex. To this end, we established a frontal polar coordinate system with the intersection of precentral sulcus and Sylvian fissure as the origin (MNI coordinate: [-55, 15, -8]). A smaller radius (d, mm) indicates a more ventral position, whereas a larger radius corresponds to a more dorsal location (Fig. 4A).

Encoding model partitioning

We employed a partitioning procedure to obtain the unique effect of a specific feature space [13]. For instance, to quantify the unique effects of the low- and high-level audiovisual features (Fig. 4C), we trained and evaluated models using either low- or high-level features (M_l and M_h), and compared them with models trained using both feature spaces ($M_{l \cup h}$). Then, the pure contribution of the low-level features was computed as the difference between the combined model and the high-level model ($M_{l \cup h} - M_h$), and that of high-level features was determined as the discrepancy between the combined model and the low-level model ($M_{l \cup h} - M_l$). We fitted all mTRF models mentioned above with 15 dimensions using PCA to ensure comparable model complexity.

Similarly, to quantify the unique contributions of audiovisual assignment and features, we trained separate models using either audiovisual assignment (M_a) or audiovisual features (M_f), as well as a combined model incorporating both ($M_{a \cup f}$). The unique effect of the modality assignment was calculated as the difference between the combined model and the audiovisual features model ($M_{a \cup f} - M_f$), while the unique contribution of audiovisual features was determined conversely ($M_{a \cup f} - M_a$; Fig. S8A).

In addition, to determine the influences of the subtitles, we built a full model that combined both the visual and subtitle features. Then, the unique subtitle effect was obtained by calculating the decrease in r-values when subtitle features were removed from the full model (Fig. S7; only for the Foreign Language condition).

Data and Code availability

The dataset of the current study will be made available from the authors upon request and documentation is provided that the data will be strictly used for research purposes and will comply with the terms of our study IRB. The code is available upon publication at https://github.com/flinkerlab/.

Acknowledgements

We thank A. Ferrari, M. Landy, and S. Michelmann for their comments on an early version of the manuscript, A. Morton for sharing the movie license for academic use, and other members in Flinker lab for extensive discussion. This work was supported by National Institutes of Health grants R01NS109367, R01NS115929, and R01DC018805 (A.F.) and National Science Foundationshum IIS-2309057.

Author Contributions

F.Z., A.K.-G and A.F. conceived and designed the research; P.D., D.F., W.D., and P.R. conducted data collection and performed neurosurgery; A.M. and O.D. conducted data collection and critically revised the article; F.Z. and A.F. wrote the paper.

Declarations

The authors declare no competing interests.

Supplementary Information

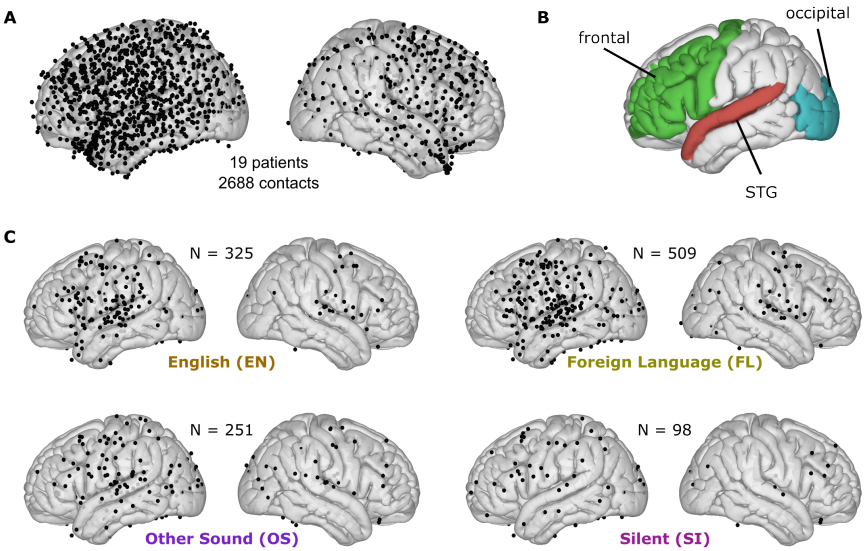

Fig. S1 (A) The electrode coverage for all 19 participants during the movie watching task (2688 contacts in total). (B) Three ROIs used in this study, including the superior temporal gyrus (STG), frontal regions (comprising the inferior frontal gyrus, middle frontal gyrus, and pre-central gyrus), and the occipital regions (comprising lateral occipital gyri, lingual gyrus, and cuneus gyrus). (C) The brain maps of the active electrodes for all four conditions that survived the permutation test (Fig. S2; see Methods: Active electrodes selection).

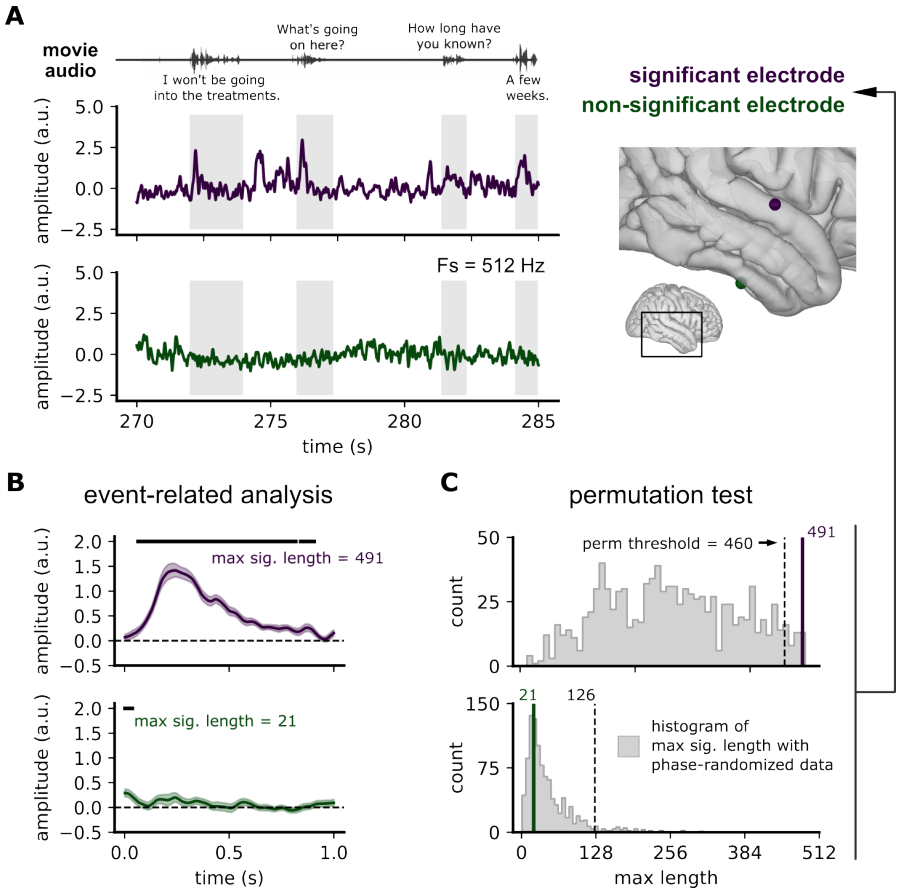

Fig. S2 Permutation procedure for selecting the active electrodes. (A) Representative signals from two electrodes (right panel): one located in the STG (purple) and the other located in the ITG (green). The STG electrode exhibited responses to auditory events (left panel), whereas the ITG electrode showed no response. (B) Event-related analysis. The neural activity during the events was aligned to the onset and averaged across time (the temporal window was set as 1 seconds). Then, the statistical significance was determined for each time point using a two-sided t-test with $p < 0.05$. The maximum length of the consecutive significant time points (i.e., maximum significant length (L_{max})) is recorded. (C) Permutation test. We repeated the event-related analysis on the phase-randomized signals for 1000 times, and computed the maximum significant length for each permutation (L^p_{max}). This procedure generated a null distribution of the L^p_{max}. The real length L_{max} was then compared against this null distribution, with significance determined at the top 5th percentile threshold (FDR corrected).

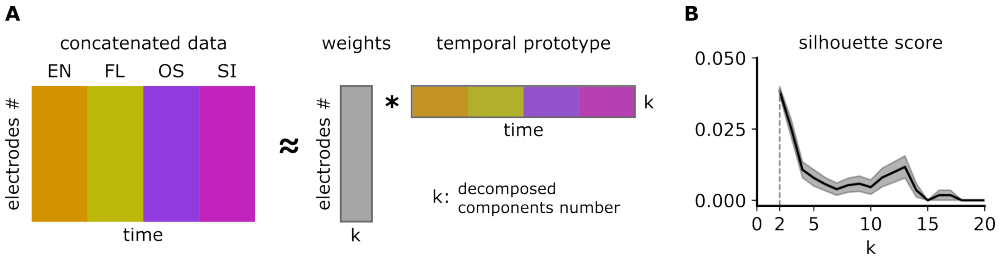

Fig. S3 Unsupervised functional clustering. (A)The data for all four conditions were concatenated and then decomposed into a weight matrix and a temporal prototype matrix (see Methods: Functional clustering analysis). (B) The result of silhouette method used to determine the optimal number of clusters (k). Higher scores indicate better clustering performance. The shaded area around the curve represents the standard error (SE).

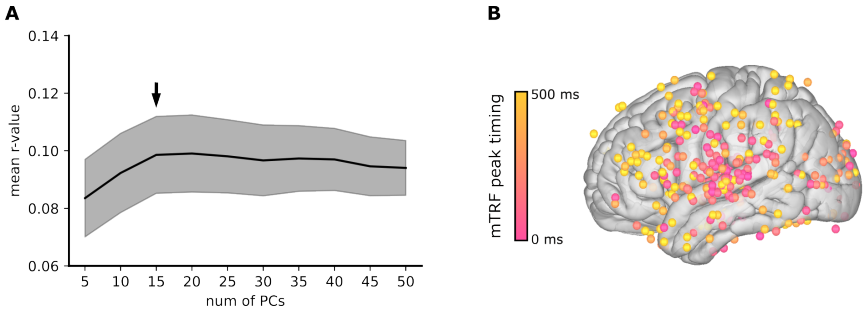

Fig. S4 (A) The result of incremental principal components (PCs) test across all the audiovisual mTRF models. The elbow point was identified using kneed Python toolbox (indicated by a downward arrow), which employs a rotation-based algorithm to detect the maximum curvature [69]. (B) Timing of peak weights across electrodes that showed significant responses in the auditory and visual mTRF models.

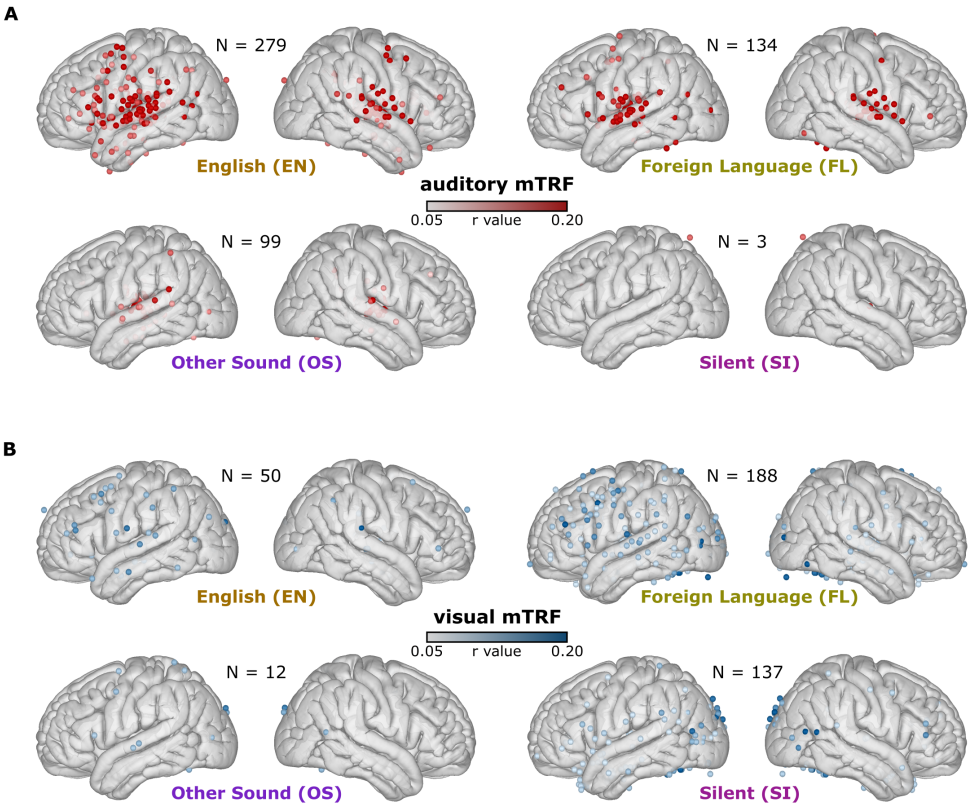

Fig. S5 Results of encoding models for auditory (A) and visual (B) information in four conditions. N represents the number of significant electrodes.

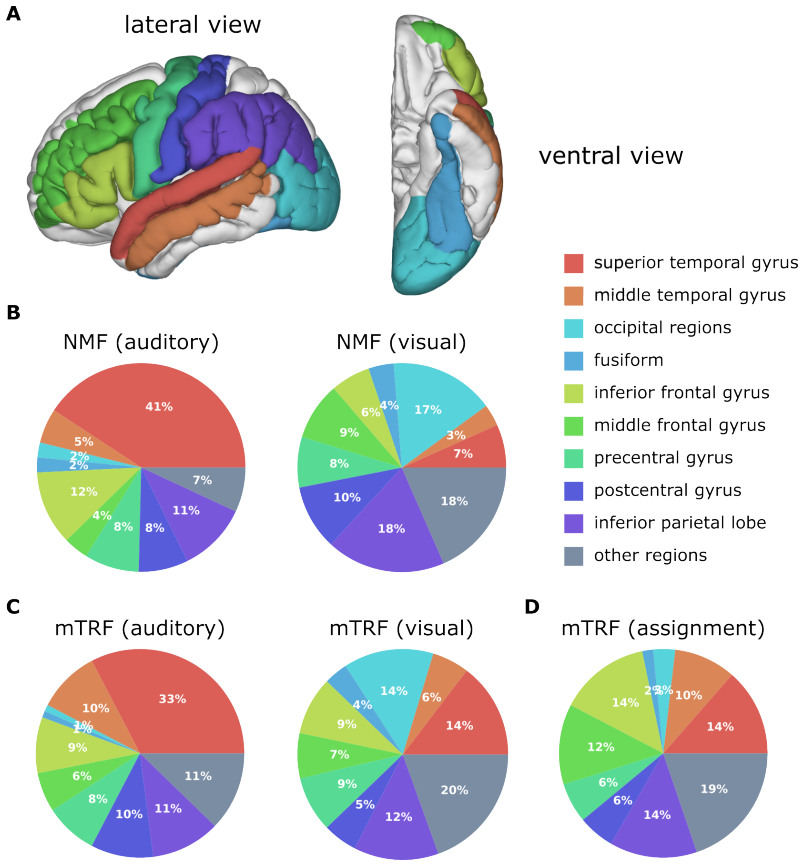

Fig. S6 Anatomy distributions for the mTRF models. (A) Nine ROIs illustrated on the brain. (B) Pie charts showing the anatomical distribution of electrodes from the auditory and visual functional clusters (NMF approach). (C) Pie charts depicting the anatomical distribution of significant electrodes from the auditory and visual encoding models (mTRF approach). (D) Pie chart showing the anatomical distribution of significant electrodes responsive to the audiovisual assignment (i.e., the global modality).

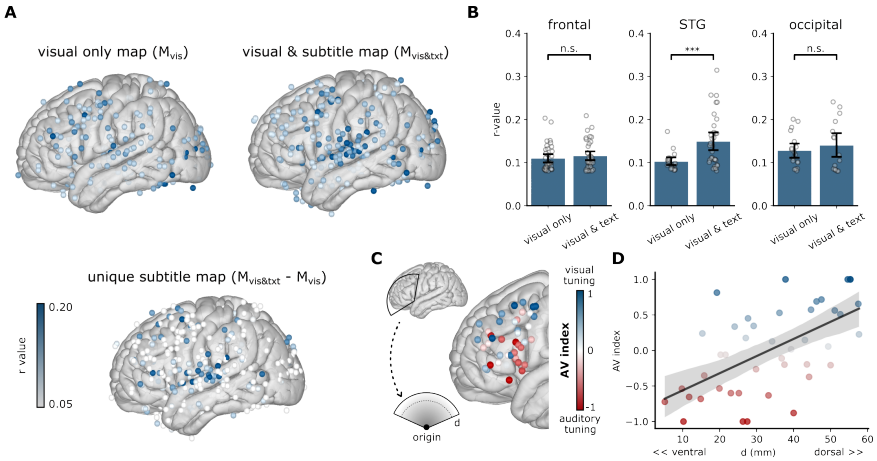

Fig. S7 The subtitle effect in Foreign Language condition. (A) Significant neural encoding maps for visual features only (M_{vis}; top left panel) and for both the visual and subtitle features ($M_{vis\&txt}$; top right panel). The unique subtitle effect map is the difference between the two maps above ($M_{vis\&txt} - M_{vis}$; bottom panel). (B) ROI analysis. The x-axis shows model types (M_{vis} and $M_{vis\&txt}$), while the y-axis indicates r-values for electrodes within each ROI. No difference was observed in the frontal (Mann-Whitney U-Test, $U(72) = 763$, $p = 0.399$, $r = -0.115$; FDR corrected) or the occipital regions (Mann-Whitney U-Test; $U(24) = 137$, $p = 0.957$, $r = -0.015$; FDR corrected). However, a significant increase was found in the STG (Mann-Whitney U-Test, $U(43) = 502$, $p = 0.002$, $r = -0.521$; FDR corrected). The error bars represent 95% CI. (C-D) Frontal gradient analysis incorporating the subtitle embeddings in the visual models (see Fig. 4A-B and Methods: Frontal ventral-dorsal gradient analysis).The correlation between radius and AV index remained significant ($r(47) = 0.594$, $p < 0.001$). Significance levels are set as $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and $p \geq 0.05$ (n.s.).

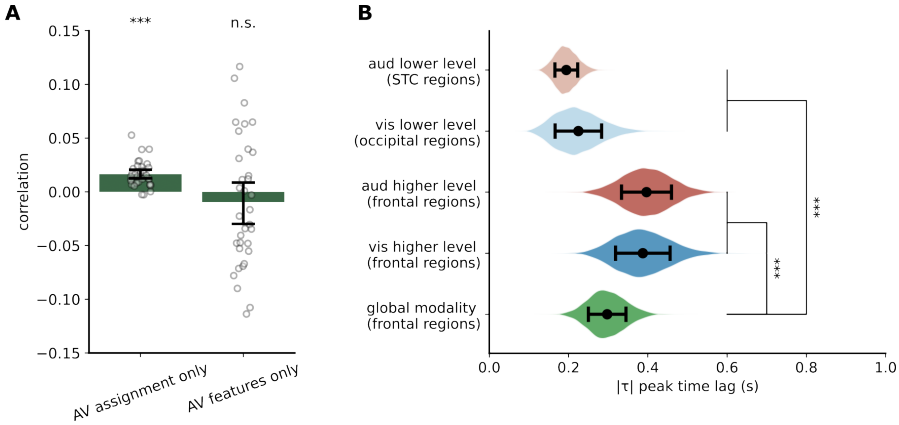

Fig. S8 (A) The unique contributions for modality assignment and audiovisual features in frontal electrodes shown in Fig. 6A (see Methods: Encoding model partitioning). A significant effect was observed only for the audiovisual assignment (one sample t-test, $t(33) = 7.968$, $p < 0.001$, $d = 1.366$, FDR corrected) but not for the audiovisual features (one sample t-test, $t(33) = -0.927$, $p = 0.361$, $d = 0.159$, FDR corrected). The error bars represent 95% CI. (B) Timing analysis of audiovisual (AV) assignment. The timing of AV assignment in the frontal cortex (297 ± 47 ms) was significantly later than the neural representation of AV features in the perceptual regions. A significant difference was also observed between the timing of encoding the AV features and the modality assignment in the frontal regions. The distributions were estimated based on a bootstrap procedure, and the statistical significance between regions was assessed via a permutation test (see Methods: Timing analysis of encoding models). The error bars represent the standard deviation (SD). Significance levels are set as $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), and $p \geq 0.05$ (n.s.).

Table S1 Patients information

ID	Coverage	Gender	Handedness	Age at Surgery (years)
1	Right	Male	Right	29
2	Left	Female	Right	45
3	Left	Male	Right	24
4	Left	Female	Right	30
5	Left	Female	Right	56
6	Bilateral	Male	Right	26
7	Right	Female	Right	21
8	Left	Male	Right	22
9	Left	Male	Right	44
10	Left	Female	Right	33
11	Left	Female	Right	23
12	Left	Female	Right	18
13	Left	Female	Right	61
14	Bilateral	Female	Right	33
15	Left	Female	Right	27
16	Right	Male	Right	28
17	Left	Male	Right	26
18	Bilateral	Female	Right	33
19	Bilateral	Female	Right	30

References

[1] French, R.L., DeAngelis, G.C.: Multisensory neural processing: from cue integration to causal inference. Current Opinion in Physiology **16**, 8–13 (2020) https://doi.org/10.1016/j.cophys.2020.04.004 . Accessed 2023-06-06

[2] Noppeney, U.: Perceptual Inference, Learning, and Attention in a Multisensory World. Annual Review of Neuroscience **44**(1), 449–473 (2021) https://doi.org/10.1146/annurev-neuro-100120-085519 . Accessed 2024-05-09

[3] Campbell, R.: The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1493), 1001–1010 (2007) https://doi.org/10.1098/rstb.2007.2155 . Publisher: Royal Society. Accessed 2024-02-09

[4] Cappe, C., Rouiller, E.M., Barone, P.: Multisensory anatomical pathways. Hearing Research **258**(1-2), 28–36 (2009) https://doi.org/10.1016/j.heares.2009.04.017 . Accessed 2025-01-09

[5] Gao, C., Green, J.J., Yang, X., Oh, S., Kim, J., Shinkareva, S.V.: Audiovisual integration in the human brain: a coordinate-based meta-analysis. Cerebral Cortex **33**(9), 5574–5584 (2023) https://doi.org/10.1093/cercor/bhac443 . Accessed 2024-02-02

[6] Sugihara, T., Diltz, M.D., Averbeck, B.B., Romanski, L.M.: Integration of Auditory and Visual Communication Information in the Primate Ventrolateral Prefrontal Cortex. The Journal of Neuroscience **26**(43), 11138–11147 (2006) https://doi.org/10.1523/JNEUROSCI.3550-06.2006 . Accessed 2024-02-09

[7] Noppeney, U., Ostwald, D., Werner, S.: Perceptual Decisions Formed by Accumulation of Audiovisual Evidence in Prefrontal Cortex. The Journal of Neuroscience **30**(21), 7434–7446 (2010) https://doi.org/10.1523/JNEUROSCI.0455-10.2010 . Accessed 2024-09-12

[8] Romanski, L.M.: Integration of faces and vocalizations in ventral prefrontal cortex: Implications for the evolution of audiovisual speech. Proceedings of the National Academy of Sciences **109**(supplement_1), 10717–10724 (2012) https://doi.org/10.1073/pnas.1204335109 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-02-13

[9] Bushara, K.O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., Hallett, M.: Neural correlates of cross-modal binding. Nature Neuroscience **6**(2), 190–195 (2003) https://doi.org/10.1038/nn993 . Accessed 2024-09-12

[10] Ozker, M., Yoshor, D., Beauchamp, M.S.: Frontal cortex selects representations of the talker's mouth to aid in speech perception. eLife **7**, 30387 (2018) https://doi.org/10.7554/eLife.30387 . Accessed 2023-09-07

[11] Sonkusare, S., Breakspear, M., Guo, C.: Naturalistic Stimuli in Neuroscience: Critically Acclaimed. Trends in Cognitive Sciences **23**(8), 699–714 (2019) https://doi.org/10.1016/j.tics.2019.05.004

[12] Nastase, S.A., Goldstein, A., Hasson, U.: Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. NeuroImage **222**, 117254 (2020) https://doi.org/10.1016/j.neuroimage.2020.117254 . Accessed 2024-01-03

[13] Khosla, M., Ngo, G.H., Jamison, K., Kuceyeski, A., Sabuncu, M.R.: Cortical response to naturalistic stimuli is largely predictable with deep neural networks. Science Advances **7**(22) (2021) https://doi.org/10.1126/sciadv.abe7547

[14] Setti, F., Handjaras, G., Bottari, D., Leo, A., Diano, M., Bruno, V., Tinti, C., Cecchetti, L., Garbarini, F., Pietrini, P., Ricciardi, E.: A modality-independent proto-organization of human multisensory areas. Nature Human Behaviour (2023) https://doi.org/10.1038/s41562-022-01507-3 . Accessed 2023-01-24

[15] Parvizi, J., Kastner, S.: Promises and limitations of human intracranial electroencephalography. Nature Neuroscience **21**(4), 474–483 (2018) https://doi.org/10.1038/s41593-018-0108-2

[16] Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A.Y., Tal, I., Haegens, S., Melloni, L., Knight, R.T., Schroeder, C.E.: Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. Science Advances (2020)

[17] Ray, S., Crone, N.E., Niebur, E., Franaszczuk, P.J., Hsiao, S.S.: Neural Correlates of High-Gamma Oscillations (60–200 Hz) in Macaque Local Field Potentials and Their Potential Implications in Electrocorticography. The Journal of Neuroscience **28**(45), 11526–11536 (2008) https://doi.org/10.1523/JNEUROSCI.2848-08.2008 . Accessed 2025-01-13

[18] Hermes, D., Miller, K.J., Vansteensel, M.J., Aarnoutse, E.J., Leijten, F.S.S., Ramsey, N.F.: Neurophysiologic correlates of fMRI in human motor cortex. Human Brain Mapping **33**(7), 1689–1699 (2012) https://doi.org/10.1002/hbm.21314 . Accessed 2023-05-17

[19] Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., Malach, R.: Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. Science **309**(5736), 951–954 (2005) https://doi.org/10.1126/science.1110913 . Accessed 2025-01-13

[20] Yu, L., Dugan, P., Doyle, W., Devinsky, O., Friedman, D., Flinker, A.: A left-lateralized dorsolateral prefrontal network for naming. Cell Reports (2025) https://doi.org/10.1101/2024.05.15.594403 . Accessed 2025-02-06

[21] Khalilian-Gourtani, A., Wang, R., Chen, X., Yu, L., Dugan, P., Friedman, D.,

Doyle, W., Devinsky, O., Wang, Y., Flinker, A.: A Corollary Discharge Circuit in Human Speech. Proceedings of the National Academy of Sciences **121**(50) (2024) https://doi.org/10.1101/2022.09.12.507590 . Accessed 2022-09-29

[22] Hamilton, L.S., Edwards, E., Chang, E.F.: A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Current Biology **28**(12), 1860 (2018) https://doi.org/10.1016/j.cub.2018.04.033

[23] Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E.: Encoding and Decoding Models in Cognitive Electrophysiology. Frontiers in Systems Neuroscience **11** (2017) https://doi.org/10.3389/fnsys.2017.00061

[24] Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.: Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current Biology **21**(19), 1641–1646 (2011) https://doi.org/10.1016/j.cub.2011.08.031 . Accessed 2023-12-07

[25] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in Vision: A Survey. ACM Computing Surveys **54**(10s), 1–41 (2022) https://doi.org/10.1145/3505244 . Accessed 2025-01-31

[26] Gillioz, A., Casas, J., Mugellini, E., Khaled, O.A.: Overview of the Transformer-based Models for NLP Tasks, pp. 179–183 (2020). https://doi.org/10.15439/2020F20 . https://annals-csis.org/Volume$_2 1/drp/20.htmlAccessed 2025-01-31$

[27] Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv. arXiv:2006.11477 [cs, eess] (2020). http://arxiv.org/abs/2006.11477 Accessed 2022-11-09

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. arXiv:2010.11929 [cs] (2021). http://arxiv.org/abs/2010.11929 Accessed 2023-07-17

[29] Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C.: The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience **10** (2016) https://doi.org/10.3389/fnhum.2016.00604 . Accessed 2023-01-25

[30] Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F.: Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science **343**(6174), 1006–1010 (2014) https://doi.org/10.1126/science.1245994

[31] Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C.: Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Current Biology **28**(5), 803 (2018) https:

//doi.org/10.1016/j.cub.2018.01.080

[32] Nentwich, M., Leszczynski, M., Russ, B.E., Hirsch, L., Markowitz, N., Sapru, K., Schroeder, C.E., Mehta, A.D., Bickel, S., Parra, L.C.: Semantic novelty modulates neural responses to visual change across the human brain. Nature Communications **14**(1), 2910 (2023) https://doi.org/10.1038/s41467-023-38576-5 . Accessed 2023-12-09

[33] Körding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., Shams, L.: Causal Inference in Multisensory Perception. PLoS ONE **2**(9), 943 (2007) https://doi.org/10.1371/journal.pone.0000943 . Accessed 2023-09-15

[34] Fetsch, C.R., Pouget, A., DeAngelis, G.C., Angelaki, D.E.: Neural correlates of reliability-based cue weighting during multisensory integration. Nature Neuroscience **15**(1), 146–154 (2012) https://doi.org/10.1038/nn.2983 . Accessed 2023-10-04

[35] Cao, Y., Summerfield, C., Park, H., Giordano, B.L., Kayser, C.: Causal Inference in the Multisensory Brain. Neuron **102**(5), 1076–10878 (2019) https://doi.org/10.1016/j.neuron.2019.03.043 . Accessed 2023-01-24

[36] Toneva, M., Wehbe, L.: Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., ??? (2019). https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf

[37] Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., King, J.-R.: Toward a realistic model of speech processing in the brain with self-supervised learning. 36th Conference on Neural Information Processing Systems (NeurIPS 2022) (2022). arXiv:2206.01685 [cs, q-bio]. Accessed 2022-11-17

[38] Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K.A., Devinsky, O., Hasson, U.: Shared computational principles for language processing in humans and deep language models. Nature Neuroscience **25**(3), 369–380 (2022) https://doi.org/10.1038/s41593-022-01026-4

[39] Wang, A.Y., Kay, K., Naselaris, T., Tarr, M.J., Wehbe, L.: Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nature Machine Intelligence **5**(12), 1415–1426 (2023)

https://doi.org/10.1038/s42256-023-00753-y . Accessed 2024-10-20

[40] Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Nastase, S.A., Hasson, U.: A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. Neuron, 0896627324004604 (2024) https://doi.org/10.1016/j.neuron.2024.06.025 . Accessed 2024-08-20

[41] Huntenburg, J.M., Bazin, P.L., Margulies, D.S.: Large-Scale Gradients in Human Cortical Organization. Trends in Cognitive Sciences **22**(1), 21–31 (2018) https://doi.org/10.1016/j.tics.2017.11.002

[42] Bernhardt, B.C., Smallwood, J., Keilholz, S., Margulies, D.S.: Gradients in brain organization. NeuroImage **251**, 118987 (2022) https://doi.org/10.1016/j.neuroimage.2022.118987 . Accessed 2025-04-12

[43] Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature **415**(6870), 429–433 (2002) https://doi.org/10.1038/415429a . Accessed 2024-10-01

[44] Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining Sensory Information: Mandatory Fusion Within, but Not Between, Senses. Science **298**(5598), 1627–1630 (2002) https://doi.org/10.1126/science.1075396 . Accessed 2025-03-12

[45] Rohe, T., Noppeney, U.: Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. PLOS Biology **13**(2), 1002073 (2015) https://doi.org/10.1371/journal.pbio.1002073 . Accessed 2023-09-15

[46] Kringelbach, M.L., Perl, Y.S., Tagliazucchi, E., Deco, G.: Toward naturalistic neuroscience: Mechanisms underlying the flattening of brain hierarchy in movie-watching compared to rest and task. Science Advances **9**(2), 6049 (2023) https://doi.org/10.1126/sciadv.ade6049 . Accessed 2023-04-30

[47] Ten Oever, S., Romei, V., Van Atteveldt, N., Soto-Faraco, S., Murray, M.M., Matusz, P.J.: The COGs (context, object, and goals) in multisensory processing. Experimental Brain Research **234**(5), 1307–1323 (2016) https://doi.org/10.1007/s00221-016-4590-z . Accessed 2025-01-07

[48] Boyle, S.C., Kayser, S.J., Kayser, C.: Neural correlates of multisensory reliability and perceptual weights emerge at early latencies during audio-visual integration. European Journal of Neuroscience **46**(10), 2565–2577 (2017) https://doi.org/10.1111/ejn.13724 . Accessed 2025-02-27

[49] Aller, M., Noppeney, U.: To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. PLOS Biology **17**(4), 3000210 (2019) https://doi.org/10.1371/journal.pbio.3000210 . Accessed 2025-02-19

[50] Rohe, T., Ehlis, A.-C., Noppeney, U.: The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. Nature Communications **10**(1), 1907 (2019) https://doi.org/10.1038/s41467-019-09664-2 . Accessed 2023-06-06

[51] Song, H., Finn, E.S., Rosenberg, M.D.: Neural signatures of attentional engagement during narratives and its consequences for event memory. Proceedings of the National Academy of Sciences **118**(33), 2021905118 (2021) https://doi.org/10.1073/pnas.2021905118

[52] Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., Hasson, U.: Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience **20**(1), 115–125 (2017) https://doi.org/10.1038/nn.4450

[53] Hamilton, L.S., Oganian, Y., Hall, J., Chang, E.F.: Parallel and distributed encoding of speech across human auditory cortex. Cell **184**, 1–14 (2021)

[54] Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (2007) https://doi.org/10.1016/j.neuroimage.2007.07.007 . Accessed 2025-05-06

[55] Nastase, S.A., Gazzola, V., Hasson, U., Keysers, C.: Measuring shared responses across subjects using intersubject correlation. Social Cognitive and Affective Neuroscience **14**(6), 669–687 (2019) https://doi.org/10.1093/scan/nsz037

[56] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999) https://doi.org/10.1038/44565 . Accessed 2024-04-30

[57] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987) https://doi.org/10.1016/0377-0427(87)90125-7 . Accessed 2023-11-09

[58] McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and Music Signal Analysis in Python, Austin, Texas, pp. 18–24 (2015). https://doi.org/10.25080/Majora-7b98e3ed-003 . https://doi.curvenote.com/10.25080/Majora-7b98e3ed-003 Accessed 2025-02-13

[59] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30. Advances in Neural Information Processing Systems, vol. 30. Neural Information Processing Systems (Nips), La Jolla (2017). ://WOS:000452649406008

[60] Vaidya, A.R., Jain, S., Huth, A.G.: Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech. Proceedings of the 39 th

International Conference on Machine Learning (2022)

[61] Zhou, Q., Du, C., He, H.: Exploring the Brain-like Properties of Deep Neural Networks: A Neural Encoding Perspective. Machine Intelligence Research **19**(5), 439–455 (2022) https://doi.org/10.1007/s11633-022-1348-x

[62] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv **abs/1907.11692** (2019)

[63] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://www.aclweb.org/anthology/N19-1423 http://dx.doi.org/10.18653/v1/N19-1423

[64] Anderson, A.J., Kiela, D., Binder, J.R., Fernandino, L., Humphries, C.J., Conant, L.L., Raizada, R.D.S., Grimm, S., Lalor, E.C.: Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning. The Journal of Neuroscience **41**(18), 4100 (2021) https://doi.org/10.1523/JNEUROSCI.1152-20.2021

[65] Huang, J., Tang, D., Zhong, W., Lu, S., Shou, L., Gong, M., Jiang, D., Duan, N.: WhiteningBERT: An Easy Unsupervised Sentence Embedding Approach. arXiv. arXiv:2104.01767 [cs] (2021). http://arxiv.org/abs/2104.01767 Accessed 2024-01-17

[66] Yu, L., Ettinger, A.: Assessing Phrasal Representation and Composition in Transformers. arXiv. arXiv:2010.03763 [cs] (2020). http://arxiv.org/abs/2010.03763 Accessed 2024-02-06

[67] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. arXiv:1912.01703 [cs] (2019). https://doi.org/10.48550/arXiv.1912.01703 . http://arxiv.org/abs/1912.01703 Accessed 2025-02-13

[68] Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I.: Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience **12**(5), 535–540 (2009) https://doi.org/10.1038/nn.2303 . Accessed 2024-01-19

[69] Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171. IEEE, Minneapolis, MN, USA (2011). https://doi.org/10.1109/ICDCSW.2011.20 . http://ieeexplore.ieee.org/document/5961514/ Accessed 2023-06-01