Beyond the Canonical HRF: Flexible Temporal Modeling Reveals Feature-Specific BOLD Profiles During Naturalistic Viewing

Xin Di [1*], George B. Hanna [1], Bharat B. Biswal [1*]

1. Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, 07102, USA

* Corresponding author:
Xin Di, Ph.D.
604 Fenster Hall, University Height
Newark, NJ, 07102, USA
xin.di@njit.edu

Bharat B. Biswal, Ph.D.
607 Fenster Hall, University Height
Newark, NJ, 07102, USA
bbiswal@yahoo.com

Abstract

Naturalistic stimuli, such as movies and narratives, are increasingly used in cognitive neuroscience to map cognitive and affective processes onto brain activity measured with functional MRI (fMRI). Features extracted from movies span multiple levels, from computational descriptions of visual and auditory input to physiological signals and subjective ratings. However, the temporal alignment between these features and the blood-oxygen-level-dependent (BOLD) response can vary considerably, and the commonly used canonical hemodynamic response function (HRF) with temporal derivatives may not adequately capture these delays. In this study, we analyzed three movie-watching datasets and examined a range of features, including visual luminance and contrast, auditory pitch, a physiological measure of pupil size, and subjective ratings of theory of mind. We performed cross-correlation analyses and finite impulse response (FIR) deconvolution between feature time series and fMRI responses across the whole brain. Our results show that the canonical HRF captures sensory features well, but misaligns with slower signals, such as pupil size and subjective ratings. Notably, raw pupil size and subjective ratings exhibited stronger correspondence with delayed hemodynamic responses. Lastly, using FIR deconvolution, we systematically characterized the true BOLD response profiles of these diverse features, revealing distinct inter-regional temporal hierarchies and peak latencies that systematically deviate from canonical assumptions. These findings highlight the critical role of temporal alignment in mapping naturalistic features to brain activity. Accounting for feature-specific delays using flexible temporal modeling may improve the interpretability and accuracy of fMRI studies using complex, real-world stimuli.

Keywords: Blood-oxygen-level dependent, Finite Impulse Response, Deconvolution, Naturalistic stimuli, Subjective rating, System response function, Timing.

1. Introduction

Naturalistic stimuli — such as movies or spoken narratives — have become increasingly popular in functional MRI (fMRI) research because they offer tools to investigate brain function under ecologically valid conditions (Di et al., 2026; Hasson et al., 2004; Vanderwal et al., 2019). Unlike traditional, highly controlled experimental paradigms, naturalistic stimuli capture the richness and temporal continuity of real-world experiences, thereby engaging a wide range of perceptual, cognitive, and affective processes simultaneously. This shift has opened new avenues for studying how the brain integrates complex, dynamic information over time.

A seminal line of work demonstrated that when multiple individuals watch the same movie or listen to the same story, their brain activity exhibits significant inter-subject correlations (ISC), revealing shared neural responses to naturalistic events (Hasson et al., 2004). Such data-driven analyses have been instrumental in identifying brain regions involved in high-level perception, attention, and social cognition (Kauppi et al., 2010). However, while these approaches effectively capture synchronized patterns of activity across individuals, they do not directly reveal which specific features of the stimulus drive these neural responses. To address this, researchers have increasingly focused on extracting quantitative features from naturalistic stimuli and mapping their temporal dynamics onto brain activity measured by fMRI (Bartels et al., 2008; Bartels and Zeki, 2004; Huth et al., 2016; Lahnakoski et al., 2012a; Raz et al., 2016).

A wide variety of stimulus-derived features can be used for such mapping. These range from low-level sensory features — such as visual luminance, contrast, and auditory pitch — to higher-level variables derived from physiological signals (e.g., pupil size) or subjective behavioral ratings (e.g., perceived emotional valence or theory of mind, ToM). In recent years, feature extraction has also benefited from advances in deep neural networks, where model-based representations capture complex visual, auditory, or semantic content (McMahon et al., 2023; Samara et al., 2025; Sohn et al., 2024). These features can be analyzed individually

using general linear models (GLMs) or jointly in encoding-model frameworks to predict regional brain activity.

Crucially, naturalistic stimuli are not monolithic; they contain deeply nested, hierarchical temporal structures. For example, visual streams progress from rapid, high-frequency shifts in luminance to slower transitions between objects and scenes, while auditory streams build from transient phonemes to extended narratives. A prominent theoretical framework posits that the brain processes these nested features through a topography of hierarchical cortical timescales, or temporal receptive windows (TRWs) (Lerner et al., 2011; Meer et al., 2020). In this framework, primary sensory regions possess short TRWs to process rapid physical properties, whereas higher-order associative networks possess long TRWs to integrate semantic and narrative information over extended durations.

Beyond extrinsic stimulus features, naturalistic viewing also engages internally generated, observer-dependent dynamics. Physiological responses, such as pupil size, act as markers of autonomic arousal and cognitive load, reflecting an integration of sensory input and internal state (Joshi and Gold, 2020). Similarly, subjective behavioral ratings—such as ToM annotations—capture conscious, high-level social-cognitive evaluations of characters' mental states over time (Saarimäki, 2021). As illustrated in our conceptual framework (Figure 1), appropriately mapping naturalistic viewing to brain activity requires integrating three distinct temporal levels: 1) the nested intrinsic hierarchies of the stimulus features; 2) the biological and cognitive integration delays inherent to the observer's physiological and behavioral responses; and 3) the region-dependent hemodynamic brain responses.

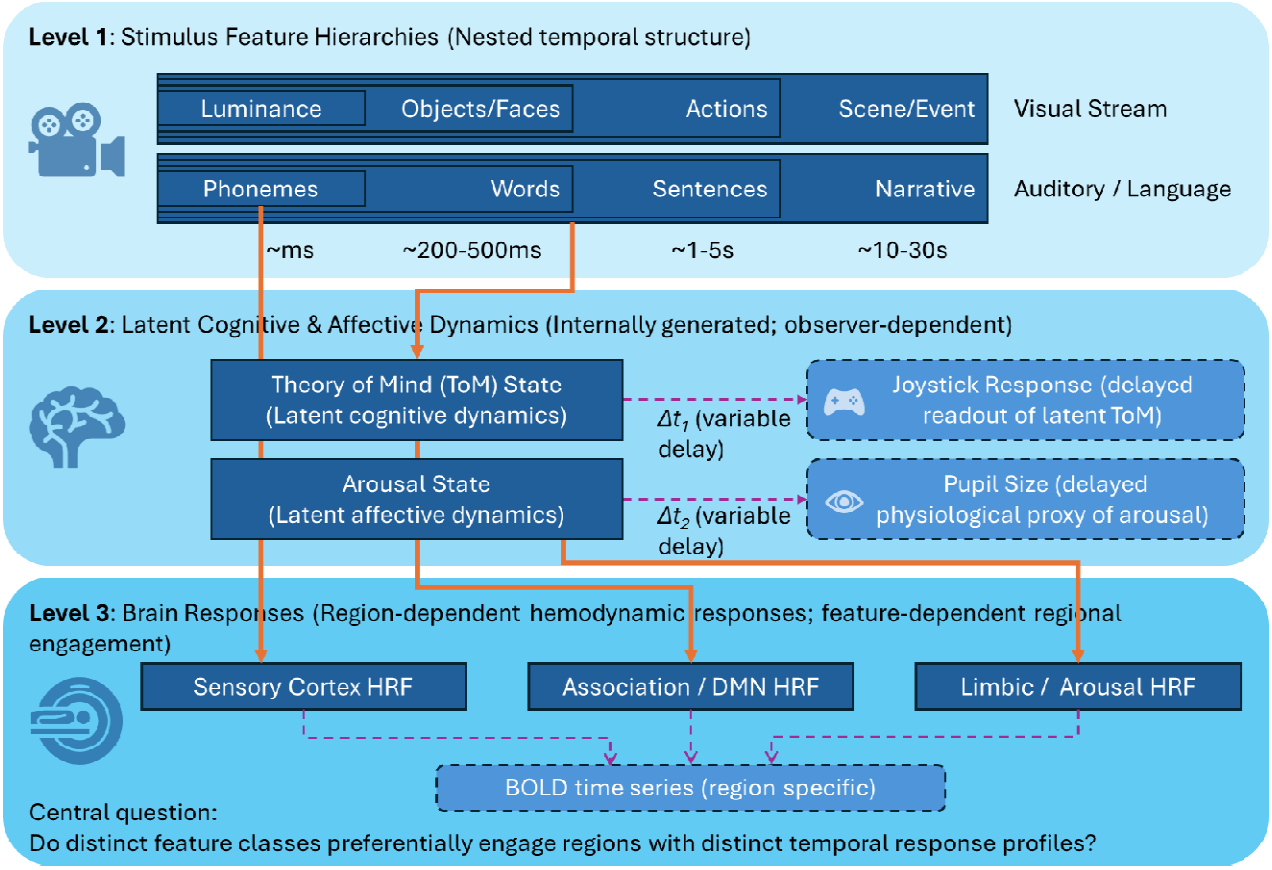


Figure 1. Conceptual framework of nested temporal hierarchies and their mapping to brain responses during naturalistic viewing.

Level 1: Intrinsic stimulus hierarchies. Naturalistic movies possess a deeply nested temporal structure across modalities. Visual features (e.g., luminance, objects, actions, scenes) and auditory/language features (e.g., pitch, phonemes, words, narratives) unfold over timescales ranging from milliseconds to tens of seconds. The brain processes these continuous streams through a topography of temporal receptive windows (TRWs), ranging from short integration windows in primary sensory areas to extended timescales in higher-order networks.

Level 2: Observer-dependent delays. Extrinsic stimulus features elicit internally generated latent cognitive and affective dynamics, which are not directly observable. Higher-order states, such as Theory of Mind (ToM) engagement and autonomic arousal, evolve continuously and are captured via delayed behavioral (joystick ratings) and physiological (pupil size) readouts. The

integration and motor/autonomic delays between the true latent neural states and these observable proxies (Δt_1, Δt_2) introduce critical temporal misalignments.

Level 3: Feature-specific BOLD dynamics. These multi-level features engage distinct brain regions, each possessing specific neurovascular coupling properties. Because different feature classes preferentially recruit different neural systems (e.g., sensory, associative/Default Mode, and limbic networks) and operate on different intrinsic timescales, the resulting BOLD signals exhibit highly diverse temporal response profiles. Consequently, mapping these features to fMRI requires flexible temporal modeling, as applying a uniform canonical hemodynamic response function (HRF) to the already-delayed readouts from Level 2 risks compounding delays.

Note: Solid arrows indicate intrinsic neural propagation and hypothesized pathways linking stimulus features to latent cognitive states and regional brain responses. Dashed arrows indicate the delayed measurement of these latent states via observable behavioral or physiological proxies. Δt denotes the variable temporal delays inherent to these readouts.

A critical challenge in linking these features to fMRI data lies in the temporal mismatch between stimulus features and the hemodynamic response. The fMRI signal—via the blood-oxygen-level-dependent (BOLD) effect—typically lags behind underlying neural activity by about 5–6 seconds (Friston et al., 1998; Handwerker et al., 2004). To account for this, most studies either apply a fixed delay (e.g., 4 s) or convolve the feature time-series with a canonical hemodynamic response function (HRF). Some approaches further include temporal derivatives of the HRF to allow limited flexibility in response timing (i.e., ≈1 s adjustment). However, the shape and latency of the HRF vary considerably across brain regions, across individuals, and even within the same region under different physiological conditions (Aguirre et al., 1998; Handwerker et al., 2004). Such intrinsic biological variability complicates efforts to accurately align continuous stimulus features with corresponding neural activity.

This temporal uncertainty becomes especially pronounced when analyzing multiple types of features that occupy different levels of the naturalistic processing hierarchy. Applying a uniform canonical HRF across all feature classes fundamentally ignores the diverse intrinsic frequencies and delays inherent to these distinct signals. Fast sensory inputs, delayed physiological signals, and slow cognitive annotations possess vastly different temporal relationships to neural and vascular processes. Because pupil size and subjective behavioral ratings are themselves delayed responses relative to the eliciting neural activity, convolving them with a standard HRF risks compounding delays. This conventional approach can result in "double-smoothing" temporal artifacts that obscure true brain-behavior coupling (Lloyd et al., 2023).

To overcome the limitations of the canonical HRF and accurately map hierarchical processing timescales, more flexible temporal modeling is required. Finite Impulse Response (FIR) deconvolution provides a robust, non-parametric alternative (Glover, 1999). Rather than imposing a predetermined temporal shape and fixed latency, FIR models utilize a set of time-lagged predictors to directly estimate the true, feature-specific system response function from the continuous fMRI data. By allowing the response shape and peak latency to vary freely, FIR deconvolution can capture the diverse temporal profiles of naturalistic processing—ranging from rapid sensory responses to the extended, diffuse hemodynamic delays that characterize slower, higher-order cognitive features.

In the present study, we analyzed three movie-watching fMRI datasets to systematically characterize how different classes of stimulus, physiological, and behavioral features are expressed in their true BOLD response shapes. We extracted continuous features spanning the hierarchical levels of naturalistic processing: low-level sensory features (frame luminance, frame contrast, and auditory pitch), a physiological measure (pupil size), and a subjective behavioral measure (ToM ratings). Moving beyond conventional cross-correlation analyses, we applied formal FIR deconvolution to estimate the BOLD system response functions across the whole

brain. We hypothesize that these diverse features will exhibit distinct response profiles that systematically deviate from the canonical HRF, and that quantifying the peak latencies of these responses will empirically reveal the underlying inter-regional temporal hierarchies involved in naturalistic viewing.

2. Materials and Methods

2.1. Movie-watching fMRI Data

We analyzed three movie-watching fMRI datasets previously used in our work (Di et al., 2023; Di and Biswal, 2022): two from the Healthy Brain Network (HBN) project (Alexander et al., 2017), and one from the Partly Cloudy dataset (Richardson et al., 2018).

2.1.1. Healthy Brain Network (HBN) Dataset

The HBN dataset was obtained from the project repository (http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/). For the current study, we included a subset of participants without diagnosed psychiatric disorders who underwent fMRI scanning while watching two video clips: *The Present* (3 min 21 s; Filmakademie Baden-Wuerttemberg, 2014) and a 10-minute segment from *Despicable Me* (Illumination, 2010).

In total, 87 participants watched *The Present* and 83 watched *Despicable Me*, with 66 participants overlapping, yielding 104 unique individuals (61 males, 43 females; age range = 5.0–21.9 years, mean = 12.0 ± 4.1). Additional inclusion and exclusion details are provided in Di et al. (2023).

MRI data were collected across two imaging centers—Rutgers University Brain Imaging Center (RUBIC) using a 3T Siemens Trio scanner, and the Citigroup Biomedical Imaging Center (CBIC) using 3T Siemens Prisma scanners. Scanning protocols were harmonized across sites. Functional MRI parameters were: TR = 800 ms, TE = 30 ms, flip angle = 31°, voxel size = 2.4 × 2.4 × 2.4 mm^3, and multiband acceleration factor = 6. Each run contained 250 volumes for *The Present* and 750 volumes for *Despicable Me*. High-resolution T1-weighted anatomical scans

were acquired using either Human Connectome Project (HCP) or Adolescent Brain Cognitive Development (ABCD) protocols and were used solely for preprocessing. Additional acquisition details can be found on the HBN website and in Alexander et al. (2017).

2.1.2. Partly Cloudy Dataset

The *Partly Cloudy* dataset was accessed from OpenNeuro (https://openneuro.org/datasets/ds000228; Richardson et al., 2018). It included two groups: adults (n = 29; 17 females, 12 males; age range = 18–39 years, mean = 24.6 ± 5.3) and children (n = 53; 28 females, 25 males; age range = 3.5–12.3 years, mean = 7.0 ± 2.5).

Participants watched a silent version of the Pixar short film *Partly Cloudy* (5.6 minutes; https://www.pixar.com/partly-cloudy) during scanning on a 3T Siemens Tim Trio system. Younger children were scanned using one of two 32-channel custom head coils, while older children and adults used the standard Siemens 32-channel coil. Functional images were acquired using a gradient-echo EPI sequence (TR = 2000 ms, TE = 30 ms, flip angle = 90°) with 32 interleaved near-axial slices (EPI factor = 64).

Voxel sizes and slice gaps varied slightly across subgroups: (a) 3.13 mm isotropic, no gap; (b) 3.13 mm isotropic, 10% gap; (c) 3 mm isotropic, 20% gap; and (d) 3 mm isotropic, 10% gap. All functional images were resampled to 3 mm isotropic voxels during preprocessing. Each run contained 168 volumes, with the first four discarded to allow for steady-state magnetization. T1-weighted anatomical images were collected in 176 interleaved sagittal slices (1 mm isotropic; GRAPPA = 3; FOV = 256 mm). Further methodological details are provided in Richardson et al. (2018).

2.2. Eye tracking data

Although eye-tracking data were not collected during the MRI sessions, a subset of participants from the HBN project participated in a separate EEG session conducted on a different day

outside the scanner (Langer et al., 2017). During this session, participants watched several videos, including *The Present*, a shorter segment of *Despicable Me*, *Fun with Fractals* (an educational video), and a trailer for *Diary of a Wimpy Kid*, while simultaneous EEG and eye-tracking data were recorded. Eye position and pupil size were measured using an infrared video-based eye tracker (iView-X Red-m, SensoMotoric Instruments GmbH) operating at sampling rates of 30, 60, or 120 Hz. In the current study, only pupil size data were analyzed. All available recordings were included, but participants were excluded if more than 10% of time points were missing, typically due to blinks or signal loss. The final sample consisted of 1,312 participants (665 males, 387 females, 260 sex not reported), ranging in age from 5 to 21 years (mean = 9.94, SD = 3.24).

2.3. Subjective Rating Data Collection

Subjective rating data were collected from a separate group of participants at the New Jersey Institute of Technology (NJIT) to capture continuous, moment-to-moment fluctuations in perceived mental states of movie characters. A total of 34 participants (age range: 18–27 years, mean ± SD = 20.9 ± 2.3 years; 17 female) took part in the study. All participants reported normal or corrected-to-normal vision and hearing, and none had a history of neurological or psychiatric disorders. Informed consent was obtained from all participants in accordance with the Institutional Review Board (IRB) of NJIT.

Participants viewed the same three video clips used in the fMRI experiment. Each session took place in a quiet, darkened room to minimize visual and auditory distractions. Stimuli were presented on a computer monitor with synchronized sound through a speaker.

Continuous subjective ratings were acquired using CARMA software (Continuous Affect Rating and Media Annotation; (Girard, 2014) with a gaming joystick as the input device. Participants were instructed to continuously indicate the extent to which they could perceive the characters' emotions, thoughts, or intentions while watching each clip. This measure was

designed to capture participants' theory-of-mind engagement—that is, the degree to which they inferred or empathically understood the internal mental states of the characters throughout the narrative. Before the experiment, participants completed a short practice session to familiarize themselves with the rating procedure.

The joystick's vertical position corresponded to a continuous numerical scale ranging from -10 (not at all) to 10 (very strongly), representing the perceived strength of mental-state understanding at each time point. Rating data were sampled at 4 Hz (every 250 ms) to provide a high-resolution temporal record of evolving subjective experience. Between clips, participants were given brief rest periods to minimize fatigue.

2.4. Video feature extraction

We extracted three sensory features from the videos: mean pixel intensity, standard deviation of pixel intensity, and pitch. For the two visual features, each video frame was first converted to grayscale, and the mean and standard deviation of pixel intensity across the two-dimensional matrix were computed. These yielded two time series corresponding to frame luminance and frame contrast, respectively. For the auditory feature, audio signals from all channels were averaged, and pitch was computed using a 100-ms window with no overlap, resulting in a 10 Hz pitch time series. Pitch was extracted only for *The Present* and *Despicable Me*, as the *Partly Cloudy* video contains no sound. Figure 2 shows the time series of mean and standard deviations of pixel intensity and pitch for the video *The Present*.

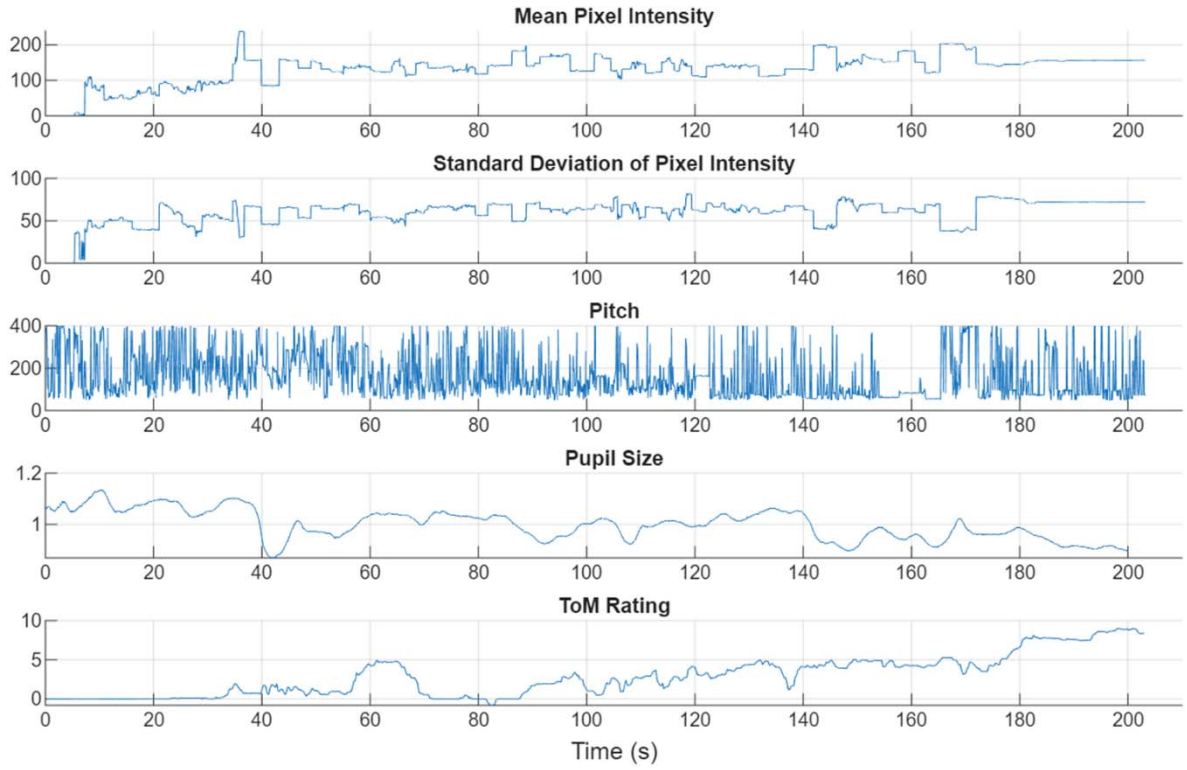


Figure 2. Continuous time series of multimodal features associated with the movie *The Present*. The waveforms illustrate the diverse macroscopic temporal properties across the five feature classes. Low-level sensory features (frame luminance, frame contrast, and auditory pitch) exhibit rapid, high-frequency fluctuations directly tied to instantaneous stimulus changes. In contrast, the physiological (pupil size) and subjective cognitive (Theory of Mind [ToM] rating) measures operate on noticeably slower, lower-frequency timescales, reflecting underlying biological and cognitive integration delays.

2.5. Pupil Size Data Processing

Pupil size data for each participant were first resampled to a uniform sampling rate of 30 Hz. Missing values (NaNs) were interpolated linearly, and blink periods—identified as time points with zero values—were also replaced using linear interpolation. Participants with more than 10% of time points classified as blinks were excluded from further analyses. To reduce inter-

individual variability in absolute pupil size, each participant's pupil time series was normalized by dividing by its mean.

To evaluate the consistency of pupil size fluctuations across participants, we computed leave-one-out inter-individual correlations. For each participant, their pupil size time series was correlated with the mean time series of all other participants, resulting in one correlation coefficient per individual. The statistical significance of the group-level mean correlation was assessed using a circular time-shift randomization procedure with 1,000 iterations to generate a null distribution (Kauppi et al., 2010). Finally, the median pupil size time series across participants was extracted and used as the representative group-level signal for subsequent analyses (Figure 2).

2.6. Rating Data Processing

The subjective rating data were analyzed using the same leave-one-out inter-individual correlation approach described above. For each video clip, the median rating time series across participants was computed and used as the group-level representative signal for further analyses (Figure 2 shows an example).

2.7. Temporal characteristics of multimodal features

Across the five selected features, we observed distinct temporal and frequency characteristics that correspond to their respective levels in the naturalistic processing hierarchy. Low-level sensory features, such as frame luminance and auditory pitch, exhibited highly transient, high-frequency fluctuations (sampled at video frame rates or 10 Hz, respectively) reflecting rapid, instantaneous changes in the external physical stimulus. In contrast, frame contrast exhibited slightly slower dynamic shifts, corresponding to sustained visual scene characteristics. Finally, the physiological (median pupil size) and behavioral (median ToM rating) features operated on markedly slower, lower-frequency timescales. These differences highlight that higher-level

features inherently incorporate sluggish biological constraints (e.g., autonomic nervous system delays in pupillary constriction) and higher-order cognitive integration over extended narrative windows. Recognizing these diverse intrinsic frequency profiles is critical, as applying a uniform, fixed-delay HRF to all features risks severe temporal misalignment, directly motivating our subsequent use of flexible FIR deconvolution.

2.8. MRI Data Processing

Preprocessing of the fMRI data was performed separately for the HBN and *Partly Cloudy* datasets, following procedures described in previous studies. Here, we provide a brief overview. Structural MRI images for each participant were first segmented and normalized to the standard Montreal Neurological Institute (MNI) space. For *The Present* and *Despicable Me* datasets, the first 10 functional volumes were discarded, resulting in 240 and 740 volumes, respectively. In contrast, all functional volumes were retained for the Partly Cloudy dataset. The remaining functional images were then realigned to the first retained image, coregistered to the participant's structural image, normalized to MNI space using the deformation fields obtained during segmentation, and spatially smoothed with an 8 mm full-width-at-half-maximum (FWHM) Gaussian kernel.

Finally, a general linear model (GLM) was constructed to regress out head motion effects using Friston's 24-parameter motion model, which also incorporated an implicit high-pass filter with a cutoff frequency of 1/128 Hz. The resulting residual time series were used in all subsequent analyses.

We adopted a region-of-interest (ROI)–based approach to achieve a balance between spatial specificity and temporal flexibility. Cortical regions were defined using the Schaefer 100-region parcellation (Schaefer et al., 2018), and subcortical regions were defined using the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). Fourteen bilateral subcortical nuclei were included: hippocampus, parahippocampus, amygdala, caudate,

putamen, pallidum, and thalamus. The Schaefer parcellation was divided into left (#1–50) and right (#51–100) hemispheres and further organized into seven large-scale functional networks: visual, somatomotor, dorsal attention, salience/ventral attention, limbic, control, and default mode.

In total, 114 ROIs were analyzed. For each preprocessed fMRI run, the mean time series was extracted from each ROI, resulting in data matrices of size 240 × 114 for *The Present*, 740 × 114 for *Despicable Me*, and 168 × 114 for *Partly Cloudy* for each participant.

2.9. Cross-correlation Analysis

Each feature time series from the three movie clips was first convolved with the canonical HRF implemented in SPM12 (using the default parameters: peak delay = 6 s, undershoot delay = 16 s, and ratio = 6). The convolved feature time series were then resampled to match the temporal resolution of the corresponding fMRI data (TR = 0.8 s for *The Present* and *Despicable Me*; TR = 2.0 s for *Partly Cloudy*).

Subsequently, cross-correlation analyses were performed between the HRF-convolved feature time series and the mean fMRI time series from each ROI for each participant. Correlations were computed across a lag range of ±16 seconds (i.e., ±20 TRs for HBN data and ±8 TRs for *Partly Cloudy*), resulting in a lag-by-ROI correlation matrix for each participant. For visualization and group-level interpretation, these matrices were averaged across participants after Fisher's *z* transformation of the correlation coefficients.

In addition, a regression analysis was conducted using the HRF-convolved feature time series as predictors in a general linear model (GLM) framework for each participant. Group-level statistical significance of the resulting parameter estimates was assessed using one-sample *t*-tests across participants, which corresponds to testing the cross-correlation at zero lag. Multiple comparisons across the 114 ROIs were controlled using the false discovery rate (FDR) at $p < 0.05$.

To further examine temporal dependencies beyond the canonical HRF assumption, we also repeated the cross-correlation and zero-lag *t*-tests using the *raw* (unconvolved) time series of pupil size and theory-of-mind (ToM) ratings. Finally, we performed a simple cross-correlation analysis between frame luminance and the group-level median pupil size time series for *The Present* video clip to explore direct visual–physiological coupling.

2.10. Finite Impulse Response Deconvolution

We performed FIR deconvolution to estimate the system response function between an input signal ($x(t)$) and an output signal ($y(t)$). Specifically, we first examined the pupillary system response using frame luminance as the input ($x(t)$) and pupil size as the output ($y(t)$). Subsequently, we applied the same approach to estimate the HRFs in different brain regions using the various video-related features described above as input signals and fMRI time series from each ROI as outputs.

In general, the output signal can be modeled as the convolution of the input signal with an unknown system response function:

$$y(t) = h(t) \otimes x(t) + \varepsilon(t)$$

where ($h(t)$) denotes the system's impulse response function, $\otimes$ represents the convolution operator, and ($\varepsilon(t)$) is the residual noise term.

The goal of FIR deconvolution is to estimate ($h(t)$) directly from the observed time series. We modeled ($h(t)$) as a finite-duration impulse response with a window of 20 seconds (equivalent to 25 TRs for *The Present* and *Despicable Me*, and 10 TRs for *Partly Cloudy*). The deconvolution was implemented using linear regression, where the input signal was expanded into a time-lagged design matrix containing shifted versions of ($x(t)$) at each time lag. The estimated regression coefficients correspond to the discrete samples of ($h(t)$).

To improve numerical stability and reduce overfitting, we first applied an autoregressive AR(1) pre-whitening step, followed by second-difference regularization with a penalty term (λ = 10). Results were qualitatively consistent across robust to a range of λ values.

For the pupil–luminance analysis, FIR deconvolution was performed on the median pupil size time series. Both the pupil and luminance signals were z-scored before analysis to remove baseline offsets and normalize amplitudes.

For the fMRI data, FIR analysis was conducted separately for each participant and ROI. Prior to deconvolution, input features were resampled to match the fMRI sampling rate, and both input and output signals were z-scored.

To determine the statistical significance of the estimated FIR models, we utilized an *F*-test at the individual participant level to evaluate the overall variance explained by the full set of time-lagged predictors. To perform group-level statistical inference, we applied Fisher's method to aggregate the individual-level *p*-values across all participants for each ROI (Fisher, 1925). The resulting group-level *p*-values were then corrected for multiple comparisons across the 114 ROIs using the FDR at $p < 0.05$, allowing us to systematically identify regions with significant FIR effects.

For regions demonstrating significant group-level responses, we obtained group-level response functions by averaging the individual FIR estimates after temporal alignment. To accurately quantify temporal hierarchies, we characterized the temporal dynamics by extracting the delay of the peak response (time-to-peak) for each ROI. To ensure that this metric captured primary stimulus-driven activations rather than secondary signal dips or undershoots, the peak delay was identified by extracting the maximum amplitude strictly from the positive values of the estimated FIR response function. These resulting temporal profiles and peak latencies were then evaluated to compare physiological and neural dynamics across brain systems.

3. Results

3.1. Sensory features: Temporal dynamics and spatial selectivity

Across all three movies, cross-correlation profiles exhibited a wide temporal spread across multiple lags (Supplementary Figure S1). This extended temporal footprint likely reflects the high inherent autocorrelation of both continuous naturalistic stimuli and the sluggish BOLD signal. Despite this temporal smearing, the strongest zero-lag alignments reflected classical functional specialization, localizing primarily to modality-specific sensory cortices.

Crucially, however, extended lag profiles revealed distinct temporal architectures depending on the complexity of the sensory feature. While rudimentary features like frame luminance exhibited tightly restricted spatial and temporal profiles mainly in the visual areas (Figure 3A), computationally richer features—specifically frame contrast and auditory pitch—demonstrated organized temporal delay structures propagating into broader cortical networks (Figure 4A). These systematic temporal offsets across non-primary associative regions indicate that fixed, uniform HRF models are insufficient for capturing the sequential dynamics of complex sensory processing. Consequently, we applied FIR deconvolution to formally quantify these temporal delays without imposing rigid hemodynamic assumptions.

3.1.1. Luminance processing aligns with canonical assumptions

For frame luminance, prominent correlation peaks near zero lag suggested that the canonical HRF adequately approximates its hemodynamic delay. Consistent with prior literature, robust zero-lag correlations localized to primary and secondary visual cortices across all three stimuli, confirming the expected coupling between luminance fluctuations and early visual processing (Figure 3B).

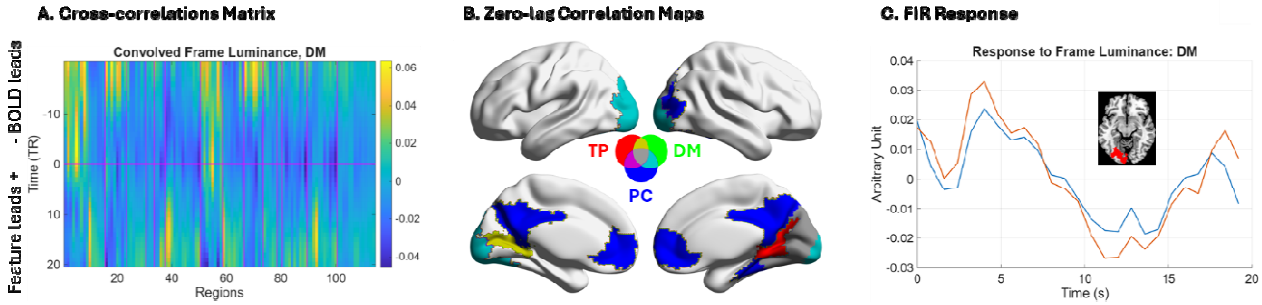


Figure 3. Neural correlates of frame luminance and temporal response dynamics. (A) Cross-correlation matrix illustrating the relationship between canonical HRF-convolved frame luminance and fMRI time series across 114 regions of interest (ROIs) for the movie *Despicable Me* (DM), plotted across multiple temporal lags. **(B)** Cortical maps displaying significant positive zero-lag correlations between canonical HRF-convolved frame luminance (mean pixel intensity) and fMRI activity for *The Present* (TP), DM, and *Partly Cloudy* (PC). Statistical maps in panel B are thresholded at a false-discovery rate (FDR) of $p < 0.05$. **(C)** Estimated FIR response functions for ROIs exhibiting significant effects in the movie DM. The inset highlights the characteristic hemodynamic profiles of two primary visual regions.

In contrast, FIR deconvolution yielded a more restricted set of significant regions compared to the zero-lag analysis (Supplementary Figure S2). This disparity likely reflects a reduction in statistical power inherent to the unconstrained FIR model relative to the more parsimonious single-regressor correlation approach. Nevertheless, significant FIR effects within the visual cortex during *Despicable Me* produced a recovered response function closely resembling the canonical hemodynamic profile (Figure 3C). Collectively, these results indicate that for rudimentary sensory features, the standard HRF model remains methodologically sufficient and potentially more sensitive.

3.1.2. Visual contrast and the complexity of naturalistic narrative

Initial zero-lag correlation analyses revealed that frame contrast elicited a predominantly modality-specific spatial profile, localized primarily within visual cortices across the stimuli (Figure 4B). However, cross-correlation profiling across extended lags demonstrated that naturalistic visual contrast is not processed in strict isolation. Specifically, the cross-correlation matrix for *The Present* (TP) revealed a widespread footprint extending well beyond primary sensory areas (Figure 4A), suggesting the sequential engagement of higher-order functional networks.

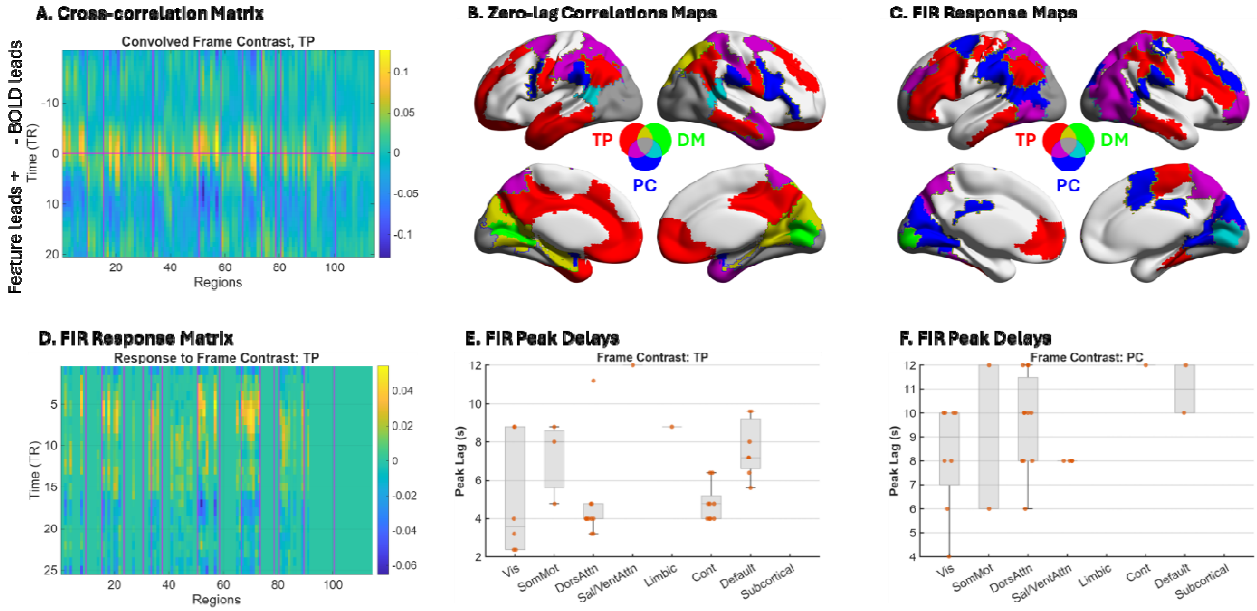


Figure 4. Neural correlates of frame contrast and associated temporal response dynamics. (A) Cross-correlation matrix illustrating the relationship between canonical HRF-convolved frame contrast and fMRI time series across 114 regions of interest (ROIs) for the movie *The Present* (TP) across multiple temporal lags. **(B)** Cortical maps displaying significant positive zero-lag correlations between canonical HRF-convolved frame contrast and fMRI activity for TP, *Despicable Me* (DM), and *Partly Cloudy* (PC). **(C)** Cortical maps illustrating significant regional responses to frame contrast estimated via Finite Impulse Response (FIR) deconvolution. **(D)** Estimated FIR response functions across all 114 ROIs during movie TP, reflecting the diversity of hemodynamic profiles. **(E, F)** Distributions of estimated temporal

delays (time-to-peak) for regions exhibiting positive FIR responses during movie TP (E) and PC (F), categorized across Yeo's seven functional networks. All statistical maps (Panels B and C) are thresholded at a false-discovery rate (FDR) of $p < 0.05$.

To evaluate this widespread temporal engagement, we applied FIR deconvolution (Figure 4C; Supplementary Figure S3). While the spatial overlap of significant FIR effects across all three movies was constrained to primary visual regions—largely due to a more limited spatial response during *Despicable Me*—the unconstrained FIR model successfully captured a distinct temporal architecture across the broader cortex for both *The Present* and *Partly Cloudy*.

The estimated FIR response functions exhibited a clear, organized temporal progression across varying ROIs (Figure 4D). To quantify this temporal hierarchy, we extracted positive peak response latencies and examined their distributions across Yeo's seven functional networks (Figure 4E, 4F). Both stimuli demonstrated a systematic hierarchical delay structure: visual networks consistently exhibited the shortest response latencies, whereas progressive delays were observed in higher-order networks, culminating in the longest latencies within the somatomotor, dorsal attention, ventral attention/salience, and default mode networks.

3.1.3. Auditory pitch and stimulus collinearity

Consistent with visual features, zero-lag correlation analysis demonstrated that auditory pitch elicited strong, modality-specific alignments within core auditory cortices. However, during *Despicable Me*, these significant correlations extended substantially into broader cortical networks (Figure 5A).

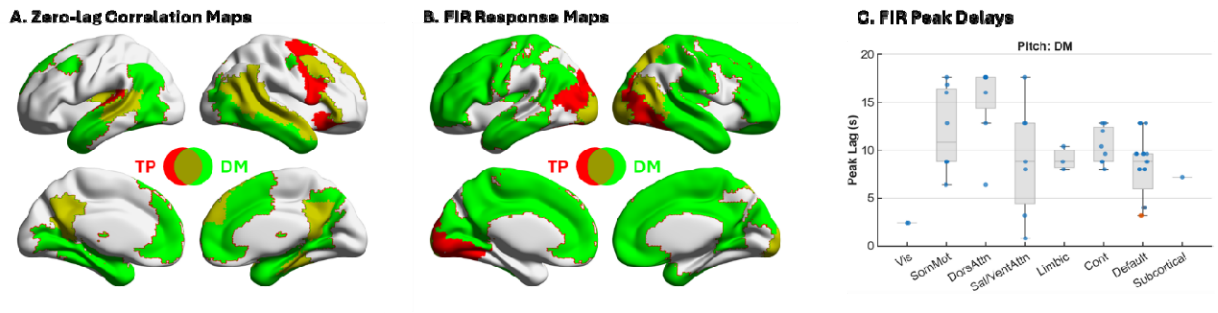


Figure 5. Neural correlates of auditory pitch and temporal delay distributions. (A) Cortical maps displaying positive zero-lag correlations between the canonical HRF-convolved auditory pitch and fMRI activity during *The Present* (TP) and *Despicable Me* (DM). **(B)** Cortical maps illustrating significant regional responses to auditory pitch estimated via Finite Impulse Response (FIR) deconvolution for both videos. Statistical maps in panels A and B are thresholded at a false-discovery rate (FDR) of $p < 0.05$. **(C)** Distributions of estimated temporal delays (time-to-peak) for regions exhibiting positive FIR responses, categorized across Yeo's seven functional networks. Red markers denote two core auditory regions that are assigned to the Default Mode Network in the Yeo parcellation.

FIR deconvolution revealed complex, stimulus-dependent temporal response profiles (Supplementary Figure S4). During *The Present*, regions exhibiting significant FIR effects were unexpectedly localized predominantly to visual cortices (Figure 5B). Extracted response functions in these visual areas revealed a distinct negative signal deflection within the first 5 seconds post-stimulus. This initial negative temporal profile indicates that auditory pitch is likely inversely correlated with concurrent salient visual features specific to this stimulus, highlighting cross-modal stimulus covariation rather than direct auditory processing in the visual cortex.

Conversely, FIR deconvolution for *Despicable Me* identified widespread significant cortical effects distributed across multiple functional networks (Figure 5B). Mirroring the temporal architecture observed for frame contrast, the latencies of positive peak responses to

auditory pitch exhibited a clear hierarchical progression. The shortest latencies localized to core auditory regions, followed by sequentially increasing delays systematically progressing through the dorsal attention, somatomotor, ventral attention/salience, limbic, executive control, and ultimately the default mode networks. (Note: Core auditory cortices are assigned to the default mode network boundary in the standard Yeo parcellation based on resting-state data; to maintain anatomical precision, these specific nodes are denoted with red markers in Figure 5C).

3.2. Pupil size: Biological delays and autonomic network coupling

The median pupil size demonstrated high inter-individual consistency (r = 0.62). However, its relationship with neural activity is complicated by inherent physiological lag. Cross-correlation between frame luminance and pupil size revealed a distinct negative peak at a 5-second lag (Supplementary Figure S5), confirming that pupil constriction is naturally delayed relative to visual input. Furthermore, FIR deconvolution of the pupil response to luminance remarkably mirrored canonical HRF peak timing (Figure 6A).

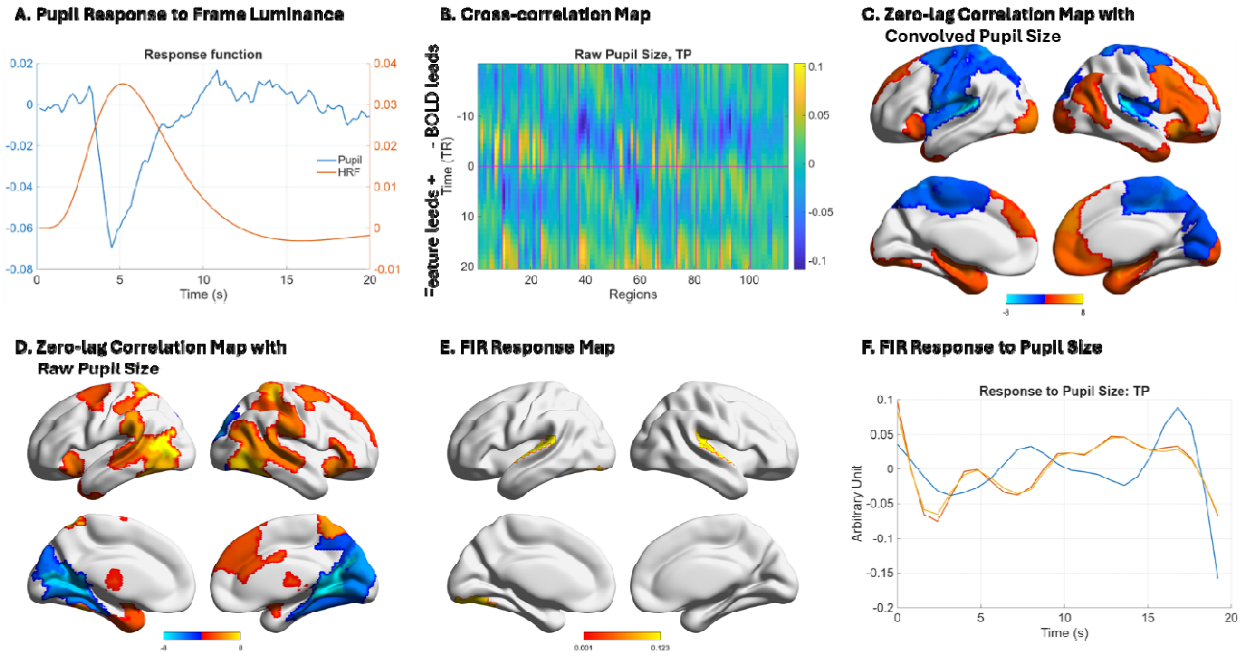


Figure 6. Neural correlates of pupil size and associated temporal response dynamics for movie *The Present* (TP). (A) Temporal response function of pupil size relative to frame luminance, shown in comparison to a canonical hemodynamic response function (HRF). **(B)** Cross-correlation matrix illustrating the relationship between the raw pupil size and fMRI time series across 114 regions of interest (ROIs) for movie TP across multiple temporal lags. **(C)** Cortical maps displaying significant positive (hot) and negative (cold) zero-lag correlations between canonical HRF-convolved pupil size and fMRI activity for movie TP. **(D)** Cortical maps displaying significant positive (hot) and negative (cold) zero-lag correlations between raw (unconvolved) pupil size and fMRI activity across the three videos. **(E)** Cortical maps illustrating significant regional responses to pupil size estimated via Finite Impulse Response (FIR) deconvolution, highlighting the peri-Sylvian autonomic hubs. **(F)** Estimated FIR response functions for significant ROIs during movie TP, demonstrating the immediate peak at lag zero. All statistical maps (Panels C, D, and E) are thresholded at a false-discovery rate (FDR) of $p < 0.05$.

Consequently, when HRF-convolved, the expected zero-lag negative correlation between pupil size and visual cortical activity disappeared, replaced by artifactual positive clusters (Figure 6C). Conversely, utilizing the raw pupil trace—which inherently contains its biological delay—restored robust, physiologically plausible zero-lag negative correlations across visual and attention networks (Figure 6B and 6D). This demonstrates that for physiological features already exhibiting HRF-like temporal properties, canonical convolution is methodologically counterproductive.

In addition, FIR deconvolution revealed robust effects localized to bilateral peri-Sylvian regions, encompassing the superior temporal gyrus, Rolandic operculum, and posterior insula (Figure 6E). Crucially, the extracted FIR response functions for these peri-Sylvian regions peaked exactly at lag zero and gradually decreased thereafter (Red and yellow lines in Figure

6F and Supplementary Figure S7). This immediate peak validates the use of the raw pupil trace, as the physically delayed peripheral signal naturally aligns with the rapid neural dynamics of these autonomic centers.

Finally, FIR analysis isolated regions with vastly different temporal profiles, underscoring the necessity of flexible modeling. Beyond the core arousal network, a distinct visual cortical region exhibited significant FIR effects characterized by a protracted temporal profile, with the response function only rising after 15 seconds before a subsequent negative dip (Blue line in Figure 6F). This delayed response likely reflects higher-order secondary processing rather than an immediate autonomic orienting reflex, demonstrating that flexible temporal modeling is critical for resolving network-specific latencies.

3.3. Theory of Mind ratings: Context-dependent dynamics of social cognition

Inter-individual correlations of ToM rating time series demonstrated high consistency across *The Present* (r = 0.58), *Despicable Me* (r = 0.41), and *Partly Cloudy* (r = 0.59; all $p < 0.001$), validating the use of the median rating time series for subsequent fMRI analyses.

Consistent with our approach for the other features, we first performed cross-correlation analyses using the HRF-convolved ToM ratings. However, the temporal alignments varied substantially across the video stimuli (Supplementary Figure S8). For *The Present*, peak correlations within the visual and dorsal attention networks occurred at positive lags (BOLD leading; Supplementary Figure S8A), suggesting that canonical convolution may have over-compensated for inherent delays in the ToM rating time series. Conversely, for *Partly Cloudy*, cross-correlation peaks in several visual areas aligned closely with the zero-lag line (Supplementary Figure S8C). The results for *Despicable Me* were more complex (Supplementary Figure S8B); while visual regions lacked clear zero-lag peaks, auditory regions exhibited massive feature-leading delays of more than 10 seconds. Because the convolved time series already incorporates an ~6-second hemodynamic delay, these extended lags cannot be

explained by neurovascular coupling alone, and instead likely reflect the collinearity of ToM with slowly unfolding narrative features. Consequently, zero-lag correlation maps using the HRF-convolved ToM ratings revealed only one consistent significant cluster in the left temporo-occipital region across all three videos (Figure 7A).

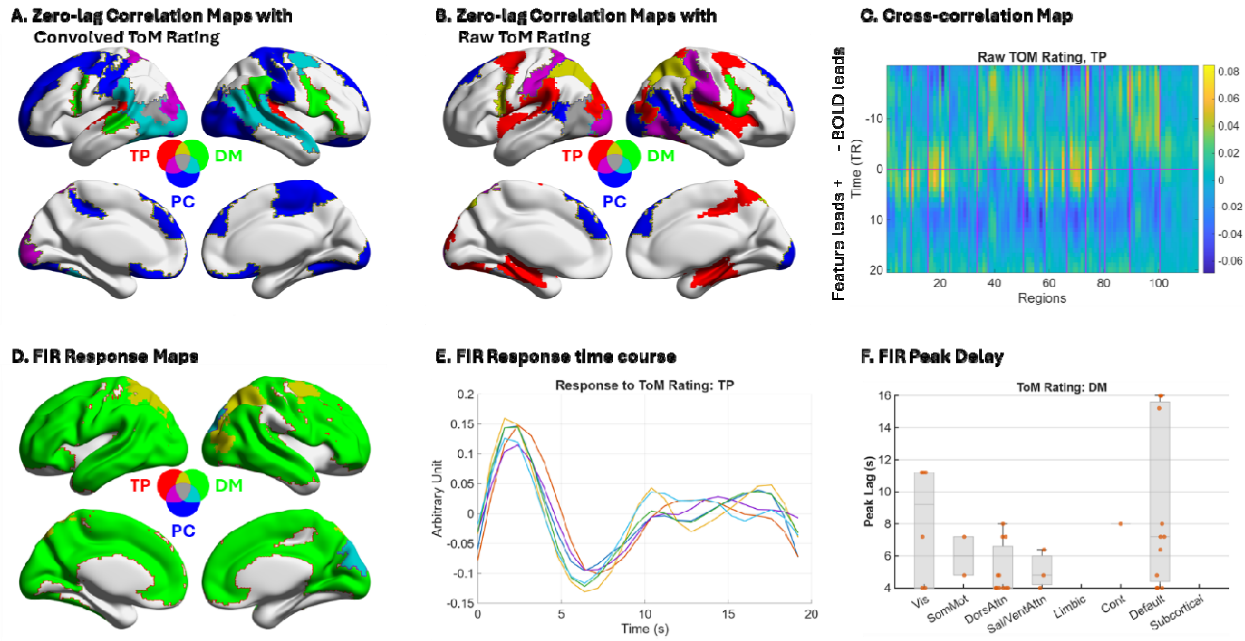


Figure 7. Neural correlates of Theory of Mind (ToM) ratings and associated temporal response dynamics. (A) Cortical maps displaying significant positive zero-lag correlations between canonical HRF-convolved ToM ratings and fMRI activity for *The Present* (TP), *Despicable Me* (DM), and *Partly Cloudy* (PC). **(B)** Cortical maps displaying significant positive zero-lag correlations between raw (unconvolved) ToM ratings and fMRI activity across the three videos. **(C)** Cross-correlation matrix illustrating the relationship between the raw ToM rating and fMRI time series across 114 regions of interest (ROIs) for movie TP across multiple temporal lags. **(D)** Cortical maps illustrating significant regional responses to ToM ratings estimated via Finite Impulse Response (FIR) deconvolution. **(E)** Estimated FIR response functions for significant ROIs during movie TP, highlighting the rapid peak latencies. **(F)** Distributions of estimated temporal delays (time-to-peak) for regions exhibiting positive FIR responses during

movie DM, categorized across Yeo's seven functional networks. All statistical maps (Panels A, B, and D) are thresholded at a false-discovery rate (FDR) of $p < 0.05$.

Given that continuous behavioral ratings intrinsically incorporate cognitive processing and motor execution delays, we hypothesized that applying the canonical HRF might artifactually "double-delay" and over-smooth the signal. To address this, we repeated the cross-correlation analyses using the raw, unconvolved ToM ratings. This approach yielded more temporally aligned cross-correlation profiles, with peaks shifting much closer to zero lag for both *The Present* (Figure 7C) and *Partly Cloudy* (Supplementary Figure S8F). Consequently, the zero-lag correlation maps using the raw ToM time series revealed a greater extent of overlapping significant clusters—particularly in the left occipito-temporal junction—across the three videos (Figure 7B).

However, while using the raw ToM time series avoids the penalty of double-smoothing, it still rigidly assumes uniform neuro-behavioral latency across all brain regions and stimuli. FIR deconvolution revealed that ToM processing dynamics are fundamentally stimulus-dependent (Figure 7D and Supplementary Figure S9). For *The Present*, significant FIR effects were primarily restricted to the dorsal visual stream and parietal regions. Crucially, the extracted response functions in these regions peaked extremely rapidly (1.6–2.4 seconds; Figure 7E), demonstrating exactly why the canonical HRF model (which assumes a ~5-second peak) was systematically inappropriate for this specific stimulus.

Conversely, FIR deconvolution for *Despicable Me* revealed widespread significant cortical responses. Unlike the distinct temporal hierarchies observed during basic sensory processing, the estimated FIR delays for ToM in *Despicable Me* did not exhibit a clear cortical feed-forward pattern (Figure 7F). This widespread, complex temporal profile likely reflects the highly multimodal nature of mentalizing required for *Despicable Me*, which demands the distributed integration of rapid dialogue, abstract social rules, and multi-character dynamics.

Together, these results demonstrate that high-level cognitive annotations cannot be accurately modeled with fixed hemodynamic assumptions; flexible approaches like FIR deconvolution are strictly essential to capture the highly context-dependent speeds of social cognition during naturalistic viewing.

4. Discussion

In the present study, we systematically investigated the temporal dynamics of BOLD responses to continuous naturalistic features during movie watching. By integrating multi-level features—ranging from low-level sensory properties (luminance, contrast, pitch) to physiological (pupil size) and subjective cognitive (ToM) signals—we evaluated how well conventional hemodynamic modeling captures the brain's nested temporal hierarchies. To overcome the limitations of fixed-delay assumptions, we utilized FIR deconvolution alongside traditional cross-correlation analyses to systematically map the true BOLD response profiles across the whole brain.

Our results demonstrate that different classes of naturalistic features exhibit distinctly different temporal BOLD profiles that reflect their position in the cortical processing hierarchy. Our main findings are threefold. First, while low-level sensory features generally aligned with the temporal expectations of the canonical HRF, slower physiological and subjective cognitive features systematically deviated from these standard models. Second, convolving inherently delayed, low-frequency signals (such as pupil size and ToM ratings) with a standard HRF introduces severe temporal artifacts—such as the "double-smoothing" effect—which led to misalignments and even inverted zero-lag correlations. Finally, by employing flexible FIR deconvolution, we successfully uncovered the feature-specific BOLD system response functions. This approach allowed us to explicitly quantify peak response latencies, revealing extensive inter-regional temporal hierarchies. These empirically derived maps demonstrated a clear progression of processing delays from primary sensory cortices to higher-order associative networks. Together, these findings underscore the necessity of feature-specific temporal

modeling in naturalistic fMRI paradigms and demonstrate that mapping the true shape of the BOLD response provides critical insights into the underlying processing timescales of the brain.

4.1. Sensory features

The present study demonstrated that for low-level sensory features—specifically frame luminance, frame contrast, and acoustic pitch—the canonical HRF implemented in standard fMRI analysis tools (e.g., SPM) effectively captured the temporal delay between stimulus dynamics and BOLD responses. Cross-correlation analyses revealed that the strongest correlations between these features and regional BOLD signals occurred at lags consistent with the canonical HRF peak (approximately 4–6 seconds). Similarly, FIR deconvolution yielded impulse response functions that closely mirrored the canonical HRF in modality specific brain regions, exhibiting the characteristic rise, peak, and undershoot phases. These results indicate that the temporal characteristics of the BOLD response to continuously varying visual and auditory inputs are largely consistent with those reported in conventional event-related paradigms, even under naturalistic viewing conditions.

This finding aligns with classic event-related fMRI studies showing that BOLD responses in early sensory regions are highly stereotyped in both shape and latency (Boynton et al., 1996; Dale and Buckner, 1997; Glover, 1999; Handwerker et al., 2004). It also resonates with more recent research using naturalistic stimuli, which has shown that BOLD activity in visual and auditory cortices remains tightly coupled to sensory fluctuations over time (Bartels and Zeki, 2004; Hasson et al., 2004; Nishimoto et al., 2011). Together, these studies support the notion that the canonical HRF offers a robust approximation of average neurovascular coupling in primary sensory areas across diverse experimental contexts.

Nonetheless, subtle deviations in the FIR-estimated responses across cortical regions were evident, suggesting regional variability in HRF dynamics. Some of these differences likely stem from intrinsic vascular properties, such as baseline cerebral blood flow and local

neurovascular coupling (Aguirre et al., 1998; Handwerker et al., 2004; Sonkusare et al., 2019; Taylor et al., 2018). However, beyond physiological vascular artifacts, our FIR delay-distribution plots revealed systematic inter-regional processing delays propagating progressively from primary sensory areas to higher-order associative and default mode networks.

This temporal progression aligns powerfully with the theoretical framework of hierarchical cortical timescales, or TRWs (Lerner et al., 2011; Meer et al., 2020). According to this framework, early sensory cortices possess short TRWs adapted to process rapid, transient physical changes in the environment, such as frame contrast or pitch variations. Consequently, these modality-specific regions exhibited sharp FIR response profiles and early peak latencies. In contrast, higher-order cortical regions possess long TRWs designed to accumulate and integrate this sensory information over extended durations (Hasson et al., 2015). Because these regions must synthesize semantic meaning and narrative context from continuous audio-visual streams, their temporal response profiles are inherently slower and more temporally diffuse (Lewis et al., 2018).

Our FIR results provide direct empirical support for this architecture, demonstrating that the "sluggishness" of the BOLD response to naturalistic sensory features in higher-order regions reflects the extended cognitive integration times required for naturalistic processing. Applying a uniform canonical HRF across the entire brain obscures this fundamental hierarchical topography, underscoring the value of flexible temporal modeling to capture true cortical timescales.

4.2. Pupil size

Previous fMRI studies have often modeled pupil diameter as a physiological regressor by convolving it with a canonical HRF before correlating it with BOLD signals (Schneider et al., 2016; Yellin et al., 2015). This approach assumes that pupil size reflects instantaneous neural or arousal-related processes that are temporally blurred by the hemodynamic response.

However, our analysis of the relationship between pupil size and frame luminance revealed an approximately 5-second delay of the pupil response to frame luminance, which directly challenges the usage of canonical HRF convolution.

When comparing correlations between BOLD activity and both HRF-convolved and raw pupil size time series, clear differences emerged. In visual cortical areas, where BOLD activity positively tracks luminance, pupil size is expected to show an inverse relationship due to the pupillary light reflex (Joshi and Gold, 2020; Mathot, 2018). Consistent with this prediction, negative correlations were observed between raw pupil size and visual cortex BOLD signals near zero lag. In contrast, when using the HRF-convolved pupil signal, these expected negative correlations disappeared and instead became artifactually inverted (positive) at zero lag, demonstrating a severe temporal misalignment between modeled and actual pupil–BOLD dynamics. Furthermore, the spatial extent of correlated regions differed markedly between the two approaches. While both revealed significant correlations in regions such as the insula and temporoparietal junction, the raw pupil signal additionally showed broader associations encompassing the lateral parietal and temporal cortices, as well as the thalamus—areas known to participate in arousal, attention, and salience processing (Menon and Uddin, 2010; Sadaghiani and D'Esposito, 2015; Shine et al., 2019).

Our FIR deconvolution analysis provided further evidence for this un-convolved temporal alignment, revealing robust effects localized to bilateral peri-Sylvian regions, encompassing the superior temporal gyrus, Rolandic operculum, and posterior insula. Crucially, the extracted FIR response functions for these peri-Sylvian regions peaked exactly at lag zero and gradually decreased thereafter. This immediate peak strongly validates the use of the raw pupil trace, as the physically delayed peripheral signal naturally aligns with the rapid neural dynamics of these autonomic centers without the need for additional temporal shifting.

From a neurophysiological perspective, pupil diameter is a reliable peripheral marker of central arousal and neuromodulatory activity, particularly reflecting the influence of the locus

coeruleus–norepinephrine (LC–NE) system (Aston-Jones and Cohen, 2005; Joshi et al., 2016; Murphy et al., 2014). The widespread correlations observed between raw pupil fluctuations and BOLD activity in the insula, temporoparietal junction, and thalamus are consistent with the distributed impact of the LC–NE network on cortical and subcortical targets (Breton-Provencher and Sur, 2019; Shine et al., 2019). These regions are frequently implicated in arousal regulation, interoceptive awareness, and adaptive cognitive control. Thus, the coupling observed here may reflect dynamic coordination between neuromodulatory arousal signals and cortical processing of salient environmental events during naturalistic viewing.

These findings suggest that applying a canonical HRF convolution to inherently delayed pupil data may distort or obscure the true temporal relationship between pupil fluctuations and neural activity. Nevertheless, this does not invalidate prior work employing HRF-convolved pupil regressors. Variability in eye-tracking systems, temporal resolution, and luminance control can influence the temporal characteristics of pupil responses (Hoeks and Levelt, 1993; McDougal and Gamlin, 2015), and in some cases—particularly with lower sampling rates—HRF convolution may remain a pragmatic approximation. However, our results underscore the importance of empirically testing temporal alignment between pupil and BOLD signals rather than assuming a fixed hemodynamic delay. Cross-correlation analyses and FIR deconvolution, as illustrated here, offer a straightforward diagnostic approach for validating these relationships and optimizing analytical strategies.

Overall, these results emphasize the need for careful temporal modeling of pupil–BOLD coupling. Because pupil responses themselves exhibit an intrinsic biological latency relative to underlying neural events (de Gee et al., 2017; Hoeks and Levelt, 1993), applying an additional HRF convolution introduces systematic misalignment (i.e., double-smoothing), reducing sensitivity to true neural–autonomic interactions. Future studies should therefore consider flexible data-driven approaches—such as empirical lag estimation, FIR modeling, or joint modeling frameworks (Cohen et al., 2016; Di and Biswal, 2019)—to better capture the temporal

dynamics linking arousal-related physiology and brain activity across resting-state, naturalistic, and controlled experimental paradigms.

4.3. Theory of Mind Rating

In this study, an independent group of participants provided continuous ratings reflecting their perceived understanding of the characters' thoughts, feelings, and intentions while viewing short movie clips. These ratings are intended to index moment-to-moment engagement of ToM—the ability to attribute mental states to others (Frith and Frith, 2006; Schurz et al., 2021). Unlike low-level sensory or physiological features, however, the temporal coupling between such introspective, socially grounded judgments and underlying neural activity is far less constrained and may vary with narrative structure, attention, and individual interpretation (Jacoby et al., 2016; Richardson et al., 2018).

Continuous introspective ratings inherently integrate information over several seconds due to cognitive evaluation and motor execution delays (Nummenmaa et al., 2012; Zahn et al., 2009). Because the generation of a continuous behavioral rating is intrinsically a temporally smoothed process, applying an additional HRF convolution to these time series is mathematically redundant. This secondary convolution compounds the temporal dynamics, creating a "double-delay" that fundamentally shifts the regressor out of phase with true neural activity, while also over-smearing the signal variance. The improved spatial consistency observed across videos when using raw (unconvolved) ToM ratings is thus likely attributable to both the adjustment of this temporal misalignment and the avoidance of this redundant smoothing. Furthermore, because the raw ToM time series already represents a delayed behavioral output relative to the initial neural event, performing FIR deconvolution directly on this trace logically yields an accelerated estimated hemodynamic peak (as the behavioral latency accounts for a portion of the typical neurovascular lag). This accelerated response profile—as empirically confirmed by our FIR estimations—explicitly demonstrates why the

canonical HRF model, which rigidly assumes a ~5-second peak, is systematically inappropriate when modeling inherently delayed and smoothed behavioral traces.

Furthermore, our FIR deconvolution results demonstrate that ToM processing dynamics are fundamentally stimulus- and context-dependent. While low-level sensory features—such as frame contrast and auditory pitch—exhibited a clear temporal hierarchy with processing delays propagating sequentially from primary sensory cortices to higher-order associative networks, ToM processing did not follow this pattern. Specifically, the peak lag distributions for ToM during *Despicable Me* did not exhibit a clear feed-forward cortical hierarchy; instead, the temporal delay distributions were remarkably similar and widespread across all the functional networks. This simultaneous, distributed temporal profile likely reflects the highly multimodal and continuous nature of the mentalizing required for this specific film, which demands the concurrent integration of rapid dialogue, abstract social rules, and dynamic multi-character interactions across the brain, rather than a simple sequential processing stream.

These complexities highlight the profound challenges of mapping continuous, high-level social cognition onto temporally sluggish BOLD signals. Our results converge with recent efforts advocating explicit examination of temporal relationships between behavioral regressors and neural responses in naturalistic paradigms (Di and Biswal, 2022; Lahnakoski et al., 2012b). Rather than assuming a canonical HRF—or relying solely on raw time series—data-driven approaches such as FIR deconvolution are strictly essential to capture the highly context-dependent and feature-specific speeds of social cognition. Future work could further disentangle these overlapping temporal profiles by integrating synchronized physiological measures (e.g., facial EMG or reaction-based time stamping) to refine lag estimation, or by employing advanced regularization techniques to account for the inherent collinearity of cinematic features.

4.4. Complexity of naturalistic stimuli: Inter-stimulus variability and feature collinearity

A striking observation in our analysis was the substantial variability in brain responses across the three films. While sensory features generally exhibited more consistent activations across datasets, there were notable exceptions—such as widespread cortical responses to frame contrast in *The Present* and to auditory pitch in *Despicable Me*. This variability was even more pronounced for high-level cognitive features like ToM. Specifically, ToM processing did not uniformly engage classical mentalizing or default mode network regions across all datasets. This divergence likely stems from profound differences in narrative complexity and the specific nature of the ToM content elicited by each film. For example, *Despicable Me* involves complex, dialogue-driven social rules and dynamic multi-character interactions, which manifested as widespread, distributed cortical engagement. In contrast, short animations like *The Present* or *Partly Cloudy* rely heavily on non-verbal, visually driven social cues, which may engage a different, more restricted sub-network of social-perceptual regions. Thus, the neural correlates of "ToM" during naturalistic viewing are not monolithic; they are highly context-dependent and intrinsically tied to the statistical and narrative structure of the specific stimulus.

Beyond narrative content, inter-stimulus variability is also driven by fundamental differences in data acquisition and technical parameters. The statistical power and temporal resolution varied considerably across datasets: *Despicable Me* contained 740 time points (TR = 800 ms), *The Present* contained 240 time points (TR = 800 ms), and *Partly Cloudy* was acquired with a significantly slower temporal resolution (TR = 2000 ms) and only 168 time points. The extended data length in *Despicable Me* likely afforded higher statistical power, which partially explains the more widespread network recruitment observed for both ToM and sensory features in this video. Furthermore, differences in TR inherently impact the temporal precision of FIR deconvolution, as higher temporal resolution allows for more accurate estimation of response peaks compared to slower acquisitions.

A final critical methodological consideration when mapping these temporal dynamics during naturalistic viewing is the inherent collinearity of stimulus features. Unlike highly

controlled, artificial experimental designs where stimuli are presented in isolation, cinematic features are continuously and simultaneously presented, frequently covarying with one another as well as with a vast array of unmeasured features (e.g., semantic meaning, emotional valence, or biological motion). In the present study, we largely evaluated the temporal relationship of individual features with regional BOLD activity one at a time. Consequently, the estimated temporal response functions capture the overall neural engagement associated with that feature's occurrence in the narrative, rather than its strictly isolated processing. For instance, the widespread cortical responses and extended temporal delays observed for auditory pitch—extending well beyond primary auditory cortex into higher-order associative networks—are likely driven by its natural covariance with human speech, music, and emotional prosody.

While combining multiple features into a single, comprehensive general linear model could theoretically isolate the unique variance of each specific feature, implementing such an approach alongside FIR deconvolution introduces severe multicollinearity. Expanding numerous correlated features into dozens of time-lagged predictors severely inflates variance and destabilizes the estimated response functions. Therefore, while our unconstrained temporal modeling reveals the complex, sequential brain states triggered by naturalistic events, future studies employing advanced regularization techniques (e.g., ridge regression) or dimensionality reduction will be necessary to disentangle the overlapping temporal profiles of highly collinear narrative features.

4.5. Broader implications for hemodynamic response

The present analysis provides robust empirical evidence that the HRF can be estimated directly from continuous naturalistic data using FIR modeling. Traditional HRF estimation approaches have primarily relied on event-related or block designs, where each stimulus or trial is modeled as a discrete impulse (Dale and Buckner, 1997; Glover, 1999). Although such methods have been instrumental in characterizing canonical HRF properties, they inherently simplify the

temporal and cognitive structure of real-world experiences. Event-based paradigms constrain the complexity and duration of modeled processes and often assume that cognitive or affective responses occur as isolated, time-locked events rather than as continuously evolving states.

In contrast, naturalistic paradigms offer an alternative framework by leveraging time-varying sensory, behavioral, or physiological features as continuous inputs to estimate neural response functions. This approach captures more ecologically valid dynamics of neural activity, encompassing sustained attention, emotional engagement, and social inference—processes that unfold gradually rather than discretely (Hasson et al., 2004; Nastase et al., 2019). Although continuous designs face challenges such as limited variability in specific features and uneven temporal sampling across frequency ranges, they allow HRF estimation in higher-order cortical and subcortical regions that are less reliably engaged by simple, transient stimuli.

The current findings demonstrate the utility of using continuous naturalistic features to perform data-driven HRF deconvolution across multiple brain regions. The FIR model, which imposes minimal assumptions about response shape or latency, revealed systematically diverse temporal profiles and hierarchical peak latencies across regions, consistent with prior evidence that HRF dynamics vary with vascular and neuronal properties (Aguirre et al., 1998; Handwerker et al., 2004; Lindquist et al., 2009). Such flexibility is valuable for uncovering region-specific or stimulus-specific variations in hemodynamic coupling that are not well captured by canonical models.

However, this flexibility comes at the cost of increased variance and reduced reliability, particularly when the number of model parameters approaches the available data length (Goutte et al., 2000; Pedregosa et al., 2015). Future research should explore hybrid or regularized approaches—such as ridge-regularized FIR models, basis function modeling (e.g., gamma or spline basis sets), or parametric HRF estimation—to balance flexibility with stability and physiological interpretability (Lindquist et al., 2009; Pedregosa et al., 2015). Comparative

evaluations across these frameworks will be critical for identifying optimal trade-offs between model adaptability and robustness in naturalistic contexts.

An additional priority is to assess the reliability and inter-individual variability of HRF estimates derived from continuous data. Individual differences in vascular dynamics, neurovascular coupling, or baseline physiology may systematically influence HRF latency and shape (Handwerker et al., 2004; West et al., 2019). Quantifying these sources of variability could enhance both within-subject reproducibility and between-subject interpretability, with potential applications in aging (Chen et al., 2024; West et al., 2019), cerebrovascular disease, and neuropsychiatric populations. Robust, data-driven HRF estimation could ultimately improve the interpretability of BOLD responses in naturalistic and task-free paradigms, bridging the gap between experimental control and ecological validity.

Finally, integrating FIR-based HRF estimation with analyses of sensory, physiological, and cognitive features—as demonstrated in the present study—illustrates the value of unified temporal modeling across multiple domains. The observation that low-level sensory features align well with the canonical HRF, whereas physiological (pupil) and introspective (ToM rating) signals exhibit more complex, delayed timing relationships, confirms that hemodynamic coupling differs systematically across cortical processing hierarchies and temporal receptive windows. Data-driven HRF estimation, therefore, provides not only a methodological advance but also a conceptual framework for understanding how distinct neural systems integrate information over time. Extending this approach across diverse stimulus types and participant populations could yield a more comprehensive characterization of the temporal architecture of brain function under naturalistic conditions.

5. Conclusion

In summary, this study systematically characterized the temporal relationships between continuous sensory, physiological, and behavioral features and fMRI BOLD signals during

naturalistic movie viewing. Using cross-correlation and flexible FIR deconvolution, we demonstrated that while low-level sensory features generally align with canonical hemodynamic expectations, slower physiological signals (pupil size) and high-level cognitive annotations (ToM) systematically deviate from these standard models. Crucially, applying a uniform canonical HRF to inherently delayed physiological and behavioral time series introduces temporal artifacts, such as double-smoothing, which obscure true brain-behavior coupling.

By employing data-driven FIR deconvolution, we successfully uncovered the feature-specific BOLD response profiles, explicitly mapping the underlying temporal hierarchies and processing timescales across the cortex. These findings highlight the necessity of flexible temporal modeling when mapping diverse, multi-level features in naturalistic paradigms. Moreover, this analysis validates the robust utility of using continuous input signals to estimate hemodynamic response functions directly from naturalistic viewing. This framework offers a powerful alternative to event-related designs for studying complex, gradually unfolding brain processes, and provides a valuable tool for investigating individual differences in neurovascular coupling across diverse populations, such as aging or clinical cohorts.

Acknowledgments

During the preparation of this revised manuscript, the authors utilized Gemini (Google) solely for language polishing and assistance with drafting technical summaries. No artificial intelligence tools were used in the generation of research data, data analysis, or the scientific interpretation of the results. The authors thoroughly reviewed and edited all AI-assisted text and take full responsibility for the final content of the manuscript.

Data and Code Availability

MRI and pupil size data were obtained from publicly available sources: the Healthy Brain Network (HBN) dataset (http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/) and the OpenNeuro repository (https://openneuro.org/datasets/ds000228).

Author Contributions

X.D. conceived and designed the study, developed the methodology, performed the analyses, and wrote the original draft of the manuscript. G.B.H. collected and curated the data and contributed to manuscript review and editing. B.B.B. and X.D. provided resources and secured funding for the project. All authors contributed to the interpretation of results and approved the final version of the manuscript.

Funding

This study was supported by grants from (US) National Institutes of Health for Xin Di (R15MH125332) and Bharat B. Biswal (5R01MH131335 and 1R01AG085665) and by (NJ) Governor's Council for Medical Research and Treatment of Autism for Xin Di (CAUT25BRP005).

Declaration of Competing Interests

The authors declare that there is no competing interest.

Reference:

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The Variability of Human, BOLD Hemodynamic Responses. NeuroImage 8, 360–369. https://doi.org/10.1006/nimg.1998.0369

Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., Chan, E., Citera, D., Clucas, J., Cohen, S., Dufek, S., Eaves, M., Fradera, B., Gardner, J., Grant-Villegas, N., Green, G., Gregory, C., Hart, E., Harris, S., Horton, M., Kahn, D., Kabotyanski, K., Karmel, B., Kelly, S.P., Kleinman, K., Koo, B., Kramer, E., Lennon, E., Lord, C., Mantello, G., Margolis, A., Merikangas, K.R., Milham, J., Minniti, G., Neuhaus, R., Levine,

A., Osman, Y., Parra, L.C., Pugh, K.R., Racanello, A., Restrepo, A., Saltzman, T., Septimus, B., Tobe, R., Waltz, R., Williams, A., Yeo, A., Castellanos, F.X., Klein, A., Paus, T., Leventhal, B.L., Craddock, R.C., Koplewicz, H.S., Milham, M.P., 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. Scientific Data 4, 1–26. https://doi.org/10.1038/sdata.2017.181

Aston-Jones, G., Cohen, J.D., 2005. AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. Annual Review of Neuroscience 28, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bartels, A., Zeki, S., 2004. Functional brain mapping during free viewing of natural scenes. Hum Brain Mapp 21, 75–85. https://doi.org/10.1002/hbm.10153

Bartels, A., Zeki, S., Logothetis, N.K., 2008. Natural Vision Reveals Regional Specialization to Local Motion and to Contrast-Invariant, Global Flow in the Human Brain. Cerebral Cortex 18, 705–717. https://doi.org/10.1093/cercor/bhm107

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. J. Neurosci. 16, 4207–4221. https://doi.org/10.1523/JNEUROSCI.16-13-04207.1996

Breton-Provencher, V., Sur, M., 2019. Active control of arousal by a locus coeruleus GABAergic circuit. Nat Neurosci 22, 218–228. https://doi.org/10.1038/s41593-018-0305-z

Chen, D.Y., Di, X., Yu, X., Biswal, B.B., 2024. The significance and limited influence of cerebrovascular reactivity on age and sex effects in task- and resting-state brain activity. Cerebral Cortex 34, bhad448. https://doi.org/10.1093/cercor/bhad448

Cohen, A.O., Breiner, K., Steinberg, L., Bonnie, R.J., Scott, E.S., Taylor-Thompson, K., Rudolph, M.D., Chein, J., Richeson, J.A., Heller, A.S., Silverman, M.R., Dellarco, D.V., Fair, D.A., Galván, A., Casey, B.J., 2016. When Is an Adolescent an Adult? Assessing Cognitive Control in Emotional and Nonemotional Contexts. Psychol Sci 27, 549–562. https://doi.org/10.1177/0956797615627625

Dale, A.M., Buckner, R.L., 1997. Selective averaging of rapidly presented individual trials using fMRI. Human brain mapping 5, 329–40. https://doi.org/10.1002/(SICI)1097-0193(1997)5:5%253C329::AID-HBM1%253E3.0.CO;2-5

de Gee, J.W., Colizoli, O., Kloosterman, N.A., Knapen, T., Nieuwenhuis, S., Donner, T.H., 2017. Dynamic modulation of decision biases by brainstem arousal systems. eLife 6, e23232. https://doi.org/10.7554/eLife.23232

Di, X., Biswal, B.B., 2022. Principal component analysis reveals multiple consistent responses to naturalistic stimuli in children and adults. Human Brain Mapping 43, 3332–3345. https://doi.org/10.1002/hbm.25568

Di, X., Biswal, B.B., 2019. Toward Task Connectomics: Examining Whole-Brain Task Modulated Connectivity in Different Task Domains. Cereb Cortex 29, 1572–1583. https://doi.org/10.1093/cercor/bhy055

Di, X., Xu, T., Miao, X., 2026. Chapter 14 - Future perspectives of functional connectivity: Advances in data acquisition and analysis, in: Rocca, M.A., Filippi, M. (Eds.), Functional Connectivity of the Human Brain. Academic Press, pp. 281–309. https://doi.org/10.1016/B978-0-443-19099-5.00020-9

Di, X., Xu, T., Uddin, L.Q., Biswal, B.B., 2023. Individual differences in time-varying and stationary brain connectivity during movie watching from childhood to early adulthood: Effects of age, sex, and behavioral associations. https://doi.org/10.1101/2023.01.30.526311

Fisher, R.A., 1925. Statistical methods for research workers Oliver and Boyd, London. Reprinted in Statistical Methods, Experimental Design and Scientific Inference.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-Related fMRI: Characterizing Differential Responses. NeuroImage 7, 30–40. https://doi.org/10.1006/nimg.1997.0306

Frith, C.D., Frith, U., 2006. The Neural Basis of Mentalizing. Neuron 50, 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Girard, J.M., 2014. CARMA: Software for Continuous Affect Rating and Media Annotation | Journal of Open Research Software. https://doi.org/10.5334/jors.ar

Glover, G.H., 1999. Deconvolution of Impulse Response in Event-Related BOLD fMRI1. NeuroImage 9, 416–429. https://doi.org/10.1006/nimg.1998.0419

Goutte, C., Nielsen, F.A., Hansen, K.H., 2000. Modeling the hemodynamic response in fMRI using smooth FIR filters. IEEE Transactions on Medical Imaging 19, 1188–1201. https://doi.org/10.1109/42.897811

Handwerker, D.A., Ollinger, J.M., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. NeuroImage 21, 1639–1651. https://doi.org/10.1016/j.neuroimage.2003.11.029

Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends in Cognitive Sciences 19, 304–313. https://doi.org/10.1016/j.tics.2015.04.006

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science (New York, N.Y.) 303, 1634–40. https://doi.org/10.1126/science.1089506

Hoeks, B., Levelt, W.J.M., 1993. Pupillary dilation as a measure of attention: a quantitative system analysis. Behavior Research Methods, Instruments, & Computers 25, 16–26. https://doi.org/10.3758/BF03204445

Huth, A.G., Lee, T., Nishimoto, S., Bilenko, N.Y., Vu, A.T., Gallant, J.L., 2016. Decoding the Semantic Content of Natural Movies from Human Brain Activity. Front. Syst. Neurosci. 10. https://doi.org/10.3389/fnsys.2016.00081

Jacoby, N., Bruneau, E., Koster-Hale, J., Saxe, R., 2016. Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. NeuroImage 126, 39–48. https://doi.org/10.1016/j.neuroimage.2015.11.025

Joshi, S., Gold, J.I., 2020. Pupil Size as a Window on Neural Substrates of Cognition. Trends in Cognitive Sciences 24, 466–480. https://doi.org/10.1016/j.tics.2020.03.005

Joshi, S., Li, Y., Kalwani, R.M., Gold, J.I., 2016. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. Neuron 89, 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kauppi, J.-P., Jääskeläinen, I.P., Sams, M., Tohka, J., 2010. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front. Neuroinform. 4. https://doi.org/10.3389/fninf.2010.00005

Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012a. Naturalistic fMRI Mapping Reveals Superior Temporal Sulcus as the Hub for the Distributed Brain Network for Social Perception. Front. Hum. Neurosci. 6. https://doi.org/10.3389/fnhum.2012.00233

Lahnakoski, J.M., Salmi, J., Jääskeläinen, I.P., Lampinen, J., Glerean, E., Tikka, P., Sams, M., 2012b. Stimulus-Related Independent Component and Voxel-Wise Analysis of Human Brain Activity during Free Viewing of a Feature Film. PLOS ONE 7, e35215. https://doi.org/10.1371/journal.pone.0035215

Langer, N., Ho, E.J., Alexander, L.M., Xu, H.Y., Jozanovic, R.K., Henin, S., Petroni, A., Cohen, S., Marcelle, E.T., Parra, L.C., Milham, M.P., Kelly, S.P., 2017. A resource for assessing information processing

in the developing brain using EEG and eye tracking. Sci Data 4, 170040. https://doi.org/10.1038/sdata.2017.40

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lewis, L.D., Setsompop, K., Rosen, B.R., Polimeni, J.R., 2018. Stimulus-dependent hemodynamic response timing across the human subcortical-cortical visual pathway identified through high spatiotemporal resolution 7T fMRI. NeuroImage 181, 279–291. https://doi.org/10.1016/j.neuroimage.2018.06.056

Lindquist, M.A., Meng Loh, J., Atlas, L.Y., Wager, T.D., 2009. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. NeuroImage, Mathematics in Brain Imaging 45, S187–S198. https://doi.org/10.1016/j.neuroimage.2008.10.065

Lloyd, B., de Voogd, L.D., Mäki-Marttunen, V., Nieuwenhuis, S., 2023. Pupil size reflects activation of subcortical ascending arousal system nuclei during rest. eLife 12, e84822. https://doi.org/10.7554/eLife.84822

Mathot, S., 2018. Pupillometry: Psychology, Physiology, and Function. Journal of Cognition 1. https://doi.org/10.5334/joc.18

McDougal, D.H., Gamlin, P.D., 2015. Autonomic Control of the Eye. Comprehensive Physiology 5, 439–473. https://doi.org/10.1002/j.2040-4603.2015.tb00603.x

McMahon, E., Bonner, M.F., Isik, L., 2023. Hierarchical organization of social action features along the lateral visual pathway. Current Biology 33, 5035-5047.e8. https://doi.org/10.1016/j.cub.2023.10.015

Meer, J.N. van der, Breakspear, M., Chang, L.J., Sonkusare, S., Cocchi, L., 2020. Movie viewing elicits rich and reliable brain state dynamics. Nat Commun 11, 5004. https://doi.org/10.1038/s41467-020-18717-w

Menon, V., Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. Brain structure & function 214, 655–67. https://doi.org/10.1007/s00429-010-0262-0

Murphy, P.R., O'Connell, R.G., O'Sullivan, M., Robertson, I.H., Balsters, J.H., 2014. Pupil diameter covaries with BOLD activity in human locus coeruleus. Human Brain Mapping 35, 4140–4154. https://doi.org/10.1002/hbm.22466

Nastase, S.A., Gazzola, V., Hasson, U., Keysers, C., 2019. Measuring shared responses across subjects using intersubject correlation. Soc Cogn Affect Neurosci 14, 667–685. https://doi.org/10.1093/scan/nsz037

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current Biology 21, 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I.P., Hari, R., Sams, M., 2012. Emotions promote social interaction by synchronizing brain activity across individuals. Proceedings of the National Academy of Sciences 109, 9599–9604. https://doi.org/10.1073/pnas.1206095109

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., Gramfort, A., 2015. Data-driven HRF estimation for encoding and decoding models. NeuroImage 104, 209–220. https://doi.org/10.1016/j.neuroimage.2014.09.060

Raz, G., Touroutoglou, A., Wilson-Mendenhall, C., Gilam, G., Lin, T., Gonen, T., Jacob, Y., Atzil, S., Admon, R., Bleich-Cohen, M., Maron-Katz, A., Hendler, T., Barrett, L.F., 2016. Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. Cogn Affect Behav Neurosci 16, 709–723. https://doi.org/10.3758/s13415-016-0425-4

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., Saxe, R., 2018. Development of the social brain from age three to twelve years. Nat Commun 9, 1–12. https://doi.org/10.1038/s41467-018-03399-2

Saarimäki, H., 2021. Naturalistic Stimuli in Affective Neuroimaging: A Review. Front. Hum. Neurosci. 15. https://doi.org/10.3389/fnhum.2021.675068

Sadaghiani, S., D'Esposito, M., 2015. Functional Characterization of the Cingulo-Opercular Network in the Maintenance of Tonic Alertness. Cereb Cortex 25, 2763–2773. https://doi.org/10.1093/cercor/bhu072

Samara, A., Zada, Z., Vanderwal, T., Hasson, U., Nastase, S.A., 2025. Cortical language areas are coupled via a soft hierarchy of model-based linguistic features. https://doi.org/10.1101/2025.06.02.657491

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral cortex (New York, N.Y.: 1991) 28, 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schneider, M., Hathway, P., Leuchs, L., Sämann, P.G., Czisch, M., Spoormaker, V.I., 2016. Spontaneous pupil dilations during the resting state are associated with activation of the salience network. NeuroImage 139, 189–201. https://doi.org/10.1016/j.neuroimage.2016.06.011

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neuroscience & Biobehavioral Reviews 42, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Radua, J., Tholen, M.G., Maliske, L., Margulies, D.S., Mars, R.B., Sallet, J., Kanske, P., 2021. Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychological Bulletin 147, 293–327. https://doi.org/10.1037/bul0000303

Shine, J.M., Hearne, L.J., Breakspear, M., Hwang, K., Müller, E.J., Sporns, O., Poldrack, R.A., Mattingley, J.B., Cocchi, L., 2019. The Low-Dimensional Neural Architecture of Cognitive Complexity Is Related to Activity in Medial Thalamic Nuclei. Neuron 104, 849-855.e3. https://doi.org/10.1016/j.neuron.2019.09.002

Sohn, W., Di, X., Liang, Z., Zhang, Z., Biswal, B.B., 2024. Explorations of using a convolutional neural network to understand brain activations during movie watching. https://doi.org/10.1101/2024.01.20.576341

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

Taylor, A.J., Kim, J.H., Ress, D., 2018. Characterization of the hemodynamic response function across the majority of human cerebral cortex. NeuroImage 173, 322–331. https://doi.org/10.1016/j.neuroimage.2018.02.061

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–89. https://doi.org/10.1006/nimg.2001.0978

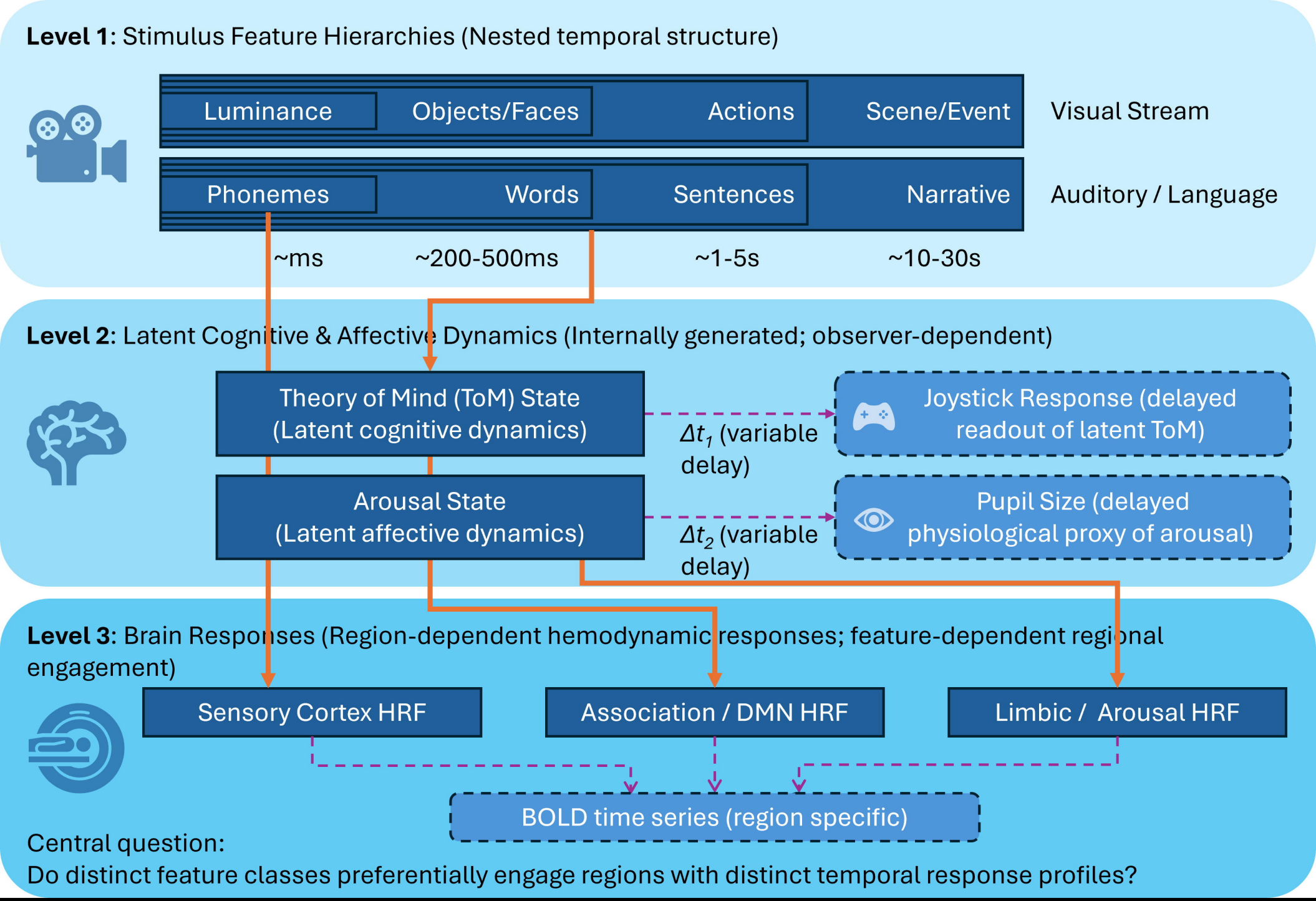
Vanderwal, T., Eilbott, J., Castellanos, F.X., 2019. Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. Developmental Cognitive Neuroscience 36, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

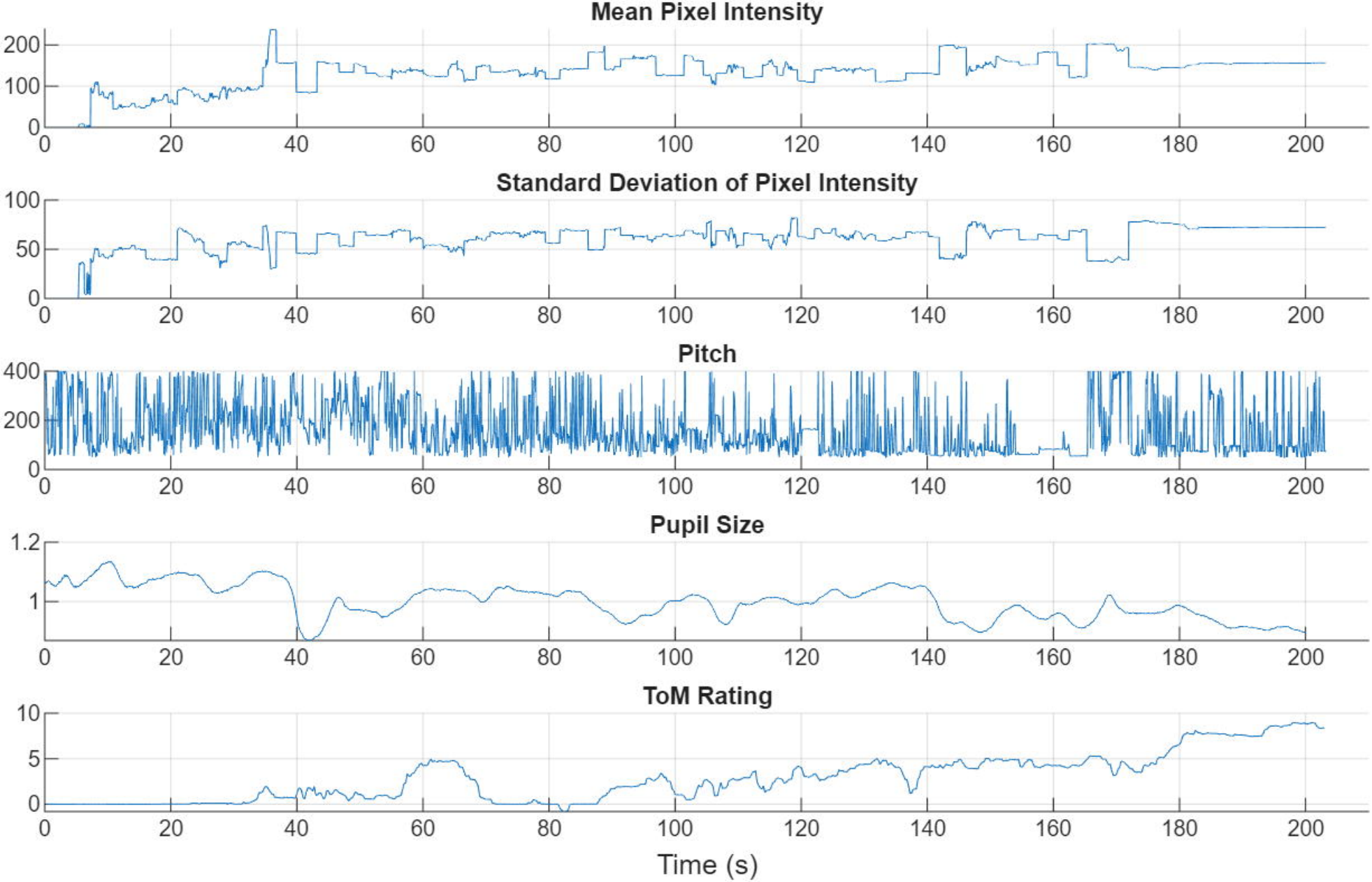
West, K.L., Zuppichini, M.D., Turner, M.P., Sivakolundu, D.K., Zhao, Y., Abdelkarim, D., Spence, J.S., Rypma, B., 2019. BOLD hemodynamic response function changes significantly with healthy aging. NeuroImage 188, 198–207. https://doi.org/10.1016/j.neuroimage.2018.12.012

Wheatley, T., Milleville, S.C., Martin, A., 2007. Understanding Animate Agents: Distinct Roles for the Social Network and Mirror System. Psychol Sci 18, 469–474. https://doi.org/10.1111/j.1467-9280.2007.01923.x

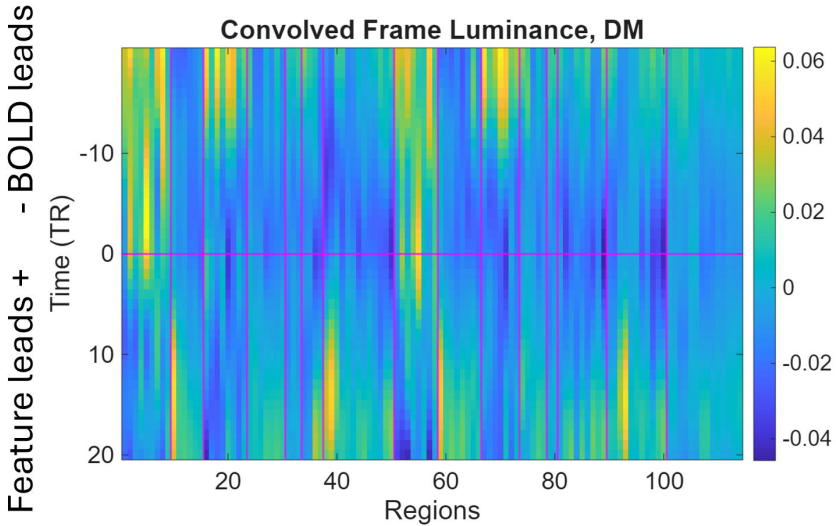
Yellin, D., Berkovich-Ohana, A., Malach, R., 2015. Coupling between pupil fluctuations and resting-state fMRI uncovers a slow build-up of antagonistic responses in the human cortex. NeuroImage 106, 414–427. https://doi.org/10.1016/j.neuroimage.2014.11.034

Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E.D., Grafman, J., 2009. The Neural Basis of Human Social Values: Evidence from Functional MRI. Cereb Cortex 19, 276–283. https://doi.org/10.1093/cercor/bhn080

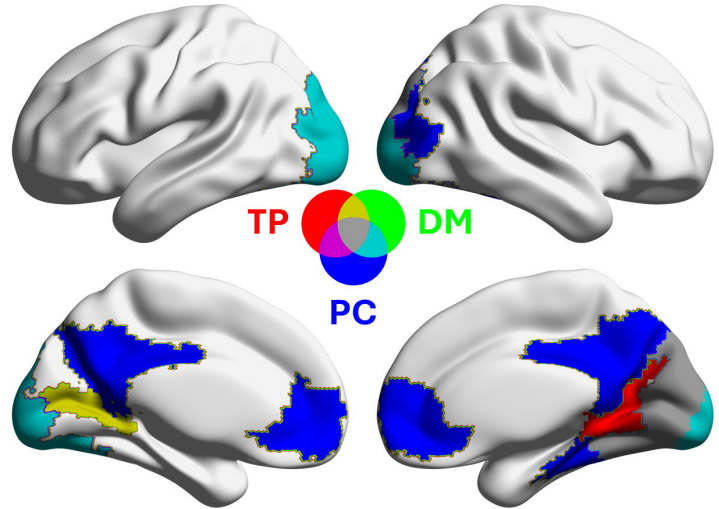

Level 1: Stimulus Feature Hierarchies (Nested temporal structure)
Luminance
Objects/Faces
Actions
Scene/Event
Visual Stream
Phonemes
Words
Sentences
Narrative
Auditory / Language
~ms
~200-500ms
~1-5s
~10-30s
Level 2: Latent Cognitive & Affective Dynamics (Internally generated; observer-dependent)
Theory of Mind (ToM) State
(Latent cognitive dynamics)
Arousal State
(Latent affective dynamics)
Δt1 (variable delay)
Δt2 (variable delay)
Joystick Response (delayed readout of latent ToM)
Pupil Size (delayed physiological proxy of arousal)
Level 3: Brain Responses (Region-dependent hemodynamic responses; feature-dependent regional engagement)
Sensory Cortex HRF
Association / DMN HRF
Limbic / Arousal HRF
BOLD time series (region specific)
Central question:
Do distinct feature classes preferentially engage regions with distinct temporal response profiles?


Mean Pixel Intensity
200
100
0
0 20 40 60 80 100 120 140 160 180 200
Standard Deviation of Pixel Intensity
100
50
0
0 20 40 60 80 100 120 140 160 180 200
Pitch
400
200
0
0 20 40 60 80 100 120 140 160 180 200
Pupil Size
1.2
1
0 20 40 60 80 100 120 140 160 180 200
ToM Rating
10
5
0
0 20 40 60 80 100 120 140 160 180 200
Time (s)

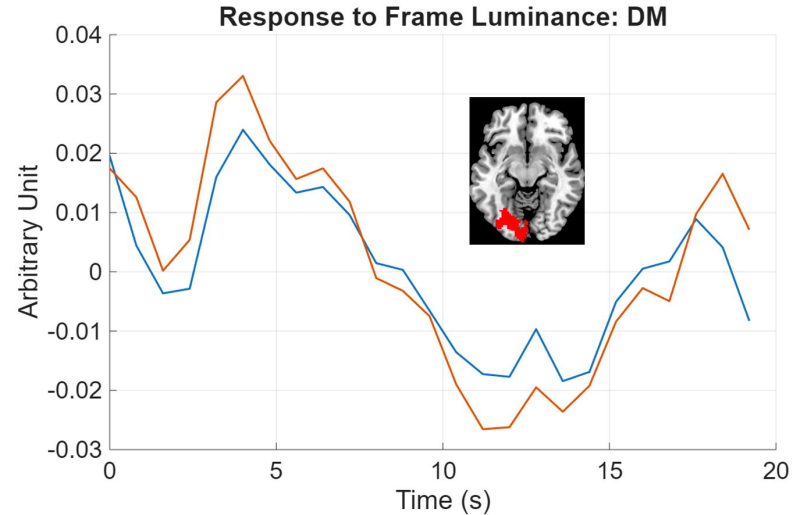
A. Cross-correlations Matrix

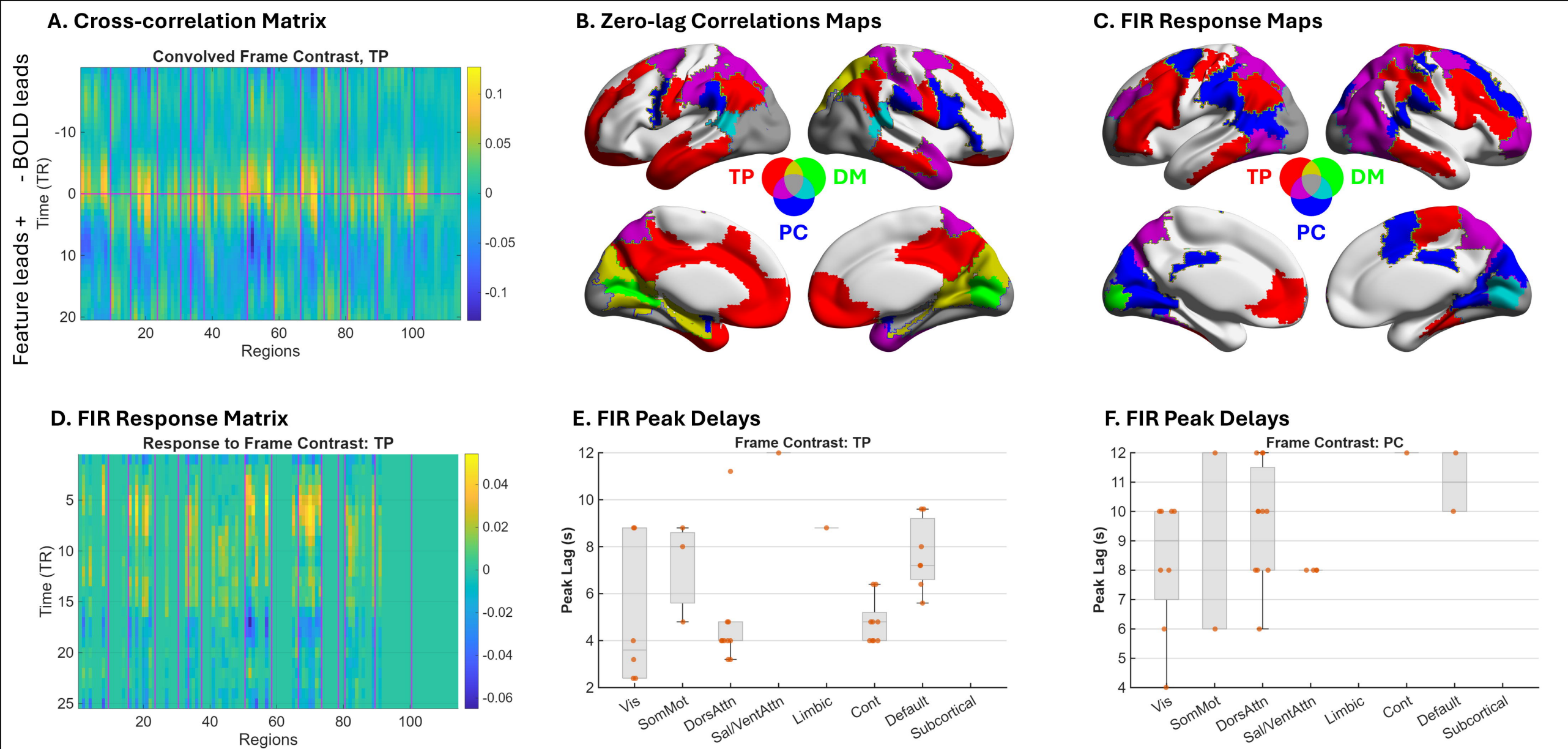


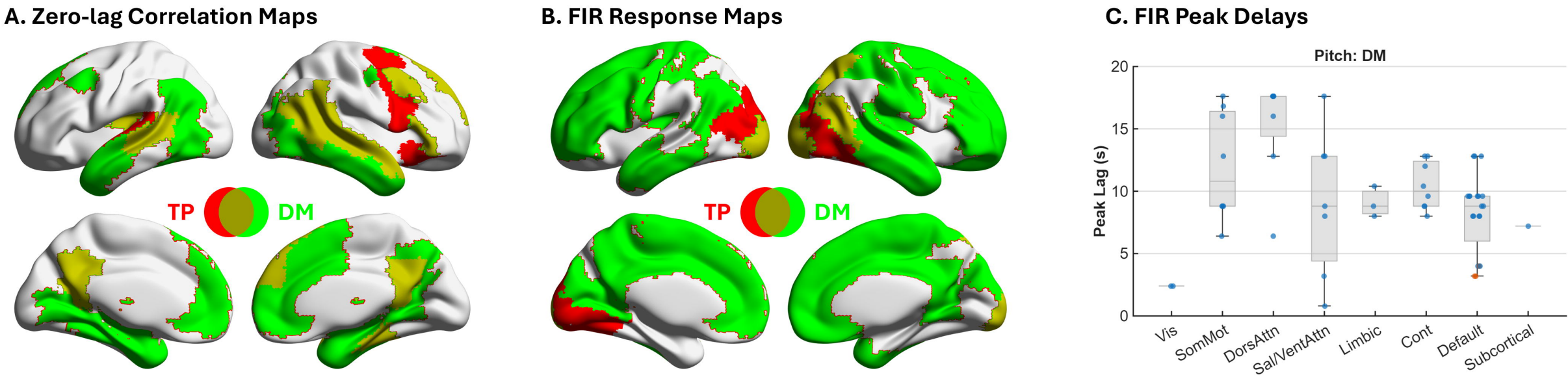
B. Zero-lag Correlation Maps



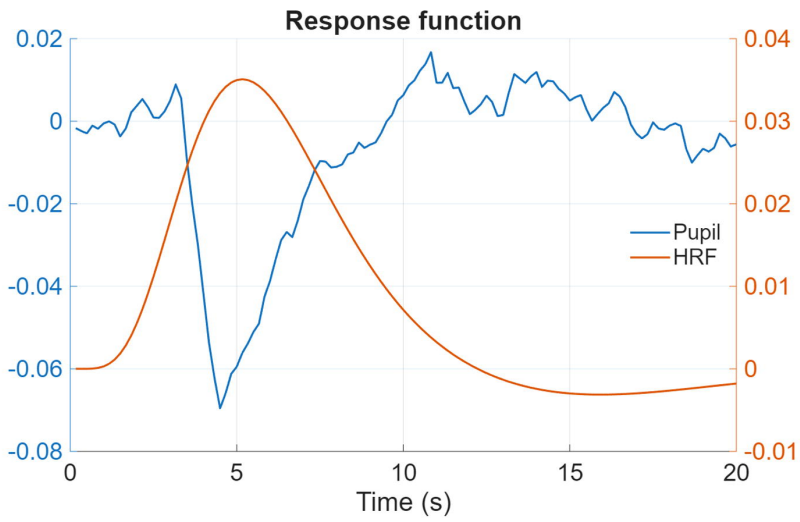
C. FIR Response



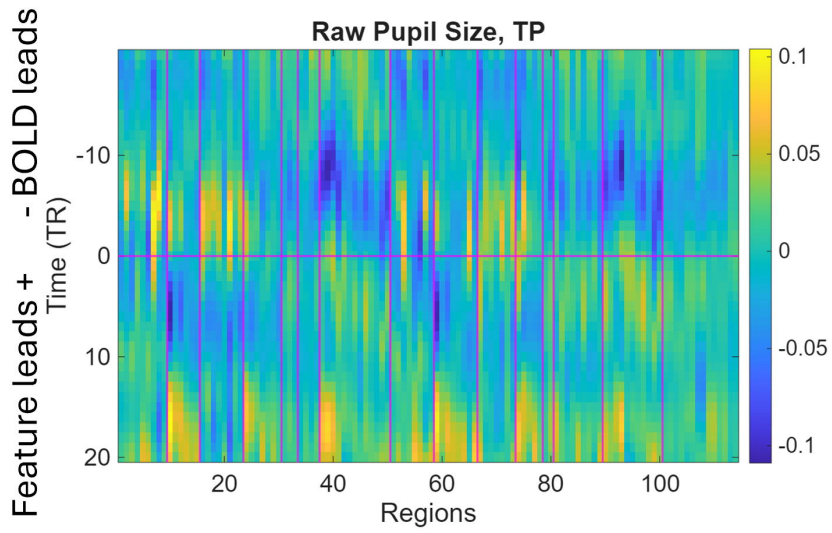

A. Cross-correlation Matrix
Convolved Frame Contrast, TP
Feature leads +
- BOLD leads
Time (TR)
Regions
B. Zero-lag Correlations Maps
TP
DM
PC
C. FIR Response Maps
TP
DM
PC
D. FIR Response Matrix
Response to Frame Contrast: TP
Time (TR)
Regions
E. FIR Peak Delays
Frame Contrast: TP
Peak Lag (s)
Vis
SomMot
DorsAttn
Sal/VentAttn
Limbic
Cont
Default
Subcortical
F. FIR Peak Delays
Frame Contrast: PC
Peak Lag (s)
Vis
SomMot
DorsAttn
Sal/VentAttn
Limbic
Cont
Default
Subcortical


A. Zero-lag Correlation Maps
TP
DM
B. FIR Response Maps
TP
DM
C. FIR Peak Delays
Pitch: DM
Peak Lag (s)
0
5
10
15
20
Vis
SomMot
DorsAttn
Sal/VentAttn
Limbic
Cont
Default
Subcortical

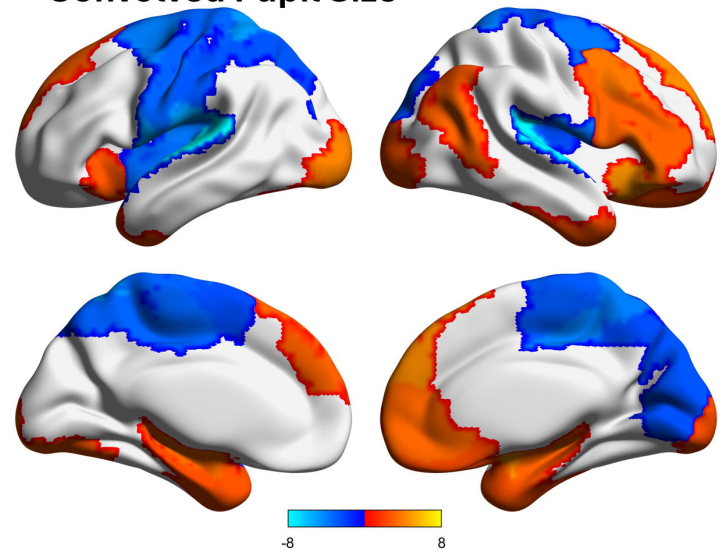
A. Pupil Response to Frame Luminance



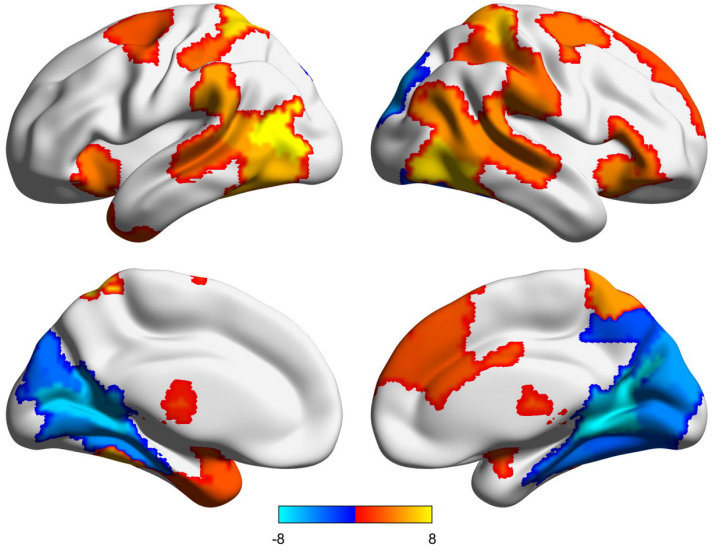
B. Cross-correlation Map



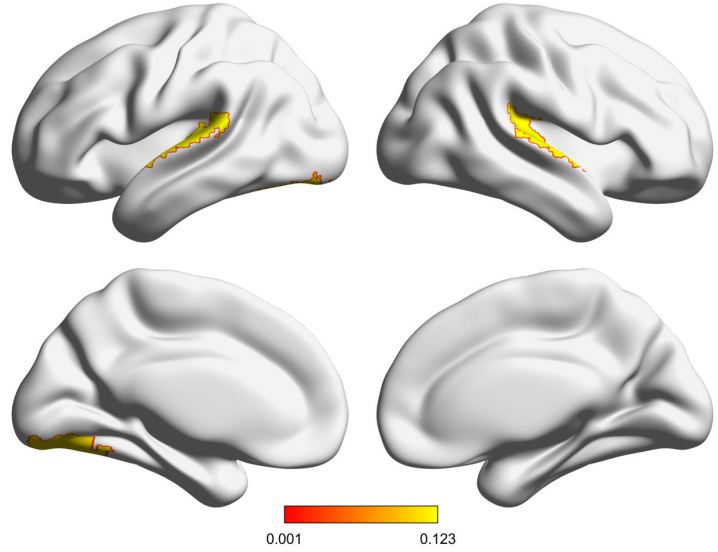
C. Zero-lag Correlation Map with Convolved Pupil Size



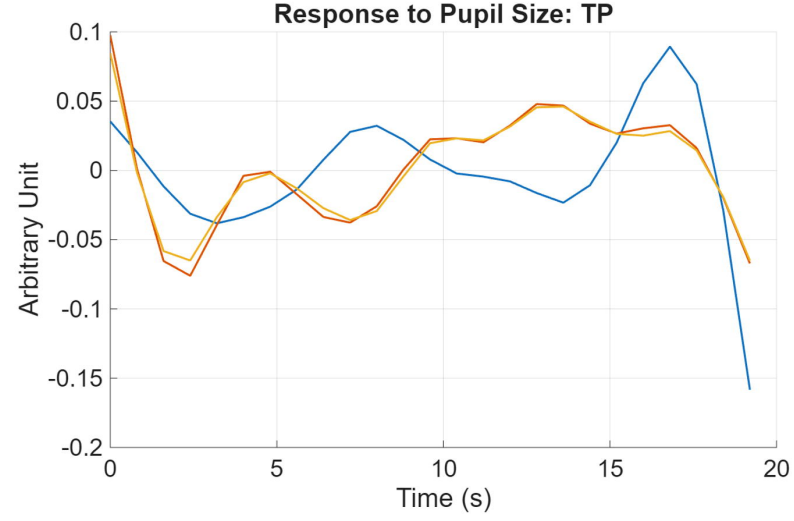
D. Zero-lag Correlation Map with Raw Pupil Size

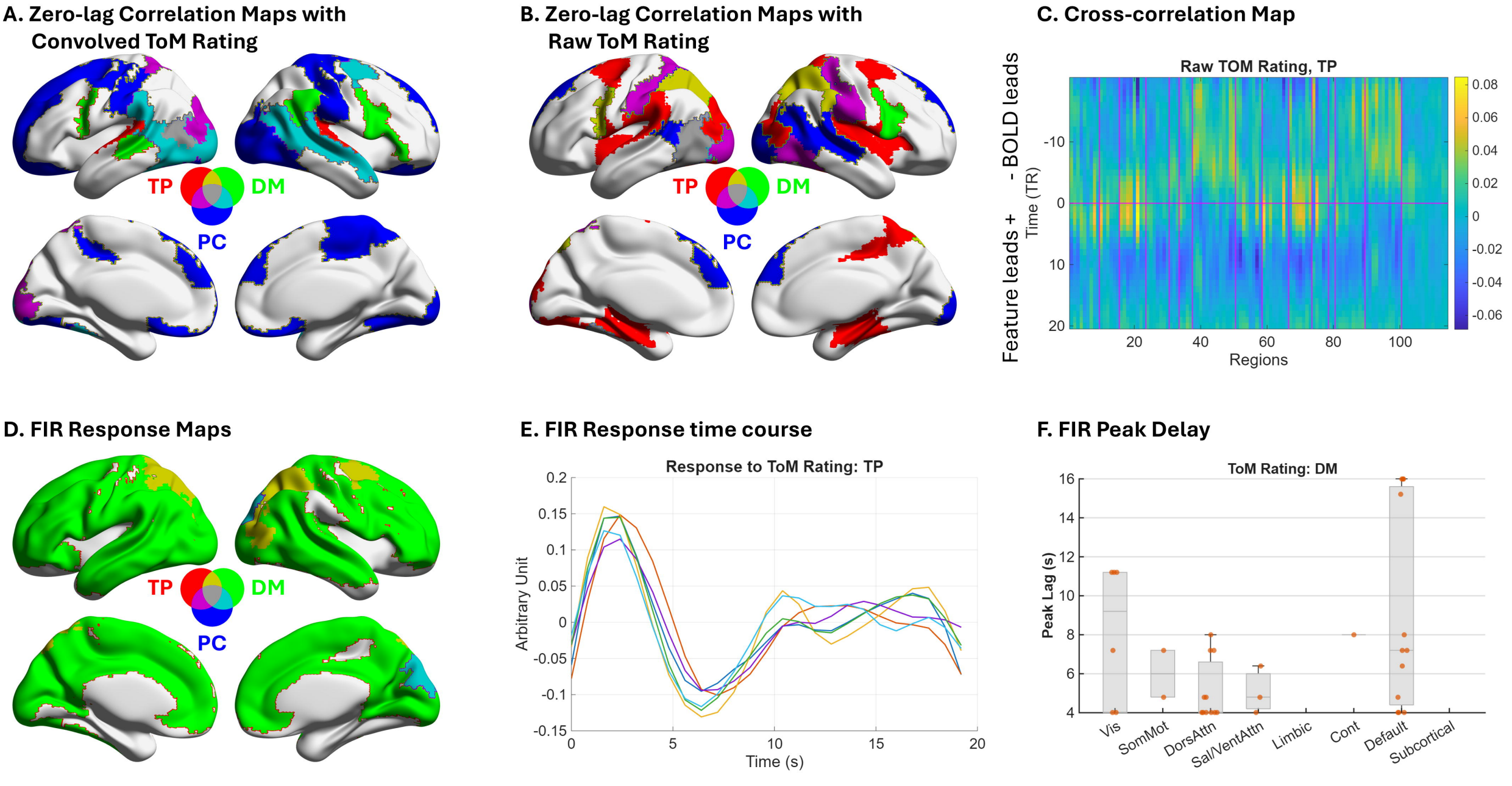


E. FIR Response Map



F. FIR Response to Pupil Size




A. Zero-lag Correlation Maps with Convolved ToM Rating
TP
DM
PC
B. Zero-lag Correlation Maps with Raw ToM Rating
TP
DM
PC
C. Cross-correlation Map
Raw TOM Rating, TP
Feature leads + - BOLD leads
Time (TR)
Regions
D. FIR Response Maps
TP
DM
PC
E. FIR Response time course
Response to ToM Rating: TP
Arbitrary Unit
Time (s)
F. FIR Peak Delay
ToM Rating: DM
Peak Lag (s)
Vis
SomMot
DorsAttn
Sal/VentAttn
Limbic
Cont
Default
Subcortical