HOW DOES LONG CONTEXT ENHANCE MULTIMODAL NARRATIVE VIDEO PROCESSING IN THE BRAIN?

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how humans and artificial intelligence systems process complex narrative videos is a fundamental challenge at the intersection of neuroscience and machine learning. We investigate how the temporal context length of video clips (3–24 s long) and the narrative-task prompting shape brain-model alignment during naturalistic movie watching. Using fMRI recordings from participants viewing full-length movies, we examine how brain regions sensitive to narrative context dynamically represent information over varying timescales and how these neural patterns align with model-derived features. We find that increasing clip duration substantially improves brain alignment for multimodal large language models (MLLMs), whereas unimodal audio/video models show little to no gain. Further, shorter temporal windows align with perceptual and early language regions, while longer windows preferentially align higher-order integrative regions, mirrored by a layer-to-cortex hierarchy in MLLMs. Finally, experiments with four narrative-task prompts show that they elicit task-specific, region-dependent brain alignment patterns and context-dependent shifts in clip-level tuning in higher-order regions. Our work positions long-form narrative movies as a principled testbed for studying long-timescale temporal integration in long-context MLLMs and its relationship to cortical responses during narrative comprehension.

1 INTRODUCTION

Brain–model correspondence studies show that contextual Transformer representations better predict cortical activity than static embeddings across language and speech (Toneva & Wehbe, 2019; Vaidya et al., 2022), with multimodal models further improving brain alignment (Nakagi et al., 2024; Subramaniam et al., 2024; Dong & Toneva, 2023a; Oota et al., 2025c; Sartzetaki et al., 2025). However, the role of extended temporal context in naturalistic multimodal settings remains unclear. Narrative movie comprehension requires integrating visual, auditory, and contextual information over long timescales, yet prior brain-encoding studies with multimodal models have largely relied on short fixed windows, often under 2 s (Subramaniam et al., 2024; Dong & Toneva, 2023a; Oota et al., 2025c). This is notable because higher-order cortical regions are known to accumulate narrative information over tens of seconds and support story-level understanding (Lerner et al., 2011; Simony et al., 2016; Baldassano et al., 2018; Khosla et al., 2021). These findings motivate evaluating MLLMs on naturalistic movies while systematically increasing temporal context.

Using fMRI recordings from participants watching naturalistic movies with audio (Boyle et al., 2020), we examine the brain alignment of pretrained MLLMs under varying temporal context. As shown in Fig. 1, we evaluate Qwen-2.5-Omni (Xu et al., 2025), which jointly processes video, audio, and a textual task instruction, and DATE (Yuan et al., 2025), which processes video and a textual instruction without audio. As cross-family benchmarks, we also evaluate unimodal-video models (TimeSFormer (Bertasius et al., 2021), VideoMAE (Tong et al., 2022)), and the audio-only Audio Spectrogram Transformer (Baade et al., 2022). Across four narrative video understanding tasks, we vary temporal context using clip duration (3, 6, 9, 12, 18, and 24 s) and measure brain alignment across tasks and context lengths.

We address four research questions: (1) How does temporal context length (3–24 s) influence the brain predictivity of MLLMs and unimodal audio/video model representations during naturalistic movie viewing? (2) Which brain regions show the largest gains or shifts in optimal context length,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

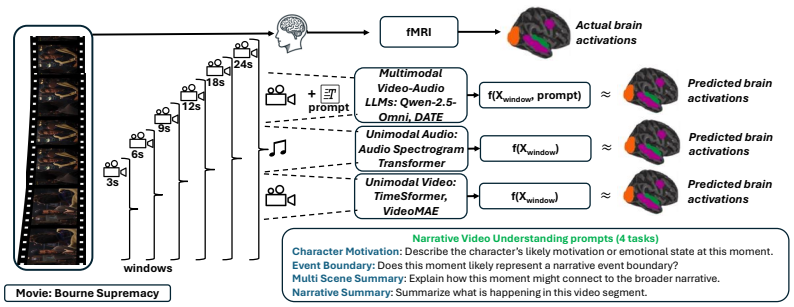


Figure 1: Leveraging temporal video context of different durations (X_{windows}) with unimodal and multimodal models for brain encoding with a diverse set of instructions (prompts). We experiment with 4 narrative video understanding tasks: character motivation, event boundary detection, multi-scene summarization, narrative summarization.

and how do these effects relate to layer-wise MLLM representations? (3) Which narrative-task instructions (character motivation, event boundary detection, multi-scene summary, and narrative summary) yield the strongest brain alignment, and do they exhibit distinct ROI-specific (region of interest) patterns? (4) Which video clips are most predictive of voxel responses across context lengths and narrative tasks, and how do these patterns vary across ROIs?

Our analyses yield four main findings. (1) Increasing temporal context (3–24 s clips) consistently improves brain predictivity for MLLMs, whereas unimodal audio and video models benefit little from longer contexts. (2) We observe an ROI-specific temporal hierarchy. Short contexts (3–6 s) best align with perceptual and early language regions (e.g., PTL or posterior temporal lobe), while longer contexts preferentially align with higher-order semantic regions (e.g., PCC or posterior cingulate cortex), mirroring the layer hierarchy of MLLMs. (3) Narrative-task prompting acts as a functional probe. *Narrative Summary* and *Multi-scene Summary* best explain activity in higher-order language regions, whereas *Character Motivation* aligns more strongly with localized temporal-language areas. (4) Maximally predictive clips remain largely stable across temporal windows in visual cortex but shift with context length in higher-order language regions, indicating context-dependent narrative representations.

Overall, our results show that temporal context and narrative-task prompting provide complementary probes of long-form narrative understanding in both the brain and MLLMs. We provide one of the first systematic studies of extended temporal context and narrative prompting in MLLMs, demonstrating distinct context- and task-dependent brain-alignment patterns across cortical regions and model layers. Code is included in the supplementary material. Related work on contextual representations, multimodal brain encoding, and prior evaluation settings is reviewed in App. A, with representative stimulus modalities and temporal windows summarized in Table 1.

2 Dataset Curation and Tasks

Brain imaging dataset. We use Movie10 (Boyle et al., 2020), a naturalistic multimodal fMRI dataset from Courtois NeuroMod, comprising five publicly available participants (s1, s2, s3, s5 and s6) watching four movies: *The Bourne Supremacy*, *The Wolf of Wall Street*, *Hidden Figures*, and *Life*. *Hidden Figures* and *Life* were each viewed twice, yielding repeated measurements. fMRI volumes were acquired every 1.49 s (1 TR). Since brain-encoding models are trained separately per participant, the large number of samples per subject is more important than the number of subjects. The dataset has been used in prior multimodal brain-encoding studies (Oota et al., 2025c; d'Ascoli et al., 2026). We use the preprocessed fsaverage6 surface data and report ROI analyses using multimodal parcellation of the human cerebral cortex based on Glasser atlas (180 ROIs per hemisphere) (Glasser et al., 2016). Language ROIs are defined using Fedorenko language parcels (Milton et al., 2021; Desai et al., 2023). We show the flatmap with labeled ROIs in Appendix Fig. 7. Additional dataset, preprocessing, and sub-ROI details are provided in App. B.

Estimating cross-subject prediction accuracy. To account for intrinsic noise in biological measurements, we adapt the method of Schrimpf et al. (2021); Oota et al. (2024) to estimate the cross-subject prediction accuracy for a model's performance for the Movie10 fMRI dataset. Each subject

Table 1: Overview of multimodal model evaluation settings in brain encoding studies.

Study	Model Type	Stimulus Modality	Brain Data	Dataset	Video Len
Popham et al. (2021)	Vision-Only CNNs vs. Vision-Language	Unimodal (Silent Videos)	fMRI	Gallant lab short video clips	2s
Tang et al. (2024)	Non-instruction-tuned multimodal model (BridgeTower)	Unimodal (Silent Videos), Unimodal (listening stories)	fMRI	Gallant lab short video clips	2s
Subramaniam et al. (2024)	Non-instruction-tuned multimodal models (SLIP-CLIP, SimCLR, BLIP, BEiT)	Image frame-text pairs (Movies)	SEEG	AMMT	4s
Lahner et al. (2024)	ResNet-50 with Temporal Shift Module	Video clips	fMRI	BOLD Moments	3s
Dong & Toneva (2023a)	Non-instruction-tuned multimodal models (MERLOT Reserve)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Friends dataset	35s
Oota et al. (2025c)	Non-instruction-tuned multimodal models (TVLT and ImageBind)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	1.49s
Sartzetaki et al. (2025)	Object and Action recognition models	Image frames to CNNs, Video clips to video models	fMRI	BOLD Moments	3s
Oota et al. (2025b)	Instruction-tuned video and audio MLLMs	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	1.49s
Ours	Unimodal Audio/Video models and Video-based MLLMs	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	3s - 24s

$s \in$ (s1, s2, s3, s5, s6) is chosen as the prediction target and the other four are used to predict this target. Note that the estimated cross-subject prediction accuracy is based on the assumption of a perfect model, which might differ from real-world scenarios, yet offers valuable insights into model's performance. We report complementary cross-subject and repeat-based within-participant reliability analyses in App. C. Across the 14 ROIs and five participants, 70 participant–ROI combinations, all cross-subject correlations are positive (mean $r = 0.104$; 69 above 0.05), with 38.9% of vertices retained; these values are comparable to established naturalistic-fMRI datasets (Oota & Toneva, 2023) (Table 4). Repeat-based reliability provides converging evidence (mean EV $= 0.172$; 69/70 combinations above 0.05; 57.41% of vertices retained; Table 5). Importantly, the same retained vertex set is used for every model, window, and prompt, so normalization rescales alignment by a fixed positive ceiling at each vertex and cannot change which condition wins.

Narrative Tasks. We focus on four core narrative tasks (Character Motivation, Event Boundary Detection, Multi-Scene Summary, and Narrative Summary; see the task instructions in App. D) because together they capture the complementary and irreducible components of narrative understanding. This choice rests on three mild assumptions. First, narrative comprehension can be decomposed into local semantics, agent-centric inference, temporal segmentation, and cross-scene integration (Bruner, 1991; Abbott, 2008). Second, a task-conditioned multimodal model is capable of recovering each component (Zellers et al., 2021; Gupta et al., 2022). Third, the outputs of these tasks can be linearly, or through a low-dimensional nonlinear readout, composed into downstream predictions. Empirical evidence from neuroimaging studies demonstrates that different cortical regions align with these narrative components, and that combining task outputs improves voxel-level prediction of brain activity during story comprehension (Huth et al., 2016; Wehbe et al., 2014).

3 METHODOLOGY

Multimodal Large Language Models (MLLMs). To study long-form narrative video understanding, we evaluate two 36-layer pretrained MLLMs: (i) Qwen-2.5-Omni (Xu et al., 2025), which jointly processes video, audio, and task instructions, and (ii) DATE (Yuan et al., 2025), built on Qwen-2.5-VL, which processes video (without audio) and task instructions and is designed for long-video understanding and temporal reasoning. Narrative-task prompts elicit task-specific representations used for brain encoding. To isolate modality and prompting effects without changing model architecture, weights, stimuli, or the encoding pipeline, we evaluate Qwen-2.5-Omni with audiovisual, video-only, and audio-only inputs, and DATE with and without task-text conditioning. Each condition is processed independently and used to fit a separate voxel-wise encoding model.

Extraction of Temporal Context from Long Videos. We extract overlapping video windows of length $W \in \{3, 6, 9, 12, 18, 24\}$ s with a 1.49 s stride. For each window, we sample 9, 9, 12, 16, 24, and 30 frames respectively, pair them with a narrative task instruction, and feed them to the MLLMs. Qwen-2.5-Omni receives frames, synchronized audio, and the task instruction, whereas DATE receives frames and the instruction only. A constant-density (1 fps) ablation confirms robustness to frame sampling (App. I). Hidden states from all Transformer layers are averaged across generated tokens, yielding one layer-wise representation per window. All representations are extracted using pretrained model weights.

Unimodal Baselines. We compare against 12-layer unimodal video models (VideoMAE, TimeSFormer), and audio model (AST). These serve as cross-family benchmarks, differing from MLLMs

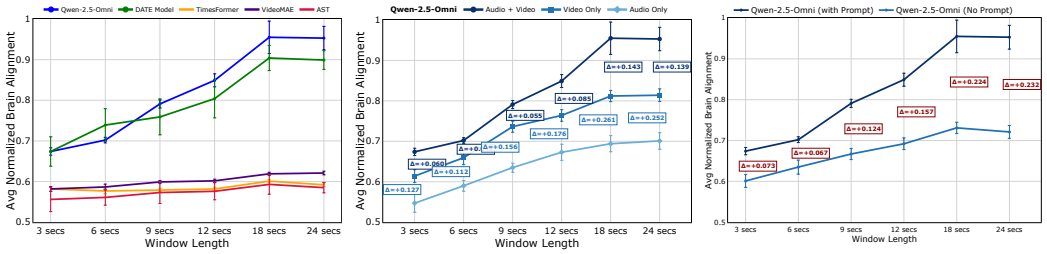


Figure 2: Average normalized brain alignment as a function of temporal window length (3 to 24s). (a) Cross-family model comparison, including the audio-only AST benchmark (left), (b) Within-modality ablation with Qwen-2.5-Omni comparing video+audio, video-only, and audio-only inputs (middle), (c) Prompting ablation within Qwen-2.5-Omni and "Video included with audio" setting (right). Panels (b) and (c) disentangle the contributions of modality and prompting from temporal context, showing that both audio and prompting gains grow with window length, with prompting the largest contributor at longer windows. Within-DATE, prompting ablation comparing video+text prompt with video-only input is in App. H Fig. 9. Error bars denote mean±SEM across subjects.

in architecture, scale, depth, and pretraining. Using the same sliding-window pipeline, we extract layer-wise representations from video-only inputs (without task prompts), average hidden states within each window, and use the resulting embeddings for brain encoding.

4 Experiments

Encoding model. We train voxel-wise fMRI encoding models using bootstrap ridge regression (Tikhonov & Arsenin, 1977) to predict voxel responses from stimulus representations obtained from the 4 narrative task-specific instructions for both MLLMs. To detect and remove outliers, stimulus representations and fMRI responses are z-scored separately for training and testing. To account for hemodynamic delay, we model it using a finite impulse response (FIR) per voxel with 5 temporal delays (TR) corresponding to $\sim$7.5 s (Huth et al., 2022). At time step t, let stimulus encoding be $X_t \in \mathbb{R}^D$ and voxel responses be $Y_t \in \mathbb{R}^V$. D is dimension of the concatenation of 5 delayed TRs, and V is number of voxels.

Train-test setup. For each participant, independent encoding models are trained on *The Bourne Supremacy* (4,024 TRs) and *The Wolf of Wall Street* (6,993 TRs) and tested on the disjoint *Life* movie (2,028 TRs), yielding a stringent cross-movie generalization setting without train–test overlap. As *Life* contains narration, agents, goal-directed behavior, causal events, and multi-scene structure, all four narrative prompts are applicable. Because all conditions use the same test film, domain shift does not confound comparisons across models, context windows, or prompts. We additionally evaluate Qwen-2.5-Omni on *Hidden Figures*, a second held-out film, using the same protocol. Hyperparameters are provided in App. E.

Evaluation metrics. We evaluate brain alignment using Pearson correlation (PC), a standard metric in brain-encoding studies (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022). For each voxel, PC measures the correlation between predicted responses $\hat{Y}$ and observed neural activity Y over the test set. Correlations are normalized by the estimated cross-subject prediction accuracy and averaged across voxels, ROIs, and participants, yielding normalized brain alignment. Following prior work (Popham et al., 2021; La Tour et al., 2022), we retain voxels with cross-subject prediction accuracy ≥ 0.05.

5 Results

[A1]: Longer time context boosts brain alignment for MLLMs but not for unimodal models

(a) Avg normalized brain alignment vs. temporal context length. Fig. 2a compares average normalized brain alignment across context lengths using the best-performing layer of each model (layer 36 for MLLMs, layer 12 for unimodal baselines; layer-wise profile is in Fig. 4). (i) As context increases from 3 s to 24 s, Qwen-2.5-Omni and DATE show substantial gains in alignment

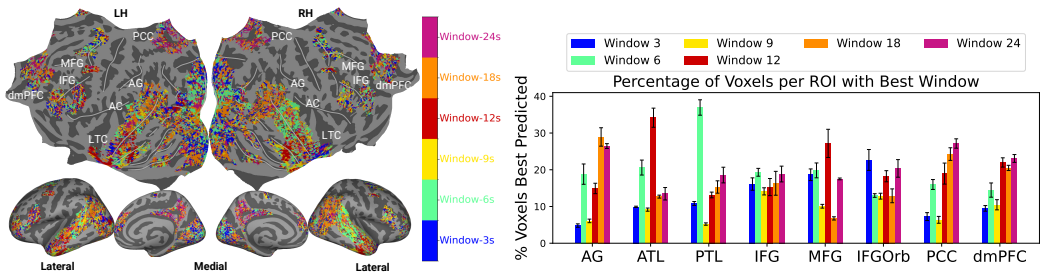


Figure 3: Each voxel is color-coded with the video duration (window length) that led to the highest normalized brain alignment with the Qwen-2.5-Omni model. (Left) The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. (Right): Percentage of best predicted voxels whose brain encoding performance is higher corresponding to each window length within language-selective regions. Error bars denote SEM. Results for visual-selective regions are in App. G Fig. 8. Results for DATE are in App. H.

(0.67 to 0.95, +41%; 0.67 to 0.90, +33%), whereas VideoMAE, TimeSFormer, and AST change little (0.56–0.62 range). (ii) Paired two-sided t-tests across subjects ($df = 4$) show Qwen-2.5-Omni significantly outperforms unimodal baselines at every context length, including 3 s (all $p \leq 0.002$); DATE significantly outperforms unimodal baselines for contexts $\geq$6 s ($p < 0.023$), but not at 3 s ($p = 0.079$); No significant difference is observed between Qwen-2.5-Omni and DATE at any window length (all $p \geq 0.435$). Subject-level analyses (App. G Table 8) show the long-context benefit for Qwen-2.5-Omni and DATE in all participants, with optimal windows at 18–24 s, while unimodal models exhibit smaller, non-monotonic changes. (iii) The effect generalizes across movies. On the held-out *Hidden Figures*, Qwen-2.5-Omni increases from 0.670 ± 0.026 at 3 s to 0.942 ± 0.017 at 18 s, remaining stable (0.941 ± 0.009) at 24 s. Overall, extended temporal context consistently benefits MLLMs, with or without audio input, but not unimodal audio or video models.

(b) Within-modality ablation: Audio+Video vs. Video-only vs. Audio-only. Holding Qwen-2.5-Omni fixed, we compare Audio+Video, Video-only, and Audio-only inputs across context lengths (Fig. 2b). Audio+Video consistently achieves the highest alignment, exceeding Video-only and Audio-only by $\Delta = +0.060/+0.127$ at 3 s and $\Delta = +0.139/+0.252$ at 24 s. Thus, audio and video provide complementary information, with benefits increasing at longer contexts.

(c) Prompting ablation: With vs. without prompts. Holding Qwen-2.5-Omni and Audio+Video inputs fixed, we compare prompted and unprompted representations (Fig. 2c). Prompting yields larger gains than adding audio, increasing from $\Delta = +0.073$ at 3 s to $\Delta = +0.232$ at 24 s, indicating that narrative instructions effectively steer representations toward brain-relevant features. Corresponding DATE results are shown in App. H, Fig. 9.

(d) Frame-sampling ablation: constant frame density vs. scaled sampling. To test whether long-context gains arise from temporal span or increased frame count (in our default scaled-sampling scheme), we repeated analyses with a constant 1 fps rate (Table 11, App. I). Brain alignment increased monotonically with context length under both sampling schemes, with only small differences ($\sim$0.03–0.06), indicating that gains are driven by temporal span rather than frame density.

[A2](A): LONG WINDOWS PREFERENTIALLY BENEFIT HIGHER-ORDER SEMANTIC REGIONS, WHILE INTERMEDIATE WINDOWS ARE OPTIMAL FOR SEVERAL MID-LEVEL TEMPORAL ROIS.

We next examine how shorter vs. longer temporal windows differentially affect brain alignment across language, visual, and auditory regions using a voxel-wise window-preference analysis. Specifically, for each voxel, we select the temporal context (clip duration) that results in the highest normalized brain alignment (averaged across subjects) and apply the window-specific color code to the voxel in Fig. 3 (left). We show brain maps for Qwen-2.5-Omni across the six temporal windows.

We observe a temporal hierarchy across ROIs. (i) Posterior temporal lobe (PTL) prefers intermediate context (6 s), consistent with its role in local syntactic processing (Friederici et al., 2003; Friederici, 2012; Lerner et al., 2011); (ii) Higher-order language ROIs (anterior temporal lobe (ATL), and middle frontal gyrus (MFG)) prefer longer context (12 s), consistent with their roles in semantic composition and lexical processing (Blank & Fedorenko, 2020; Toneva & Wehbe, 2019); (iii) Higher-order

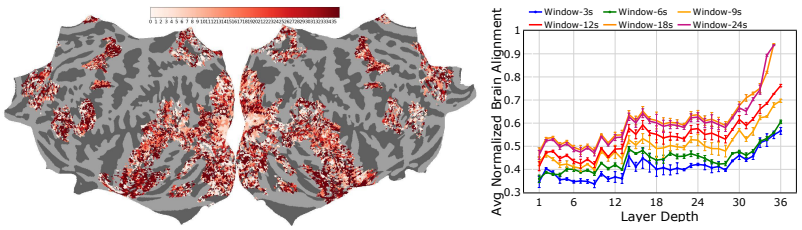


Figure 4: Layer-wise alignment results for Qwen-2.5-Omni: Each voxel is color coded with the MLLM layer number (out of 36) that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. Results for DATE are in App. J.

narrative associative regions (angular gyrus (AG), posterior cingulate cortex (PCC) and dorsomedial prefrontal cortex (dmPFC)) peak at longer contexts (18-24 s), consistent with their roles in narrative integration, episodic memory, and discourse-level processing (Hasson et al., 2008; Lerner et al., 2011); (iv) Frontal regions inferior frontal gyrus (IFG) and orbital IFG (IFGOrb) show strong alignment for both shorter and longer context, suggesting sensitivity across multiple timescales, consistent with their role in hierarchical syntactic and semantic processing (Friederici, 2012); (v) Across ROIs, although 9 s is often competitive (e.g., runner-up in PTL; Table 2), it rarely dominates. Overall, shorter contexts favor lower-order regions and longer contexts favor higher-order areas, broadly matching cortical hierarchy of temporal receptive windows (Hasson et al., 2008; Lerner et al., 2011).

Fig. 3 (right) shows ROI-specific preferences across the language network. Higher-order regions (AG, PCC, dmPFC) favor long contexts (18–24 s), while ATL and MFG peak at 12 s. PTL is best predicted at 6 s and EVC favors short windows (3–6 s), revealing a progression from short-context sensory/language regions to long-context narrative regions. Other ROIs, including IFG and IFGOrb, exhibit intermediate or heterogeneous preferences. Subject-level analyses largely mirror these trends (App. G Table 9): 9/12 ROIs show agreement in at least four of five participants, with strongest consistency in PCC, dmPFC, and EVC, whereas IFG, IFGOrb, and OFA exhibit greater variability. We therefore interpret the group-level maps as dominant tendencies rather than universal subject-level preferences.

Table 2 shows ROI-level margins between the best and second-best windows (median $\Delta = 0.064$, range: 0.021–0.159). Small margins in OFA/PPA and participant-level variability in IFG/IFGOrb warrant cautious interpretation. DATE shows similar but smaller effects (median $\Delta = 0.038$; range: 0.008–0.066; Appendix Table 10). Overall, results reveal a cortical temporal gradient. Shorter windows favor perceptual and temporal regions, whereas longer windows maximize alignment in higher-order semantic and narrative regions.

Consistent with the pattern in Fig. 3, DATE shows a broadly similar temporal gradient to Qwen-2.5-Omni (App. H, Fig. 10, Table 10): PCC and dmPFC prefer 24 s in all five participants, AG prefers 18 s, ATL and MFG prefer 12 s, and PTL and EVC prefer 6 s, as for Qwen-2.5-Omni. The models differ in several ROIs: DATE's preferences are more distributed across windows, with LOC peaking at 9 s (18 s for Qwen-2.5-Omni) and PPA at 12 s (24 s for Qwen-2.5-Omni), and its margins are smaller (median Δ=0.038 vs. 0.064), indicating weaker temporal selectivity. In contrast, the unimodal video baselines show little to no gain in average normalized brain alignment as temporal context increases. App. H Fig. 11 shows the brain map for the unimodal baselines. The voxel-wise maps indicate that longer temporal windows tend to yield higher alignment in higher-order language/association regions, whereas shorter windows are more predictive in early sensory regions.

Table 2: Temporal-window effect magnitudes for Qwen-2.5-Omni. Best and runner-up cells report the window and mean±SD across five participants. Δ denotes the best–runner-up difference. The final row reports the participant-level preference and agreement.

	AG	ATL	PTL	IFG	MFG	IFGOrb	PCC	dmPFC	EVC	LOC	OFA	PPA
Best	18 s .968±.013	12 s .791±.072	6 s .779±.042	18 s .908±.149	12 s .892±.196	24 s .845±.140	24 s .902±.016	24 s .957±.073	6 s .671±.063	18 s .852±.053	18 s .873±.074	24 s .900±.057
Second Best	24 s .908±.032	6 s .672±.007	9 s .712±.029	12 s .821±.157	24 s .832±.182	3 s .686±.158	18 s .855±.086	18 s .851±.097	3 s .568±.055	12 s .794±.051	24 s .852±.068	18 s .866±.080
Δ	.060	.119	.067	.087	.060	.159	.047	.106	.103	.058	.021	.034
Participant preference	18 s (3/4)	12 s (3/4)	6 s (3/4)	Heterogeneous	12 s (3/4)	3 s (2/4)	24 s (4/4)	24 s (4/4)	3–6 s (4/4)	18 s (3/4)	18 s (2/4)	24 s (3/4)

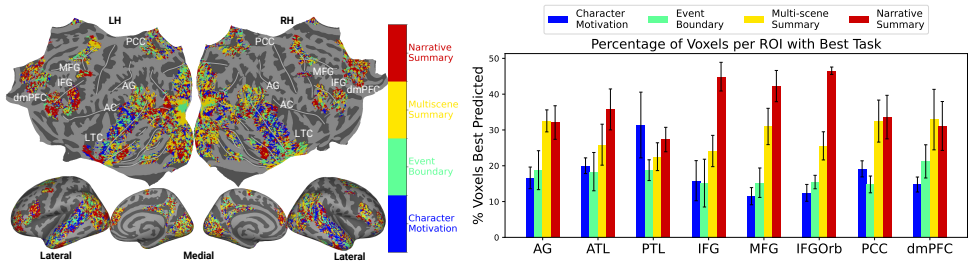


Figure 5: (Left) Each voxel is color-coded with the instruction that led to the highest normalized brain alignment for Qwen-2.5-Omni model. The color bar highlights color codes for each instruction. The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. (Right) Percentage of voxels in each ROI corresponding to each instruction. Error bars denote SEM. Results for visual-selective regions are in App. K Fig. 15. Results for DATE are in App. K.

[A2](B): LAYER-WISE MLLM REPRESENTATIONS FORM A CORTICAL LANGUAGE HIERARCHY AND SHIFT WITH TEMPORAL CONTEXT.

We investigate how MLLM brain-encoding performance varies across model layers as the temporal context window increases. Unlike the temporal-window comparison, which fixes one representative layer per model family (Sec. [A1]), Fig. 4 evaluates brain alignment across all model layers. Across all temporal windows, Fig. 4 (right) shows a consistent stratification into three layer groups: early layers (1–12), middle layers (13–28), and late layers (29–36). Normalized brain alignment generally improves from early to late layers across temporal windows, suggesting that later-layer representations are more predictive of the measured brain responses.

Fig. 4 (left) visualizes voxel-wise layer preferences, projected onto the `fsaverage` subject. We observe that (i) early sensory regions (early visual and auditory cortex) prefer lower layers, consistent with shallow representations capturing low-level sensory features; (ii) higher-level visual regions such as lateral occipital complex (LOC) and parahippocampal place area (PPA), and temporal/parietal language ROIs (PTL and AG) tend to align with middle-to-late layers; and (iii) language-related regions such as inferior frontal gyrus (IFG), anterior temporal lobe (ATL), and angular gyrus align most strongly with the deepest layers. Together, these results suggest that MLLMs exhibit a layered representational hierarchy that broadly mirrors cortical processing hierarchies. Layer-wise brain maps for each subject for Qwen-2.5-Omni are in App. J Fig. 13. Layer-wise results for DATE, including where they differ from Qwen-2.5-Omni, are in App. J Fig. 12.

[A3]: NARRATIVE TASK-SPECIFIC REPRESENTATIONS SHOW ROI-SPECIFIC DISSOCIATIONS WHILE ENCODING.

To investigate which task instructions are more effective in predicting brain activity and whether MLLMs differentiate task-specific representations and provide clear separation in brain regions, we assign each voxel the task with highest normalized brain alignment and color-code accordingly (Fig. 5, left). For each ROI, we report fraction of voxels where each task is optimal (Fig. 5, right).

From Fig. 5 (left): (i) *Narrative and multiscene summary* tasks explain a large fraction of voxels across widespread association cortex, consistent with these tasks emphasizing long-range semantic integration over extended narrative context. (ii) *Character Motivation* contributes most in posterior lateral temporal cortex (PTL; 3/5 participants), suggesting that character-centric inference aligns with regions supporting person- and situation-level representations, although this prompt is the least matched to the animal-centered *Life* stimulus (preferred by only 17% of vertices).

Table 3: Narrative-task effect magnitudes for Qwen-2.5-Omni. Best and runner-up cells report the task and mean±SD across five participants. Δ denotes their difference. CM: Character Motivation; EB: Event Boundary Detection; MSS: Multi-Scene Summary; NS: Narrative Summary.

	AG	ATL	PTL	IFG	MFG	IFGOrb	PCC	dmPFC	EVC	LOC	OFA	PPA
Best	MSS .993±.091	NS .921±.082	CM .790±.090	NS .906±.036	NS .911±.275	NS .918±.133	NS .963±.301	NS .970±.100	MSS .792±.175	EB .920±.221	MSS .804±.121	MSS .860±.089
Second Best	NS .973±.178	MSS .804±.040	NS .745±.090	MSS .844±.161	MSS .906±.376	MSS .805±.016	MSS .904±.064	MSS .884±.160	NS .701±.070	NS .832±.065	CM .780±.190	NS .843±.261
Δ	.020	.117	.045	.062	.005	.113	.059	.086	.091	.088	.024	.017
Participant preference	MSS (4/5)	NS (5/5)	CM (3/5)	NS (5/5)	NS (4/5)	NS (5/5)	Tie (3/5 each)	NS (4/5)	MSS (4/5)	EB (3/5)	MSS (3/5)	MSS (4/5)

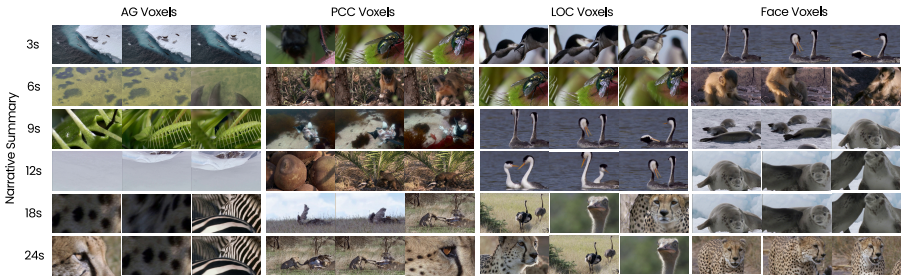


Figure 6: Temporal window conditioned top-K video clips driving voxel responses for Narrative summary task: Results for ROI selective voxels for Subject-1. We identify the top-10 videos from the test stimulus with highest average activation in each temporal window of ROI selective voxels for the AG, PCC, LOC and Face regions. *For visualization, we show only the single highest-ranked video segment in each condition; the three columns display three frames sampled in temporal order from the same (W)-second segment; they do not represent three different clips.*

(iii) *Event Detection* is sparse and localized, indicating that boundary/event-focused representations explain fewer voxels overall and may be confined to more specialized subregions. (iv) The variation in winning prompts across voxels indicates that different task-conditioned representations preferentially predict different voxels within the same cortical region. Thus, task-instruction prompts serve as functional probes of the narrative information represented in voxel responses, revealing heterogeneous representational preferences.

Fig. 5 (right) summarizes task preferences across language ROIs. *Narrative Summary* dominates frontal and temporal language regions (IFG, MFG, IFGOrb, ATL, PTL) and dmPFC, indicating that global narrative integration prompts yield representations most predictive of these regions. *Multi-scene Summary* is most prominent in integrative hubs AG and PCC. *Character Motivation* contributes relatively more in temporal language ROIs (ATL and PTL), whereas *Event Boundary Detection* is rarely the plurality winner, with comparatively higher contributions in dmPFC (and to a lesser extent AG). Table 3 quantifies task separation; the median best–runner-up margin is 0.061 (0.005–0.117), largest for ATL and IFGOrb and smallest for MFG, PPA, AG, and OFA. DATE shows similar effects (median margin 0.060; range of 0.016–0.135; Appendix Table 13).

The Subject-level task preferences are reported in the Appendix Table 12. Eight of the 12 reported ROIs show agreement in at least three of five participants. Narrative Summary is preferred unanimously in ATL, IFG, MFG, IFGOrb and dmPFC, whereas PTL, PCC, LOC, and OFA show heterogeneous or less consistent preferences. Overall, integration-focused prompts (Narrative/Multi-scene Summary) dominate higher-order ROIs, while agent-centric prompts (Character Motivation) contribute more in temporal language areas. DATE also shows ROI-dependent task effects, but it does not uniformly reproduce the Qwen pattern and more frequently favors *Multi-scene Summary* where Qwen favors *Narrative Summary* (App. K, Fig. 14). Further, interpretability analyses of voxel-wise encoding weights across tasks and context windows are in App. M Fig. 17.

Robustness of task-instruction prompts. To examine whether our task prompts are sensitive to wording or reflect broader task-conditioned representational structure, we performed two additional analyses using MLLM-derived task-conditioned embeddings: (i) paraphrased rewrites preserving meaning, and (ii) generic prompts stripped of task-specific content (see App. D Table 7). We computed cosine similarity between the original prompt-conditioned embeddings and those obtained under the rephrased and generic prompts from MLLMs. Original-to-rephrased similarities are uniformly high (>0.93), while original-to-generic similarities drop substantially (0.67-0.75). This confirms that our task prompts encode meaningful task-specific semantic structure that is preserved under paraphrasing but lost when content is genericized, supporting the interpretation that observed brain-alignment patterns reflect genuine task-conditioning rather than surface text effects.

[A4]: MAXIMALLY ACTIVATING VIDEO CLIPS ARE STABLE IN VISUAL ROIS BUT SHIFT WITH TEMPORAL CONTEXT IN HIGHER-ORDER LANGUAGE ROIS.

To select the video clips that elicit the highest response in a given voxel, we used the voxelwise estimated encoding-model weights across temporal windows. For each temporal window w and voxel v, let $\beta_v^{(w)} \in \mathbb{R}^F$ be the learned weight vector and $x_c^{(w)} \in \mathbb{R}^F$ be the feature vector for clip

c. We scored every clip by dot product $s_v^{(w)}(c) = (\beta_v^{(w)})^\top x_c^{(w)}$, and selected largest scoring K=10 clips. Top-ranked clips are those predicted by encoding model to elicit the strongest responses in voxel v for window w, offering an interpretable view of voxel tuning in the feature space.

Window-conditioned retrieval. We analyze 1000 best-predicted voxels across 6 temporal windows to identify which video clips drive neural responses during long-form movie processing. Fig. 6 shows retrieved clips for representative voxels in visual (LOC, face-selective cortex) and higher-order language regions (AG, PCC) across context windows. We quantify temporal consistency using Jaccard similarity (J) of retrieved videos across window pairs. Visual ROIs show high consistency (LOC: J=0.50; Face: J=0.38) with category-coherence (object-/shape-dominated for LOC and face/head-dominated for face-selective voxels). But, AG (Win3 to Win24: ΔJ = 0.094) and PCC (Win3 to Win24: ΔJ = -0.153) exhibit larger shifts in retrieved clips as context length increases, suggesting greater sensitivity to broader semantic and narrative context.

Task-conditioned retrieval. We next examine whether the identity of the narrative-task instruction changes which video clips are predicted to maximally activate voxels within a given ROI. Using the same scoring-and-retrieval procedure as above, we select the top-K clips for ROI-selective voxels under each task instruction and visualize the resulting clips for AG, PCC, LOC, and face-selective voxels. Fig. 16 in App. L shows retrieved clips under different narrative-task prompts. To quantify task-dependent modulation, we compute Jaccard similarity between retrieved video sets for each ROI under different task instructions. Similar to Window-conditioned retrieval, visual ROIs remain stable across tasks (LOC: J = 0.42, Face: J = 0.38), whereas AG and PCC vary more across prompts (AG: J = 0.04, PCC: J = 0.13), indicating that task instructions modulate the semantic/contextual aspects of representations that best align with higher-order regions.

We observe that (i) in category-selective visual ROIs (LOC and face-selective cortex), the retrieved clips remain visually coherent and category-consistent across tasks (e.g., object/texture- and face/animal-head-dominated clips), suggesting that these regions are driven primarily by stable visual content rather than task framing; (ii) in higher-order regions (AG and PCC), the retrieved clips vary more substantially across task instructions, indicating that task prompting modulates the semantic/contextual aspects of the representation that best align with these ROIs. For example, *Character Motivation* tends to retrieve agent-centric clips with salient animals/faces, whereas *Event Boundary* and summarization-style prompts retrieve more context-rich exemplars.

Together with the window-based retrieval (Fig. 6), these results suggest that visual ROIs show stable tuning across both temporal contexts and tasks, whereas higher-order ROIs exhibit stronger task- and context-dependent shifts in the stimuli predicted to drive voxel responses.

6 Conclusion

We studied how temporal context and narrative-task prompting affect brain–model alignment during naturalistic movie watching using voxel-wise encoding on Movie10, comparing two video–audio MLLMs with unimodal baselines, and obtained these findings. **First**, increasing clip duration (3–24 s) substantially improves brain alignment for both MLLMs but not for unimodal models (Fig. 2). This suggests gains arise from temporal and multimodal integration rather than simply adding frames, highlighting long-form narrative video as a probe of brain-relevant timescales in MLLMs. **Second**, context-length preferences reveal an ROI-specific temporal gradient: shorter windows (3–6 s) align with perceptual and early language regions (PTL), whereas longer windows align with higher-order integrative regions (PCC, dmPFC, ATL); this pattern is mirrored in model depth, with deeper layers better predicting higher-level cortex (Figs. 3, 4), indicating a representational hierarchy paralleling cortical processing during movie watching. **Third**, narrative-task instructions produce task-specific, ROI-dependent representations rather than a single generic embedding: different instructions yield distinct voxel/ROI alignment patterns (Fig. 5), and the top-ranked video clips driving voxels differ by ROI and context (Fig. 6, 16). **Fourth**, visual ROIs (LOC, face-selective cortex) show stable clip preferences across windows, while higher-order regions (AG, PCC) exhibit strong context-dependent shifts. Overall, MLLM representations show a layer hierarchy and temporal-context sensitivity that track cortical processing during narrative processing. Task conditioning provides a useful functional probe of brain-aligned representations. Our results bridge neuroscience studies of long-timescale narrative processing with interpretable evaluations of long context MLLMs. We discuss limitations in App. N.

AI USE STATEMENT

We did not use generative AI tools for any of the tasks requiring mandatory disclosure under the ICLR AI policy. We acknowledge the use of AI tools (Grammarly, OpenAI GPT-5.5, Claude Opus 4.8) for grammar correction, refining phrasing for better readability, and refining the presentation of figures. All AI-assisted edits were verified and finalized by the authors.

ETHICS STATEMENT

We did not create any new neural recordings data as part of this work. We used the Movie10 dataset which is publicly available without any restrictions. Movie10 dataset can be downloaded from `https://github.com/courtois-neuromod/movie10/tree/33a97c01503315e5e09b3ac07c6ccadb8b887dcf`. Please read their terms of use[1] for more details.

We do not foresee any harmful uses of this technology.

REPRODUCIBILITY STATEMENT

We provide clear details on the fMRI dataset (Sec. 2), the implementation details E, and the training procedure in the Methodology. We release the code in the supplementary.

REFERENCES

H. Porter Abbott. *The Cambridge Introduction to Narrative*. Cambridge University Press, 2008.

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

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

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

Christopher Baldassano, Uri Hasson, and Kenneth A Norman. Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699, 2018.

Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

Idan A Blank and Evelina Fedorenko. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, 219:116925, 2020.

Julie A. Boyle, Basile Pinsard, Amal Boukhdhir, Sylvie Belleville, Simona Brambatti, Jeni Chen, Julien Cohen-Adad, André Cyr, Adrian Fuente, Pierre Rainville, and Pierre Bellec. The courtois project on neuronal modelling - first data release. In *26th OHBM annual meeting*. Organization for Human Brain Mapping (OHBM), 2020. URL `https://publications.polymtl.ca/50613/`.

Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.

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

[1] `https://docs.cneuromod.ca/en/latest/ACCESS.html`

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction. In *International Conference on Learning Representations*, volume 2026, pp. 152742–152759, 2026.

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

Rutvik H Desai, Usha Tadimeti, and Nicholas Riccardi. Proper and common names in the semantic system. *Brain Structure and Function*, 228(1):239–254, 2023.

Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023a.

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

Angela D Friederici. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268, 2012.

Angela D Friederici, Shirley-Ann Rüschemeyer, Anja Hahne, and Christian J Fiebach. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex*, 13(2):170–177, 2003.

Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451):691–703, 2000.

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

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

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*, 2022.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of neuroscience*, 28(10):2539–2550, 2008.

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

Alexander G Huth, Shinji Nishimoto, An T Vu, and T Dupre La Tour. Gallant lab natural short clips 3t fmri data. *G-Node doi*, 10, 2022.

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

Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, 2021.

Tom Dupré La Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, 31(8): 2906–2915, 2011.

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

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

Subba Reddy Oota and Mariya Toneva. What aspects of nlp models and brain datasets affect brain-nlp alignment? In *2023 Conference on Cognitive Computational Neuroscience*, 2023.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.462.

Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). In *The Thirteenth International Conference on Learning Representations*, 2025a.

Subba Reddy Oota, Khushbu Pahwa, Prachi Jindal, Satya Sai Srinath Namburi, Maneesh Singh, Tanmoy Chakraborty, Bapi S Raju, and Manish Gupta. Instruction-tuned video-audio models elucidate functional specialization in the brain. *arXiv preprint arXiv:2506.08277*, 2025b.

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, and Bapi Raju Surampudi. Multi-modal brain encoding models for multi-modal stimuli. In *The Thirteenth International Conference on Learning Representations*, 2025c.

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

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Aniketh Janardhan Reddy and Leila Wehbe. Can fmri reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34:9843–9856, 2021.

Christina Sartzetaki, Gemma Roig, Cees GM Snoek, and Iris IA Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=LM4PYXBId5.

Oliver Schoppe, Nicol S Harper, Ben DB Willmore, Andrew J King, and Jan WH Schnupp. Measuring the performance of neural models. *Frontiers in computational neuroscience*, 10:10, 2016.

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

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

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

V Subramaniam, C Wang, A Barbu, G Kreiman, and B Katz. Revealing vision-language integration in the brain with multimodal networks. In *International Conference on Machine Learning*. International Conference on Machine Learning (ICML), 2024.

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

Andreĭ Nikolaevich Tikhonov and VIAK Arsenin. Solutions of ill-posed problems. *(No Title)*, 1977.

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

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

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

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Chao Yuan, Yang Yang, Yehui Yang, and Zach Cheng. Date: Dynamic absolute time enhancement for long video understanding. *arXiv preprint arXiv:2509.09263*, 2025.

Jeffrey M. Zacks and Khena M. Swallow. Event segmentation. *Current Directions in Psychological Science*, 16(2):80–84, 2007. doi: 10.1111/j.1467-8721.2007.00480.x. URL `https://doi.org/10.1111/j.1467-8721.2007.00480.x`. PMID: 22468032.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021.

Overview of Appendix Sections

App. A Related Work
App. B Dataset and Detailed sub-ROIs of language, visual and auditory regions
App. C Cross-subject prediction accuracy
App. D Description of Narrative video understanding tasks
App. E Implementation Details for Reproducibility
App. F Statistical Significance
App. G Subject-level Normalized Brain Alignment Across Temporal Windows
App. H Window-wise Voxel Alignment for DATE and TimeSformer
App. I Robustness to Frame Sampling.
App. J Layer-wise Analysis for DATE
App. K Task-wise Voxel Alignment for DATE
App. L Maximally activated video clips
App. M Interpretability Analysis for Qwen2.5-Omni and DATE
App. N Limitations
App. O Ethics Statement

A Related work

Contextual Representations and Brain Alignment. Our work is most closely related to studies by Jain & Huth (2018); Toneva & Wehbe (2019); Aw & Toneva (2023); Vaidya et al. (2022); Oota et al. (2024), who investigate how high-level contextual representations from modern language models map onto human brain activity, and examine how contextual dynamics affect brain-model alignment in language-related ROIs. Jain & Huth (2018); Toneva & Wehbe (2019); Aw & Toneva (2023) focus on text-based language models with context lengths ranging from 1 to 500 words, whereas Vaidya et al. (2022); Oota et al. (2024) study speech-based models with context windows ranging from 1.49 to 64 seconds. Beyond text and speech, a growing body of work models brain responses during naturalistic movie viewing using representations from unimodal and multimodal models, and more recently from MLLMs, providing strong baselines for video-to-brain encoding (Dong & Toneva, 2023a;b; Oota et al., 2025a;c). Motivated by these studies, we study how long-range visual context in naturalistic videos affects brain–model alignment, comparing unimodal encoders with MLLMs.

Multimodal Brain Encoding. Our work also relates to a growing literature on aligning AI model representations with human brain activity under naturalistic stimuli. Several studies have used unimodal video models and multimodal models (e.g., video+audio) to predict movie-evoked brain activity with strong predictive performance (Dong & Toneva, 2023a;b; Oota et al., 2025c). We summarize representative evaluation settings, particularly stimulus modality and temporal window length in Table 1. However, much of this prior multimodal work relies on short clips or limited temporal windows, which can underrepresent the role of extended temporal context in naturalistic video understanding. Complementary to these efforts, our work builds on and extends this line of work by proposing an architecture-agnostic evaluation protocol to characterize context- and instruction-dependent representational changes across varying time windows and tasks, and relate these changes to brain alignment.

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

Brain imaging dataset. We experiment with Movie10 (Boyle et al., 2020), a multimodal naturalistic fMRI dataset, obtained from the Courtois NeuroMod databank. This dataset was collected while five human subjects (s1, s2, s3, s5, s6; data for s4 is not public) passively watched four different movies: *The Bourne supremacy (∼100 mins)*, *The wolf of wall street (∼170 mins)*, *Hidden figures (∼120 mins)* and *Life (∼50 mins)*. Among these, *Hidden figures* and *Life* are repeated twice, with the repeats used for testing and the remaining movies for training. In this work, we use *Life* movies for testing where we average the two repetitions to reduce noise in brain data. This dataset is one of the largest publicly available multimodal fMRI datasets in terms of the number of samples per participant. It includes 4024 TRs (Repetition Time) of *The Bourne supremacy* and 6993 TRs of *The wolf of wall street* for training and 2028 TRs of *Life* as test data. We build encoding models where the train and test sets are totally disjoint. The fMRI data is collected every 1.49 seconds (= 1 TR).

The dataset is already preprocessed and projected onto the surface space ("fsaverage6"). We use the multimodal parcellation of the human cerebral cortex based on the Glasser Atlas (which consists of 180 regions of interest in each hemisphere) to report the ROI (region of interest) analysis for the brain maps (Glasser et al., 2016). This includes four visual processing regions (early visual cortex (EVC), object-related areas (LOC), face-related areas (OFA) and scene-related areas (PPA)), one early auditory area (AC), and eight language-relevant regions, encompassing broader language regions: angular gyrus (AG), anterior temporal lobe (ATL), posterior temporal lobe (PTL), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), middle frontal gyrus (MFG), posterior cingulate cortex (PCC) and dorsomedial prefrontal cortex (dmPFC), based on the Fedorenko lab's language parcels (Milton et al., 2021; Desai et al., 2023).

The data covers seven brain regions of interest (ROIs) in the human brain with the following subdivisions: (i) early visual (EV: V1, V2, V3, V3B, and V4); (ii) object-related areas (LO1 and LO2); (iii) face-related areas (OFA), (iv) scene-related areas (PPA), (v) middle temporal (MT: MT, MST, LO3, FST and V3CD), (vi) late language regions, encompassing broader language regions: angular gyrus (AG: PFm, PGs, PGi, TPOJ2, TPOJ3), lateral temporal cortex (LTC: STSda, STSva, STGa,

TE1a, TE2a, TGv, TGd, A5, STSdp, STSvp, PSL, STV, TPOJ1), inferior frontal gyrus (IFG: 44, 45, IFJa, IFSp) and middle frontal gyrus (MFG: 55b) (Baker et al., 2018; Milton et al., 2021; Desai et al., 2023).

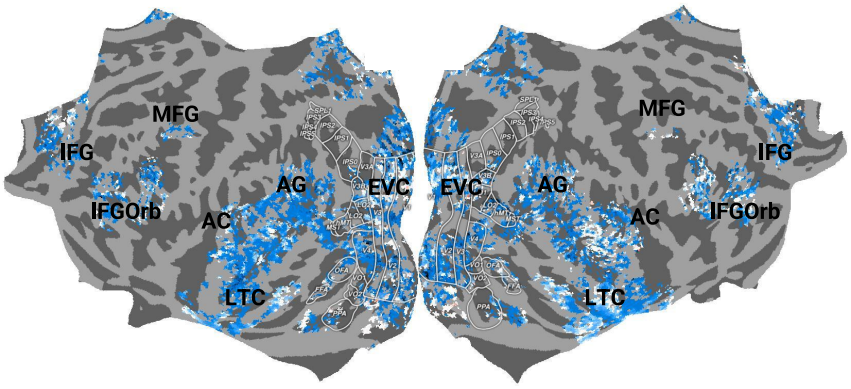


Figure 7: Flattened cortical surfaces for language-, visual- and auditory-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants.

C Cross-subject prediction accuracy

We follow the method introduced by Schrimpf et al. (2021) to estimate how well brain activity in one individual can be predicted from others, using the Movie10 fMRI dataset. Starting with data from n participants (e.g., $n = 5$), for each subject $s \in$ (s1, s2, s3, s5, s6) is chosen as the prediction target and the other four are used to predict this target, we use a voxel-wise encoding model (see Sec. 4) to predict one participant's response from others. For every combination, one participant was randomly chosen as the target, and the model was trained to predict their brain responses using data from the remaining $n-1$ participants. This gave us an average prediction score (correlation) for each voxel at each participant. To extrapolate to infinitely many humans and thus to obtain the highest possible (most conservative) estimate, as suggested by Schrimpf et al. (2021), we fit the equation $v = v_0 \times \left(1 - e^{-\frac{x}{\tau_0}}\right)$ where x is each subsample's number of participants, v is each subsample's correlation score and v_0 and τ_0 are the fitted parameters. This fitting was performed for each voxel independently with 100 bootstraps each to estimate the variance where each bootstrap draws x and v with replacement. The final ceiling value was the median of the per-voxel ceilings v_0.

Table. 4 shows the estimated cross-subject prediction accuracy for all subjects for the naturalistic movie watching across language and visual ROIs. Across the 14 ROIs and five participants, all 70 participant$\times$ROI mean cross-subject correlations are positive (mean r=0.104; 69/70 exceed r=0.05), and 38.9% of cortical vertices pass the $r \geq 0.05$ retention threshold used for normalization (Table 4). These values fall within the range reported for established naturalistic-fMRI datasets (r=0.075–0.139; Oota & Toneva, 2023). Importantly, the same retained vertex set is used for every model, window, and prompt, so normalization rescales alignment by a fixed positive ceiling at each vertex and cannot change which condition wins.

D Description of Narrative video understanding tasks

Each task addresses a unique aspect of narrative meaning. Character Motivation captures implicit beliefs and goals of agents Rabinowitz et al. (2018). Event Boundary Detection identifies fine-grained change points critical for temporal segmentation Zacks & Swallow (2007). Multi-Scene Summary integrates local events across broader spans. Narrative Summary consolidates long-range relational structure and thematic coherence Bartlett (1995). Removing any one task leaves certain forms of information unrecoverable from the others (e.g., agent beliefs, change points, relational links, or fine event details), establishing necessity. Conversely, under the decomposition and representability assumptions, the four tasks together span the relevant semantic space, up to a simple readout, establishing sufficiency.

Table 4: Cross-subject predictivity and vertex retention by participant and ROI. Each cell reports the ROI-mean Pearson correlation r and the percentage of vertices satisfying the $r \geq 0.05$ retention criterion.

ROI	S1 r/%	S2 r/%	S3 r/%	S5 r/%	S6 r/%	Mean r/%
AG	0.078/32.42	0.094/36.67	0.102/31.07	0.081/37.97	0.085/41.22	0.088/35.87
ATL	0.092/39.99	0.085/31.19	0.101/35.44	0.109/29.84	0.088/36.74	0.095/34.64
PTL	0.104/47.29	0.108/50.54	0.101/41.74	0.117/45.99	0.125/40.39	0.111/45.19
IFG	0.137/37.23	0.116/44.13	0.120/47.38	0.113/38.58	0.129/42.83	0.123/42.03
MFG	0.042†/19.39	0.121/43.69	0.196/63.14	0.161/59.99	0.125/47.74	0.129/46.79
IFGOrb	0.074/29.28	0.090/33.53	0.098/27.93	0.077/34.83	0.081/38.08	0.084/32.73
PCC	0.089/37.04	0.072/28.24	0.076/32.49	0.064/26.89	0.074/33.79	0.075/31.69
dmPFC	0.079/34.11	0.083/37.36	0.076/28.56	0.092/32.81	0.100/27.21	0.086/32.01
EVC	0.144/41.01	0.123/47.91	0.127/51.16	0.120/42.36	0.136/46.61	0.130/45.81
LOC	0.111/37.76	0.119/32.16	0.098/39.06	0.102/42.31	0.095/33.51	0.105/36.96
OFA	0.081/32.88	0.097/37.13	0.105/31.53	0.084/38.43	0.088/41.68	0.091/36.33
PPA	0.080/40.31	0.073/31.51	0.089/35.76	0.097/30.16	0.076/37.06	0.083/34.96
MT	0.124/46.31	0.128/49.56	0.121/40.76	0.137/45.01	0.145/39.41	0.131/44.21
AC	0.143/39.99	0.122/46.89	0.126/50.14	0.119/41.34	0.135/45.59	0.129/44.79
Mean	**0.098/36.79**	**0.102/39.32**	**0.110/39.73**	**0.105/39.04**	**0.106/39.42**	**0.104/38.86**

†The S1 MFG ROI mean is below 0.05; vertices within this ROI that individually satisfy the retention criterion remain included.

Table 5: Within-participant reliability by participant and ROI. Each cell reports the ROI-mean repeat-based explainable variance (EV) and the percentage of vertices satisfying the EV ≥ 0.05 retention criterion.

ROI	S1 EV/%	S2 EV/%	S3 EV/%	S5 EV/%	S6 EV/%	Mean EV/%
AG	0.182/53.15	0.132/49.30	0.216/66.85	0.158/65.00	0.152/57.70	0.168/58.40
ATL	0.215/62.28	0.134/48.08	0.158/55.63	0.129/60.43	0.179/50.98	0.163/55.48
PTL	0.184/61.66	0.132/47.91	0.177/69.96	0.123/57.06	0.149/57.21	0.153/58.76
IFG	0.190/49.78	0.138/57.98	0.165/63.33	0.113/47.03	0.144/55.03	0.150/54.63
MFG	0.147/64.48	0.171/47.93	0.207/59.38	0.140/46.03	0.160/53.33	0.165/54.23
IFGOrb	0.132/45.65	0.182/49.50	0.216/63.20	0.158/61.35	0.152/54.05	0.168/54.75
PCC	0.138/56.80	0.042†/23.85	0.140/59.10	0.085/45.75	0.105/48.00	0.102/46.70
dmPFC	0.246/65.54	0.194/51.79	0.239/73.84	0.185/60.94	0.211/61.09	0.215/62.64
EVC	0.245/57.19	0.193/65.39	0.220/70.74	0.168/54.44	0.199/62.44	0.205/62.04
LOC	0.436/87.10	0.117/40.85	0.299/76.55	0.252/70.85	0.256/68.40	0.272/68.75
OFA	0.121/46.88	0.171/50.73	0.205/64.43	0.147/62.58	0.141/55.28	0.157/55.98
PPA	0.150/52.12	0.069/37.92	0.093/45.47	0.064/50.27	0.114/40.82	0.098/45.32
MT	0.265/70.92	0.213/57.17	0.258/79.22	0.204/66.32	0.230/66.47	0.234/68.02
AC	0.201/53.16	0.149/61.36	0.176/66.71	0.124/50.41	0.155/58.41	0.161/58.01
Mean	**0.204/59.05**	**0.146/49.27**	**0.198/65.31**	**0.146/57.03**	**0.168/56.37**	**0.172/57.41**

†ROI-mean EV is below 0.05. Repeat-based EV (Schoppe et al., 2016) captures reliable participant-specific responses and complements cross-subject predictivity. Across the 70 participant–ROI combinations, 69 have a mean EV above 0.05. The overall mean EV is 0.172, and 57.41% of vertices satisfy EV ≥ 0.05.

Table 6: Instructions for multimodal narrative understanding tasks.

Task	Description
Character Motivation	Based on this given video segment, describe the character's likely motivation or emotional state at this moment. Justify your answer using visual or contextual cues.
Event Boundary Detection	You are shown a video segment from a larger movie. Based only on this segment, does this moment likely represent a narrative event boundary, such as a scene change?
Multi Scene Summary	You are shown a video segment from a larger movie. Based only on this clip, explain how this moment might connect to the broader narrative.
Narrative Summary	Summarize what is happening in this video segment. Clearly describe the observed events, characters, and any inferences about the context.

Table 7: Paraphrased and generic instructions for multimodal narrative understanding tasks.

Task	Description
Character Motivation	Watch this video clip from a movie. What is the character feeling or trying to achieve in this moment? Support your interpretation using cues from the clip
Event Boundary Detection	Watch this video clip from a movie. Does this moment appear to mark a transition in the story, such as a change in scene, time, or topic? Explain your answer using cues from the clip.
Multi Scene Summary	Watch this video clip from a movie. How does this moment fit into the broader story? Describe the narrative connection you can infer from the clip
Narrative Summary	Watch this video clip from a movie. Describe what happens in this moment, including the key actions, characters, and situation shown in the clip.
Generic Prompt	You are shown a video segment from a larger movie. Summarize what is happening in this segment, clearly describing the observed events, characters, and any inferences about the context. Describe the character's likely motivation or emotional state at this moment, justifying your answer using visual or contextual cues. Explain how this moment might connect to the broader narrative. Finally, indicate whether this moment likely represents a narrative event boundary, such as a scene change.

E IMPLEMENTATION DETAILS FOR REPRODUCIBILITY

All feature extraction experiments were conducted on a machine equipped with an NVIDIA A100 GPU with 80 GB of GPU RAM, partitioned into two devices of 40 GB each. The voxelwise encoding models were trained on NVIDIA GeForce RTX 3050 GPU with 4GB of GPU RAM. We used bootstrap ridge-regression with the following parameters: MSE loss function; L2-decay (λ) varied from 10^{-1} to 10^{3}; the best λ was chosen by tuning on validation data that comprised a randomly chosen 10% subset from the train set used only for hyper-parameter tuning.

F STATISTICAL SIGNIFICANCE

To determine if normalized predictivity scores are significantly higher than chance, we run a permutation test using blocks of 10 contiguous fMRI TRs (considering the slowness of hemodynamic response) rather than individual TRs. By permuting predictions 5000 times, we create an empirical distribution for chance performance, from which we estimate p-value of the actual performance. The choice of these specific permutation test configurations is based on established methodologies in previous research (Deniz et al., 2019; Reddy & Wehbe, 2021; Oota et al., 2024). To estimate the statistical significance of performance differences, such as between the model's predictions and chance or residual predictions and chance, we utilized the Wilcoxon signed-rank test (Conover, 1999), applying it to the mean normalized predictivity for the participants. Finally, the Benjamini-Hochberg False Discovery Rate (FDR) correction for multiple comparisons (Benjamini & Hochberg, 1995) is used for all the tests (appropriate because fMRI data is considered to have positive dependence (Genovese, 2000)).

G SUBJECT-LEVEL NORMALIZED BRAIN ALIGNMENT ACROSS TEMPORAL WINDOWS

We report all five subject-level trajectories and provide their numerical values in Table 8. The long-context benefit for Qwen-2.5-Omni and DATE is observed in every participant, although the peak window varies, whereas TimeSFormer shows substantially smaller, non-monotonic changes.

Table 8: Subject-level normalized brain alignment across temporal windows. The final column reports mean±SEM across five participants.

Model	Window	S1	S2	S3	S5	S6	Mean±SEM
Qwen-2.5-Omni	3 s	0.645	0.660	0.698	0.678	0.689	0.674 ± 0.009
	6 s	0.681	0.699	0.695	0.721	0.714	0.702 ± 0.007
	9 s	0.784	0.785	0.807	0.762	0.819	0.791 ± 0.010
	12 s	0.803	0.833	0.844	0.893	0.872	0.849 ± 0.016
	18 s	0.854	0.891	1.006	1.077	0.946	0.955 ± 0.040
	24 s	0.871	0.903	1.035	0.963	0.991	0.953 ± 0.029
DATE	3 s	0.639	0.666	0.788	0.571	0.705	0.674 ± 0.036
	6 s	0.691	0.732	0.874	0.632	0.768	0.739 ± 0.040
	9 s	0.695	0.751	0.909	0.649	0.792	0.759 ± 0.044
	12 s	0.726	0.796	0.969	0.694	0.833	0.804 ± 0.047
	18 s	0.831	0.841	0.932	1.000	0.918	0.904 ± 0.031
	24 s	0.837	0.869	0.971	0.897	0.923	0.899 ± 0.023
TimeSFormer	3 s	0.587	0.582	0.568	0.590	0.575	0.580 ± 0.004
	6 s	0.576	0.577	0.562	0.592	0.586	0.579 ± 0.005
	9 s	0.570	0.579	0.581	0.585	0.574	0.578 ± 0.003
	12 s	0.574	0.582	0.591	0.580	0.588	0.583 ± 0.003
	18 s	0.611	0.591	0.592	0.610	0.604	0.602 ± 0.004
	24 s	0.605	0.577	0.584	0.602	0.596	0.593 ± 0.005

Table 9 reports the ROI-wise temporal-window preferences for Qwen-2.5-Omni separately for each participant. Nine of the 12 reported ROIs show agreement in at least three of five participants, while the remaining ROIs show heterogeneous or less consistent preferences. Thus, the principal temporal-window patterns are reproducible across participants but are not universal.

H WINDOW-WISE VOXEL ALIGNMENT FOR DATE AND TIMESFORMER

Within-model DATE prompting ablation. We perform a corresponding within-model ablation for DATE by comparing video with a narrative-task text prompt against video-only input without

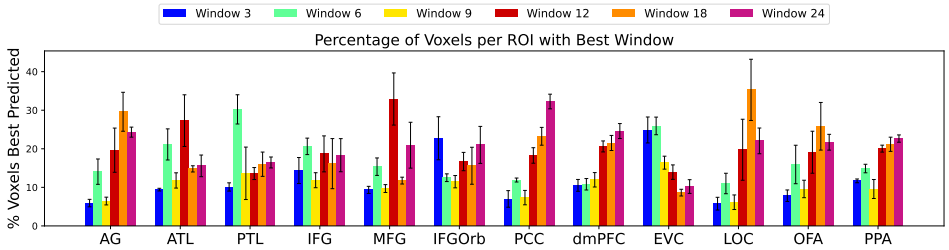


Figure 8: Qwen-2.5-Omni: Percentage of best predicted voxels whose brain encoding performance is higher corresponding to each window length within language-selective, and visual regions.

Table 9: Subject-level temporal-window preferences by ROI for Qwen-2.5-Omni. Each participant cell reports the preferred window and the percentage of retained ROI vertices for which that window is optimal.

ROI	S1	S2	S3	S5	S6	Agreement
AG	18 s (36.78%)	12 s (34.59%)	18 s (34.14%)	18 s (32.81%)	18 s (35.26%)	18 s, 4/5
ATL	12 s (33.62%)	6 s (33.05%)	12 s (24.44%)	12 s (41.13%)	12 s (36.48%)	12 s, 4/5
PTL	6 s (33.17%)	9 s (33.95%)	6 s (32.61%)	6 s (36.03%)	6 s (34.72%)	6 s, 4/5
IFG	24 s (26.76%)	6 s (20.64%)	12 s (27.43%)	18 s (32.76%)	18 s (25.91%)	Heterogeneous
MFG	12 s (40.98%)	24 s (38.71%)	12 s (29.37%)	12 s (45.82%)	12 s (37.64%)	12 s, 4/5
IFGOrb	24 s (30.82%)	18 s (26.32%)	3 s (37.47%)	3 s (24.83%)	3 s (28.94%)	3 s, 3/5
PCC	24 s (31.89%)	24 s (32.27%)	24 s (27.80%)	24 s (37.08%)	24 s (34.65%)	24 s, 5/5
dmPFC	24 s (28.41%)	24 s (24.60%)	24 s (19.28%)	24 s (26.11%)	24 s (25.37%)	24 s, 5/5
EVC	6 s (32.39%)	6 s (25.94%)	3 s (29.06%)	3 s (30.19%)	6 s (27.88%)	6 s, 3/5
LOC	18 s (40.66%)	12 s (43.05%)	18 s (40.85%)	18 s (47.64%)	18 s (42.18%)	18 s, 4/5
OFA	18 s (34.84%)	12 s (29.18%)	18 s (38.09%)	24 s (26.82%)	18 s (31.76%)	18 s, 3/5
PPA	24 s (25.00%)	18 s (23.38%)	24 s (22.27%)	24 s (22.91%)	24 s (24.63%)	24 s, 4/5

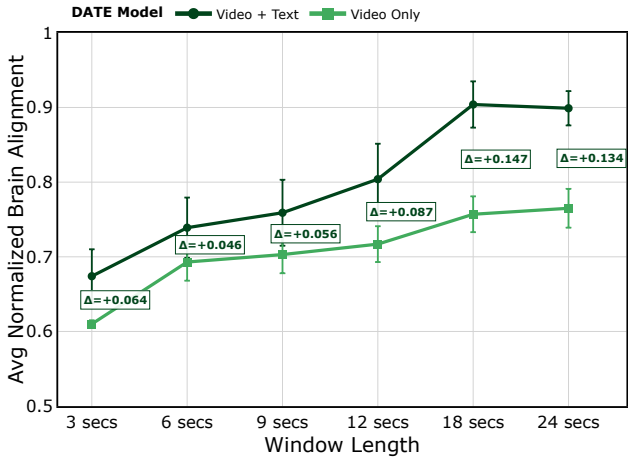


Figure 9: Average normalized brain alignment for DATE as a function of temporal window length (3–24 s). The within-model comparison contrasts video with task-text conditioning against video-only input without prompting. Error bars denote mean±SEM across participants.

prompting, as shown in Fig. 9. From the Fig. 9, we observe that the prompted condition yields higher normalized brain alignment at every temporal window, with improvements of 0.064, 0.046, 0.056, 0.087, 0.147, and 0.134 at 3, 6, 9, 12, 18, and 24 s, respectively. This result shows that task-text conditioning contributes additional brain-relevant information while holding the DATE architecture and video input fixed.

To test whether the temporal-context effects generalize beyond Qwen-2.5-Omni, we repeat the same window-wise voxel analysis using the DATE model. As in the main analysis, for each voxel we select the temporal window length (3 s, 6 s, 9 s, 12 s, 18 s, or 24 s) that yields the highest normalized brain alignment, and visualize the resulting best-window map on the `fsaverage` surface along with the ROI-wise distribution of best windows (Appendix Fig. 10).

Qualitatively, as shown in Fig. 10 (top), DATE exhibits a similar temporal gradient to Qwen-2.5-Omni: shorter-to-intermediate windows (3–6 s) account for a larger share of best-predicted voxels

Table 10: Temporal-window effect magnitudes for DATE. Values are mean±SD across five participants. Δ denotes the best–runner-up difference.

ROI	Best window	Runner-up window	Δ	Participant preference
AG	18 s (0.953 ± 0.064)	24 s (0.915 ± 0.118)	0.038	18 s (4/5)
ATL	12 s (0.805 ± 0.316)	24 s (0.750 ± 0.330)	0.055	12 s (4/5)
PTL	6 s (0.722 ± 0.137)	18 s (0.674 ± 0.179)	0.048	6 s (4/5)
IFG	24 s (0.913 ± 0.260)	18 s (0.890 ± 0.243)	0.023	24 s (2/5)
MFG	12 s (0.814 ± 0.393)	24 s (0.781 ± 0.350)	0.033	12 s (4/5)
IFGOrb	24 s (0.791 ± 0.132)	18 s (0.756 ± 0.103)	0.035	24 s (3/5)
PCC	24 s (0.934 ± 0.211)	18 s (0.883 ± 0.142)	0.051	24 s (5/5)
dmPFC	24 s (0.951 ± 0.167)	18 s (0.908 ± 0.071)	0.043	24 s (5/5)
EVC	6 s (0.678 ± 0.192)	3 s (0.653 ± 0.194)	0.025	6 s (3/5)
LOC	9 s (0.854 ± 0.210)	12 s (0.824 ± 0.164)	0.030	9 s (4/5)
OFA	18 s (0.885 ± 0.258)	24 s (0.844 ± 0.232)	0.041	18 s (3/5)
PPA	12 s (0.880 ± 0.132)	24 s (0.814 ± 0.146)	0.066	12 s (4/5)
AC	3 s (0.634 ± 0.025)	6 s (0.626 ± 0.067)	0.008	3 s (3/5)

in more perceptual and "local" regions, while longer windows (18-24 s) dominate in higher-order semantic/integrative regions. In particular, regions such as PCC and dmPFC show a preference for the longest window, consistent with longer-timescale narrative integration. Likewise, several language-selective ROIs (e.g., ATL, IFG, MFG, and IFGOrb) show a clear shift toward longer context windows, whereas PTL retains a comparatively stronger contribution from shorter windows, while AG shows both shorter and longer windows dominance, suggesting that parts of the temporal language system are optimally predicted at shorter timescales. Visual and auditory ROIs show weaker monotonic shifts overall, with early visual cortex appearing comparatively mixed across windows.

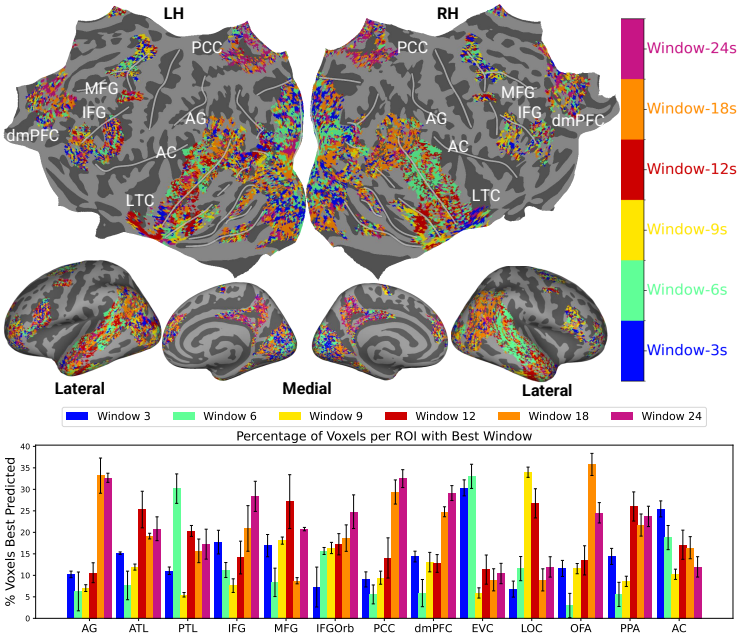


Figure 10: DATE model results: Each voxel is color-coded with the video duration (window length) that led to the highest normalized brain alignment with the DATE model. The color bar highlights color codes for each window length. (Top) The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. (Bottom): Percentage of voxels in each ROI corresponding to each window length.

The ROI-wise histogram (Fig. 10, bottom) further supports these observations, showing that 18-24 s is the most frequent best window across many higher-order ROIs, while 3–6 s windows contribute more substantially in PTL/AG and several sensory-selective regions. Together, these results indicate that DATE captures heterogeneous temporal integration across cortex in a manner consistent with the main findings, reinforcing that long-context gains in brain alignment reflect model representations that better support long-timescale narrative processing rather than simply increased visual input length.

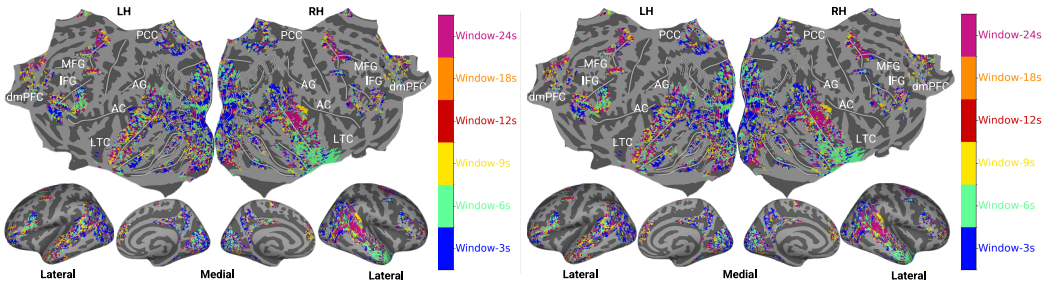


Figure 11: Unimodal video baselines (TimeSFormer: Left, and VideoMAE: Right): Each voxel is color-coded with the video duration (window length) that led to the highest normalized brain alignment with the Unimodal video baselines. The color bar highlights color codes for each window length. (Left) The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject.

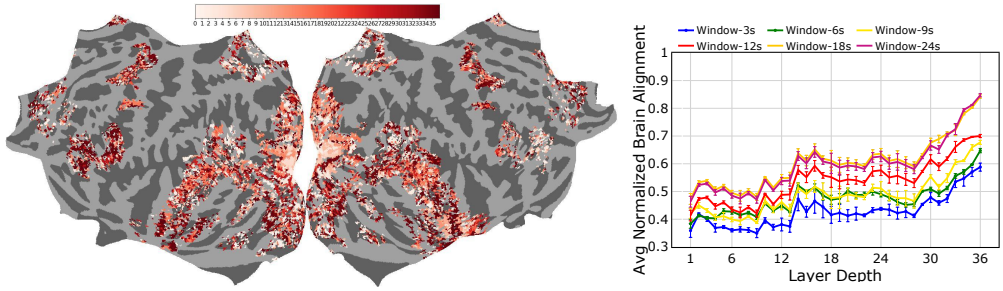


Figure 12: Layer-wise alignment results for DATE: Each voxel is color coded with the MLLM layer number that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject.

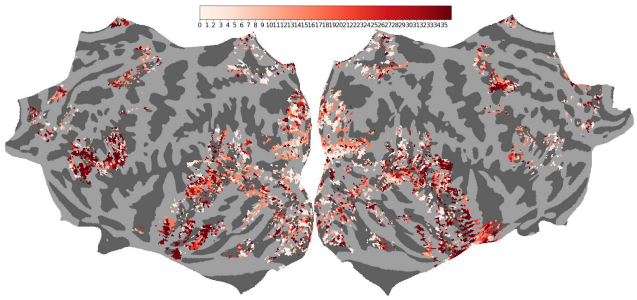
I Robustness to Frame Sampling.

Our default protocol uses a window-length-dependent frame count (9–30 frames; see Methodology in Sec. 4), so effective frame density varies modestly across windows. To verify that the gains we report with longer context are not an artifact of this scheme, we re-ran the analysis under constant frame density (1 fps), in which the per-window frame count scales strictly with temporal span. Brain alignment increases monotonically with temporal span under both sampling strategies (Table 11), and absolute differences between the two settings are small (∼0.03–0.06). This indicates that the observed gains are driven primarily by temporal span rather than by frame count or density, and our main conclusion about the benefit of longer temporal context is not an artifact of the frame-sampling strategy.

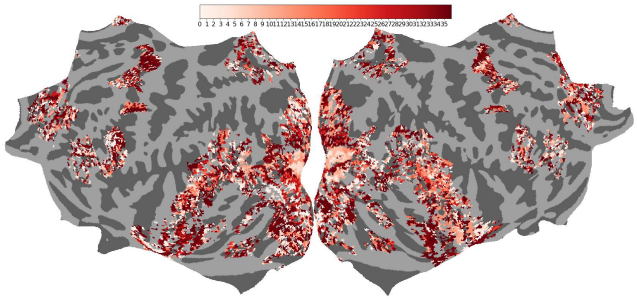
Table 11: Brain alignment under constant frame density (1 fps) vs. our default scaled sampling (9, 9, 12, 16, 24, and 30 frames for 3, 6, 9, 12, 18, 24 s). At 3 s under the 1 fps condition we use 4 frames, the minimum required by the MLLM input pipeline. The monotonic increase with temporal span is preserved under both strategies.

Window	1 fps†	Default sampling
3 s	0.636	0.674
6 s	0.692	0.702
9 s	0.752	0.791
12 s	0.818	0.849
18 s	0.957	0.955
24 s	0.943	0.953

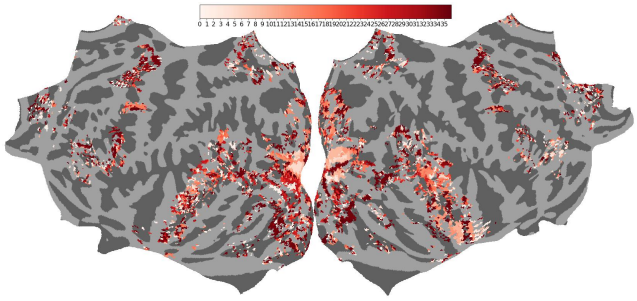
†3 s uses 4 frames (model minimum); 6, 9, 12, 18, 24 s use 6, 9, 12, 18, 24 frames.



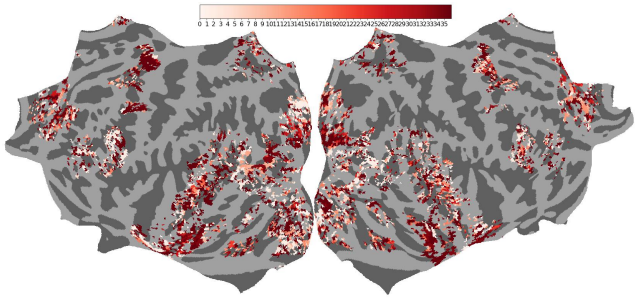
(a) Subject-01



(b) Subject-02



(c) Subject-03



(d) Subject-05

Figure 13: Qwen2.5-Omni: Each voxel is color coded with the video MLLM layer number (out of 36) that led to the highest normalized brain alignment. The color bar highlights color codes for each layer. The voxels are projected onto the flattened cortical surface of each subject on 'fsaverage' surface.

J LAYER-WISE ANALYSIS FOR DATE

We investigate how MLLM brain-encoding performance varies across model layers as the temporal context window increases, as shown in Fig. 12 (right). Across all temporal windows, we observe a consistent stratification into three layer groups: early layers (1–12), middle layers (13–28), and late layers (29–36). Overall, normalized brain alignment increases with longer sliding-window context, indicating that MLLMs benefit from long-range temporal information compared to shorter windows (e.g., 3 s and 6 s). We also find a layer-wise hierarchy: alignment generally improves from early

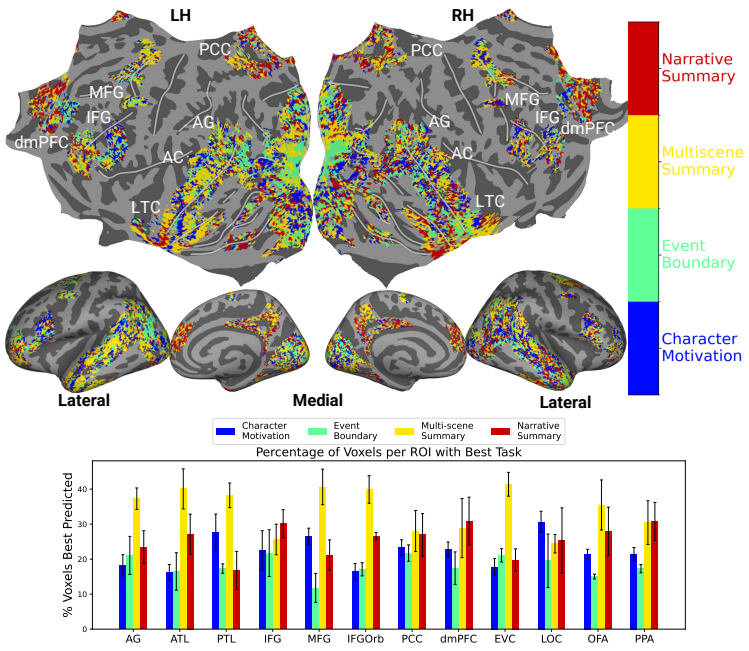


Figure 14: DATE model results: Each voxel is color-coded with the instruction that led to the highest normalized with the DATE model. The color bar highlights color codes for each window length. (Top) The voxels are projected onto the flattened cortical surface of the 'fsaverage' subject. (Bottom): Percentage of voxels in each ROI corresponding to each window length.

to deeper layers across windows, suggesting later-layer representations are more predictive of the measured brain responses.

Fig. 12 (left) visualizes voxel-wise layer preferences, projected onto the `fsaverage` subject. We observe that (i) early sensory regions (early visual and auditory cortex) align best with lower layers, consistent with shallow representations capturing low-level sensory features; (ii) higher-level visual regions such as lateral occipital complex (LOC) and parahippocampal place area (PPA), and temporal/parietal language ROIs (PTL and AG) tend to align with middle-to-late layers; and (iii) language-related regions such as inferior frontal gyrus (IFG), anterior temporal lobe (ATL), and angular gyrus align most strongly with the deepest layers. Together, these results suggest that MLLMs exhibit a layered representational hierarchy that broadly mirrors cortical processing hierarchies.

K Task-wise Voxel Alignment for DATE

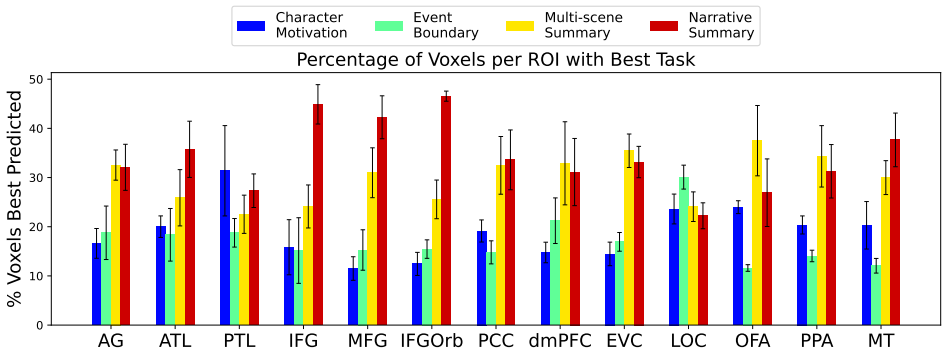


Figure 15: Qwen-2.5-Omni: Percentage of voxels in each ROI corresponding to each instruction.

To test whether the task-specific dissociations generalize beyond Qwen-2.5-Omni, we repeat the voxel-wise winner-task analysis using the DATE model. For each voxel, we select the narrative-task

Table 12: Subject-level narrative-task preferences by ROI for Qwen-2.5-Omni. Each participant cell reports the preferred task and the percentage of retained ROI vertices for which that task is optimal. CM: Character Motivation; EB: Event Boundary Detection; MSS: Multi-Scene Summary; NS: Narrative Summary.

ROI	S1	S2	S3	S5	S6	Agreement
AG	MSS (35.39%)	MSS (32.54%)	MSS (35.82%)	NS (40.55%)	MSS (34.76%)	MSS, 4/5
ATL	NS (37.14%)	NS (35.74%)	NS (32.55%)	NS (46.84%)	NS (39.62%)	NS, 5/5
PTL	MSS (32.31%)	CM (35.85%)	NS (46.80%)	CM (31.00%)	CM (33.48%)	CM, 3/5
IFG	NS (37.67%)	NS (44.89%)	NS (45.49%)	NS (51.50%)	NS (43.28%)	NS, 5/5
MFG	MSS (41.01%)	NS (42.26%)	NS (44.54%)	NS (48.41%)	NS (45.17%)	NS, 4/5
IFGOrb	NS (47.19%)	NS (46.55%)	NS (44.54%)	NS (47.92%)	NS (45.83%)	NS, 5/5
PCC	MSS (41.95%)	NS (33.60%)	MSS (33.74%)	NS (44.27%)	MSS (36.18%)	MSS, 3/5
dmPFC	NS (29.62%)	MSS (32.90%)	NS (48.89%)	NS (44.78%)	NS (41.36%)	NS, 4/5
EVC	MSS (39.56%)	MSS (35.45%)	MSS (38.09%)	NS (39.53%)	MSS (36.92%)	MSS, 4/5
LOC	MSS (33.23%)	EB (28.89%)	EB (31.73%)	NS (47.42%)	EB (30.64%)	EB, 3/5
OFA	CM (48.40%)	MSS (37.50%)	MSS (40.07%)	NS (40.48%)	MSS (38.26%)	MSS, 3/5
PPA	MSS (46.80%)	MSS (34.31%)	MSS (37.04%)	NS (36.37%)	MSS (39.22%)	MSS, 4/5

Table 13: Narrative-task effect magnitudes for DATE. Values are mean±SD across five participants. Δ denotes the best–runner-up difference. CM: Character Motivation; EB: Event Boundary Detection; MSS: Multi-Scene Summary; NS: Narrative Summary.

ROI	Best task	Runner-up task	Δ	Participant preference
AG	MSS (0.887 ± 0.137)	NS (0.752 ± 0.223)	0.135	MSS (4/5)
ATL	MSS (0.869 ± 0.281)	NS (0.853 ± 0.259)	0.016	MSS (5/5)
PTL	MSS (0.872 ± 0.063)	CM (0.771 ± 0.097)	0.101	MSS (3/5)
IFG	NS (0.929 ± 0.271)	MSS (0.826 ± 0.121)	0.103	Heterogeneous
MFG	MSS (0.950 ± 0.170)	CM (0.824 ± 0.308)	0.126	MSS (4/5)
IFGOrb	MSS (0.905 ± 0.112)	NS (0.872 ± 0.009)	0.033	MSS (5/5)
PCC	MSS (0.939 ± 0.136)	NS (0.881 ± 0.212)	0.058	MSS (5/5)
dmPFC	NS (0.977 ± 0.218)	MSS (0.952 ± 0.115)	0.025	NS (5/5)
EVC	MSS (0.755 ± 0.203)	EB (0.627 ± 0.187)	0.128	MSS (4/5)
LOC	CM (0.833 ± 0.308)	NS (0.698 ± 0.240)	0.135	CM (3/5)
OFA	MSS (0.897 ± 0.346)	NS (0.881 ± 0.455)	0.016	MSS (3/5)
PPA	MSS (0.897 ± 0.424)	NS (0.848 ± 0.330)	0.049	MSS (4/5)
AC	MSS (0.725 ± 0.209)	NS (0.665 ± 0.188)	0.060	MSS (4/5)

instruction that yields the highest normalized brain alignment and visualize the resulting winner-task maps on the `fsaverage` surface (Fig. 14) (top). We additionally quantify, for each ROI, the percentage of voxels for which each task is optimal (Fig. 14 (bottom)).

Fig. 14 (top) shows that DATE exhibits a qualitatively similar task-based organization: (i) Both *Narrative Summary* and *Multi-scene Summary* explain a large fraction of voxels across widespread association cortex, consistent with these prompts emphasizing long-range semantic integration; (ii) *Character Motivation* more frequently emerges in lateral temporal and occipito-temporal regions, suggesting stronger alignment with regions supporting person- and situation-level inference grounded in audiovisual content; and (iii) *Event Boundary* is comparatively sparse overall, but appears more prominently in select higher-order territories, indicating more specialized and localized boundary-related tuning rather than broad dominance.

The ROI-wise histogram (Fig. 14 (bottom)) further supports these observations. *Multi-scene Summary* is the plurality winner across many ROIs (including several language/association ROIs such as AG/ATL/MFG/IFGOrb, and multiple visual ROIs), indicating that multi-scene integration prompts capture representations that are broadly predictive across networks. In contrast, *Narrative Summary* shows comparatively stronger contributions in some integrative ROIs (e.g., PPA and several frontal/temporal language ROIs), suggesting that global narrative compression preferentially matches higher-order semantic integration demands. *Character Motivation* accounts for the largest share of best-predicted voxels in more "local" regions such as LOC (and a substantial fraction in PTL), consistent with character/agent-centric prompts aligning with regions sensitive to entity-level cues. Finally, *Event Boundary* contributes relatively more in dmPFC and PCC than in most other ROIs, consistent with boundary-related structure being distributed but not the dominant explanatory factor in most regions.

Overall, DATE reproduces the main qualitative conclusion that narrative-task prompting acts as a functional probe: different prompts yield distinct, ROI-specific patterns of brain alignment rather than a single task dominating everywhere.

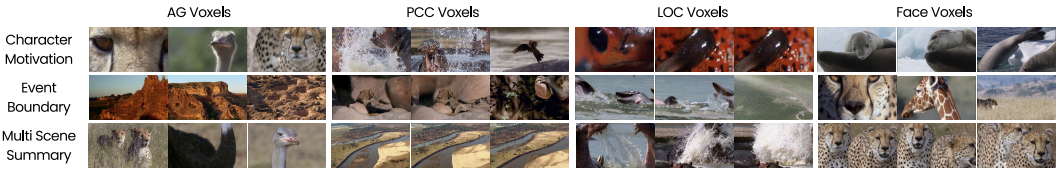


Figure 16: Task conditioned top-K video clips driving voxel responses for longer temporal window (18 s) for other tasks; results for narrative summary task are in Fig. 6.

L MAXIMALLY ACTIVATED VIDEO CLIPS

We examine whether the identity of the narrative-task instruction changes which video clips are predicted to maximally activate voxels within a given ROI. Using the same scoring-and-retrieval procedure as above, we select the top-K clips for ROI-selective voxels under each task instruction and visualize the resulting clips for AG, PCC, LOC, and face-selective voxels. Fig. 16 shows retrieved clips under different narrative-task prompts. To quantify task-dependent modulation, we compute Jaccard similarity between retrieved video sets for each ROI under different task instructions. Similar to Window-conditioned retrieval, visual ROIs remain stable across tasks (LOC: J = 0.42, Face: J = 0.38), whereas AG and PCC vary more across prompts (AG: J = 0.04, PCC: J = 0.13), indicating that task instructions modulate the semantic/contextual aspects of representations that best align with higher-order regions.

M INTERPRETABILITY ANALYSIS FOR QWEN2.5-OMNI AND DATE

Fig. 17 shows a 16×16 similarity matrix comparing the voxel-wise encoding weight patterns across 4 tasks × 4 temporal window lengths within language and visual voxels. We make the following observations: (i) Task identity drives the largest differences in voxel-wise tuning (cross-task similarity is low). (ii) Increasing temporal context tends to make weight patterns more consistent within a task, especially for multi-scene integration and narrative, while event boundary remains relatively weakly consistent across windows. (iii) Even within a task, correlations are not high, implying that context length meaningfully reshapes the learned voxel-feature mapping rather than merely scaling it.

N LIMITATIONS

Our study has several limitations that scope its claims and motivate future work. One possible limitation our study is missing lies in the maximum temporal context: Prior brain-MLLM alignment work uses windows of only up to ∼1.49 s, whereas we substantially extend this to 3–24 s, sufficient to establish the ROI-specific temporal gradient in Sec. 5 and consistent with psycholinguistic MEG findings that mPFC and IFG integrate context for story-level meaning over longer timescales, a pattern we also observe in our long temporal-context experiments. Naturalistic narrative comprehension is argued to further engage windows of tens of seconds to a minute, which we leave to future work for compute reasons rather than scientific ones: feature extraction with two MLLMs scales rapidly with sequence length and currently bottlenecks minute-scale brain-alignment analyses. Future work will focus on memory-efficient long-context MLLMs and hierarchical chunking are a clear next step that our findings directly motivate. A second limitation is dataset scope: we use Movie10 from the Courtois NeuroMod project (St-Laurent et al., 2023), which provides high-quality fMRI but is restricted to English-language Hollywood-style films. Extending the analysis to other languages, narrative genres, and higher-temporal-resolution acquisition protocols (e.g., MEG/ECoG) is a natural follow-up.

O ETHICS STATEMENT

We did not create any new neural recording data as part of this work. We used the Movie10 dataset which is publicly available without any restrictions. Movie10 dataset can be downloaded from `https://github.com/courtois-neuromod/movie10/tree/`

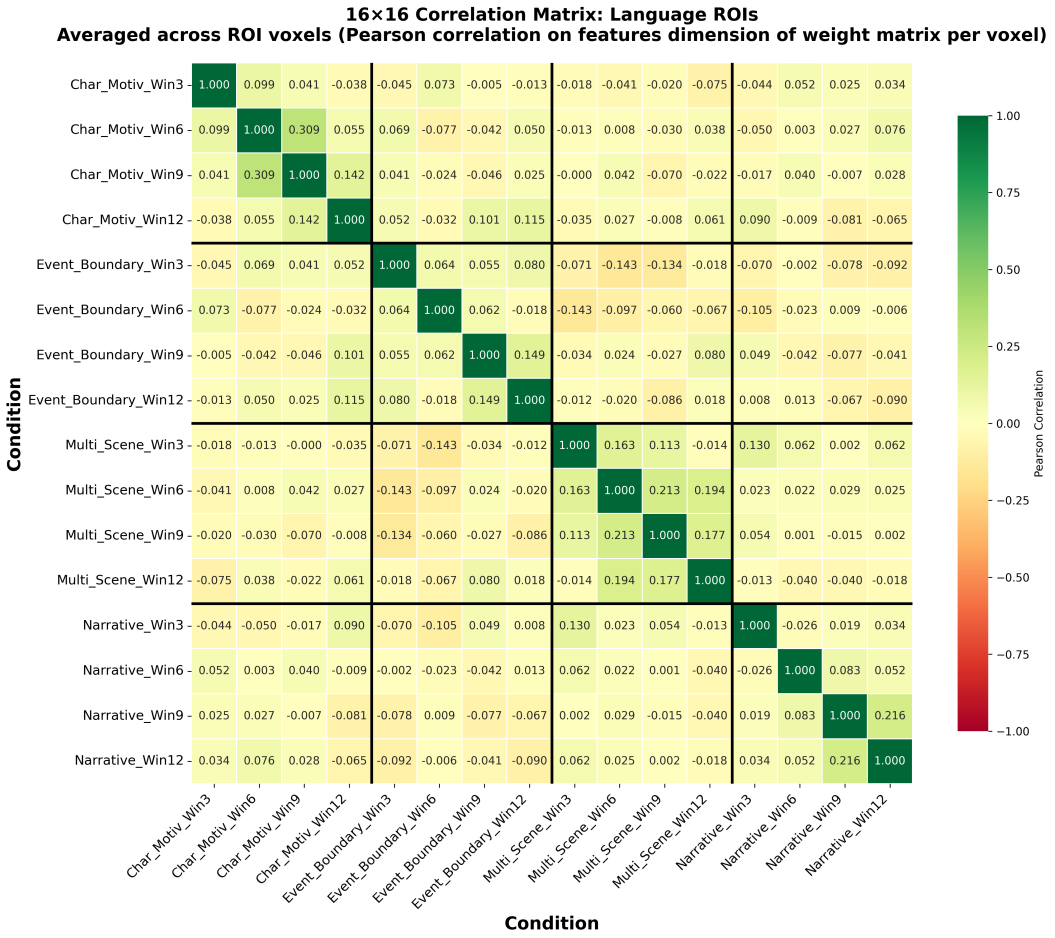


Figure 17: A 16×16 similarity matrix comparing the voxel-wise encoding weight patterns across 4 tasks × 4 temporal windows within language and visual voxels.

33a97c01503315e5e09b3ac07c6ccadb8b887dcf. Please read their terms of use[2] for more details.

We do not foresee any harmful uses of this technology.

[2] https://docs.cneuromod.ca/en/latest/ACCESS.html