DECODING NATURALISTIC SPEECH FROM FMRI WITH TEMPORAL MODELING AND MULTIMODAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing what a person perceives from their brain activity is a stringent test of whether neural recordings carry meaningful information. Naturalistic speech offers a rich testbed for this challenge: it unfolds over multiple timescales, but fMRI's noisy, temporally smeared measurements make this structure hard to recover. We introduce NOEMA, a temporal convolutional decoder that aligns windows of consecutive fMRI measurements with multimodal stimulus embeddings. These embeddings combine complementary text, audio, and speech representations of the same stimulus segment, enabling its retrieval directly from brain activity. On two naturalistic speech datasets, *Cephalonauts One* and *LeBel*, NOEMA retrieves the correct segment among 2,000 candidates within the top 10 for more than 42% of held-out queries. This substantially outperforms baseline decoders, an advantage that persists when the candidate pool is extended with thousands of distractors absent from training. Controlled ablations establish the contributions of temporal modeling and complementary stimulus representations.

1 INTRODUCTION

Can one gain access to the cognitive experience of a living being from recordings of their brain activity? A growing body of work shows that machine learning models can reconstruct what a participant is viewing, hearing, or even imagining from brain activity alone (Horikawa & Kamitani, 2017). This *decoding* endeavor is tightly linked to representation learning: if a model can precisely predict cognitive states from brain activity, it should have extracted meaningful structures from the signals at hand, opening a window into a better understanding of the complexity of the living brain.

Seminal work from Haxby et al. (2001) showed that fMRI measurements of the ventral visual stream contain sufficient information to distinguish the category of a viewed image from a small set of categories. Recent data collection efforts like the *Natural Scenes Dataset* (Allen et al., 2022) have since scaled up the number of image categories, enabling more accurate decoders (Ozcelik & VanRullen, 2023). Yet, since these stimuli are static, these decoders mostly rely on spatial information - where the activations are happening on the cortex - to perform their task. Given the inherently dynamic nature of the brain, it is crucial to develop decoding approaches that can leverage, and therefore reveal, the temporal structure of the underlying neural processes.

To this end, naturalistic speech stimuli are particularly promising: unlike static images, they unfold over time, giving decoders temporal structure to exploit. Yet fMRI trades temporal precision for spatial resolution, while M/EEG offer the reverse trade-off. Pereira et al. (2018) showed that fMRI can reconstruct the semantic content of words and sentences, primarily by leveraging spatial patterns of brain activity, while Défossez et al. (2023) showed that M/EEG support speech decoding through their finer temporal dynamics. This suggests that spatial and temporal information may offer complementary routes to decoding naturalistic speech. Seeking to leverage both, we introduce Neural Ordered Embeddings with Multimodal Alignment (NOEMA), a temporal convolutional decoder that maps windows of consecutive fMRI measurements to multimodal representations of the corresponding speech segments (Figure 1). On two large, densely-sampled speech datasets, NOEMA substantially outperforms existing decoding baselines on held-out sessions and extended retrieval sets.

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.

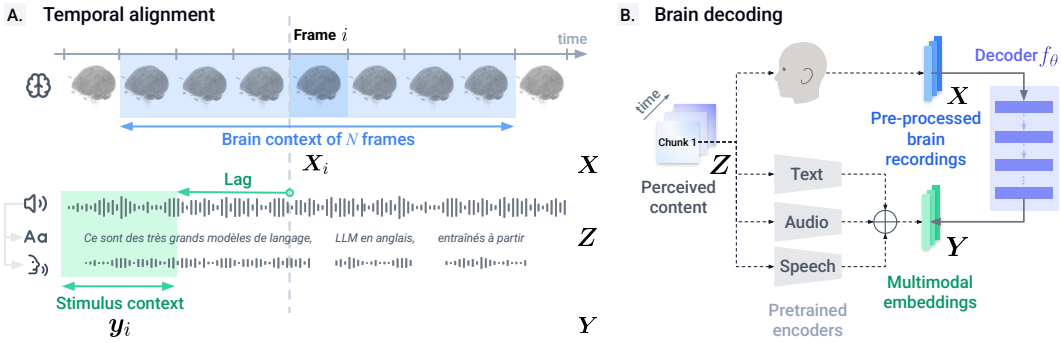


Figure 1: **Temporal alignment and stimulus decoding. A.** A temporal offset (lag) aligns each stimulus-context endpoint with the central fMRI frame to account for the hemodynamic delay. **B.** A decoder maps brain recordings to vector representations of segments produced by pretrained stimulus encoders. Timing values are schematic.

2 Related work

Natural language decoding from fMRI. Pereira et al. (2018) trained semantic decoders on individual concepts and generalized to sentences, while Brain2Word (Affolter et al., 2020) decoded words from a fixed vocabulary. For naturalistic listening, Huth et al. (2016) modeled cortical responses to semantic features; Tang et al. (2023) combined an encoding model with language-model-guided beam search to reconstruct continuous text. BrainLLM (Ye et al., 2025) instead feeds fMRI-derived embeddings directly into a language model.

Temporal integration of fMRI frames. Prior methods integrate temporal context in different ways. Jalouzot et al. (2025) average consecutive volumes before decoding text embeddings. *UniCoRN* (Xi et al., 2023) combines snapshot and sequence reconstruction pretraining with text generation. BP-GPT (Chen et al., 2025) encodes fMRI windows into language-model prompts, using contrastive alignment with text prompts alongside a generation loss. *CogReader* (Lu et al., 2025) combines spatiotemporal encoding with semantic summaries carried between generated segments. UniCoRN reports lower generation scores under stimulus-disjoint than random splits; BP-GPT evaluates on a held-out story, and CogReader uses stimulus-disjoint splits from naturalistic stories. These generation results are not directly comparable to segment retrieval. Here, we compare learned temporal convolutions with frame-wise decoders and an MLP receiving the same context.

Cross-modal alignment with contrastive learning. CLIP (Radford et al., 2021) aligns paired modalities for retrieval in a shared space. Défossez et al. (2023) aligned MEG and EEG with heard-audio representations, while MindEye (Scotti et al., 2023) used contrastive alignment with CLIP image representations for fMRI decoding. Language-model representations predict neural responses to language (Schrimpf et al., 2021; Caucheteux & King, 2022; Pasquiou et al., 2022), and can guide stimulus selection to modulate language-network responses (Tuckute et al., 2024). Beyond text, *Cephalonauts One* (Collas et al., 2026) combines text and audio targets for single-frame decoding. Distinct cortical contributions of acoustic, articulatory, and semantic features (de Heer et al., 2017) motivate our evaluation of complementary targets with temporal decoding.

Contributions. We introduce NOEMA, a temporal convolutional decoder mapping fMRI windows to multimodal speech representations. Our experiments show that NOEMA (1) outperforms the evaluated frame-wise baselines and a Context MLP receiving the same seven-frame input on two naturalistic listening datasets with held-out sessions and unseen stimuli, exceeding 42% top-10 accuracy among approximately 2,000 candidates; (2) benefits from temporal context and frame order, as tested by context-length and shuffle ablations; (3) retains its retrieval advantage when candidates from unseen episodes of the same podcast are added; and (4) improves with complementary text, audio, and speech representations.

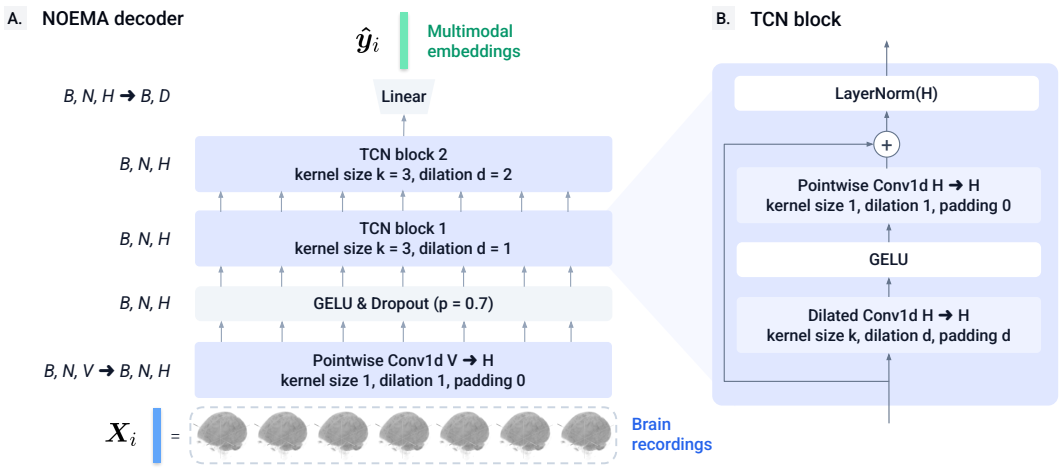


Figure 2: **NOEMA architecture. A.** Information flow from a window of brain recordings to the predicted multimodal embedding. Tensor dimensions are shown beside each stage; only the central-frame representation is projected into the multimodal embedding space. B, N, V, D, H denote the batch size, number of frames, cortical features, target dimension, and hidden dimension respectively. **B.** Structure of a residual temporal convolutional block (TCN block), with a skip connection that bypasses both convolutions and adds the block input before layer normalization. These convolutions use stride one, include biases, and preserve the hidden dimension; symmetric padding preserves the temporal length.

3 Method

3.1 Problem formulation and retrieval objective

Decoding task. We frame naturalistic speech decoding as a retrieval problem: given fMRI measurements, the decoder must identify the corresponding perceived speech segment among candidate segments. As illustrated in Figure 1, we learn a participant-specific decoder f_θ that maps $\boldsymbol{X}_i \in \mathbb{R}^{N \times V}$, a window of N fMRI frames with V cortical features, to $\boldsymbol{y}_i \in \mathbb{R}^D$, the embedding of the corresponding speech segment. At evaluation, we rank the cosine similarities s_{ij} between the predicted embedding $\hat{\boldsymbol{y}}_i \triangleq f_\theta(\boldsymbol{X}_i)$ and each of M candidate segment embeddings $\boldsymbol{y}_j$ from a held-out session. In particular, at inference, the decoder receives only fMRI; candidate audio and transcripts are used solely to precompute the fixed stimulus embeddings.

Contrastive alignment. We train the fMRI decoder by aligning its predictions with the multimodal stimulus targets using the CLIP (Radford et al., 2021) contrastive loss defined over a batch of B samples:

$$\mathcal{L}_{\text{CLIP}} \triangleq -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}/\tau)}\right]$$

In this loss, i, j index paired examples within the batch, and s_{ii} scores a matched brain–stimulus pair. The temperature $\tau > 0$ is optimized per model as described in Section 4.2.

3.2 NOEMA Architecture

Temporal window. NOEMA instantiates f_θ as a temporal convolutional decoder that maps $N = 7$ consecutive fMRI frames, each lasting Δ seconds, to the stimulus embedding associated with the central frame (Figure 2A). For a given stimulus context ending at time t, the associated fMRI window is centered at $t + \delta$ (where $\delta = 2.5$ seconds), and lasts $N \times \Delta$ seconds (see Section 4.1 for values). We use an alignment offset of 2.5 s, which yields the highest accuracy for the MLP and near-peak accuracy for NOEMA (Appendix C).

Spatial projection. A pointwise (1×1) convolution maps V cortical features to a hidden width of $H = 4096$, followed by GELU and dropout with probability 0.7. More simply, this convolution applies the same linear projection independently to each frame, preserving temporal order and yielding an $N \times H$ representation per sample.

Temporal mixing. Two residual temporal convolutional blocks (Bai et al., 2018) integrate consecutive frames using kernel size $K = 3$ and dilations $d_1 = 1$ and $d_2 = 2$, measured in frames. Each TCN block applies a dilated convolution, GELU, and a pointwise convolution, followed by residual addition and layer normalization (Figure 2B). Together, the blocks give the central representation a seven-frame receptive field that includes both earlier and later measurements, making the decoder noncausal.

Central-frame prediction. A final linear layer maps the central representation of dimension H to the $D = 8704$-dimensional multimodal target space (Section 3.3). Only this prediction contributes to the training loss and retrieval score. Appendix B.1 summarizes the decoder architecture.

3.3 Multimodal stimulus representations

Stimulus encoders. We construct multimodal stimulus targets from three frozen encoders operating on text, audio, and speech, extracted at a sequence of time markers t spaced by Δ seconds. Text and audio share a longer context $(t - T, t]$ with $T = 10$ seconds: transcript words are selected by their end times, and audio uses the corresponding waveform segment. `Qwen3-Embedding-8B` (Zhang et al., 2025) embeds the transcripts and supplies 4096 text features from the last non-padding token of its final hidden layer. `LCO-Embedding-Omni-7B` (Xiao et al., 2025) supplies 3584 audio features from the audio waveform. Finally, the ASR encoder `Parakeet-TDT-0.6B-v3` (Sekoyan et al., 2025) uses a window of size Δ seconds, $(t - \Delta, t]$, producing 1024 speech features by averaging its penultimate encoder-layer states within that interval. This shorter context provides a temporally precise speech target that complements the broader text and audio contexts. Since this window matches the marker spacing, successive speech targets tile the stimulus without overlap, giving them a sequential, window-specific organization. We evaluate the effect of changing the speech context in Appendix D. The encoders are used only to precompute targets for training and evaluation; their parameters remain fixed while the participant-specific fMRI decoder is trained.

Target embedding. We ℓ_2-normalize each encoder output $\boldsymbol{e}_i^{\text{text}}$, $\boldsymbol{e}_i^{\text{audio}}$, and $\boldsymbol{e}_i^{\text{speech}}$, then concatenate them into a 8704-dimensional target:

$$\boldsymbol{y}_i \triangleq \left[\frac{\boldsymbol{e}_i^{\text{text}}}{\|\boldsymbol{e}_i^{\text{text}}\|_2} ; \frac{\boldsymbol{e}_i^{\text{audio}}}{\|\boldsymbol{e}_i^{\text{audio}}\|_2} ; \frac{\boldsymbol{e}_i^{\text{speech}}}{\|\boldsymbol{e}_i^{\text{speech}}\|_2} \right]$$

Each modality is normalized separately before concatenation; cosine similarity then normalizes the complete predicted and target vectors.

4 Experimental setup

4.1 Datasets, preprocessing, and splits

Datasets. We evaluate NOEMA on the naturalistic speech decoding task across two datasets, training and evaluating a separate decoder for each participant in each dataset independently:

- *Cephalonauts One* (Collas et al., 2026): fMRI recordings from three participants listening to podcasts in their native language, French. Each participant completed approximately 30 hours across 24 sessions; sessions are shared across participants, but no stimulus is repeated within a participant.
- *LeBel* (LeBel et al., 2023): fMRI recordings from eight participants listening to 6 hours of English stories. We focus on the three participants who completed an additional 10 hours, for sufficient training data, and exclude localizer runs and the repeatedly presented story *Where There's Smoke*.

Table 1 in Appendix A reports the number of sessions, runs, and acquisition hours for each participant.

Preprocessing. For both datasets, we use DeepPrep (Ren et al., 2025) to process fMRI images and reconstruct `fsaverage5` cortical surfaces, concatenate vertex data from both hemispheres, and detrend and standardize each run's vertex time series. *Cephalonauts One* recordings have a native repetition time (TR) of 0.52 seconds. Following the Cephalonauts One preprocessing protocol (Collas et al., 2026), we apply a four-TR moving average and retain every fourth frame, yielding a frame spacing of $\Delta = 2.08$ seconds. *LeBel*'s native TR already matches the stimulus marker spacing ($\Delta = 2.0$ seconds), so we keep the original frame spacing without additional smoothing. Using this preprocessed data, we form the seven-frame fMRI windows described in Section 3.2.

Dataset splits. Consecutive fMRI frames are temporally correlated and share stimulus context, so random splits would leak information between training and evaluation. We therefore split by recording session, which, given the datasets' design, also ensures that no stimulus is shared across training, validation, and test sets. Each model is evaluated on five splits, each holding out one session for testing, one for validation, and using the rest for training. For *Cephalonauts One*, the five test sessions are those of the dataset benchmark (Collas et al., 2026).

4.2 Baselines, Training, and Evaluation

Baselines. We compare NOEMA with four baselines: (1) `Ridge`, a center-frame Ridge regressor with L2 regularization; (2) `MLP`, a single-frame MLP with one hidden layer and GELU activation, from Collas et al. (2026); (3) `MindEye MLP`, a deeper frame-wise MLP with four residual blocks and GELU activation, adapted from Scotti et al. (2023); and (4) `Context MLP`, which applies `MLP`'s architecture to concatenated fMRI frames, controlling for access to the same temporal input window. All baselines predict the same concatenated target embedding as NOEMA, using identical training and evaluation splits. Unless stated otherwise, NOEMA and Context MLP receive seven-frame windows while the remaining baselines receive only the central frame; Context MLP keeps its name even when this window size is varied in the temporal-context ablation (Section 5), to distinguish it from the single-frame MLP.

Training. The Ridge baseline is fitted in closed form using $\alpha = 4.5 \times 10^3$, selected by the hyperparameter search described in Appendix B.2. All other baselines and NOEMA are trained with AdamW (Loshchilov & Hutter, 2017) for up to 200 epochs with an early stopping patience of 10 epochs, using hyperparameters (learning rate, weight decay, batch size, temperature) tuned through a Bayesian search with Optuna (Akiba et al., 2019). For NOEMA, this yields a learning rate of 1.1×10^{-4}, a weight decay of 6.8×10^{-3}, a batch size of 4096, and a temperature of 6.2×10^{-3}; the complete search space and final settings for every model are given in Appendix B.2. Training uses a single NVIDIA H100 GPU (80GB) or AMD MI250X GPU.

Retrieval settings. For model comparisons and ablations on both datasets, the retrieval set consists of all stimulus embeddings from the held-out test session. The retrieval-set extension experiment (Section 5.3) instead evaluates generalization beyond the original stimulus collection: for *Cephalonauts One*, we progressively augment each test-session retrieval set with up to 8,494 embeddings corresponding to ten-second segments from 50 additional episodes of the same radio podcast, absent from the original dataset, reaching a retrieval set roughly five times larger than the original at the largest extension.

Evaluation metrics. We evaluate retrieval by ranking M candidate stimulus embeddings by cosine similarity to each predicted embedding. Let r_i denote the rank of the correct target for test sample i, with rank 1 corresponding to the highest similarity. Our primary metric is Top-10 accuracy, the fraction of test samples for which the correct target appears among the ten highest-ranked candidates:

$$\text{Top-10@}M = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[r_i \leq 10],$$

where n is the number of test samples and $\mathbb{1}[\cdot]$ is the indicator function. The shorthand @2000 denotes approximately 2,000 candidates per test session. To characterize retrieval beyond this fixed cutoff, we also report median rank and its normalized form, median relative rank:

$$\text{Median rank} = \text{median}(r_1, \ldots, r_n), \qquad \text{Median relative rank} = \frac{\text{Median rank}}{M}.$$

Lower values indicate better retrieval for both rank metrics. We use median relative rank to compare retrieval sets of different sizes.

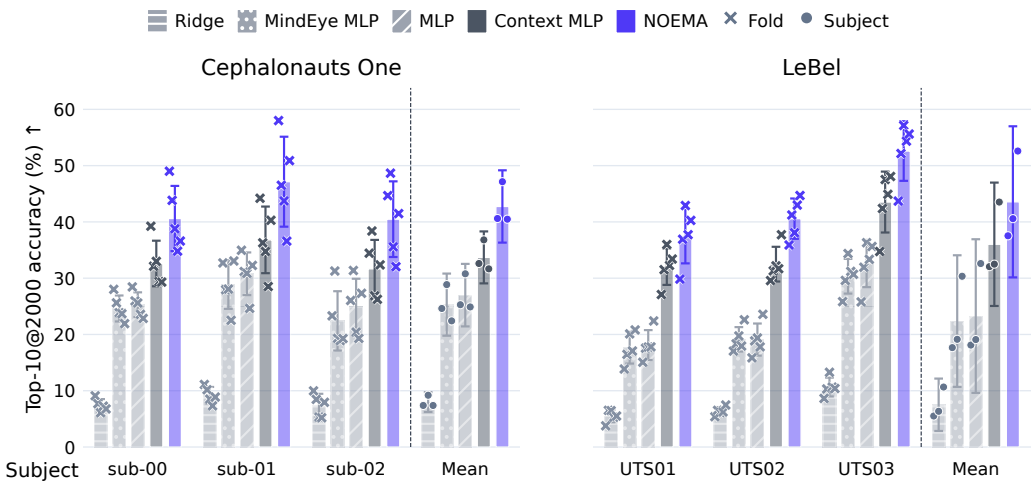


Figure 3: **Top-10@2000 stimulus retrieval across decoder configurations.** Baselines are shown in light gray - with horizontal, dotted and diagonal hatchings for Ridge, MindEye MLP, and MLP respectively - Context MLP in dark gray, and NOEMA in indigo. Subject columns show mean accuracy and standard deviation across five folds, with folds indicated by crosses. In the Mean column, circles report subject means, while bars and error bars show the mean and 90% Student-t confidence interval across the three subject means.

5 Results

5.1 Retrieval performance

We first evaluate how accurately each decoder identifies the speech segment that elicited a given neural response. Figure 3 compares NOEMA with the baseline decoders on *Cephalonauts One* and *LeBel* datasets. All models predict the same multimodal (text, audio, speech) targets and are tuned according to the hyperparameter optimization procedure described in Appendix B.2. NOEMA consistently outperforms all baselines across subjects, retrieving the correct speech segment within the top 10 candidates for 42.7% of queries on *Cephalonauts One* and 43.6% on *LeBel*, far above the chance level of 0.5%. Notably, NOEMA also surpasses the Context MLP, even though both models receive the same input context and the Context MLP has more parameters (Appendix F.4). NOEMA's advantage therefore does not appear to only arise from additional context or greater model capacity. One plausible explanation is its use of temporal convolutions, which explicitly model the temporal structure of the neural signal. This advantage holds across alternative retrieval metrics and rank thresholds: Top-1@2000 accuracy, median relative rank, cumulative rank distributions, and rank histograms are reported in Appendix E.

5.2 Stimulus representations and temporal context

Stimulus representations. To assess the contribution of each stimulus representation, we trained NOEMA, MLP, and Context MLP on targets built from cumulative combinations of text (T), audio (A), and speech (S) representations. On *Cephalonauts One*, adding audio to a text-only target raises NOEMA's Top-10@2000 accuracy from 27.6% to 35.5% (Figure 4a), and further adding speech to the text-and-audio target raises it to 42.7% (Figure 4b). Adding audio and then speech improves retrieval for all three models in every subject and fold, with the largest mean gains for NOEMA. These gains exceed those of Context MLP, which receives the same seven-frame input, by 1.58 and 2.34 percentage points, respectively. The same pattern holds on *LeBel*, where adding speech has an even larger effect (Appendix F.1). Absolute accuracies for each model under the T, T+A, and T+A+S targets are reported in Appendix F.2.

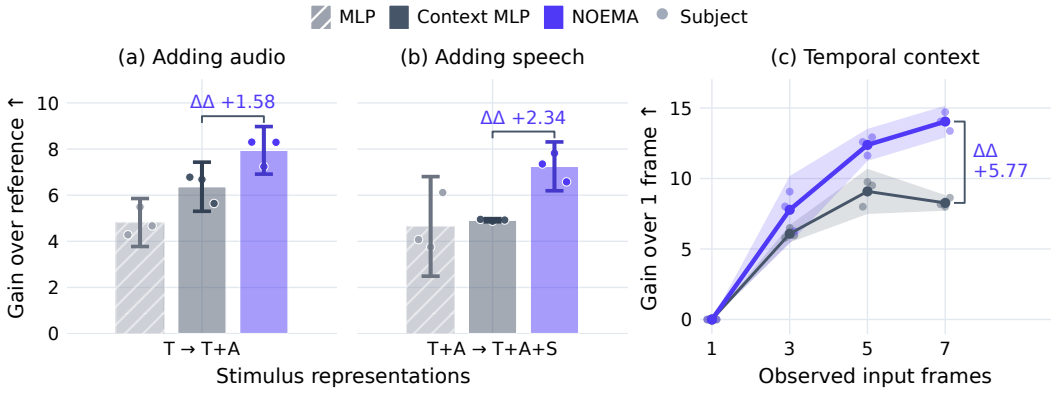


Figure 4: **Effect of stimulus representations and temporal context (*Cephalonauts One*).** All panels show the gain in Top-10@2000 accuracy (percentage points) over a reference configuration, computed within each model, subject, and fold. **(a)** Adding audio to a text-only target (T→T+A). **(b)** Adding speech to a text and audio target (T+A→T+A+S). T: Qwen text; A: LCO audio; S: Parakeet speech (ASR). ΔΔ is the mean NOEMA gain minus the mean Context MLP gain. **(c)** Gain from increasing the input from one frame to 3, 5, or 7 frames, for Context MLP and NOEMA. Here, ΔΔ compares the two models' gains from one to seven frames. Circles show subject means over five folds. Bars in (a, b) and lines in (c) show the mean across three subjects. Error bars and shaded bands show 90% Student-t confidence intervals. See Appendix F.2 and F.3 for absolute accuracies.

Temporal context. To assess how effectively each architecture exploits information spread across neighboring fMRI frames, we train Context MLP and NOEMA with 1, 3, 5, or 7 input frames. Adding frames increases Context MLP's parameter count, while NOEMA's remains fixed. Both models benefit from additional frames, with larger accuracy gains for NOEMA (Figure 4c). On *Cephalonauts One*, NOEMA improves steadily with each added frame, rising from 28.7% Top-10@2000 accuracy with one frame to 42.7% with seven. Context MLP, in contrast, rises from 25.4% to a peak of 34.5% at five frames and declines slightly to 33.7% at seven. On *LeBel*, both models improve at every window length, and again NOEMA benefits more (Appendix F.1). NOEMA achieves this higher accuracy with fewer parameters: 253.8M trainable parameters, compared with 623.0M for the seven-frame Context MLP (Appendix F.4). These results suggest that NOEMA's temporal convolutions make better use of additional fMRI frames than the tested Context MLP. Absolute accuracies are reported in Appendix F.3. To probe the role of this temporal inductive bias, we retrain and evaluate both models with their input frames permuted. A permutation shared across samples can be offset by reordering the Context MLP's input weights, but it disrupts the temporal neighborhoods on which NOEMA's convolutions rely. Accordingly, NOEMA's accuracy drops while the Context MLP's remains essentially unchanged, suggesting that NOEMA's inductive bias is well matched to the chronological order of fMRI frames (Appendix G).

5.3 Extended retrieval sets

Retrieval beyond the original dataset. Retrieval within a single test session uses a closed set of about 2,000 candidates. A practical decoder, however, must identify the correct segment within a much larger, open-ended pool. To test whether our conclusions hold in this setting, we extend each *Cephalonauts One* test-session retrieval set with up to 8,494 ten-second segments drawn from 50 episodes of the same radio podcast that are not part of the dataset (Appendix I). Drawing distractors from the same show limits broad differences in recording conditions and stimulus domain; this reduces, but does not eliminate, acoustic and stylistic differences between original and added candidates. All models are evaluated without retraining, and we track the median rank of the correct segment as the retrieval set grows. If the added distractors were exactly as confusable as the original candidates, the median rank would grow in proportion to the retrieval-set size, and the median relative rank would stay constant.

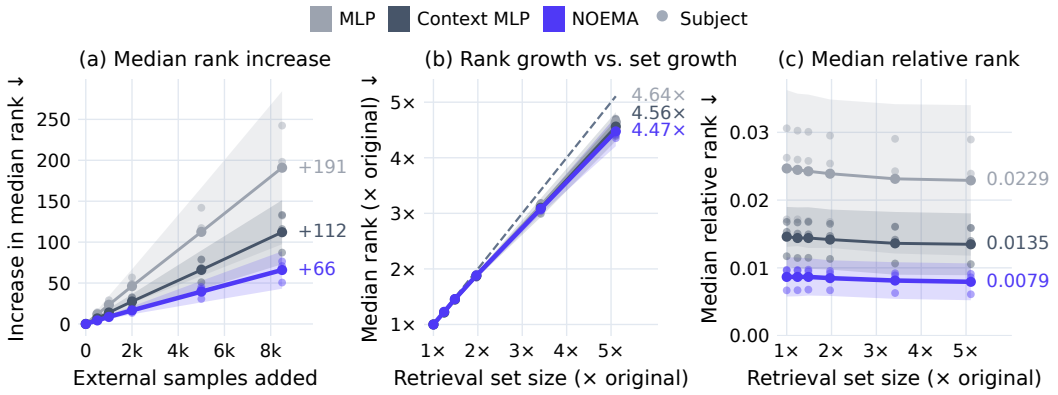


Figure 5: **Extended retrieval sets (*Cephalonauts One*).** Up to 8,494 segments from unseen episodes are added to each test-session retrieval set. **(a)** Increase in median rank versus number of added samples. **(b)** Median rank relative to the original set versus retrieval-set size; the dashed line marks proportional growth, expected if external samples were as confusable as in-set ones. **(c)** Median rank divided by retrieval-set size. Lower is better. Circles show subject means over five folds; lines and bands show the mean and 90% Student-t confidence interval across three subjects.

Effect of retrieval-set size. NOEMA shows the smallest absolute increase in median rank as distractors are added. At the largest extension, its median rank increases by 66 positions, compared with 112 and 191 for the Context MLP and MLP baselines (Figure 5a). Relative to their original median rank, the ranks of all three models grow by a similar factor (4.5-4.6× for a 5.1× larger set; Figure 5b). Because NOEMA already ranks the correct target within the top 0.9% of candidates, compared with 2.5% for the MLP, it moves fewer positions as the pool grows. Across all three models, median rank increases slightly less than proportionally to retrieval-set size, yielding a small decrease in median relative rank (Figure 5c). NOEMA maintains the lowest median relative rank throughout the expansion.

5.4 Model Interpretation

Finally, we examine the models' sensitivity to their fMRI inputs using feature-importance maps derived from integrated gradients (Sundararajan et al., 2017). Figure 6 shows NOEMA's left and right lateral views. Details on the method, other views for NOEMA, and maps for the standard and Context MLP baselines can be found in Appendix J. Among the seven-frame models, only NOEMA shows a bias toward the central frames which is consistent with its temporal convolutions and central-frame readout. Across all three models, importance is higher in lateral temporal cortex, with smaller patches in frontal cortex, while the medial views show no distinct high-importance regions.

These frontotemporal regions are consistent with the *language network* areas identified with individual-subject localizers in Fedorenko et al. (2010) and by sensitivity to sentence constituent structure in Pallier et al. (2011). This method only describes model sensitivity and does not establish causal relationship between these regions and the decoding task. Nonetheless, the prominence of such plausible regions across the models is still encouraging given that they were trained without any prior on language-responsive regions.

Also, those maps do not show a clear left-right asymmetry that could be expected given the well-established left lateralization of language processing (Broca, 1865). However, while linearly aligning language-model representations with fMRI responses, Bonnasse-Gahot & Pallier (2024) found that this asymmetry re-appears for larger language models. Therefore, despite the different setup, larger scale models might be necessary for the same re-emergence to occur.

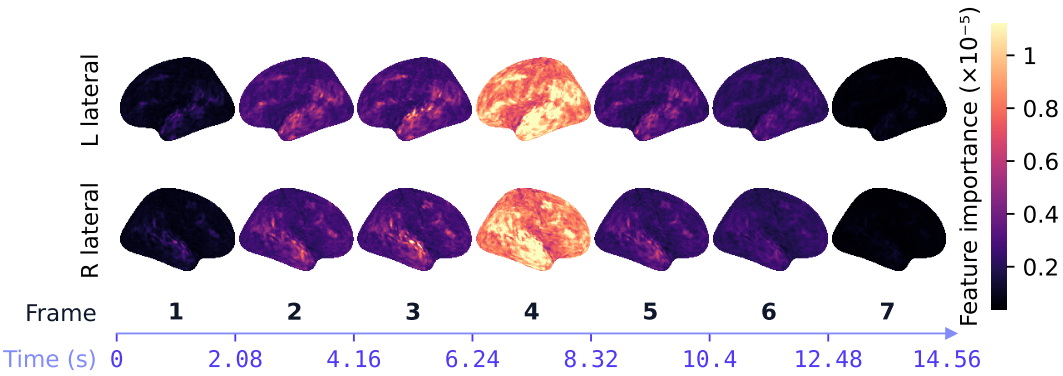


Figure 6: **Feature importance across input frames.** Absolute integrated-gradients attributions with respect to the 7 fMRI frames which NOEMA takes as input. The figure shows the left and right lateral views for *Cephalonauts One*'s `sub-01`. Other views and more details on the method can be found in Appendix J.

6 Conclusion

NOEMA combines temporal modeling of fMRI with alignment to text, audio, and speech representations to improve naturalistic speech retrieval. On entirely held-out recording sessions containing speech stimuli absent from training, NOEMA outperforms all baselines across two naturalistic listening datasets, demonstrating generalization to new sessions and stimuli within participants. This advantage is preserved when the retrieval set is extended with external distractors.

Our ablations show that both temporal context and the choice of target representations contribute to retrieval performance, supporting the view that the temporal structure of brain measurements and the representations used for supervision are complementary design choices for speech decoders. NOEMA benefits more from additional fMRI frames than an MLP receiving the same context, despite having fewer parameters, and its performance depends on chronological frame order.

Limitations and future work. Our evaluation is limited to six participants across two listening datasets, with a separate decoder trained for each participant; generalization to new individuals and to other tasks remains untested, and extending NOEMA across participants is a natural next step. Our analyses also do not establish what underlies the benefit of temporal context: it may reflect accommodation of the delayed hemodynamic response, stimulus-related temporal structure, or both. Moreover, the temporal window is bounded by the architecture's seven-frame receptive field, and the current results do not establish an optimal context length. Finally, NOEMA retrieves from a predefined candidate set. Although its advantage persists with external distractors, retrieval from any finite set does not demonstrate unrestricted speech reconstruction; using NOEMA's predicted embeddings to guide text generation is a promising direction toward this goal.

AI USE STATEMENT

We used generative AI tools extensively throughout the research and writing process. Their use included developing and debugging code, preparing and running experiments, analyzing and interpreting results, producing figures, searching the literature, and drafting and revising the manuscript. Coauthors reviewed AI-assisted contributions through a pull-request process to check their correctness. The authors take full responsibility for the methods, results, and final content of this work.

ETHICS STATEMENT

This paper presents a methodological study of naturalistic speech decoding from fMRI using two existing datasets: *Cephalonauts One* (Collas et al., 2026) and *LeBel* (LeBel et al., 2023). Both datasets were collected under protocols approved by the respective ethics committees, and all subjects provided written informed consent.

REPRODUCIBILITY STATEMENT

We describe the datasets, preprocessing, data splits, training procedure, and evaluation protocol in Section 4. Where available, we use the test sessions specified by the dataset benchmark, as for *Cephalonauts One* (Collas et al., 2026). Appendix B provides the detailed architecture, a standalone PyTorch implementation of NOEMA, and the hyperparameter search space and selected configurations.

ACKNOWLEDGMENTS

We thank the developers and maintainers of the open-source software used in this work. Our Python implementation uses NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), pandas (McKinney, 2010), xarray (Hoyer & Hamman, 2017), scikit-learn (Pedregosa et al., 2011), and chardet for data handling and analysis. Neuroimaging preprocessing uses DeepPrep (Ren et al., 2025), Nilearn (Abraham et al., 2014), and NiBabel, while audio processing uses TorchAudio (Yang et al., 2022), librosa (McFee et al., 2015), and SoundFile. Model training and stimulus embedding extraction rely on PyTorch (Paszke et al., 2019), Transformers (Wolf et al., 2020), Sentence Transformers (Reimers & Gurevych, 2019), and Hugging Face Hub. We use Optuna (Akiba et al., 2019), Exca, Submitit, joblib, Pydantic, and PyYAML to configure, optimize, and run experiments, and MLflow (Zaharia et al., 2018), Loguru, psutil, and tqdm (da Costa-Luis, 2019) to monitor them. Visualizations use Matplotlib (Hunter, 2007) and Plotly.

REFERENCES

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014. doi: 10.3389/fninf.2014.00014.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271, 2018. URL https://arxiv.org/abs/1803.01271.

Laurent Bonnasse-Gahot and Christophe Pallier. fMRI predictors based on language models of increasing complexity recover brain left lateralization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125231–125263. Curran Associates, Inc., 2024. doi: 10.52202/079017-3977. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e28e19d00b23fe0265f433fa05a96b06-Abstract-Conference.html.

Paul Broca. Sur le siège de la faculté du langage articulé. *Bulletins de la Société d'Anthropologie de Paris*, 6:377–393, 1865. doi: 10.3406/bmsap.1865.9495.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, 2022. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Xiaoyu Chen, Changde Du, Che Liu, Yizhe Wang, and Huiguang He. Bp-gpt: Auditory neural decoding using fmri-prompted llm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Antoine Collas, Louis Jalouzot, Géraud Ilinca, Corentin Caris, Romain Valabrègue, Ahmed Hassayoune, David Goncalves, Madeleine Hueber, Thaddée Delebarre, Julien Savatovsky, Clara Fonteneau, Charles Maussion, Bertrand Thirion, and Alexis Thual. Cephalonauts One: A deep fMRI dataset for decoding naturalistic speech in the human brain, 2026.

Casper O. da Costa-Luis. tqdm: A fast, extensible progress meter for Python and CLI. *Journal of Open Source Software*, 4(37):1277, 2019. doi: 10.21105/joss.01277.

Wendy A. de Heer, Alexander G. Huth, Thomas L. Griffiths, Jack L. Gallant, and Frédéric E. Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5511884/.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5. URL https://www.nature.com/articles/s42256-023-00714-5.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010. ISSN 0022-3077. doi: 10.1152/jn.00032.2010. URL https://journals.physiology.org/doi/full/10.1152/jn.00032.2010.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293(5539):2425–2430, September 2001. ISSN 0036-8075. doi: 10.1126/science.1063736.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017. URL https://www.nature.com/articles/ncomms15037.

Stephan Hoyer and Joe Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1):10, 2017. doi: 10.5334/jors.148.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

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. doi: 10.1038/nature17637. URL https://www.nature.com/articles/nature17637.

Louis Jalouzot, Alexis Thual, Yair Lakretz, Christophe Pallier, and Bertrand Thirion. Optimizing fMRI Data Acquisition for Decoding Natural Speech with Limited Participants, 2025. URL http://arxiv.org/abs/2505.21304.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL https://arxiv.org/abs/2009.07896.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10:555, 2023. doi: 10.1038/s41597-023-02437-z. URL https://www.nature.com/articles/s41597-023-02437-z.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Wentao Lu, Dong Nie, Pengcheng Xue, Zheng Cui, Piji Li, Daoqiang Zhang, and Xuyun Wen. Brain-inspired fMRI-to-text decoding via incremental and wrap-up language modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/dca63f2650fe9e88956c1b68440b8ee9-Abstract-Conference.html.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pp. 18–25, 2015. doi: 10.25080/Majora-7b98e3ed-003.

Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion, June 2023. URL http://arxiv.org/abs/2303.05334. arXiv:2303.05334 [cs, q-bio].

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108 (6):2522–2527, 2011. doi: 10.1073/pnas.1018711108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1018711108.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. In *ICML 2022 - 39th International Conference on Machine Learning*, pp. 18, Baltimore, United States, July 2022. URL https://inria.hal.science/hal-03704504.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9:963, 2018. doi: 10.1038/s41467-018-03068-4. URL https://doi.org/10.1038/s41467-018-03068-4.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Jianxun Ren, Ning An, Cong Lin, Youjia Zhang, Zhenyu Sun, Wei Zhang, Shiyi Li, Ning Guo, Weigang Cui, Qingyu Hu, et al. Deepprep: an accelerated, scalable and robust pipeline for neuroimaging preprocessing empowered by deep learning. *Nature Methods*, 22(3):473–476, 2025.

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*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4ddab70bf41ffe5d423840644d3357f4-Abstract-Conference.html.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and high-performance models for multilingual ASR and AST. *arXiv preprint arXiv:2509.14128*, 2025. URL https://arxiv.org/abs/2509.14128.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. URL https://www.nature.com/articles/s41562-023-01783-7.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13277–13291, 2023.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e839f274b3d3d682a0a82e5129d2cc4a-Paper-Conference.pdf.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, et al. TorchAudio: Building blocks for audio and speech processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6982–6986, 2022.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8:346, 2025. doi: 10.1038/s42003-025-07731-7.

Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Fen Xie, and Corey Zumar. Accelerating the machine learning lifecycle with MLflow. *IEEE Data Engineering Bulletin*, 41(4), 2018. URL https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A Dataset size

Table 1: **Number of sessions, runs, and hours per participant.** Counts cover all training, validation, and test sessions, with each run counted once. LeBel localizer runs and repeated presentations of *Where There's Smoke* are excluded. Hours are computed from the number of native fMRI frames multiplied by the TR, before temporal downsampling and window selection.

Dataset	Participant	Sessions	Runs	Hours
Cephalonauts One	sub-00	24	120	29.94
	sub-01	24	121	30.24
	sub-02	24	117	29.13
LeBel	UTS01	15	82	16.14
	UTS02	15	83	16.37
	UTS03	15	83	16.37

B Detailed configurations

B.1 Model architecture

Table 2 lists each operation of the NOEMA decoder with its output shape, parameter count, and per-example compute.

Table 2: **NOEMA decoder architecture.** Shapes are time $\times$ features, with $N = 7$ frames, $V = 20{,}484$ cortical vertices, $H = 4096$, and $D = 8704$. $\mathrm{Conv}^{(\ell)}_{K,d,C}$ is a temporal convolution with kernel size K, dilation d, and C output channels, in TCN block ℓ (blank for the spatial projection); Linear_C is a linear layer with C output channels. Parameter counts and per-example compute are given as orders of magnitude, omitting constants and biases.

Operation	Shape	Params	Compute
Input window	$N \times V$	–	–
$\mathrm{Conv}_{1,1,H}$	$N \times H$	VH	NVH
GELU; dropout $p = 0.7$	$N \times H$	–	NH
$\mathrm{Conv}^{(1)}_{3,1,H}$	$N \times H$	H^2	NH^2
GELU; $\mathrm{Conv}^{(1)}_{1,1,H}$	$N \times H$	H^2	NH^2
Residual; LayerNorm	$N \times H$	H	NH
$\mathrm{Conv}^{(2)}_{3,2,H}$	$N \times H$	H^2	NH^2
GELU; $\mathrm{Conv}^{(2)}_{1,1,H}$	$N \times H$	H^2	NH^2
Residual; LayerNorm	$N \times H$	H	NH
Select central frame	H	–	–
Linear_D	D	HD	HD
Total	–	$VH + H^2 + HD$	$NVH + NH^2 + HD$

Figure 7 gives the full model definition in PyTorch (Paszke et al., 2019): a spatial projection applied independently to each frame, two TCN blocks, and a linear layer applied to the central frame.

```
import torch.nn as nn


class TCNblock(nn.Module):
    def __init__(self, hidden_dim, dilation):
        super().__init__()
        self.layers = nn.Sequential(
            nn.Conv1d(hidden_dim, hidden_dim, kernel_size=3,
                      padding=dilation, dilation=dilation),
            nn.GELU(),
            nn.Conv1d(hidden_dim, hidden_dim, kernel_size=1),
        )
        self.norm = nn.LayerNorm(hidden_dim)

    def forward(self, x):
        # x: (batch, frames, hidden_dim)
        return self.norm(
            x + self.layers(x.transpose(1, 2)).transpose(1, 2)
        )


class NOEMA(nn.Module):
    def __init__(self, input_dim, output_dim=8704,
                 hidden_dim=4096, dropout=0.7):
        super().__init__()
        self.spatial = nn.Sequential(
            nn.Linear(input_dim, hidden_dim),
            nn.GELU(),
            nn.Dropout(dropout),
        )
        self.temporal = nn.Sequential(
            TCNblock(hidden_dim, dilation=1),
            TCNblock(hidden_dim, dilation=2),
        )
        self.output = nn.Linear(hidden_dim, output_dim)

    def forward(self, x):
        # x: (batch, frames, input_dim).
        x = self.temporal(self.spatial(x))

        # project the central frame
        return self.output(x[:, x.shape[1] // 2])
```

Figure 7: **PyTorch implementation of NOEMA.** The spatial projection is applied independently to each frame, and two TCN blocks with dilations 1 and 2 give the central frame a seven-frame receptive field. Only the central frame is projected to the target space. Default dimensions follow Table 2.

B.2 Hyperparameter Optimization

We use Optuna (Akiba et al., 2019) to optimize the training hyperparameters for NOEMA and the baselines. Specifically, we perform a Bayesian search over the learning rate, weight decay, batch size, temperature, and hidden width of the neural models, and over the penalty α of the Ridge decoder. Other architecture choices, dropout rates, and early stopping patience are fixed to the values used in the experiments; for MindEye-MLP, dropout rates are the default values from Scotti et al. (2023). Test sessions are excluded entirely from this search. We use five session-based training/validation splits of the remaining sessions, each holding out one session for validation and using the others for training. The search is performed only on *Cephalonauts One* subject `sub-01`, using the concatenated text–audio–speech target representation. The resulting search space is summarized in Table 3. The best hyperparameters found for each model are reported in Table 4.

Table 3: **Hyperparameter search space.** Bayesian search with Optuna on *Cephalonauts One* sub-01, using five session-based training/validation splits of the non-test sessions. Each trial minimizes the validation median relative rank averaged over folds. Neural models are NOEMA, both MLP variants, and MindEye-MLP; hyperparameters not listed are fixed to the values in Section 4.2.

Model	Hyperparameter	Search space	Sampling
Neural models	Batch size	$\{2048, 4096, 8192\}$	categorical
	Hidden width	$\{1024, 2048, 4096, 8192\}$	categorical
	Temperature	$[10^{-3}, 1]$	log-uniform
	Learning rate	$[10^{-5}, 10^{-2}]$	log-uniform
	Weight decay	$[10^{-5}, 10^{-2}]$	log-uniform
Ridge	Penalty α	$[10^{-1}, 10^{5}]$	log-uniform

Table 4: **Selected hyperparameters.** Best Optuna trial for each decoder, from the search space in Table 3. The last column gives the validation median relative rank (mean $\pm$ standard deviation over five folds; lower is better). Ridge uses $\alpha = 4.50 \times 10^3$. Experiments use temperature, learning rate, weight decay, and Ridge penalty rounded to two significant figures; batch sizes and hidden widths are used exactly as listed.

Model	Batch size	Width	τ	Learning rate	Weight decay	Val. rel. rank $\downarrow$
NOEMA	4096	4096	6.23×10^{-3}	1.13×10^{-4}	6.79×10^{-3}	0.0077 ± 0.0024
Context MLP	4096	4096	8.79×10^{-3}	3.68×10^{-5}	5.33×10^{-3}	0.0128 ± 0.0029
MLP	8192	8192	9.12×10^{-3}	1.30×10^{-4}	4.26×10^{-5}	0.0193 ± 0.0047
MindEye-MLP	4096	2048	8.83×10^{-3}	6.65×10^{-4}	2.95×10^{-5}	0.0203 ± 0.0046
Ridge	–	–	–	–	–	0.0919 ± 0.0186

C SENSITIVITY TO LAG

We vary the alignment offset δ (Section 3.2) from -5 to 10 seconds in 0.75-second steps on *Cephalonauts One* sub-01, comparing NOEMA with the MLP. Both models receive the same seven-frame windows, but the MLP uses only the central frame. Each model is retrained at every offset on five session folds, for 210 fits in total. NOEMA is more accurate than the MLP at every offset (Figure 8) and is also less sensitive to the choice of offset: its Top-10@2000 accuracy stays within 10% of its value at the default $\delta = 2.5$ seconds for offsets from -1.25 to 4 seconds, compared with 1 to 3.25 seconds for the MLP. This robustness is expected, since NOEMA's receptive field spans approximately ± 6 seconds around the central frame. NOEMA's accuracy peaks at $\delta = 1$ second (47.75%), 1.04 percentage points above the default (46.71%), whereas the MLP peaks at the default (29.90%). Because these curves come from the test sessions of a single participant, we do not use them to select δ. The offset is measured from the end of the stimulus context, so it is not an estimate of the hemodynamic response peak.

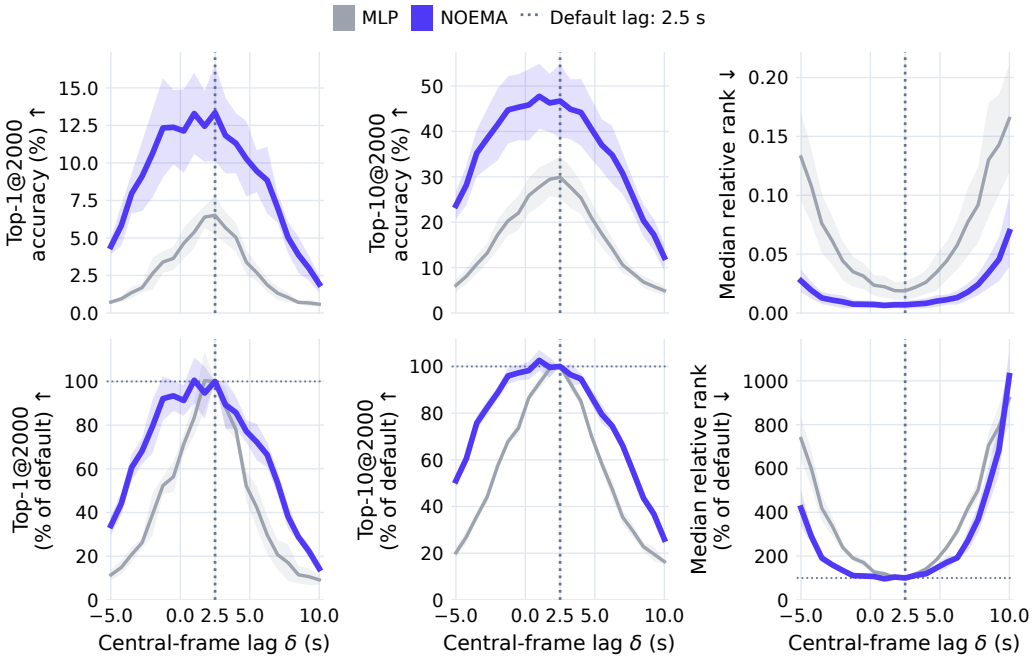


Figure 8: **Sensitivity to the alignment offset (*Cephalonauts One*, sub-01).** Columns: Top-1@2000 and Top-10@2000 accuracy (higher is better), and median relative rank (lower is better). Top row: absolute values. Bottom row: values as a percentage of each model's value at the default $\delta = 2.5$ seconds (dotted lines), computed within each fold before averaging. Lines show the mean over five folds; shaded bands show one standard deviation.

D ASR CONTEXT WINDOW

Section 3.3 sets the trailing context window of the ASR encoder equal to the marker spacing Δ (2.08 seconds for *Cephalonauts One*, 2.0 seconds for *LeBel*), so that successive speech targets tile the stimulus without overlap. Here we test whether a different window improves retrieval. We vary the Parakeet window over 1 second, Δ, and 3 to 10 seconds in 1-second steps, keeping the marker spacing Δ and the text and audio context $T = 10$ seconds fixed. NOEMA and the MLP are retrained at each window with the hyperparameters of Appendix B.2 on five session folds, for one subject per dataset (*Cephalonauts One* sub-01 and *LeBel* UTS01), for 200 fits in total. Accuracy follows an inverted-U shape for both models on both datasets (Figure 9). NOEMA's Top-10@2000 accuracy peaks at 5 seconds on *Cephalonauts One* (48.19%, compared with 43.27% at 1 second and 42.55% at 10 seconds) and at 3 seconds on *LeBel* (39.84%, compared with 33.18% and 32.13%). The default window lies 1.70 and 2.52 percentage points below these peaks, respectively; for the MLP, the gaps are 1.03 and 0.82 points. Validation accuracy peaks at the same windows, so tuning the ASR window per dataset could bring further gains. We keep the window equal to Δ because it gives each target a distinct, non-overlapping speech segment, and leave this tuning to future work.

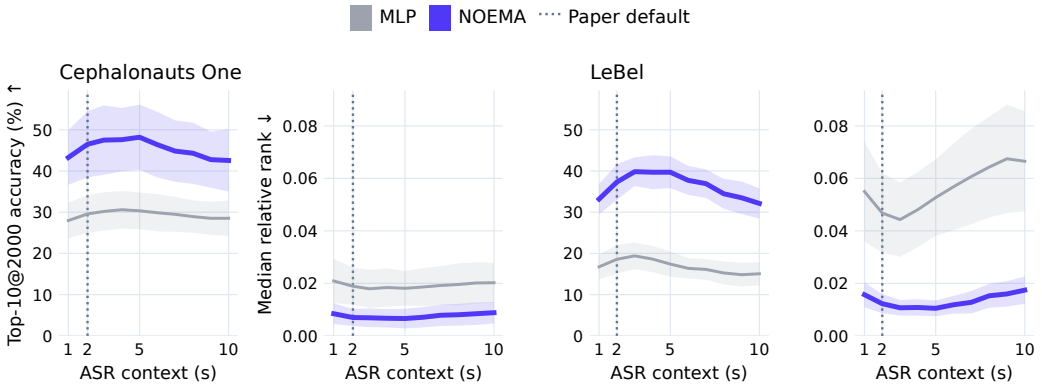


Figure 9: **Effect of the ASR context window.** Top-10@2000 accuracy (higher is better) and median relative rank (lower is better) as a function of the trailing Parakeet window, for *Cephalonauts One* sub-01 (left pair) and *LeBel* UTS01 (right pair). Dotted lines mark the default window, equal to the marker spacing Δ (2.08 and 2.0 seconds). Lines show the mean over five folds; shaded bands show one standard deviation.

E Evaluation with additional metrics

E.1 Top-1 accuracy and median relative rank

Figure 10 complements the Top-10@2000 accuracy in Figure 3 with two further metrics. Top-1@2000 accuracy is the strictest criterion: the correct segment must rank first among approximately 2,000 candidates. Median relative rank instead summarizes the typical position of the correct segment divided by the retrieval-set size, including queries where it falls outside the top ten. NOEMA performs best on both metrics for every subject. It ranks the correct segment first for 12.0% of queries on *Cephalonauts One* and 13.6% on *LeBel*, compared with 7.7% and 10.0% for the strongest baseline, the seven-frame MLP. It also reduces the mean median relative rank from 0.0146 to 0.0087 on *Cephalonauts One* and from 0.0145 to 0.0093 on *LeBel*.

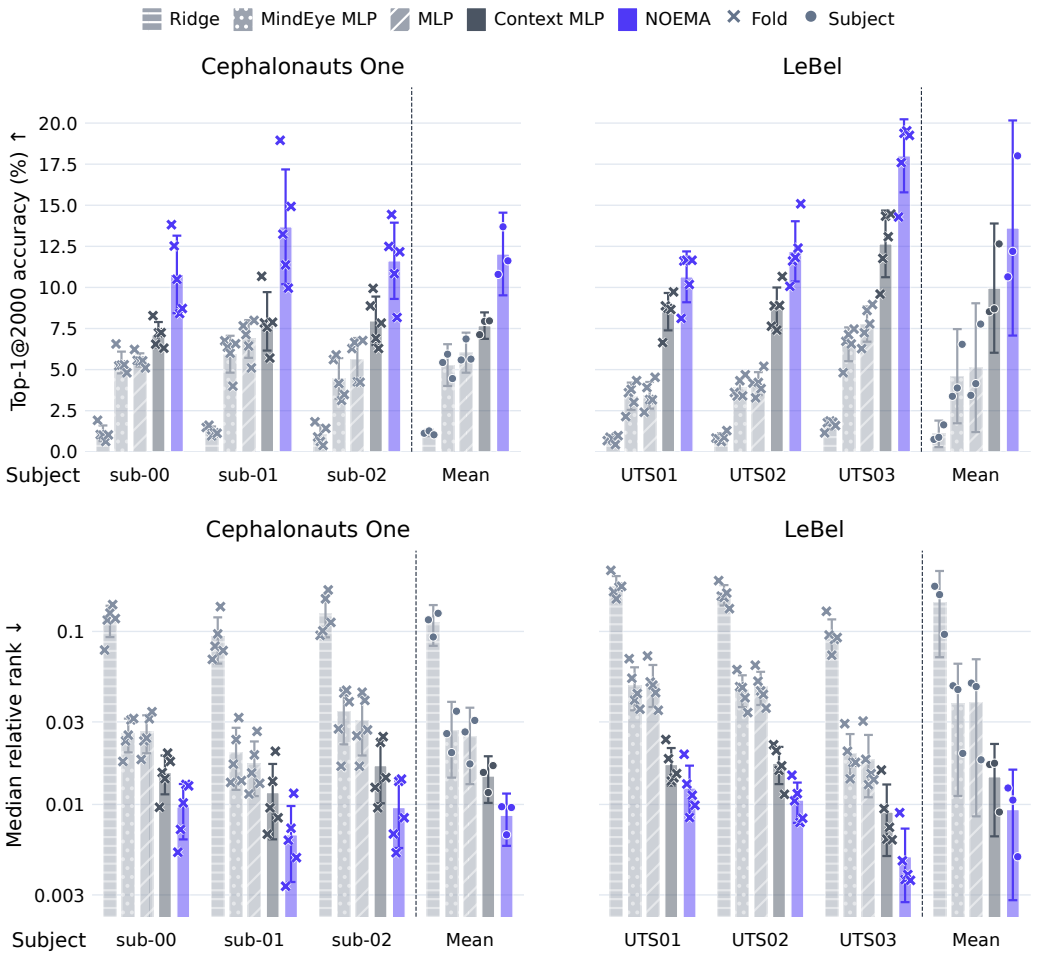


Figure 10: **Top-1@2000 accuracy and median relative rank. Top:** Top-1@2000 accuracy (higher is better). **Bottom:** Median relative rank of the correct segment, on a log scale (lower is better). Models and colors follow Figure 3. In subject columns, bars and error bars show the mean and standard deviation over five folds (crosses); in Mean columns, they show the mean and 90% Student-t confidence interval over three subjects (circles).

E.2 RANK DISTRIBUTIONS

Top-k accuracy and median rank each summarize the retrieval ranks with a single number. Figures 11 and 12 show the full distribution of the correct segment's rank, for the models in Figure 3. The cumulative distributions show that NOEMA's advantage is not limited to a particular threshold. For every subject, its curve lies on or above that of every baseline across nearly the entire rank range, with small crossings as all models approach 100%. Averaged over subjects, NOEMA places half of all queries within rank 16 on both datasets, compared with ranks 28 and 25 for the seven-frame MLP on *Cephalonauts One* and *LeBel*, respectively. It places 90% of queries within rank 328 on *Cephalonauts One* and 263 on *LeBel*, compared with 445 and 390.

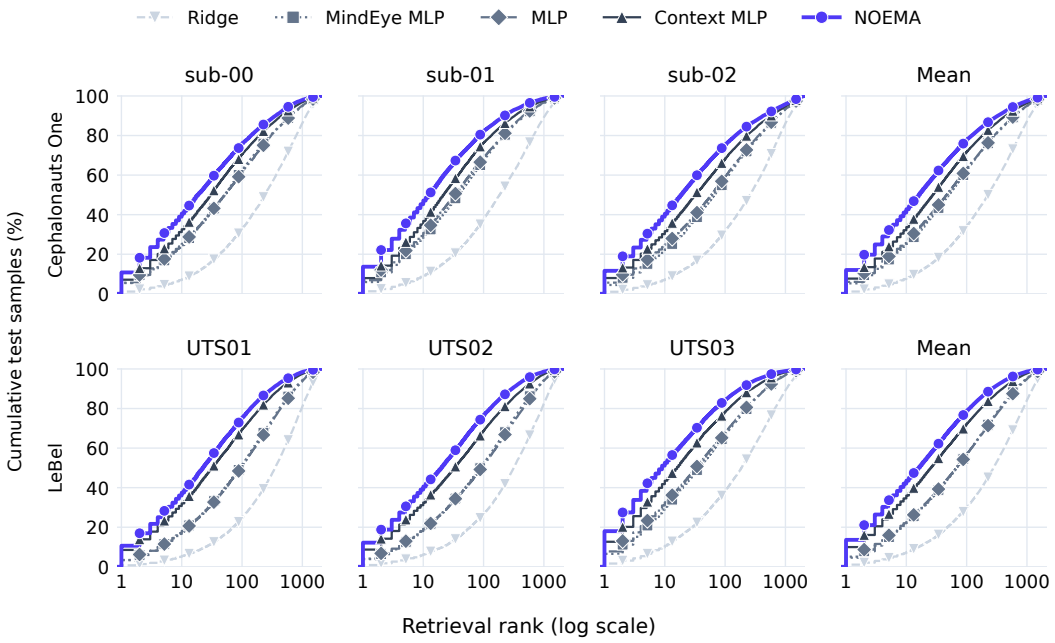


Figure 11: **Cumulative distribution of retrieval ranks.** Percentage of test queries whose correct segment ranks at or above a given position, for the models in Figure 3; ranks are on a log scale. Values at ranks 1 and 10 equal Top-1@2000 and Top-10@2000 accuracy. Rows show datasets; columns show individual subjects and the mean over subjects, with folds weighted equally. Higher curves indicate better retrieval.

The histograms show where this advantage comes from. NOEMA's density is highest among the lowest ranks and decreases steadily beyond rank 100, so most of its queries fall within the top few percent of the approximately 2,000 candidates. The baselines instead peak at intermediate ranks, around 100 for the MLPs, and Ridge places most of its mass near the top of the range, where a decoder that ranked candidates at random would also place most of its mass. Moving from Ridge to the MLPs, the seven-frame MLP, and NOEMA, probability mass shifts steadily toward lower ranks, the same ordering as in Figure 3.

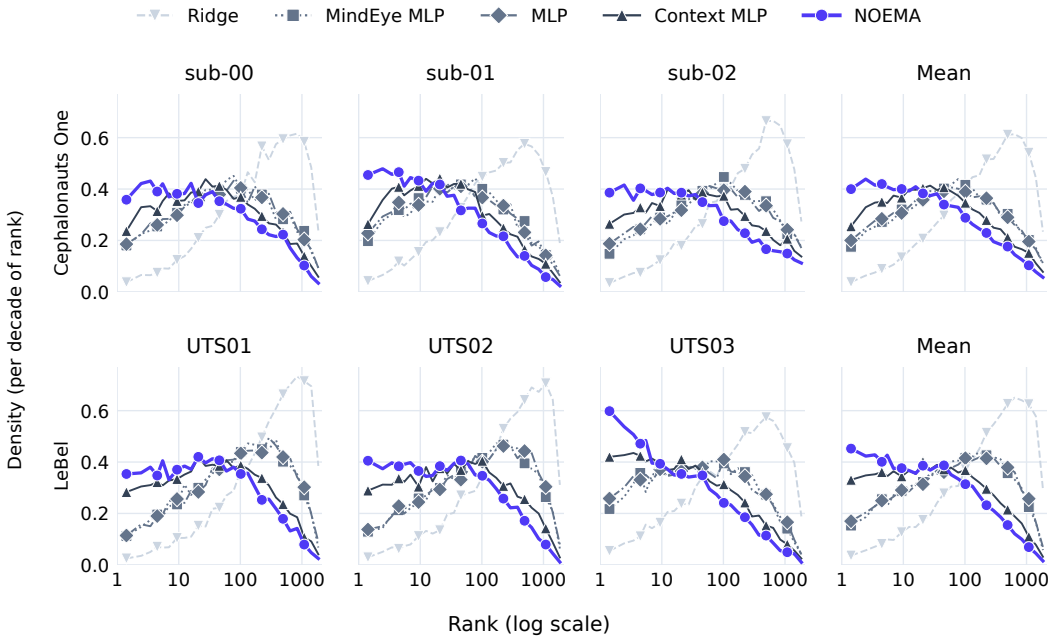


Figure 12: **Distribution of retrieval ranks.** Density of the correct segment's rank for the models in Figure 3, with ranks binned on a log scale (rank k falls in $[k, k{+}1)$). Density is per decade of rank, so the area under each curve over $\log_{10}$ rank is one. Panels follow Figure 11. Curves shifted to the left indicate better retrieval.

F STIMULUS REPRESENTATIONS AND TEMPORAL CONTEXT

F.1 RESULTS ON *LeBel* DATASET

Figure 13 repeats the experiments of Figure 4 on *LeBel*. Adding audio to a text-only target improves NOEMA's Top-10@2000 accuracy by 8.52 percentage points, and adding speech improves it by a further 12.06 points, exceeding the Context MLP's gains by 2.01 and 3.90 points, respectively. The speech addition thus has a larger effect on *LeBel* than on *Cephalonauts One* (7.25 points). Extending the input window from one to seven frames improves NOEMA by 18.53 points and the Context MLP by 12.76 points; unlike on *Cephalonauts One*, the Context MLP keeps improving up to seven frames.

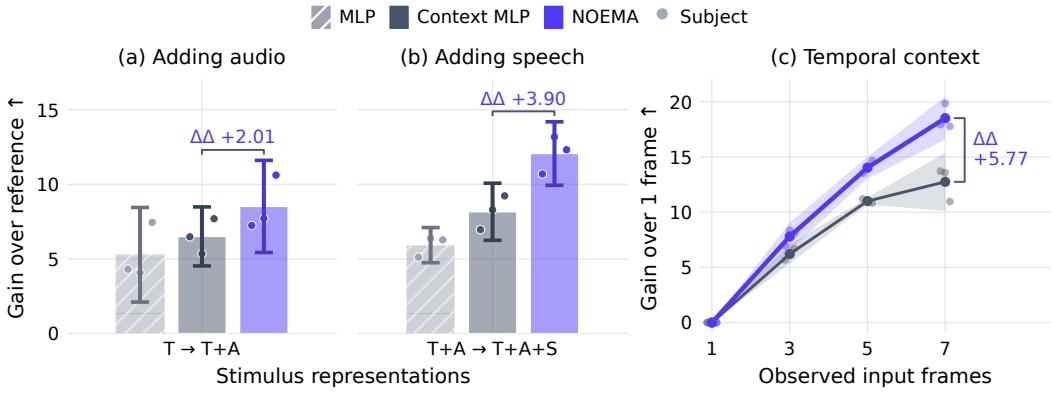


Figure 13: **Effect of stimulus representations and temporal context (*LeBel*).** Gain in Top-10@2000 accuracy (percentage points) over a reference configuration, computed within each model, subject, and fold. **(a)** Adding audio to a text-only target (T→T+A). **(b)** Adding speech to a text-and-audio target (T+A→T+A+S). **(c)** Increasing the input window from one frame to three, five, and seven frames (Context MLP and NOEMA only). T: Qwen text; A: LCO audio; S: Parakeet speech. ΔΔ is the mean NOEMA gain minus the mean Context MLP gain, at seven frames in (c). Circles show subject means over five folds; bars in (a, b) and lines in (c) show the mean over three subjects, with 90% Student-t confidence intervals.

F.2 ABSOLUTE RETRIEVAL ACCURACIES ACROSS STIMULUS REPRESENTATIONS

Figure 14 reports the absolute accuracies underlying the gains in Figures 4 and 13. NOEMA achieves the highest mean accuracy for every target combination on both datasets, and every model is most accurate with the full T+A+S target.

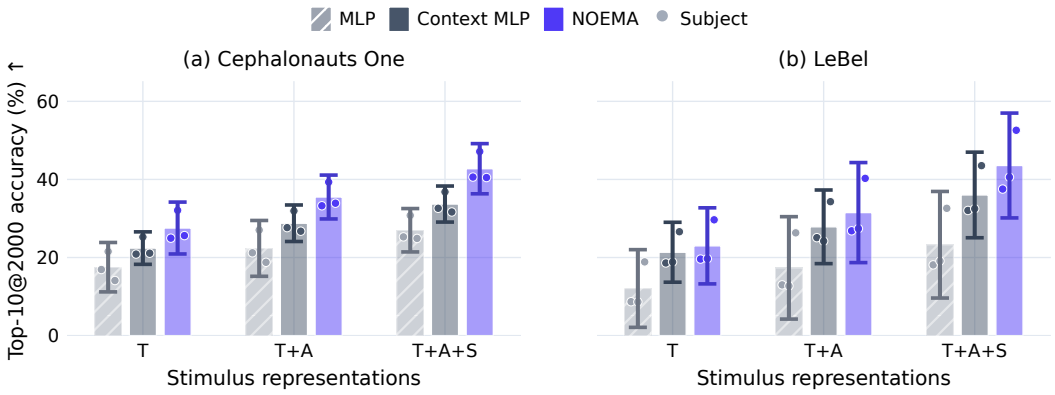


Figure 14: **Top-10@2000 accuracy across stimulus representations.** Absolute accuracies underlying panels (a) and (b) of Figures 4 and 13, for **(a)** *Cephalonauts One* and **(b)** *LeBel*. T: text; A: audio; S: speech. MLP uses one input frame; Context MLP and NOEMA use seven. Circles show subject means over five folds; bars and error bars show the mean and 90% Student-t confidence interval over three subjects.

F.3 Absolute retrieval accuracies with increasing temporal context

Figure 15 reports the absolute accuracies underlying the temporal-context gains. On average, NOEMA is more accurate than the Context MLP at every window length on both datasets, and the gap widens with more frames: from 3.3 to 9.1 percentage points on *Cephalonauts One* and from 1.8 to 7.6 points on *LeBel* between one and seven frames.

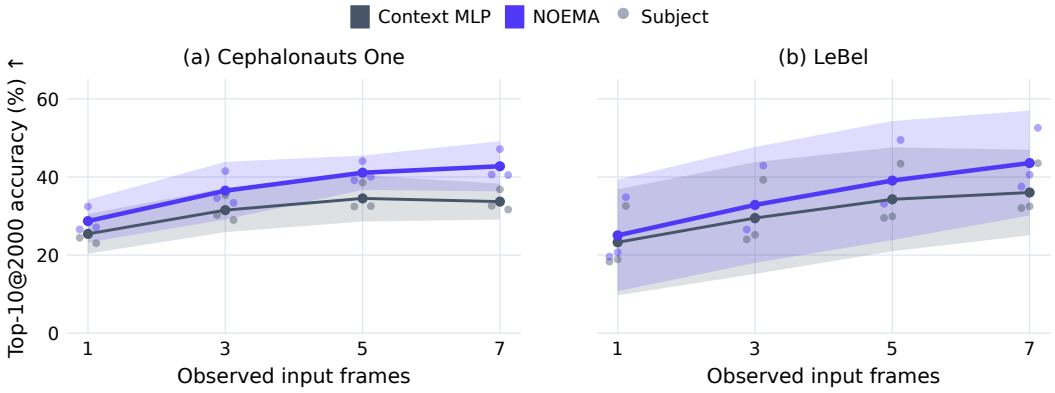


Figure 15: **Top-10@2000 accuracy with increasing temporal context.** Absolute accuracies underlying panel (c) of Figures 4 and 13, for **(a)** *Cephalonauts One* and **(b)** *LeBel*. Circles show subject means over five folds; lines and shaded bands show the mean and 90% Student-t confidence interval over three subjects. The wide bands on *LeBel* reflect differences in overall accuracy between subjects; the within-subject gains in Figure 13c are the more direct comparison.

F.4 Retrieval accuracy and model size

Figure 16 relates retrieval accuracy to the number of decoder parameters. Context MLP grows from 119.6M to 623.0M parameters as its input increases from one to seven frames. NOEMA uses 253.8M parameters independently of the input window length; its seven-frame configuration achieves higher mean accuracy than all four Context MLP configurations on both datasets. This comparison shows that NOEMA's advantage does not come from a larger parameter count.

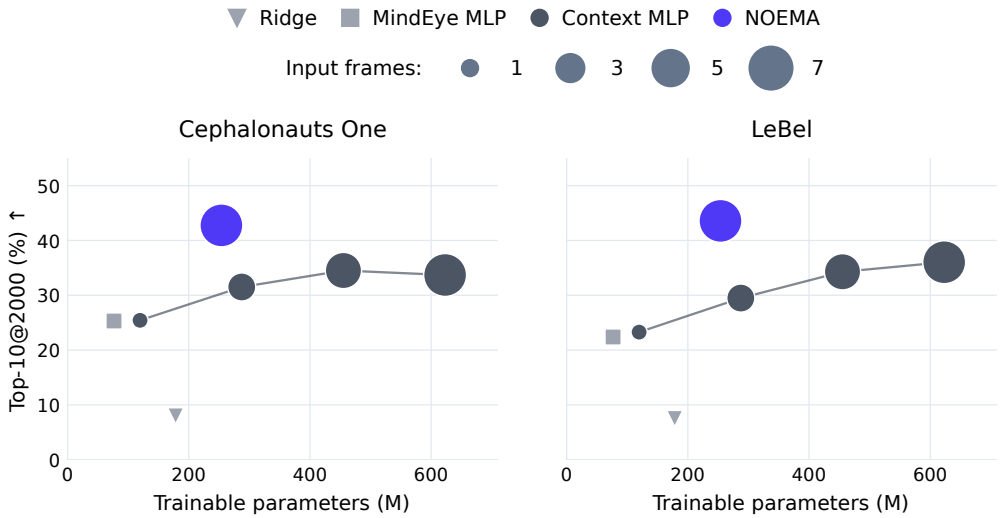


Figure 16: **Retrieval accuracy versus decoder parameter count.** Context MLP uses one, three, five, or seven input frames; NOEMA is shown with seven frames only. MindEye MLP and Ridge use one frame. Marker area is proportional to the number of input frames. Points show mean Top-10@2000 accuracy across three subjects after averaging five folds per subject. Counts include decoder weights, biases, and normalization parameters, and exclude frozen stimulus encoders. All decoders use 20,484 input vertices and 8,704 output features. Hidden widths are 4,096 for Context MLP and NOEMA, and 2,048 for MindEye MLP.

G SENSITIVITY TO FMRI FRAME ORDER

Figure 17 compares seven-frame Context MLP and NOEMA after retraining and testing with shuffled frame orders. Each input retains exactly the same frame values and prediction target; only the assignment of frames to input positions changes. Because the shuffling rule is applied during both training and evaluation, this experiment measures the benefit of the original temporal organization after adaptation, rather than robustness to a test-time perturbation.

A shared permutation probes the architectural bias toward temporal order. Context MLP is nearly unchanged ($+0.33$ percentage points): its first dense layer operates on concatenated frames and can absorb any fixed input permutation by reordering its weights. Thus, the model can learn position-specific temporal information without requiring neighboring input positions to correspond to neighboring acquisition times. NOEMA instead loses 4.26 points despite retraining. Its local, shared temporal filters impose structure on how frames interact; arbitrary permutations alter temporal neighborhoods and their relation to the central readout. This result suggests that NOEMA benefits from the chronological organization of fMRI frames, which its temporal inductive bias is designed to exploit.

Per-sample permutations probe the value of consistent temporal positions. Shuffling each sample differently breaks the temporal ordering across samples. Accuracy falls by 10.90 points for Context MLP and 18.46 for NOEMA. Both models therefore benefit from consistent temporal positions, with a larger observed dependence for NOEMA.

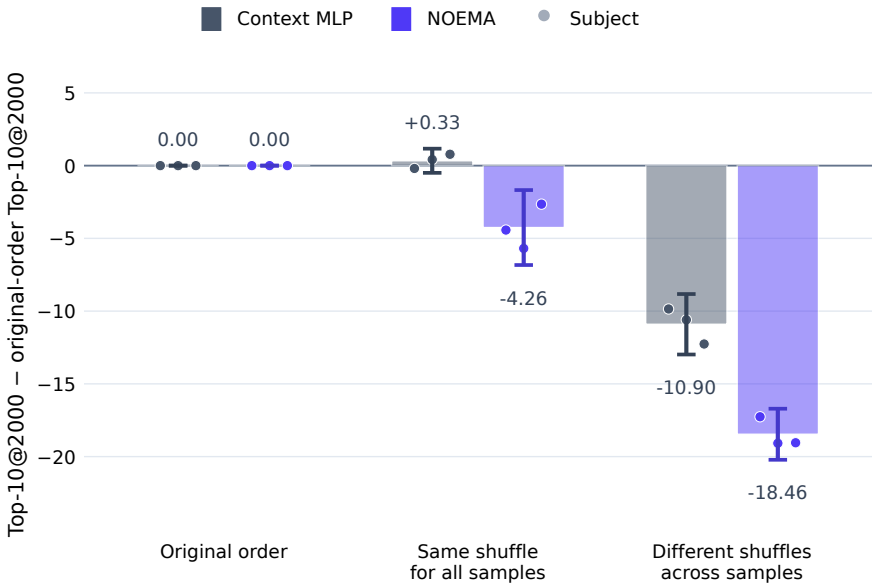


Figure 17: **Sensitivity of Context MLP and NOEMA to fMRI frame order on *Cephalonauts One*.** We randomly permute the temporal dimension of each seven-frame fMRI input. Global shuffling applies a single fixed permutation to all samples across subjects, folds, and data splits. Per-sample shuffling assigns each sample its own permutation, fixed across epochs. Context MLP and NOEMA are trained from scratch for each condition, with the same shuffling rule applied during training, validation, and testing. We report the change in Top-10@2000 accuracy relative to training and evaluating on the original frame order, paired by model, subject, and fold. Each circle represents one subject: paired differences are first averaged over available folds within each permutation seed, then equally across permutation seeds within the subject. Bars show means across three subjects, with 90% Student-t confidence intervals.

H TRAINING-DATA SCALING

We compare seven-frame Context MLP and NOEMA as the number of training sessions increases. Both models use the same nested subsets of whole sessions, with validation and test sessions held fixed within each fold. The six training budgets are 1, 2, 3, 6, 12, and 22 sessions for *Cephalonauts One*, and 1, 2, 3, 5, 8, and 13 for *LeBel*.

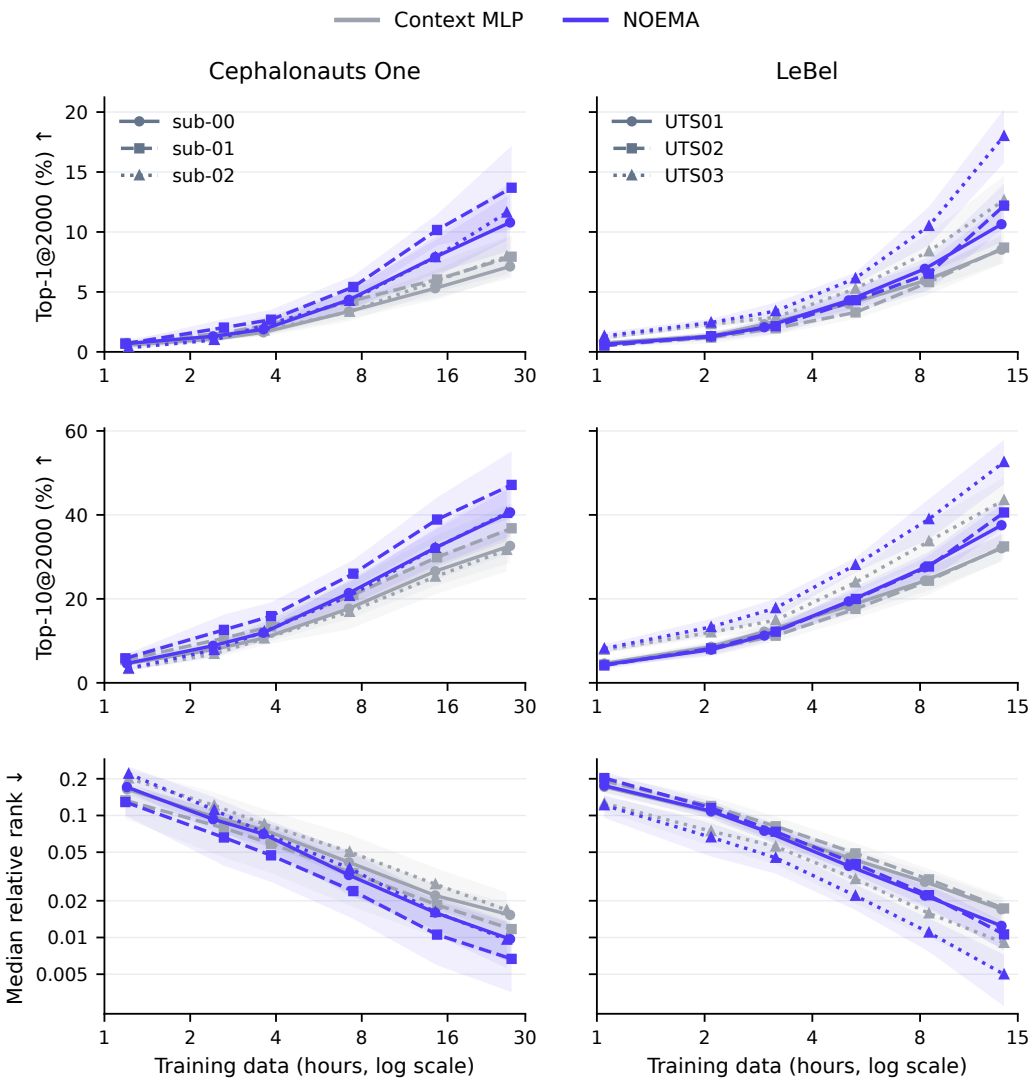


Figure 18: **Retrieval performance versus training-data duration.** Rows show Top-1@2000 accuracy, Top-10@2000 accuracy, and median relative rank (lower is better). Each curve represents one model and subject. Points average five session folds; shaded bands show ± 1 standard deviation across folds. Color identifies the model; line style and marker identify the subject. All horizontal axes and the median-relative-rank vertical axes are logarithmic.

I Extended retrieval sets

The extended retrieval sets in Section 5.3 draw on 50 additional episodes of the podcast show used in *Cephalonauts One*, none of which appear in the dataset. We transcribe these episodes with the AssemblyAI `universal-3-pro` speech-to-text model, as for the original stimuli (Collas et al., 2026), and split them into non-overlapping ten-second segments. Treating the end of each segment as a time marker t, we compute its text, audio, and speech embeddings exactly as for the original stimuli (Section 3.3), yielding 8,494 additional candidates. For each test session, we add random subsets of 500, 1,000, 2,000, and 5,000 candidates, or all 8,494, to the original retrieval set of approximately 2,000 segments. We evaluate the MLP, Context MLP, and NOEMA, trained as in Section 4.2, on each extended set without retraining.

J Feature-importance maps

For the feature-importance analysis, we compute attributions for NOEMA, the one-frame MLP, and the seven-frame MLP on the held-out fold 0 test set of *Cephalonauts One* `sub-01`. We attribute each model's CLIP loss to every input vertex in each of its fMRI frames using 50-step integrated gradients from a zero baseline as implemented in Captum (Sundararajan et al., 2017; Kokhlikyan et al., 2020). The choice of zero baseline is motivated by the fact that the input fMRI vertex time series are standardized within each run during preprocessing. We average attributions over held-out samples and for better visualization we show absolute values of the attributions and the color scale spans from the 1st to the 99th percentiles over all the values. Figure 6 in the main text shows NOEMA's lateral views, while Figure 19 adds the medial views. The MLP maps appear in Figures 20 and 21.

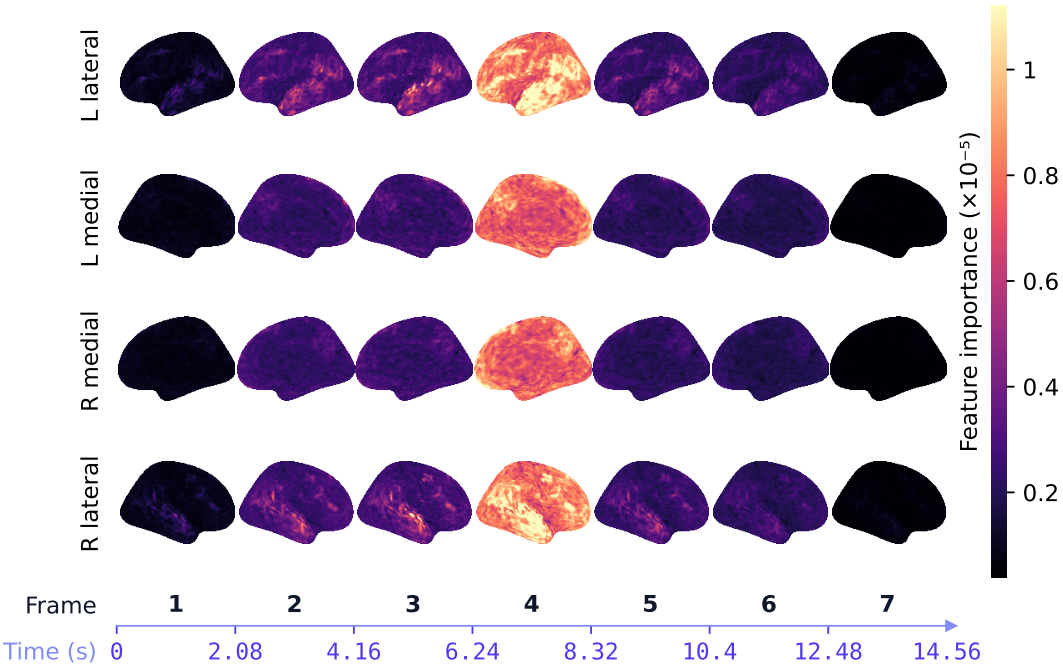


Figure 19: **NOEMA feature-importance maps.** Rows show the left lateral, left medial, right medial, and right lateral views and columns show frames 1–7.

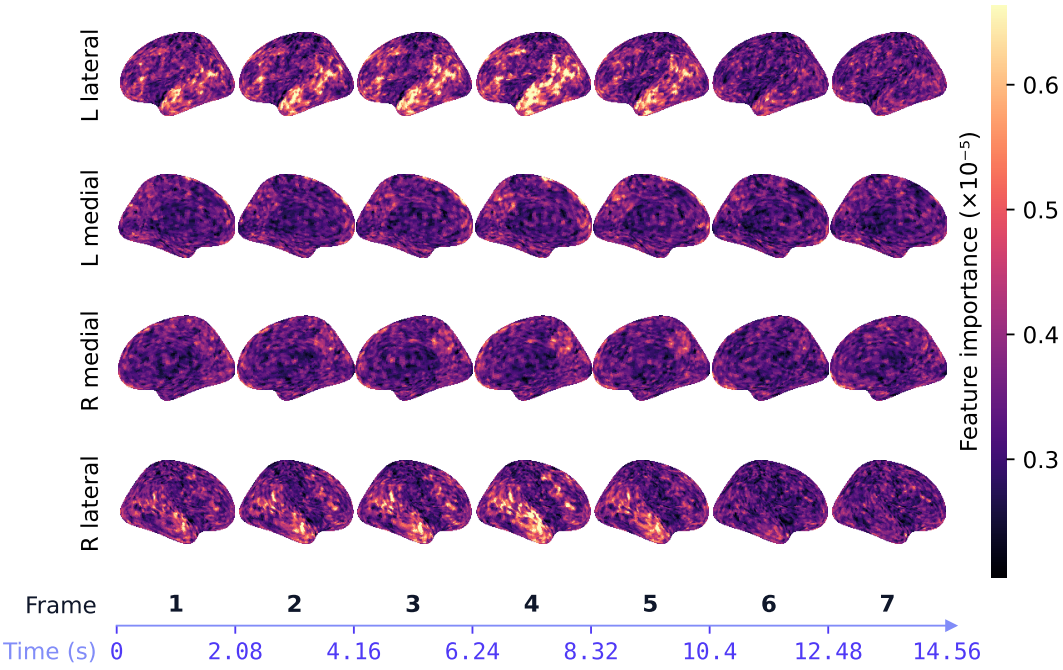


Figure 20: **Context MLP feature-importance maps.** Rows show the left lateral, left medial, right medial, and right lateral views and columns show frames 1–7.

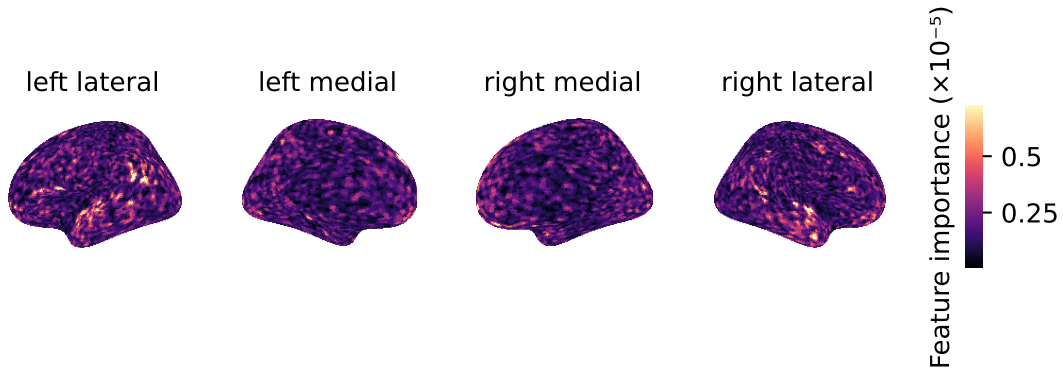


Figure 21: **One-frame MLP feature-importance maps.** From left to right: left lateral, left medial, right medial, and right lateral views.