BRAIN-ALIGNED INTERMEDIATE TRANSFORMER STATES DEPEND ON MODALITY, STIMULUS, AND POSITIONAL ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment between Transformers and brain activity has mainly been studied at the layer level, but each layer contains computational states with different properties. Previous work identified intermediate states that align more strongly with brain activity than the commonly used block output, yet the generality of this finding and the mechanisms behind the advantage of particular states remain unclear. Here, we examined brain alignment across within-block states in text, audio, music-generation, and vision Transformers. We quantified this alignment by using these states to predict brain responses to speech, music, and images. We also compared models that share the same basic architecture but use different positional-encoding schemes. We found that intermediate states generally yielded better predictions than the conventional block output, but that the strongest correspondences varied with input modality, stimulus, and positional encoding. Text and audio models favored distinct state families even for the same brain responses, and the preferred states changed between speech and music within the same audio model. In vision, positional encoding reshaped the pattern of state-specific alignment. We further examined the computations and representational content underlying these differences, providing a mechanistic interpretation of why particular states align with auditory and visual cortex. These results suggest that the correspondence between brain responses and states is more diverse than previously assumed, and that this diversity is rooted in the computation and the representational content of each state.

1 INTRODUCTION

Predicting brain responses to natural stimuli with encoding models provides a way to compare computational representations with information processing in the brain (Kay et al., 2008; Naselaris et al., 2011; Huth et al., 2016). With deep neural networks, this comparison became hierarchical: successive layers of deep neural networks map onto successive stages of the cortical hierarchy (Yamins et al., 2014; Güçlü & van Gerven, 2015; Kell et al., 2018). Transformers are now the dominant models in these comparisons, and their layers likewise predict cortical responses to speech, music, images, and language (Schrimpf et al., 2021; Caucheteux & King, 2022; Millet et al., 2022; Li et al., 2023; Denk et al., 2026; Conwell et al., 2024; Goldstein et al., 2025).

Yet a Transformer layer is not a single computational stage: attention, feed-forward transformations, and residual updates produce functionally distinct intermediate states whose relationship to brain responses has been less systematically examined. Recent studies have begun to reveal substantial variation at this finer scale (Kumar et al., 2024; Chen & Sivakumar, 2026). In particular, Chen & Sivakumar (2026) showed that states within the same block can differ markedly in their brain-prediction performance, with intermediate states often outperforming the block output (Figure 1a). These findings suggest that comparing layers alone can overlook important differences in the computations that support brain alignment.

These findings raise two questions: whether the patterns of alignment between intermediate states and brain activity observed in text-input language models generalize to other modalities, and which properties of intermediate representations account for differences in brain alignment. Chen &

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.

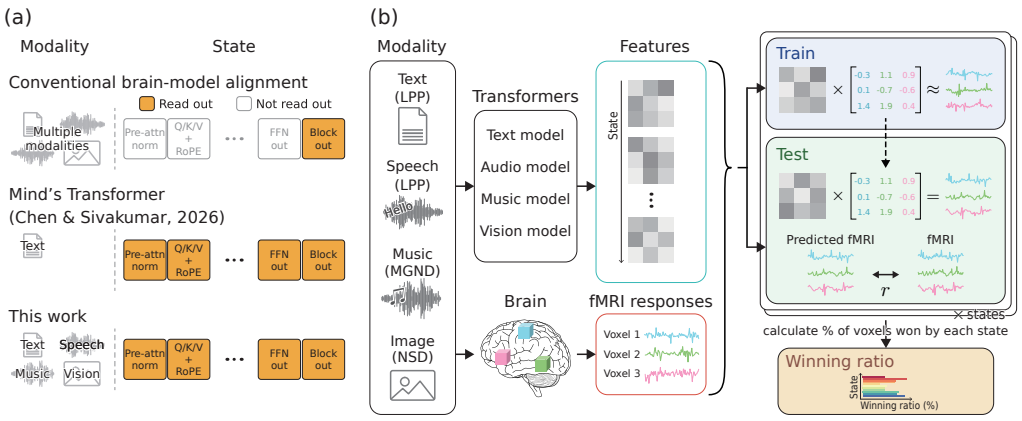

Figure 1: **Overview of the analyses.** (a) Previous studies mainly compared layer outputs, whereas recent studies (Chen & Sivakumar, 2026) compared within-layer states in a particular modality. We extend these comparisons to speech, music, and images. (b) voxel-wise encoding models use features from each model, layer, and state to predict held-out brain responses.

Sivakumar (2026) compared queries with and without rotary transformation and reported that post-rotation queries provided the most accurate predictions for many voxels in early auditory cortex, suggesting a role for positional encoding. However, it remains unclear whether this advantage depends on the modality of the representations to which positional encoding is applied.

To address these questions, we compare intermediate states in text, audio, music-generation, and vision Transformers using functional magnetic resonance imaging (fMRI) responses to speech, music, and natural images (Figure 1b). We examine differences associated with input modality by predicting the same responses to narrated speech from text and waveforms, and differences associated with stimulus conditions by applying the same audio models to speech and music. In both audio and vision, we also compare models with and without rotary positional embeddings (RoPE) that share the same basic architecture. Finally, we examine what information each state represents and how this depends on positional encoding and other within-block computations, providing a mechanistic interpretation of why particular states align with different brain regions. Our contributions are as follows:

- We evaluated intermediate states from text, audio, music-generation, and vision models using a common voxel-wise encoding framework, identifying the best-predicting state for each layer and voxel and quantifying how often each state was selected.
- The best-predicting states depended on modality (text vs. speech for the same responses), on stimulus (speech vs. music within the same audio models), and on positional encoding (RoPE vs. absolute positional embedding in vision).
- We analyzed internal representations to interpret state-specific brain alignment, characterizing the acoustic and linguistic information available in audio states and comparing how accurately states in vision models with and without RoPE recovered independently estimated population receptive field (pRF) centers.

2 Related Work

Encoding models relate features of natural stimuli to brain responses (Kay et al., 2008; Naselaris et al., 2011; Huth et al., 2016), and deep neural networks have served as feature spaces in vision (Yamins et al., 2014; Güçlü & van Gerven, 2015), audition (Kell et al., 2018; Giordano et al., 2023), and language (Jain & Huth, 2018; Toneva & Wehbe, 2019). A recurring result is a correspondence between network depth and cortical hierarchy, with early layers predicting early sensory areas and later layers predicting higher regions (Güçlü & van Gerven, 2015; Kell et al., 2018; Jain & Huth,

2018). Transformers have since become the main object of such comparisons in language (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Antonello et al., 2023), speech and music (Millet et al., 2022; Li et al., 2023; Tuckute et al., 2023; Denk et al., 2026), and vision (Takagi & Nishimoto, 2023; Conwell et al., 2024), and the layerwise correspondence has been reproduced in them, with intermediate layers often aligning best and layer depth tracking the hierarchy and timing of cortical processing (Caucheteux & King, 2022; Millet et al., 2022; Goldstein et al., 2025). Cross-modal comparisons further reveal differences in the brain-relevant information captured by text-derived and sensory representations (Nakagi et al., 2024; Oota et al., 2024; Doerig et al., 2025). Together, these findings motivate examining how the computations that generate model representations shape their correspondence with the brain.

Beyond comparisons between models and layers, recent work has examined brain alignment at the level of individual computations within Transformer blocks. Kumar et al. (2024) related transformations computed by attention heads to cortical responses and showed that linguistic specialization across heads is reflected in their brain-prediction profiles. Mind's Transformer (Chen & Sivakumar, 2026) systematically compared intermediate states within language-model blocks and identified substantial differences in brain alignment. These studies establish the relevance of within-block computations, but leave open how their findings extend beyond text-input language models. We extend this computational perspective across modalities, and provide mechanistic interpretations of why different states align with different brain regions.

3 METHODS

3.1 DATASETS AND BRAIN REGIONS

We used three fMRI datasets with narrated speech, music, and natural images. Details of the datasets and ROI definitions are described in Appendix A.1.

For speech, we used the English subset of the Le Petit Prince (LPP) corpus (Li et al., 2022), a publicly available dataset of naturalistic story listening. The English subset contains fMRI responses from 49 native speakers listening to a 94-minute audiobook divided into nine runs. Following Mind's Transformer (Chen & Sivakumar, 2026), we averaged fMRI responses across participants. We then predicted this same group-level time series from the transcript and the corresponding waveform, allowing text and audio models to be evaluated against a common target.

For music, we used the Music Genre Neuroimaging Dataset (MGND) (Nakai et al., 2022), which contains fMRI responses from five participants who listened to 540 music excerpts spanning 10 genres over three days. Each excerpt lasts 15 seconds. The released dataset separates 480 training excerpts from 60 test excerpts, with the test stimuli presented four times.

For vision, we used all eight participants in the Natural Scenes Dataset (NSD) (Allen et al., 2022), which contains high-resolution 7T fMRI responses to thousands of natural images. Participants viewed images for three seconds per presentation while performing a continuous recognition task across repeated scanning sessions. We trained participant-wise encoding models. We also tested whether each model state could recover the visual-field location to which each voxel responds most strongly, using independently estimated pRF centers as a reference.

We compared early and higher regions of interest (ROIs) in auditory and visual cortex. Auditory ROIs were based on HCP-MMP (Glasser et al., 2016): early auditory cortex comprised A1 and the Belt regions, and higher auditory cortex comprised A4, A5, TA2, and the STS regions. For vision, we used HCP-MMP for broad early/higher comparisons and a probabilistic retinotopic atlas (Wang et al., 2015) for pathway-specific analyses.

3.2 MODELS AND INTERMEDIATE STATES

We used Llama 3.2 1B Instruct (hereafter Llama) from the Llama 3 family (Grattafiori et al., 2024) for text; wav2vec 2.0 (hereafter wav2vec2) (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) for audio; and MusicGen (Copet et al., 2023) for music. Vision models comprised DeiT III (Touvron et al., 2022), ViT (Dosovitskiy et al., 2021), and Swin (Liu et al., 2021). For the positional-encoding comparisons, we used RoPE and non-RoPE variants of DeiT III (a ViT-family model) and Swin (Su

et al., 2024; Heo et al., 2024), together with two audio models based on the Conformer architecture (Gulati et al., 2020): wav2vec2-RelPos and wav2vec2-RoPE. Each pair comprised separately pretrained models with the same basic architecture and training data. All four audio models were evaluated on both speech and music. Details are provided in Appendix A.2.

We distinguished Q/K/V, the attention context and attention output, intermediate FFN activations after the nonlinearity and the final FFN output, and residual-stream states before and after each branch. Where available, RoPE Q/K were extracted before and after rotation. Conformer states additionally covered the pre-attention FFN1 and convolution branch. Speech and music features were aggregated over 10-second windows. For vision, we averaged patch representations within each state (patch-mean) to keep the readout consistent. State definitions are given in Appendix A.3.

3.3 Encoding models and evaluation

We used ridge regression to predict voxel-wise fMRI responses from each model, layer, and state (Figure 1b). Each feature dimension was standardized using the training-set mean and standard deviation, then applied to the evaluation data. We used `RidgeCV` from `himalaya` (Dupré la Tour et al., 2022) to select regularization strength separately for each voxel by five-fold cross-validation within the training data, over 25 logarithmically spaced candidates from 10^{-12} to 10^{12}. Prediction accuracy was the Pearson correlation between predicted and observed responses. LPP correlations were computed separately for each held-out run in leave-one-run-out cross-validation and then averaged; MGND and NSD used independent evaluation sets. Details are described in Appendix A.1.

3.4 Measures of state-specific brain alignment

Following Mind's Transformer (Chen & Sivakumar, 2026), we adopted two measures of state-specific brain alignment: *winning ratio*, which quantifies how often each state provides the best prediction, and *state depth*, which characterizes where the best-predicting states occur in the within-block computational sequence.

For winning ratio, we selected the state with the highest prediction accuracy for each layer and voxel and calculated the percentage of selections for each state. Let $r_{m\ell kv}$ denote the prediction accuracy for model m, layer ℓ, state k, and voxel v, and let $\mathcal{S}_m$ be the model's candidate set. With $k^*_{m\ell v} = \arg\max_{k\in\mathcal{S}_m} r_{m\ell kv}$, the winning ratio of state k in ROI R is

$$W_{mkR} = \frac{100}{L_m|R|} \sum_{\ell=1}^{L_m} \sum_{v\in R} \mathbf{1}[k^*_{m\ell v} = k],$$

where L_m is the number of layers and $|R|$ is the number of ROI voxels. For correlations between Transformer and Conformer profiles, we restricted both candidate sets to 11 matched states (Appendix A.4).

For state depth, we ordered a subset of states by computational sequence and assigned equally spaced values from 0 to 1. We used seven stages from Q through attention and the FFN to the block output for Transformer models, and 17 stages including FFN1 and the convolution branch for Conformer-based models. For each layer and voxel, we selected the depth of the best-predicting candidate and averaged these depths across layers and ROI voxels. We used this measure to examine how the preferred computational stage changes from A1 to STS in auditory cortex. Candidate sequences and ROI ordering are described in Appendix A.5.

3.5 Mechanistic analyses of state-specific advantages

3.5.1 Linear probing of audio states

To characterize the information underlying state-specific brain alignment, we performed linear probing using the English narration from LPP. We fed the speech waveform to wav2vec2, wav2vec2-RelPos, and wav2vec2-RoPE and used the representations extracted from each layer and state as inputs to linear probes. The prediction targets were acoustic and linguistic properties of the same speech rather than brain responses: Mel spectrograms and Gabor filterbank features, phoneme identities, syntactic annotations (part-of-speech tags, dependency relations, and function-word labels),

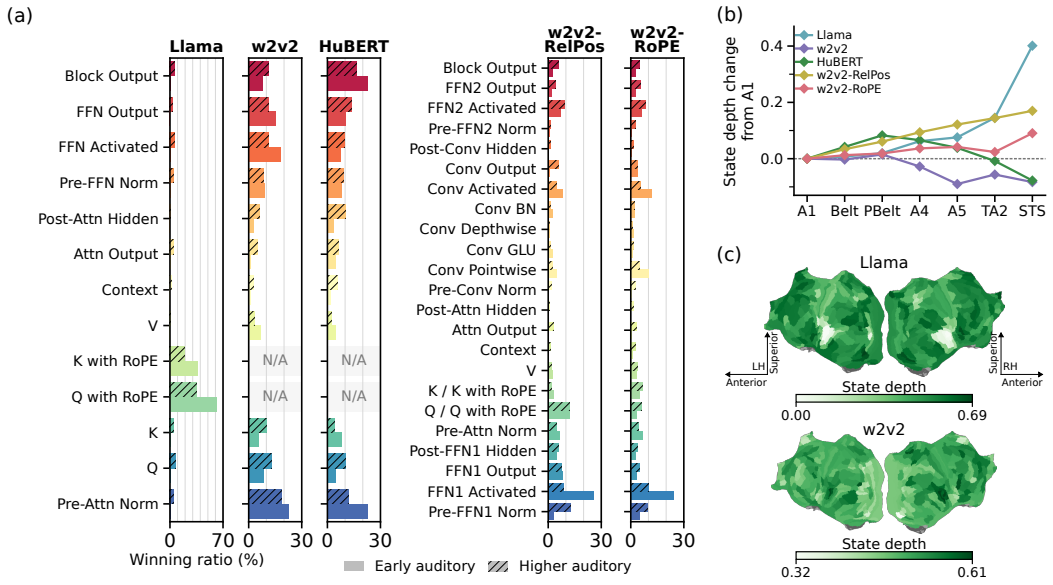

Figure 2: **Speech: input modality changes the type of best-predicting state.** Analyses use mean LPP responses from 49 participants. (a) Winning ratios for Llama and four audio models predicting responses to the same narration. Solid and hatched bars show early and higher auditory cortex, respectively. (b) Changes in state depth relative to A1, indicating whether later or earlier within-block computations are preferred along auditory cortex. (c) Cortical distributions of state depth.

and semantic features constructed from answers to a fixed set of questions about individual words or their surrounding context. These targets were adapted from the feature families in Shimizu et al. (2025). Probe inputs were aligned to the frames, phonemes, or words of each target. We fitted ridge regression for continuous targets and logistic regression for categorical targets, evaluating predictions on held-out sections using Pearson correlation and chance-corrected accuracy, respectively. Target construction, temporal alignment, and evaluation details are provided in Appendix A.6.

3.5.2 Positional encoding and visual-field position

To examine how positional encoding affects brain alignment, we compared how well corresponding states in DeiT III RoPE Mixed and DeiT III with absolute positional embedding (APE) predicted NSD fMRI responses. We examined Q/K, where RoPE rotates features according to patch position, as well as states before and after this operation.

We then tested whether states that better predict brain responses also recover each voxel's preferred visual-field location more accurately, using independently estimated pRF centers as a reference (Dumoulin & Wandell, 2008; Allen et al., 2022). For each state and voxel, the linear encoding model allowed us to decompose the prediction from patch-mean features into contributions from individual image patches. To identify the image regions whose contributions tracked the voxel's responses, we calculated the covariance between each patch's contribution and the observed responses across images, normalized to form a spatial contribution map. After averaging the maps across layers and retaining high-contribution patches, we used their contribution-weighted center as the predicted pRF center. We compared the predicted and independently estimated centers using coordinate correlations and center-to-center distances (Appendices A.7 and A.8).

4 Results

4.1 Speech: input format changes the type of best-predicting state

To test whether state-specific alignment differs with input modality, we predicted the same LPP responses to narrated speech from Llama receiving the transcript and audio models receiving the

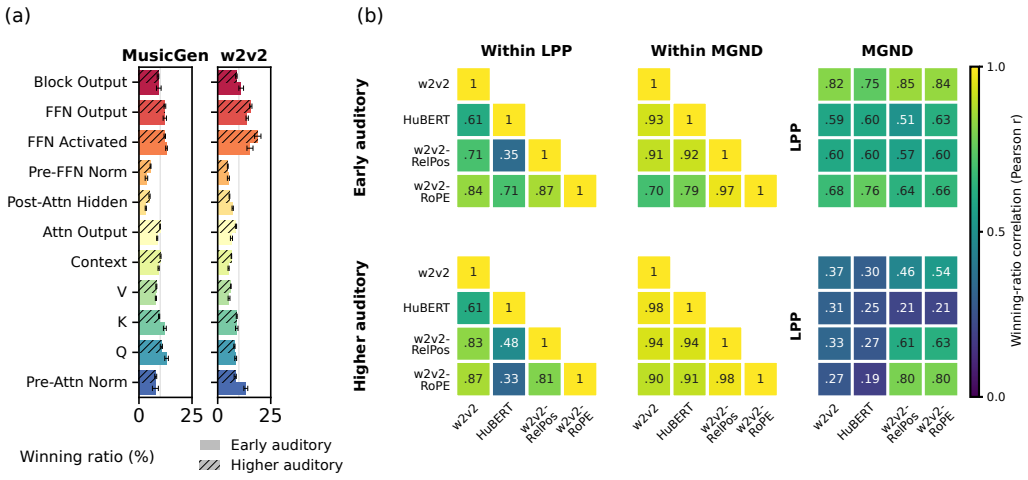

Figure 3: **Music: the preferred state changes within the same audio model.** (a) Winning ratios for MusicGen and wav2vec2 (w2v2) in early and higher auditory cortex (mean ± SEM over five participants). Solid and hatched bars show early and higher auditory cortex, respectively. Results for the other models are shown in Figure A3. (b) Pearson correlation between state-wise winning-ratio profiles. Rows show early and higher auditory cortex; columns show within-LPP, within-MGND, and cross-dataset comparisons.

waveform (Figure 2a). Consistent with the post-rotation Q/K advantage reported in Mind's Transformer (Chen & Sivakumar, 2026), Llama's post-rotation Q/K dominated in early auditory cortex, together accounting for a winning ratio of 98.5%. By contrast, across all audio models, the leading states belonged to the residual stream or FFN branches rather than the post-rotation Q/K favored by Llama.

To examine the influence of positional encoding on state-specific alignment, we compared wav2vec2-RelPos and wav2vec2-RoPE, which share the same basic architecture and training dataset but use relative positional encoding and RoPE, respectively. Intermediate FFN1 activations were selected most often in early auditory cortex in both models, with a weaker advantage in higher auditory cortex. The RoPE variant did not shift the preference toward post-rotation Q/K. Thus, the association between positional encoding and high brain alignment observed in the text modality did not extend to these audio models.

To examine how the preferred computational stage changes across auditory cortex, we used state depth, a measure of how early or late within a block the best-predicting state occurs (Section 3.4; Figure 2b,c). In Llama, the preferred states became monotonically deeper from A1 to STS. Audio models did not show a consistent direction: wav2vec2-RelPos and wav2vec2-RoPE shifted toward later computations, whereas wav2vec2 and HuBERT shifted toward earlier ones. The correspondence between within-block computational order and auditory cortical organization was therefore model-dependent rather than shared across text and audio models. Appendix B provides additional state-depth results for speech and music, together with cortical maps for all five models evaluated on speech.

4.2 Music: stimulus conditions change the best state within the same model

To test whether the state-specific alignment observed for speech generalizes to music, we applied the same four audio models to MGND music stimuli and additionally evaluated MusicGen (Figure 3a). Winning ratios for HuBERT, wav2vec2-RelPos, and wav2vec2-RoPE, together with participant-wise wav2vec2 results, are provided in Appendix C. Residual-stream and FFN-related states were frequently selected, but the preferred states changed with the stimulus. In both early and higher auditory cortex, intermediate FFN activations were selected most often for wav2vec2 and HuBERT, replacing the pre-attention residual stream and block output favored for speech. In the relative-position and RoPE variants, FFN1-related states remained among the leading predictors of early

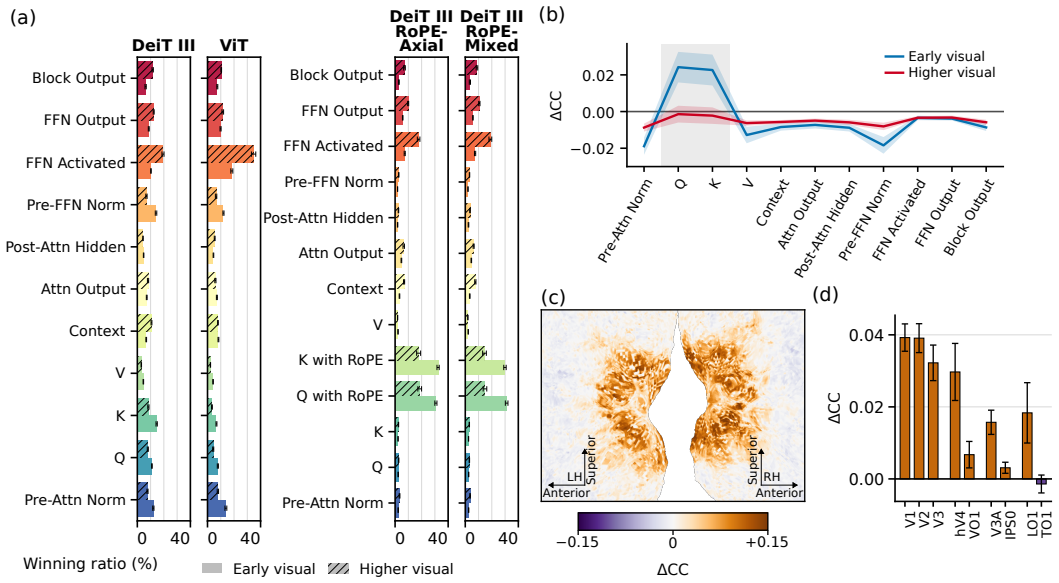

Figure 4: **Images: positional encoding changes state-specific alignment with early visual cortex.** (a) Winning ratios for the DeiT variants and ViT in early (solid bars) and higher (hatched bars) visual cortex (mean ± SEM over eight participants). (b) Differences in prediction accuracy (ΔCC; DeiT III RoPE Mixed minus DeiT III), averaged across all 24 layers. Lines and bands show mean participant-matched differences and pointwise 95% confidence intervals. Shading marks Q/K, represented by post-rotation states in RoPE. (c) Cortical distribution of the Q difference in subject01, with enlarged views of the regions around visual cortex. (d) The same Q difference across retinotopic ROIs (Wang et al., 2015): V1–V3, followed by higher visual ROIs from the ventral (hV4, VO1), dorsal (V3A, IPS0), and lateral (LO1, TO1) pathways. Values show mean ± SEM over eight participants.

auditory responses. MusicGen likewise favored FFN-related states and Q, indicating that specialization for music generation did not make the block output the preferred readout.

To quantitatively compare the overall patterns of state-specific brain alignment, we computed correlations between state-wise winning-ratio profiles across models and datasets (Figure 3b). Within datasets, correlations between different models were often 0.8 or higher, with many exceeding 0.9 in MGND. By contrast, correlations across speech and music were below 0.8 for most same-model comparisons, represented by the diagonal entries in the cross-dataset matrices. Cross-dataset correlations between different models were particularly low in higher auditory cortex, where most off-diagonal entries were below 0.4. Thus, different models often showed similar alignment profiles within the same dataset, whereas substantial differences emerged between speech and music even when the model was held fixed.

4.3 Images: positional encoding affects Q/K alignment with early visual cortex

To examine whether positional encoding shapes state-specific alignment in vision, we predicted NSD responses using vision Transformers. Figure 4a shows the DeiT variants and ViT. In early visual cortex, both RoPE-based DeiT variants favored post-rotation Q/K, and this preference was reproduced in the RoPE-based Swin models (Appendix D). Models without RoPE favored Q/K or intermediate FFN activations. Higher visual cortex, by contrast, generally favored intermediate FFN activations, with DeiT III RoPE Axial the exception. Across the vision models, the block output was never the most frequently selected state in either ROI. Participant-wise winning ratios for DeiT III are provided in Appendix D.2.

To identify where the positional-encoding differences emerged within a block, we compared corresponding states in DeiT III RoPE Mixed and DeiT III (Figure 4b). In early visual cortex, the RoPE advantage was concentrated in post-rotation Q/K: these states predicted responses better than their

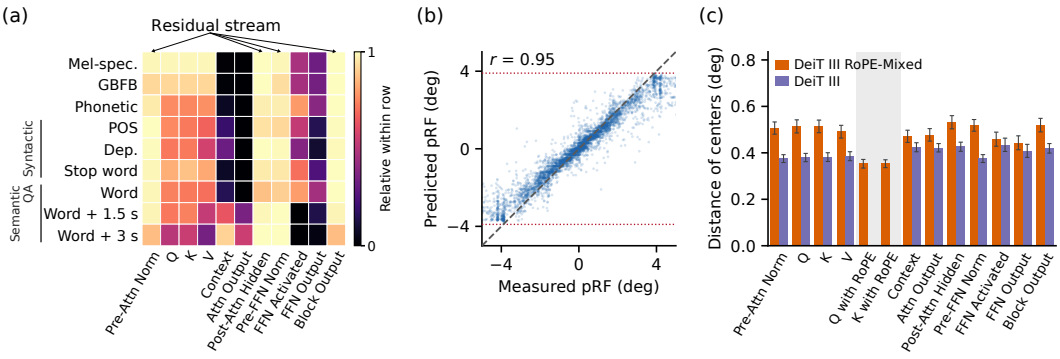

Figure 5: **Representational analyses of state-specific advantages.** (a) Linear probing of wav2vec2 (w2v2). Rows show targets and columns show within-block states. Chance-corrected classification accuracy or Pearson correlation is summarized by the median across available layers and min–max normalized within each row, indicating relative readout performance across states for a given target. (b) Predicted versus independently estimated pRF centers for post-rotation Q in DeiT III RoPE Mixed (subject01, V1–V3, averaged over 24 layers). Horizontal and vertical coordinates are pooled for the scatter and correlation. Red dotted lines indicate the range represented by the patch grid (± 3.9 deg). (c) Center-to-center distances by state (mean $\pm$ SEM over eight participants). Hatching marks post-rotation Q/K, which have no counterpart.

APE counterparts, whereas states upstream and downstream of the rotation showed no comparable improvement. This advantage was observed in all four six-layer groups and was largest in L13–18 (Appendix E).

The Q difference was also spatially selective. On the cortical surface of subject01, positive differences clustered near early visual cortex (Figure 4c). Corresponding cortical maps for all eight participants are provided in Appendix E.2. Across eight participants, ROI-wise differences decreased from V1 to V3 and declined further along the ventral, dorsal, and lateral pathways, approaching zero in their later regions (Figure 4d). The advantage associated with RoPE was therefore localized both computationally, to Q/K, and cortically, to early visual regions.

4.4 Representational information underlying state-specific advantages

Information represented in audio states. Linear probing suggested that state-specific brain-encoding advantages reflect both the breadth and the type of information represented in audio states (Figure 5a). In wav2vec2, residual-stream states ranked highest across all probing targets, suggesting that their broad advantage in brain encoding reflects the diversity of information they represent. Intermediate FFN activations and attention context, by contrast, showed opposing patterns across linguistic targets: intermediate FFN activations ranked relatively highly for phoneme identity and word-level semantics but poorly for semantics with temporal context, whereas context improved in relative rank as the semantic context window increased. This contrast was consistent with their winning-ratio profiles in Figure 2a: context was selected more often in higher than in early auditory cortex, whereas intermediate FFN activations showed the reverse pattern. In wav2vec2-RelPos and wav2vec2-RoPE, intermediate FFN1 activations ranked highest for low-level acoustic features (Appendix F), consistent with their high winning ratios in early auditory cortex.

Visual-field positions read out from vision states. To test whether the advantage of post-rotation Q/K is related to more accurate spatial readout, we compared pRF centers derived from patch contributions with independently estimated pRF centers. For post-rotation Q in DeiT III RoPE Mixed, the predicted centers closely matched the independent estimates ($r = 0.95$ in subject01; Figure 5b). This correspondence was consistent across participants, with correlations exceeding 0.90 in all eight participants (Appendix G). Within this model, post-rotation Q/K yielded the smallest pRF-center errors, whereas errors increased again at the subsequent attention-context state (Figure 5c). This selective improvement in spatial readout coincided with the Q/K advantage in predicting early visual responses (Figure 4b), supporting a mechanistic link between positional transformation and state-

specific brain alignment. By contrast, DeiT III (APE) showed a more uniform pattern of pRF-center error across states, with less pronounced specialization at particular computational stages than DeiT III RoPE Mixed (Figure 5c).

5 DISCUSSION

We investigated how brain alignment is organized within Transformer blocks across text, audio, music-generation, and vision models. By comparing intermediate states and positional-encoding variants, we found that the strongest correspondences depend on input modality, stimulus, and positional encoding. Analyses of representational content further provided a mechanistic interpretation of these differences, relating the advantages of particular states to the information available for predicting brain responses. The resulting picture is not one of a single state that universally replaces the block output, but of state-specific correspondences grounded in how models transform their inputs.

The comparisons revealed distinct patterns across modalities. For speech, Llama favored post-rotation Q/K, consistent with Mind's Transformer (Chen & Sivakumar, 2026), whereas audio models favored residual-stream or FFN-related states even when they used RoPE. For music, wav2vec2 and HuBERT instead favored intermediate FFN activations, showing that the preferred state can change with the stimulus even within the same model. In vision, RoPE-based models favored post-rotation Q/K in early visual cortex, a pattern observed in DeiT and reproduced in Swin, while higher visual cortex generally favored intermediate FFN activations. These comparisons extend the previous finding by showing that the same preference for post-rotation Q/K is absent in the audio models tested here but appears in early visual cortex for RoPE-based vision models. The model-dependent state-depth trends further indicate that within-block computational order alone does not determine correspondence with the auditory cortical hierarchy.

The representational analyses suggest why these patterns arise. In audio models, residual-stream states supported a broad range of acoustic and linguistic readouts, whereas intermediate FFN activations and attention context exhibited distinct information profiles. These differences provide a representational basis for understanding why particular computational stages predict auditory responses well. In vision, APE adds positional information to the input, whereas RoPE rotates Q/K according to patch position. This allows a linear readout to weight features differently at different image locations, even after patch averaging (Appendix A.7). In DeiT III RoPE Mixed, post-rotation Q/K also yielded the most accurate pRF-center estimates, suggesting that this spatial readout mechanism contributes to their advantage in predicting early visual responses. This account also relates to the higher configural-shape scores reported for RoPE-based ViT variants on the Object-Anagram benchmark (Doshi et al., 2025). Our analysis complements work on spatially explicit brain–model mappings, attention-based feature routing, and image-level attribution (Yang et al., 2024; Adeli et al., 2025; Luo et al., 2025) by examining how positional transformations within pretrained Transformer blocks affect spatial readout.

These results have practical implications for encoding-model studies. Feature choices should be evaluated for each modality and stimulus, rather than assuming that preferences observed in language models generalize, and intermediate states deserve consideration alongside block outputs. Because models with different representations can achieve similar brain predictivity (Conwell et al., 2024), state-wise comparisons can complement overall prediction scores by separating differences that extend across a block from those localized to particular computations, as illustrated by the post-rotation Q/K advantage in vision. Such profiles provide a basis for evaluating computational design choices that comparisons of block outputs alone may overlook. More broadly, these findings motivate a computationally explicit approach to brain–AI alignment, in which the unit of comparison is not merely a layer but a transformation and the information it makes accessible. Extending such analyses across architectures, tasks, and readout schemes could distinguish general principles of brain-relevant computation from model-specific effects, and guide the development of models that better capture the organization of sensory processing. Limitations and directions for future work are discussed in Appendix H.

AI USE STATEMENT

Generative AI tools were used to assist with coding and manuscript preparation, including English translation and editing. The authors take responsibility for the final code, analyses, and text.

ETHICS STATEMENT

This study used only publicly available human fMRI datasets: LPP, MGND, and NSD (Li et al., 2022; Nakai et al., 2022; Allen et al., 2022). Ethical approval for data collection was obtained by the respective original studies.

REPRODUCIBILITY STATEMENT

All datasets used in this study are publicly available, and the analysis code is included in the supplementary materials. The code will be released on GitHub upon acceptance.

REFERENCES

Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses, 2025.

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, 2022.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Vinamra Benara, Chandan Singh, John X. Morris, Richard J. Antonello, Ion Stoica, Alexander G. Huth, and Jianfeng Gao. Crafting Interpretable Embeddings for Language Neuroscience by Asking LLMs Questions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124137–124162, 2024.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134, 2022.

Cheng-Yeh Chen and Raghupathy Sivakumar. The mind's transformer: Computational neuroanatomy of LLM-Brain alignment. In *International Conference on Learning Representations*, 2026.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. Text-to-music generation models capture musical semantic representations in the human brain. *Nature Communications*, 17:91, 2026.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7:1220–1234, 2025.

Fenil R. Doshi, Thomas Fel, Talia Konkle, and George A. Alvarez. Visual anagrams reveal hidden differences in holistic shape processing across vision models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, 2008.

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.

Bruno L. Giordano, Michele Esposito, Giancarlo Valente, and Elia Formisano. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26:664–672, 2023.

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, Stephen M. Smith, and David C. Van Essen. 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, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369–380, 2022.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, Omer Levy, Roi Reichart, and Uri Hasson. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16:10529, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*, pp. 5036–5040, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

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: 453–458, 2016.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-Light: A benchmark for ASR with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008.

Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018.

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1):5523, 2024.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1):530, 2022.

Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H. Carney, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26:2213–2225, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Andrew F. Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M. Henderson, Leila Wehbe, and Michael J. Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. In *International Conference on Learning Representations*, 2025.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017*, pp. 498–502, 2017.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi King. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20313–20338, 2024.

Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. Music genre neuroimaging dataset. *Data in Brief*, 40:107675, 2022.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011.

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, 2024.

Qwen Team. Qwen3.8-27B: Model card. Hugging Face, 2026. Accessed September 25, 2026.

Marc René Schädler, Bernd T. Meyer, and Birger Kollmeier. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131(5):4134–4151, 2012.

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.

Riki Shimizu, Richard J. Antonello, Chandan Singh, and Nima Mesgarani. Interpretable embeddings of speech explain and enhance the brain encoding performance of audio models. arXiv:2507.16080v2, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

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

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision*, pp. 516–533, 2022.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 2023.

Liang Wang, Ryan E. B. Mruczek, Michael J. Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23030–23040, 2024.

A Supplementary Methods

A.1 Datasets, preprocessing, and brain regions

A.1.1 Narrated speech: Le Petit Prince

We used the Le Petit Prince multilingual naturalistic fMRI corpus (LPP) (Li et al., 2022). Forty-nine participants listened to an approximately 94-minute English narration of *The Little Prince*, divided into nine runs. Data were acquired using 3 T multi-echo echo planar imaging (EPI), with a voxel size of $3.75 \times 3.75 \times 3.8$ mm and a TR of 2 seconds.

We used the released ME-ICA-preprocessed fMRI time series in MNI Colin27 space. All runs for each participant were resampled to an isotropic 4 mm grid, and voxels were extracted within a symmetric mask common to all 49 participants. Each participant's time series was linearly detrended, high-pass filtered at $1/128$ Hz, and z-scored voxel-wise within each run. Responses were then averaged across the 49 participants separately for each run and z-scored again after averaging (Chen & Sivakumar, 2026).

Following Chen & Sivakumar (2026), we constructed Llama features from the narration transcript. The audio models received the waveform of the same narration. Word onsets were taken from the dataset annotations. For words split into subword tokens, token representations were averaged to obtain a word vector. Both input formats used 10-second aggregation windows with a 2-second stride: Llama features were averaged over words whose onsets fell within each window, and audio features were averaged over frames in the same interval. Hemodynamic delay was treated as a fixed window-level lag: responses were shifted forward by 2 TRs (4 seconds) within each run, and responses at TRs overlapping each window were averaged. The first and last 10 TRs (20 seconds) of every run were excluded. These features were used to predict the mean fMRI responses of the 49 participants.

In each outer leave-one-run-out fold, one run was held out for evaluation. The remaining runs were used for feature standardization, regularization selection, and encoding-model training. Prediction accuracies were computed within each held-out run and averaged across the nine runs.

A.1.2 Music: Music Genre Neuroimaging Dataset

We used all five participants in the Music Genre Neuroimaging Dataset (MGND) (Nakai et al., 2022). Participants passively listened to 540 clips, each lasting 15 seconds and drawn from 10 GTZAN genres, while maintaining central fixation. The clips had 2-second fade-in and fade-out ramps and were normalized by their root mean square. Data were collected over three days, with six 10-minute runs per day and 40 clip presentations per run. Functional images were acquired using 3 T multiband gradient-echo EPI, with isotropic 2 mm voxels and a TR of 1.5 seconds.

We used the same preprocessed voxel-wise fMRI time series as in Denk et al. (2026), in each participant's native functional image space. Head-motion correction was performed for each run, aligning all volumes to the first EPI image of each participant. Low-frequency drift was removed using a median filter with a 240-second window, and each voxel's responses were standardized to zero mean and unit variance. Cortical voxels were identified by registering FreeSurfer-derived cortical surfaces to the functional images, and all cortical voxels in the preprocessed response data were included in the analysis.

Following Denk et al. (2026), each music stimulus was presented to the four audio models and MusicGen , and features were extracted using 10-second windows with a 1.5-second stride. Each state was averaged over time within the window, and waveforms supplied to MusicGen were resampled to 32 kHz. To account for hemodynamic delay, fMRI responses were shifted by 2 TRs (3 seconds) and averaged over the interval corresponding to each feature window.

We used the predefined split of 480 training clips, presented once across 12 runs, and 60 evaluation clips, presented four times across six runs. Responses to repeated evaluation stimuli were averaged. Encoding models were fitted independently for each participant, and the predefined training and evaluation sets were used throughout. ROI summaries were computed per participant and averaged with equal weight.

A.1.3 NATURAL IMAGES: NATURAL SCENES DATASET

We used all eight participants in the Natural Scenes Dataset (NSD) (Allen et al., 2022). Each participant viewed 9,000–10,000 color natural images selected from COCO over 30–40 sessions while performing a continuous recognition task. Images subtended 8.4 degrees of visual angle and were presented for 3 seconds, with a 1-second interstimulus interval. Data were acquired using 7 T whole-brain gradient-echo EPI, with isotropic 1.8 mm voxels and a TR of 1.6 seconds.

We used the released version 3 single-trial beta estimates (`betas_fithrf_GLMdenoise_RR`) in each participant's 1.8 mm native volume space. Cortical voxels were defined by the intersection of the FreeSurfer cortical ribbon and cerebral cortical labels. Beta estimates were z-scored within each session for each voxel and averaged across repetitions for images presented three times.

Each image was bicubically resized to 224×224 pixels without cropping and normalized using the statistics associated with the corresponding checkpoint. For DeiT and ViT, patch tokens excluding the CLS token were averaged; for Swin, all spatial tokens were averaged. These features were paired with the corresponding image-averaged fMRI beta estimates.

Encoding models were fitted independently for each participant. Training sets comprised participant-specific images presented three times: 9,000 images for subject01, subject02, subject05, and subject07; 5,620 for subject03 and subject06; and 4,930 for subject04 and subject08. The evaluation set comprised 515 shared images presented three times to all eight participants. The visual-field position analysis used pRF centers estimated independently of the natural-image encoding analysis.

A.1.4 ROI DEFINITIONS

Table 1 gives the ROI definitions used for early and higher regions in each dataset, and Table 2 lists the ordered ROI sequence used for auditory state depth. The retinotopic ROIs used in Figure 4d are specified in the figure caption. The broad early/higher visual distinction in Table 1 follows the official HCP-MMP grouping (Glasser et al., 2016): Groups 1 and 2 form early visual cortex, and Groups 3, 4, and 5 (dorsal, ventral, and MT+) form higher visual cortex.

Table 1: ROI definitions used in this study. All parcels are from HCP-MMP (Glasser et al., 2016). Each ROI group includes the listed parcels in both hemispheres.

Dataset	ROI group	Constituent HCP-MMP parcels
LPP / MGND	Early auditory cortex	A1, LBelt, MBelt, PBelt
	Higher auditory cortex	A4, A5, TA2, STSda, STSdp, STSva, STSvp
NSD	Early visual cortex	V1, V2, V3, V4
	Higher visual cortex	*dorsal*: V6, V3A, V7, IPS1, V3B, V6A; *ventral*: V8, VVC, PIT, FFC, VMV1, VMV2, VMV3; *MT+*: MT, MST, V4t, FST, LO1, LO2, LO3, V3CD, PH

Table 2: Ordered ROI sequence used as the horizontal axis for auditory state depth. The sequence is based on HCP-MMP (Glasser et al., 2016); parcels in parentheses are combined into a single point.

Modality	Axis	ROI sequence (low-level → high-level)
Auditory	Auditory hierarchy	A1 → Belt (LBelt, MBelt) → PBelt → A4 → A5 → TA2 → STS (STSda, STSdp, STSva, STSvp)

A.2 MODEL CONFIGURATIONS AND PRETRAINED WEIGHTS

All models were used as fixed pretrained feature extractors. Table 3 summarizes the models, the datasets to which they were applied, and the principal configurations relevant to state comparisons. We used Llama 3.2 1B Instruct for text; wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and the Conformer-based wav2vec2-RelPos and wav2vec2-RoPE for audio; and MusicGen

(Copet et al., 2023) in addition to those audio models for music. Vision models comprised DeiT III and its RoPE variants, ViT, and Swin and its RoPE variants (Touvron et al., 2022; Liu et al., 2021; Heo et al., 2024).

For the wav2vec 2.0 family, we used Large checkpoints from the official fairseq model zoo[1] that were pretrained on 60,000 hours of Libri-Light (Kahn et al., 2020) and had not undergone ASR fine-tuning in their released form. These comprised the standard Transformer variant (LV-60) and the Conformer `rel_pos` and `rope` variants. For HuBERT, we likewise used the released Large checkpoint from the official fairseq model zoo,[2] pretrained on 60,000 hours of Libri-Light without ASR fine-tuning. Waveforms supplied to wav2vec 2.0 models and HuBERT were converted to 16 kHz mono, and each extracted 10-second waveform segment was normalized independently to zero mean and unit variance.

Table 3: Models, datasets, and principal configurations. No. layers denotes the number of Transformer layers; for Swin, brackets give the numbers of blocks in its four stages. We used the released pretrained weights as fixed feature extractors, without further fine-tuning for the downstream analyses.

Dataset	Model	No. layers	Hidden dim.	Position mechanism
LPP	Llama 3.2 1B Instruct	16	2048	1D RoPE
LPP / MGND	wav2vec 2.0 Large (LV-60)	24	1024	Convolutional position embedding
LPP / MGND	HuBERT Large (LL-60K)	24	1024	Convolutional position embedding
LPP / MGND	wav2vec 2.0 Large Conformer-RelPos	24	1024	Relative positional attention
LPP / MGND	wav2vec 2.0 Large Conformer-RoPE	24	1024	1D RoPE
MGND	MusicGen Medium	48	1536	Sinusoidal absolute position embedding
NSD	DeiT III Large	24	1024	Learned 2D APE
NSD	DeiT III Large RoPE-Axial	24	1024	2D RoPE-Axial
NSD	DeiT III Large RoPE-Mixed	24	1024	2D RoPE-Mixed
NSD	ViT Large (ImageNet-21k)	24	1024	Learned 2D APE
NSD	Swin Transformer Base	[2,2, 18,2]	[128,256, 512,1024]	Learned relative position bias
NSD	Swin Transformer Base RoPE-Axial	[2,2, 18,2]	[128,256, 512,1024]	Window-local RoPE-Axial
NSD	Swin Transformer Base RoPE-Mixed	[2,2, 18,2]	[128,256, 512,1024]	Window-local RoPE-Mixed

Positional-encoding comparisons paired variants of the same basic architecture using separate pretrained checkpoints.

[1] `https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md`

[2] `https://github.com/facebookresearch/fairseq/blob/main/examples/hubert/README.md`

A.3 MAPPING STATES TO COMPUTATIONAL STAGES

Tables 4 and 5 map the extracted within-block states to their computational stages in Transformer and Conformer models, respectively. Input-frontend features were excluded. Conformer analyses used 23 states spanning FFN1, attention, the convolution branch, and FFN2.

Table 4: Extracted states and computational stages in Transformer models. Standard Transformers used 11 states; Llama and the vision RoPE models used 13 states, adding Q with RoPE and K with RoPE.

State	Computation point
Pre-Attn Norm	Normalized representation input to self-attention
Q / K / V	Query, key, and value projections before rotary position encoding in RoPE models
Q with RoPE / K with RoPE	Query and key after rotary position encoding
Context	Attention-weighted aggregation of V before the output projection
Attn Output	Attention output after the output projection and before residual addition
Post-Attn Hidden	Residual stream after adding Attn Output
Pre-FFN Norm	Normalized representation input to the FFN
FFN Activated	Intermediate FFN representation after the nonlinear transformation
FFN Output	FFN output after the final projection and before residual addition
Block Output	Residual stream after adding FFN Output

Table 5: Extracted states and computational stages in Conformer models.

State	Computation point
Pre-FFN1 Norm	Normalized representation input to the first FFN
FFN1 Activated	Intermediate FFN1 representation after the nonlinear transformation
FFN1 Output	FFN1 output after the final projection and before residual addition
Post-FFN1 Hidden	Residual stream after adding the scaled FFN1 Output
Pre-Attn Norm	Normalized representation input to self-attention
Q / K / V	Query, key, and value projections used by wav2vec2-RelPos; V is also used by wav2vec2-RoPE
Q with RoPE / K with RoPE	Rotary-encoded query and key used by wav2vec2-RoPE
Context	Attention-weighted aggregation of V before the output projection
Attn Output	Attention output after the output projection and before residual addition
Post-Attn Hidden	Residual stream after adding Attn Output
Pre-Conv Norm	Normalized representation input to the convolution module
Conv Pointwise	Output of the first pointwise convolution before the GLU
Conv GLU	Output of the gated linear unit
Conv Depthwise	Output of the depthwise convolution
Conv BN	Output after batch normalization
Conv Activated	Output after the SiLU activation
Conv Output	Output of the second pointwise convolution before residual addition

Continued on next page

State	Computation point
Post-Conv Hidden	Residual stream after adding Conv Output
Pre-FFN2 Norm	Normalized representation input to the second FFN
FFN2 Activated	Intermediate FFN2 representation after the nonlinear transformation
FFN2 Output	FFN2 output after the final projection and before residual addition
Block Output	Output after adding the scaled FFN2 Output and applying the final normalization

A.4 Correlations between winning-ratio profiles

For Figure 3b, we formed a vector of state-wise winning ratios for each dataset–model–ROI condition and computed Pearson correlations between the resulting profiles.

When both models had the same number of candidates, their original candidate sets were used: 11 states for standard Transformers and 23 for Conformers. This also applied to comparisons of the same model across datasets. In RoPE models, Q with RoPE and K with RoPE were matched to Q and K, respectively.

When candidate counts differed, as for a standard Transformer and a Conformer, we repeated the layerwise and voxel-wise argmax over only the 11 shared states and recalculated winning ratios before computing correlations. The standard Transformer retained its original 11 states. From the Conformer, we selected Pre-Attn Norm, Q, K, V, Context, Attn Output, Post-Attn Hidden, Pre-FFN2 Norm, FFN2 Activated, FFN2 Output, and Block Output. The additional FFN1 and convolution branches were excluded, and the final FFN2 branch was matched to the standard Transformer FFN.

A.5 Details of state depth

The Transformer candidate sequence comprised seven stages: Q $\rightarrow$ Context $\rightarrow$ Attn Output $\rightarrow$ Post-Attn Hidden $\rightarrow$ FFN Activated $\rightarrow$ FFN Output $\rightarrow$ Block Output. The Conformer sequence comprised 17 stages: FFN1 Activated $\rightarrow$ FFN1 Output $\rightarrow$ Post-FFN1 Hidden $\rightarrow$ Q $\rightarrow$ Context $\rightarrow$ Attn Output $\rightarrow$ Post-Attn Hidden $\rightarrow$ Conv Pointwise $\rightarrow$ Conv GLU $\rightarrow$ Conv Depthwise $\rightarrow$ Conv BN $\rightarrow$ Conv Activated $\rightarrow$ Conv Output $\rightarrow$ Post-Conv Hidden $\rightarrow$ FFN2 Activated $\rightarrow$ FFN2 Output $\rightarrow$ Block Output. In RoPE models, Q was replaced by Q with RoPE.

Relative to the 11 standard Transformer states in Table 4, this sequence excluded four states: Pre-Attn Norm, K, V, and Pre-FFN Norm. Relative to the 23 Conformer states in Table 5, it excluded six: Pre-FFN1 Norm, Pre-Attn Norm, K, V, Pre-Conv Norm, and Pre-FFN2 Norm. K and V are parallel branches entering attention rather than successive points along the chosen computational sequence. Pre-* Norm states normalize representations at the same computational position as the preceding residual-stream point, so the pre-normalization residual-stream point was retained. Ties in prediction accuracy were resolved in favor of the earlier candidate in the sequence.

Candidate positions were equally spaced from zero to one. At each layer and voxel, the candidate with the highest prediction accuracy determined the selected depth, which was then averaged across layers and ROI voxels. Auditory changes were expressed relative to A1, following the ROI order in Table 2.

The cortical maps in Figures 2c and A2 used the same candidate sequences. The best candidate's position was selected separately for each layer and voxel and averaged across layers to obtain voxel-wise depth. Voxel-wise depths were then averaged within each ROI.

A.6 Details of linear probing for audio states

To characterize the information represented by each state independently of brain responses, we linearly predicted stimulus features or categories from every layer and state of wav2vec 2.0 LV-60, wav2vec2-RelPos, and wav2vec2-RoPE. Each speech section was supplied to the model in 10-second windows at 2.5-second intervals. We retained states from the entire first window and, for

subsequent windows, from the previously unused trailing interval of up to 2.5 seconds. Retained states were extracted or averaged at the temporal units specified in Table 6. Thus, the model received 10 seconds of context, but the probes used target-aligned representations rather than 10-second-averaged features.

Targets comprised Mel spectrograms, GBFB features, phonemes, syntactic information (part-of-speech tags, dependency relations, and function-word labels), and Semantic QA (Shimizu et al., 2025). Their definitions, temporal alignment with states, and estimators are given in Table 6. Semantic QA used QA-Emb, with each dimension corresponding to an answer to a yes/no question (Benara et al., 2024). We used the 56 fixed questions from Shimizu et al. (2025) and generated answers with Qwen3.8-27B (Qwen Team, 2026) under three conditions: the target word alone, the target word with the preceding 1.5 seconds, and the target word with the preceding 3 seconds. Accordingly, the matrices in Figure 5a and Figure A9 show nine separate rows rather than averaging across syntactic or semantic conditions.

Table 6: Prediction targets and temporal alignment for linear probing of audio states.

Probing target	Definition	Temporal alignment of state and target	Estimator
Mel spectrogram	32-dimensional features computed from the audio waveform	State and Mel features averaged within the same 100-ms interval	Ridge regression
GBFB	455-dimensional spectrotemporal modulation features obtained by applying a two-dimensional Gabor filterbank to a log-Mel spectrogram	One time-aligned state–GBFB pair every 100 ms	Ridge regression
Phoneme	39 ARPAbet classes obtained by forced alignment	State averaged over the central third of each phoneme interval	Logistic regression
Part of speech	34 classes from the LPP annotations	State averaged over the target-word interval	Logistic regression
Dependency relation	31 classes from the LPP annotations	State averaged over the target-word interval	Logistic regression
Function word	Binary function-word label	State averaged over the target-word interval	Logistic regression
Semantic QA	Responses generated by Qwen3.8-27B for 56 fixed yes/no questions under three context conditions	State averaged over the target-word interval	Ridge regression

Phoneme boundaries were obtained by forced-aligning the word-tier transcriptions in the provided TextGrid files with the corresponding audio using Montreal Forced Aligner (McAuliffe et al., 2017). GBFB features were computed using the public reference implementation that applies a two-dimensional Gabor filterbank to a log-Mel spectrogram (Schädler et al., 2012). States and GBFB features were available at 20 ms and 10 ms intervals, respectively, so we selected one pair with matching center times every 100 ms. Because the Gabor filters themselves include up to approximately 400 ms of temporal context, no additional temporal averaging was applied, avoiding attenuation of rapid modulation components. Outer evaluation used nine-fold leave-one-section-out cross-validation; inner regularization selection used four-fold splits grouped by section.

Continuous targets used ridge regression, and categorical targets used logistic regression. Regression performance was the mean Pearson correlation across target dimensions. Classification accuracy was corrected as

$$\frac{\mathrm{acc} - \mathrm{chance}}{1 - \mathrm{chance}},$$

where chance was the baseline obtained by always predicting the most frequent class in the training set. For each state, we took the median across layers and then min–max normalized each target row.

A.7 Positional rotation and linear readout from patch-mean features

We make the action of RoPE explicit to describe how positional encoding affects readout from an averaged representation. Let q_{ip} be the pre-rotation query for image i at patch position p, and let $R(p)$ be the position-dependent rotation. RoPE gives $q'_{ip} = R(p)q_{ip}$ (Su et al., 2024; Heo et al., 2024). The matrix $R(p)$ is block diagonal, grouping feature dimensions into pairs and applying a two-dimensional rotation with a pair-specific angle. In vision models, p is a two-dimensional patch coordinate. The axial variant assigns separate pairs to the vertical and horizontal axes, whereas the mixed variant learns angles that combine both coordinates.

Standardization can be absorbed into the linear readout by defining $\widetilde{w}_v = w_v \oslash \sigma_T$, where $\oslash$ denotes elementwise division and μ_T, σ_T are the training-set statistics. With P patches, centering introduces an image-independent offset $\widetilde{b}_v = -\widetilde{w}_v^\top \mu_T$; this is not an additional fitted intercept. For post-rotation query features, the prediction can be written as

$$\widehat{y}_{iv} - \widetilde{b}_v = \widetilde{w}_v^\top \left(\frac{1}{P} \sum_p R(p) q_{ip} \right) = \frac{1}{P} \sum_p \left(R(p)^\top \widetilde{w}_v \right)^\top q_{ip}.$$

Thus, applying a shared linear readout to the mean post-rotation representation is equivalent to applying position-dependent effective weights $R(p)^\top \widetilde{w}_v$ to the pre-rotation patch representations. Because $R(p)$ is orthogonal, these effective weights correspond to rotating the readout weights in the opposite direction by the position-dependent angle. The same identity holds for keys.

Each two-dimensional rotation can be interpreted as a phase rotation in the complex plane. The post-rotation patch mean therefore sums patch content after assigning position-dependent phases, and can be viewed as a phase projection over position, with the same form as a Fourier transform over spatial position. If readout phases are selected so that contributions reinforce one another at particular positions, readout sensitivity can be concentrated within a limited region of the visual field.

Two implications follow. First, without rotation, the effective readout weights are identical across patches. Position cannot then be selected by the readout weights themselves; position-dependent prediction must instead rely on positional information already encoded in the features. APE adds positional embeddings to the input, allowing position to persist through this route. RoPE additionally provides a multiplicative coupling between position and content. Second, because each pair's phase is determined by its rotation angle, the spatial detail accessible through this mechanism depends on the set of angles used by RoPE.

This identity is a mathematical description of the readout. Whether the resulting spatial sensitivity corresponds to a voxel's visual-field position is tested by comparison with independently estimated pRF centers (Appendix A.8).

A.8 Decomposing patch contributions and predicting pRF centers

Linear encoding from a state's patch mean can be decomposed into contributions from individual patches. For image i, let P be the number of patches, $f_{ip} \in \mathbb{R}^D$ the features of patch p, and $\widetilde{f}_{ip} = (f_{ip} - \mu_T) \oslash \sigma_T$ their standardized form using training-set statistics. Let $w_v \in \mathbb{R}^D$ be the weights for voxel v from an encoding model fitted to patch-mean features. Because standardization is a patch-independent affine transformation, standardizing the patch mean is equivalent to averaging the standardized patch features. Thus,

$$\widehat{y}_{iv} = w_v^\top \left(\frac{1}{P} \sum_p \widetilde{f}_{ip} \right) = \sum_p \frac{1}{P} w_v^\top \widetilde{f}_{ip} = \sum_p c_{ipv},$$

where c_{ipv} is the contribution of patch p to the predicted response. By linearity of covariance,

$$\mathrm{Cov}_i(\widehat{y}_{iv}, y_{iv}) = \sum_p \mathrm{Cov}_i(c_{ipv}, y_{iv}).$$

Contributions were estimated by five-fold cross-fitting over training images. Each image was predicted using weights fitted without that image. We computed these covariances across the training images separately for each patch and voxel, and defined the contribution map as

$$m_{pv} = \frac{\mathrm{Cov}_i(c_{ipv}, y_{iv})}{\mathrm{Cov}_i(\hat{y}_{iv}, y_{iv})}.$$

The denominator equals the sum of the numerator over patches, so $\sum_p m_{pv} = 1$. Accordingly, m_{pv} represents the fraction of the covariance between the predicted and observed responses attributable to patch p. Voxels with $\mathrm{Cov}_i(\hat{y}_{iv}, y_{iv}) \leq 0$ were excluded from center estimation.

For each voxel, we averaged m_{pv} over all 24 layers, subtracted its spatial mean across patches, and retained only patches with values at least 50% of the peak. Their weighted centroid defined the predicted pRF center. Patch coordinates formed a 14×14 grid, obtained by dividing the 224×224-pixel input into patches of size 16. Coordinates were converted to degrees so that the full image subtended 8.4 degrees. Each patch spanned 0.6 degrees, setting the spatial sampling scale of the contribution map. The 50% threshold was fixed before inspecting the empirical data.

Predicted centers were paired voxel-wise with independently estimated pRF centers (Dumoulin & Wandell, 2008; Allen et al., 2022). Voxels were selected from V1–V3 if the pRF model explained more than 10% of their response variance. Coordinate correlations were computed for each state on matched voxels and layers.

B Additional State-Depth Results

For LPP, state depth relative to A1 generally increased from A1 to STS for Llama, wav2vec2-RelPos, and wav2vec2-RoPE, but decreased for wav2vec2 and HuBERT (Figure A1a). The participant-averaged MGND results likewise showed a general increase for wav2vec2-RelPos, wav2vec2-RoPE, and MusicGen, and a decrease for wav2vec2 and HuBERT (Figure A1b).

The main text showed cortical distributions only for Llama and wav2vec2. Here, we additionally include HuBERT, wav2vec2-RelPos, and wav2vec2-RoPE to show the distributions for all five models (Figure A2).

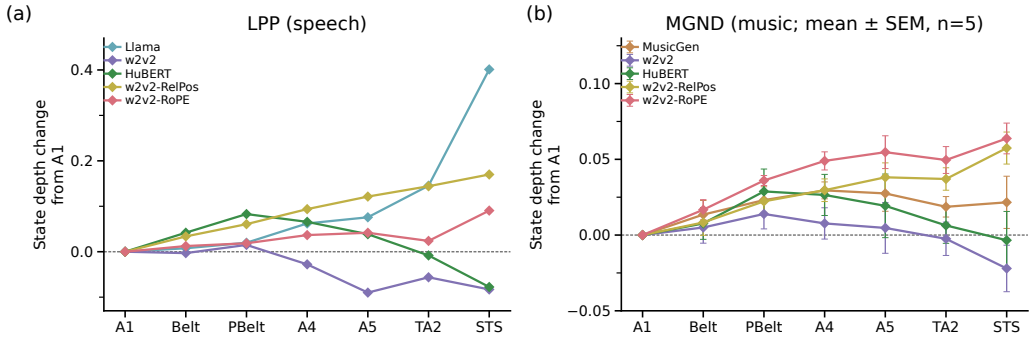

Figure A1: State depth for speech and music. (a) Differences from A1 for LPP mean responses, reproducing Figure 2b for comparison. (b) Mean ± SEM of within-participant differences for MGND (five participants).

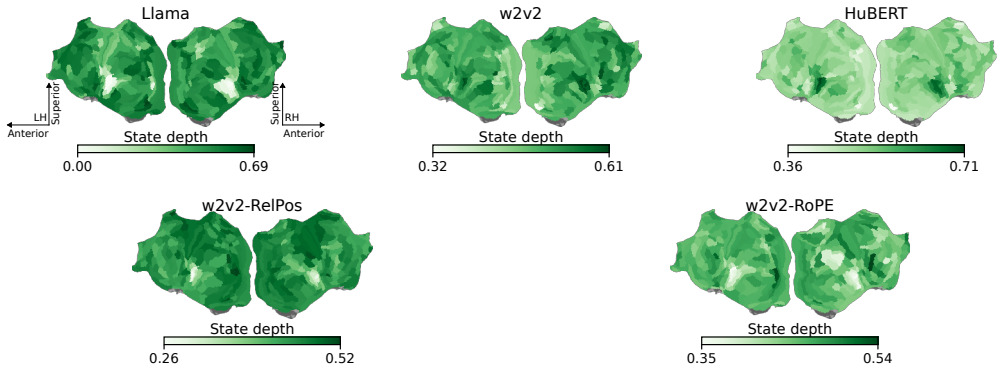

Figure A2: Cortical distributions of state depth for the five models on LPP.

C Additional Winning-Ratio Results for Music

C.1 Three models not shown in the main figure

Figure A3 shows MGND winning ratios for HuBERT, wav2vec2-RelPos, and wav2vec2-RoPE, which are not included in Figure 3a. Aggregation is identical to that in the main figure.

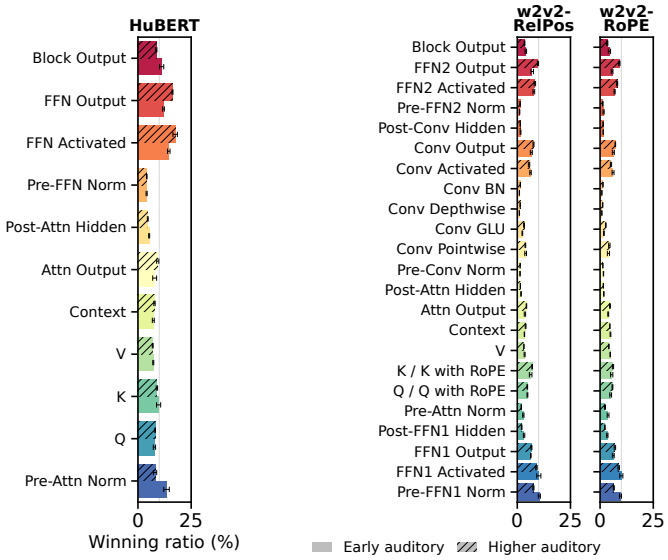

Figure A3: Winning ratios for HuBERT, wav2vec2-RelPos, and wav2vec2-RoPE in response to MGND music stimuli. Solid bars show early auditory cortex and hatched bars show higher auditory cortex. Bars and error bars indicate the mean and SEM across five participants.

C.2 Participant-wise winning ratios

Figure A4 shows the participant-wise winning ratios. The tendencies reported in the main text were maintained in the majority of participants in both early and higher auditory cortex.

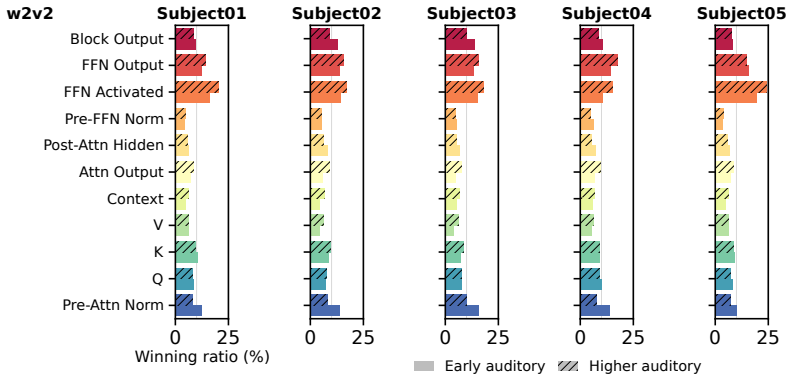

Figure A4: Participant-wise winning ratios for wav2vec2 (w2v2) in response to MGND music stimuli. Solid bars show early auditory cortex and hatched bars show higher auditory cortex.

D Additional Winning-Ratio Results for Vision

D.1 Three Swin models not shown in the main figure

Figure A5 shows winning ratios for Swin, Swin RoPE-Ax, and Swin RoPE-Mix in response to NSD image stimuli, which are not included in Figure 4a. Aggregation is identical to that in the main figure.

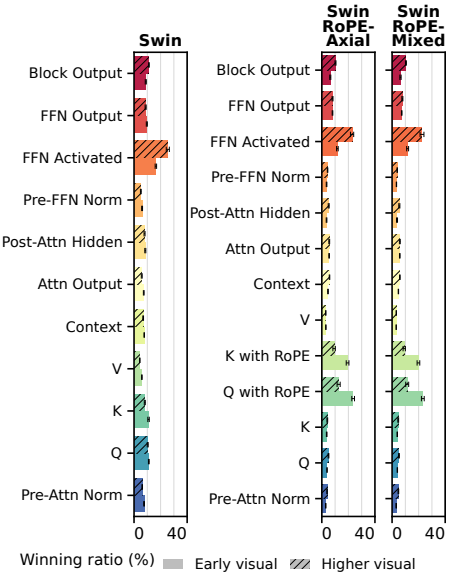

Figure A5: Winning ratios for Swin, Swin RoPE-Ax, and Swin RoPE-Mix in response to NSD image stimuli. Solid bars show early visual cortex and hatched bars show higher visual cortex. Bars and error bars indicate the mean and SEM across eight participants.

D.2 Participant-wise winning ratios

Figure A6 shows participant-wise winning ratios for DeiT III. The identities of the highest-ranking states were broadly consistent across participants both for the early and higher visual cortex.

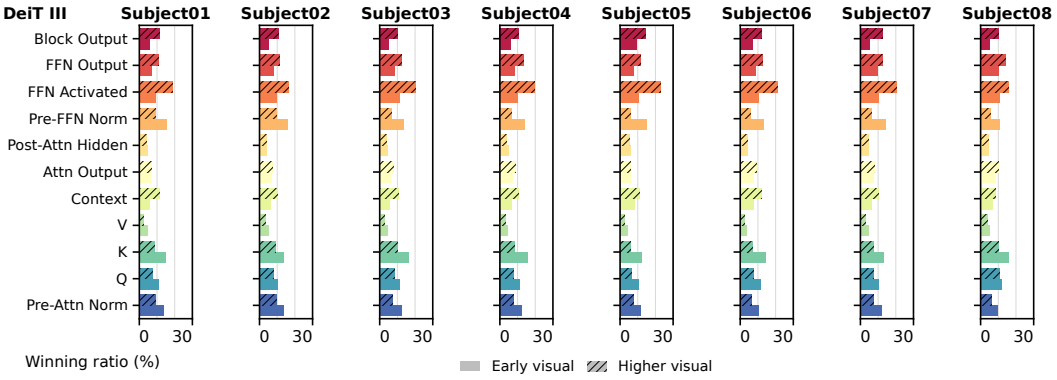

Figure A6: Participant-wise winning ratios for DeiT III in response to NSD image stimuli. Solid bars show early visual cortex and hatched bars show higher visual cortex.

E Positional-Encoding Differences in DeiT by State and Layer Group

E.1 Differences across layer groups

We aggregated the same ΔCC (prediction-accuracy difference) as in Figure 4b into four groups of six consecutive layers rather than averaging over all 24 layers (Figure A7).

In early visual cortex, the differences between post-rotation Q/K and APE Q/K were positive in all four layer groups and largest in L13–18. Other states showed no comparable positive differences across layer groups, and Q/K in higher visual cortex showed no consistent improvement.

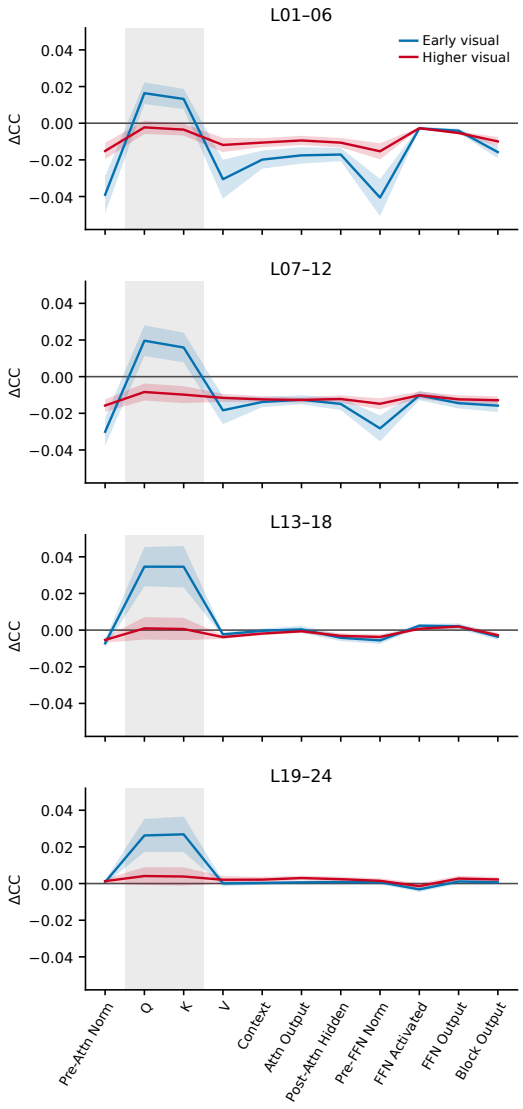

Figure A7: Prediction-accuracy differences ΔCC between DeiT III RoPE-Mix and DeiT III (APE), computed as RoPE-Mix minus APE, for groups of six layers. From top to bottom, panels show L01–06, L07–12, L13–18, and L19–24. Blue lines indicate early visual cortex and red lines indicate higher visual cortex. Bands show pointwise 95% confidence intervals, and shading marks Q/K, which undergo positional rotation.

E.2 PARTICIPANT-WISE CORTICAL DISTRIBUTIONS

Figure A8 shows the difference between post-rotation Q and APE Q for all eight participants, extending the subject01 result in Figure 4c of the main text. In individual participants, positive ΔCC values likewise clustered near early visual cortex and became smaller toward higher visual cortex. Thus, the advantage of post-rotation Q in earlier regions of the visual hierarchy was broadly shared across participants.

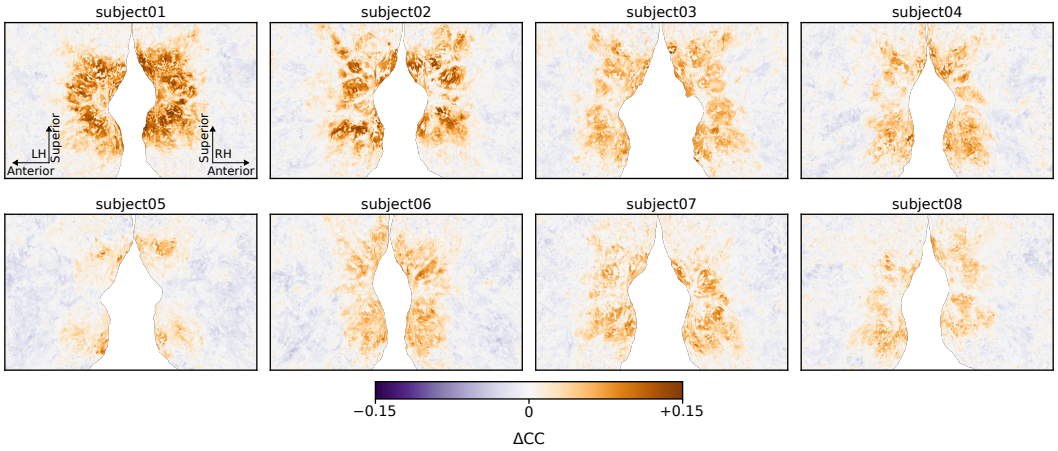

Figure A8: Cortical distributions of prediction-accuracy differences ΔCC for eight participants, computed as post-rotation Q in DeiT III RoPE-Mix minus Q in DeiT III (APE). Values are averaged over 24 layers at each voxel, with enlarged views of the regions around visual cortex. Orange indicates an advantage for RoPE-Mix and purple for APE.

F ADDITIONAL PROBING RESULTS FOR AUDIO STATES

Figure A9 shows the probing results for the two audio models, wav2vec2-RelPos and wav2vec2-RoPE. These models favored intermediate FFN1 activations for low-level acoustic features and context for semantics in a 3-second window. The relative patterns across states were broadly similar between the two models. Note that post-rotation Q/K in wav2vec2-RoPE did not rank highest for any target. This is consistent with the result in Section 4.1 that post-rotation Q/K did not show an advantage in predicting brain responses.

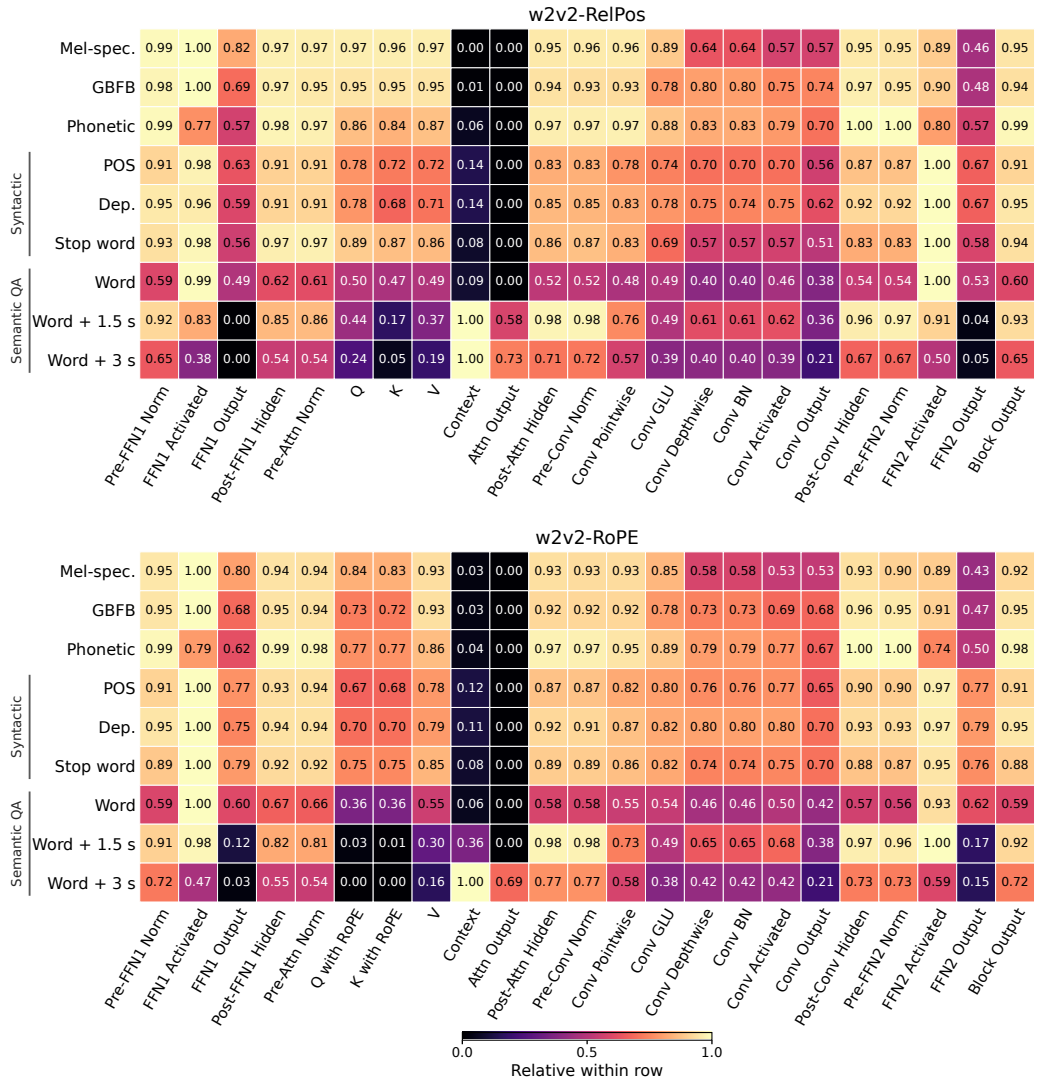

Figure A9: State-wise linear probing of the two Conformer models. The upper panel shows wav2vec2-RelPos and the lower panel shows wav2vec2-RoPE. The state columns are the same in both panels except for the two Q/K columns, which show Q/K without rotation in the upper panel and post-rotation Q/K in the lower panel. Rows show prediction targets and columns show within-block states. Colors and cell values show chance-corrected accuracy or Pearson correlation, summarized by the median across layers and min–max normalized within each row.

G Additional Results for pRF-Center Correspondence

Figure A10 presents the same comparison as in Figure 5b for all eight participants. Correlations between predicted and independently estimated pRF centers exceeded 0.90 in all eight participants. Thus, the main-text finding that visual-field position can be read out from the patch contributions of post-rotation Q was maintained across participants.

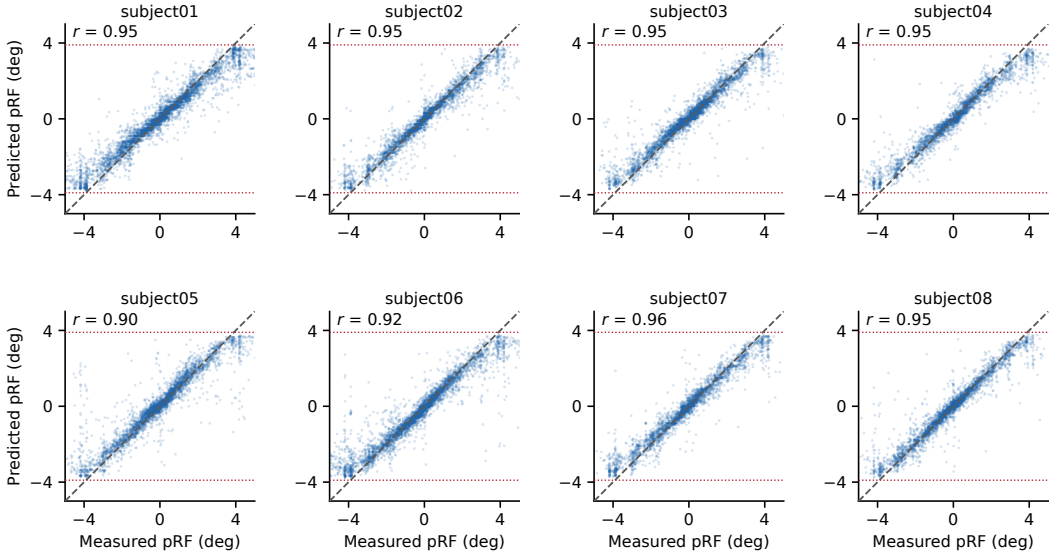

Figure A10: Correspondence between predicted pRF centers, obtained from the centroids of patch contributions, and independently estimated pRF centers for Q with RoPE in DeiT III RoPE-Mix, shown for eight participants. Each participant's V1–V3 voxels were used, with averaging over 24 layers. Horizontal and vertical coordinates are pooled into a single scatter plot and correlation r for each participant. The dashed line is the identity line, and red dotted lines indicate the range represented by the patch grid (± 3.9 deg). Axis ranges and ticks are shared across panels; axis labels are shown only in the left column and bottom row.

H Limitations

The present findings suggest several directions for further investigation. We outline four aspects of the current study that could be extended to develop a broader understanding of state-specific brain alignment.

Generalization across models and datasets. Our study examines a selected set of pretrained Transformer models using three datasets covering narrated speech, music, and natural images. Although these settings allow comparisons across several model families and stimulus types, the extent to which the observed patterns generalize to architectures beyond the Transformer family, other model scales, and additional datasets remains to be explored. Extending the analysis to a broader range of models and stimuli, including audiovisual inputs, would help establish which aspects of state-specific brain alignment are shared across settings.

Temporal dynamics of brain–model correspondence. Our analyses use fMRI responses to examine how intermediate model states correspond to different cortical regions, without directly characterizing the timing of these correspondences. Applying the framework to temporally resolved neural measurements, such as MEG or EEG, could help determine when different states best predict neural responses during stimulus processing. This would complement the spatial patterns examined here and provide a fuller picture of how within-block computations relate to the progression of sensory processing in the brain.

Individual differences. Our analyses emphasize patterns shared across participants, including group-averaged responses for narrated speech, rather than systematically examining the sources of individual variation. Future work could combine participant-specific encoding analyses with behavioral measures to investigate whether state preferences vary with factors such as language proficiency, musical expertise, or familiarity with the stimuli. This would extend the framework from identifying common patterns of brain alignment to understanding how these correspondences relate to individual experience and perceptual abilities.

Emergence of alignment during learning. We evaluated pretrained models with fixed weights, rather than tracking how their representations change during learning. The present analyses therefore leave open how state-specific brain alignment emerges over the course of training. Examining checkpoints at different training stages, as well as before and after task-specific fine-tuning, could reveal when particular states acquire brain-relevant information and how their relative advantages evolve with learning. This would complement the present comparisons by providing a developmental perspective on the organization of intermediate representations.