REPEAT-CALIBRATED NEURAL REPRESENTATIONS FOR CROSS-SUBJECT GENERALIZATION IN STEREOELECTROENCEPHALOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Model–brain alignment within recorded individuals does not establish transfer across people with different electrode layouts or across auditory domains. We introduce the Repeated Auditory Stereoelectroencephalography Dataset (RASD) and Repeat-Calibrated Neural Fields (RCNF) to investigate how response reliability and spatial aggregation affect such generalization. RASD records 30 pediatric and adolescent participants listening to one story and six music excerpts, each scheduled three times, linking 2,052 signal-valid contacts to stimulus events and providing 2,050 matched contact coordinates. RCNF uses training-repeat consistency to calibrate contacts before coordinate-based, participant-balanced aggregation, constructing a fixed-dimensional representation for a shared predictor. Four tasks evaluate held-participant prediction on new repetitions, reduced source coverage, and bidirectional story–music transfer; other participants supply neural responses to the test content. Across 21 participants and 161 auditory contacts, calibration improves all four endpoints over uniform reliability weighting after correction for 16 comparisons, with absolute gains of .00069–.00128. Temporal history improves three tasks; the full spatial representation has no corrected advantage over either spatial control. All aligned endpoints exceed time-shifted references. Together, RASD and RCNF make repeat reliability an explicit component of neural representation design, providing a data and method foundation for predictors that can be reused across people and auditory materials.

1 INTRODUCTION

Understanding what generalizes across people is essential for reusable neural predictors. Recent work evaluates artificial representations through model–brain alignment: multimodal networks predict SEEG responses associated with vision–language integration (Subramaniam et al., 2024), while local representational geometry correlates with brain alignment and task performance (Yu et al., 2026). Such correspondence, measured through participant-specific mappings, does not establish that a shared predictor transfers across people or stimulus domains. We ask how repeated auditory responses can be represented to predict another participant's neural activity, and which design choices improve transfer.

Stereoelectroencephalography (SEEG) (Cardinale et al., 2013) makes this problem concrete: clinically implanted depth electrodes directly record neural activity, but contact counts and anatomical coverage differ across participants. Corresponding channels cannot be assumed. Contacts also vary in response consistency, so spatial proximity alone need not identify useful signals; pooling all contacts equally lets densely sampled participants dominate. A transferable representation must therefore accommodate nonmatching locations, account for source reliability, and separate participant influence from contact count.

Existing intracranial datasets leave a specific experimental gap. NeuroListen (Zhang et al., 2025) targets auditory speech reconstruction, and Brain Treebank (Wang et al., 2024) provides naturalistic film recordings with linguistic annotations. Neither supplies a common repeated story-and-music design for jointly testing participant transfer, independent repetitions and auditory-domain changes. Repeated presentations expose variability relevant to model assessment (Schoppe et al., 2016). Us-

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.

ing different repetitions for calibration and evaluation separates consistency estimation from the predicted response; shared materials and contact coordinates then permit controlled tests of reliability and coverage across domains.

RASD provides this experimental basis: 30 pediatric and adolescent participants heard one Chinese story and six music excerpts, each scheduled three times. Four tasks combine participant holdout with story/music repeat generalization and bidirectional story–music transfer. The music repeat task uses known excerpts; cross-domain tasks evaluate materials from a domain excluded from fitting. Source-contact deletion probes dependence on coverage. Other participants supply responses to the test content. The held participant does not fit the shared predictor; auxiliary training-domain recordings determine only eligibility and response scaling.

We introduce Repeat-Calibrated Neural Fields (RCNF) to test whether repeat consistency can guide transferable representation design. Consistency supplies a reliability cue, anatomical proximity guides aggregation around a target location, and participant balancing prevents dense implants from dominating. These operations convert variable contact sets into fixed-dimensional inputs for a shared readout. Controlled RCNF variants test reliability, spatial scale, temporal history and coordinate dependence; time-shifted predictions test temporal alignment. An explicit pooling identity explains how calibration redistributes responses, without assuming that repeatability guarantees transfer.

Our contributions are threefold:

- **Data for generalization.** RASD couples common auditory repetitions with anatomical coordinates across 30 participants, enabling separate tests of repeat generalization, cross-subject prediction, reduced source coverage and story–music transfer within one cohort.
- **Repeat-calibrated representation.** RCNF converts repeated observations into a reliability cue for spatial aggregation. Its participant-balanced construction handles unequal contact sets without letting sampling density determine participant influence. The resulting fixed-dimensional representation supports a shared predictor, while an analytic identity specifies how and when calibration changes the pooled signal.
- **Evidence on transfer factors.** Four component contrasts and a temporal reference distinguish the roles of reliability, history, spatial aggregation and response alignment. Calibration yields small gains in all four tasks and temporal history helps three, while the full representation has no corrected advantage over single-scale or global pooling.

2 Related work

Clinical electrophysiological resources. CHB-MIT (Guttag, 2010) provides scalp EEG with seizure annotations, whereas F-TRACT (Trebaul et al., 2018) aggregates electrically evoked SEEG responses into connectivity estimates. These resources address different targets from auditory response transfer. Miller's open ECoG library (Miller, 2019) covers multiple behavioral tasks. The iEEG-BIDS specification (Holdgraf et al., 2019) standardizes recording, event and electrode metadata for intracranial data reuse. RASD links these information types through a common repeated-auditory schedule and a defined transfer benchmark.

Auditory and naturalistic language datasets. Berezutskaya et al. (2022) provide common-film iEEG recordings from 51 participants, 16 with SEEG, alongside fMRI and native-space electrode coordinates. NeuroListen (Zhang et al., 2025) records SEEG from five participants hearing Mandarin words and digits and English words for auditory reconstruction. Brain Treebank (Wang et al., 2024) provides 43 hours of intracranial film recordings from ten participants, with localization, word timings and syntactic annotations. These datasets emphasize reconstruction or naturalistic annotation rather than the repeated cross-domain protocol studied here. RASD focuses on three scheduled presentations of every common auditory material, coupling repeat calibration with participant holdout, source-contact deletion and bidirectional story–music transfer. Its contribution is this acquisition and evaluation design; its short materials cover less naturalistic content than Brain Treebank.

Shared and aligned neural representations. Hyperalignment (Haxby et al., 2011; Guntupalli et al., 2016) maps fMRI response patterns into shared representational spaces; the Shared Response Model (Chen et al., 2015) learns participant-specific projections into a common low-dimensional response space. Neural Latent Aligner (Cho et al., 2023) learns consistent representations and cross-

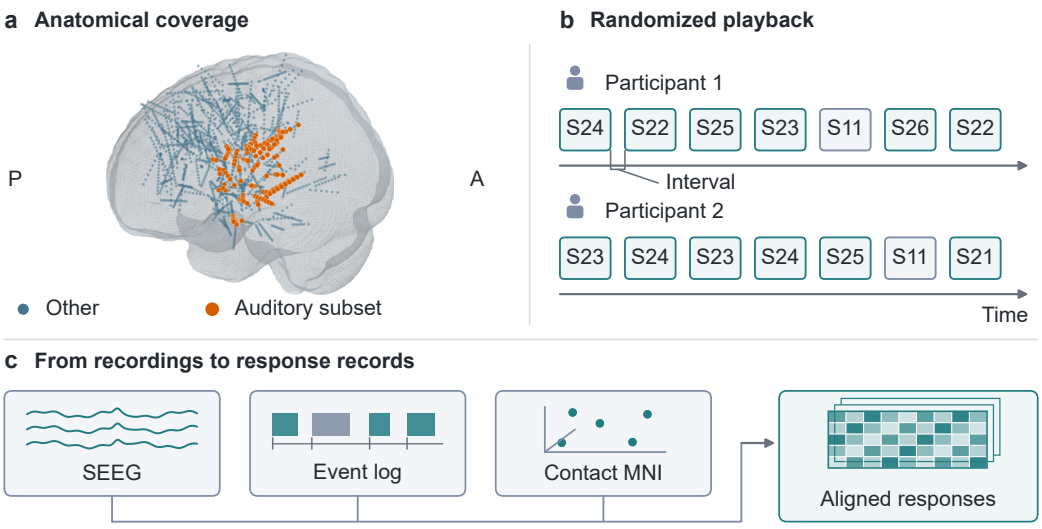

Figure 1: RASD acquisition and organization. (a) Contact coverage on an MNI template. (b) Recorded excerpts of randomized playback with 5–10-s intervals; blue denotes story S11, teal music S21–S26. Block widths and gaps are schematic. (c) SEEG recordings, event logs and contact coordinates jointly define aligned response records; waveforms and matrices are schematic.

trial temporal alignment from repeated natural speech production. CEBRA (Schneider et al., 2023) learns neural embeddings using behavioral or temporal supervision. These approaches learn correspondence between observations. RCNF instead constructs a coordinate-based representation from repeat-calibrated contacts, allowing a shared readout to query a held participant's locations without fitting a target-specific representation map.

Models of heterogeneous neural populations. Deep Sets (Zaheer et al., 2017) and Set Transformer (Lee et al., 2019) represent unordered, variable-size inputs. Population Transformer (Chau et al., 2025) aggregates sparse neural channels; POYO (Azabou et al., 2023) uses a shared decoder across recording sessions, and POYO+ (Azabou et al., 2025) extends shared decoding across tasks and neural populations. These models motivate pooling nonmatching neural observations. RCNF uses explicit spatial kernels and equal participant weighting to form fixed features. Its matched controls test reliability, temporal history and spatial aggregation within the same observation protocol.

3 RASD AND THE GENERALIZATION PROTOCOL

3.1 SHARED OBSERVATIONS ACROSS HETEROGENEOUS CONTACT LAYOUTS

RASD links three types of information needed for cross-subject response transfer: a common stimulus timeline, repeated neural observations, and the spatial location of each contact. All 30 participants listened to the same Chinese story (S11) and six music excerpts (S21–S26), each scheduled for three presentations, R1–R3. The inventory retains 628 of 630 planned presentations. All story repetitions are available; two participant–music cells lack one presentation. Figure 1 gives the dataset overview, and Table 1 distinguishes the full resource from the subset evaluated here.

Contact identity links 2,050 signal-valid contacts to finite coordinates in the supplied Montreal Neurological Institute (MNI) coordinate system. The two unmatched contacts remain outside coordinate-based analyses. The auditory subset comprises contacts labeled Heschl or Temporal_Sup that pass training-response eligibility checks, yielding 161 contacts from 21 participants. Coverage ranges from 2 to 20 contacts per participant (median 6); ten participants have at most four. The benchmark therefore retains unequal spatial coverage rather than selecting a common electrode grid. Additional anatomical views appear in Appendix A.

Table 1: Dataset inventory and the auditory subset used for the benchmark.

Item	Full resource	Auditory benchmark
Participants	30	21
Age range; median	3y 4m–17y; 11y 4m	3y 4m–17y; 8y 4m
Recorded sex	19 male / 11 female	15 male / 6 female
Contacts	2,052 signal-valid; 2,050 matched	161, all coordinate-matched
Retained presentations	628/630 planned	439/441 planned
Materials	One story and six music excerpts; three planned repeats each	
Sampling	1,000-Hz SEEG; 100-Hz modeled high-gamma power	
Analysis windows	Story: 23.5 s; music: 20 s	

Each participant retains a separate contact axis. Onset alignment supplies a shared time axis, while stimulus and repetition identifiers keep repeated presentations distinct. Missing presentations remain absent. This organization lets other predictors use the same response windows and held-out observations without adopting RCNF's pooling rule. The full resource supports broader regional analyses; the performance claims in this paper concern the defined auditory subset.

3.2 Response construction and permitted calibration

We extract high-gamma power from onset-aligned windows using a fourth-order 70–150-Hz Butterworth filter, squared Hilbert amplitude and 50-ms smoothing, then downsample to 100 Hz. Story windows span 23.5 s and music windows 20 s. Contact eligibility and response mean/standard deviation use only the selected training-domain repeats. Across domains, response scales, reliability weights, feature standardization, readout coefficients and selected bandwidth remain fixed at their training-domain values. Appendix A specifies the transformation and missing-cell rules.

3.3 Four transfer tasks

Let i denote the held participant and u a stimulus–repetition unit. A predictor receives onset-aligned responses $\widetilde{y}_{j,e}^{(u)}(t)$ at source coordinates $c_{j,e}$ from participants $j \neq i$, and predicts the held response at coordinate q:

$$\widehat{\widetilde{y}}_i^{(u)}(q,t) = f_\theta\Big(\{(c_{j,e}, \widetilde{y}_{j,e}^{(u)}(t)) : j \neq i\}, q, t\Big). \tag{1}$$

The shared parameters θ are fitted without participant i. Source and target observations are matched by material and available-repeat index and onset-aligned across sessions. Thus a story-trained model receives source music responses when predicting the held participant's music response. Transfer changes the fitted predictor's training domain while preserving access to neural observations of the test content.

Table 2: Training and evaluation for the four transfer tasks. A/D share a story-trained model; B/C share a music-trained model.

Task	Training domain	Held-target evaluation	Primary endpoint
A: story → story	Story: two repeats	Remaining story repeat	Normalized deletion AUC
B: music → music	Six music excerpts: two repeats each	Remaining repeat per excerpt	Material-balanced deletion AUC
C: music → story	Music (as in B)	All three story repeats	Full-source Pearson r
D: story → music	Story (as in A)	All available music repeats	Material-balanced full-source r

Tasks A/B separate participant transfer from repeat variability within a domain. Their three folds hold out R1, R2 and R3 in turn. The music model jointly fits all six excerpts, so B tests new repetitions of known materials. Tasks C/D then test whether a predictor fitted in one domain transfers to the other. They use the same three training-repeat pairs and evaluate all available target-domain

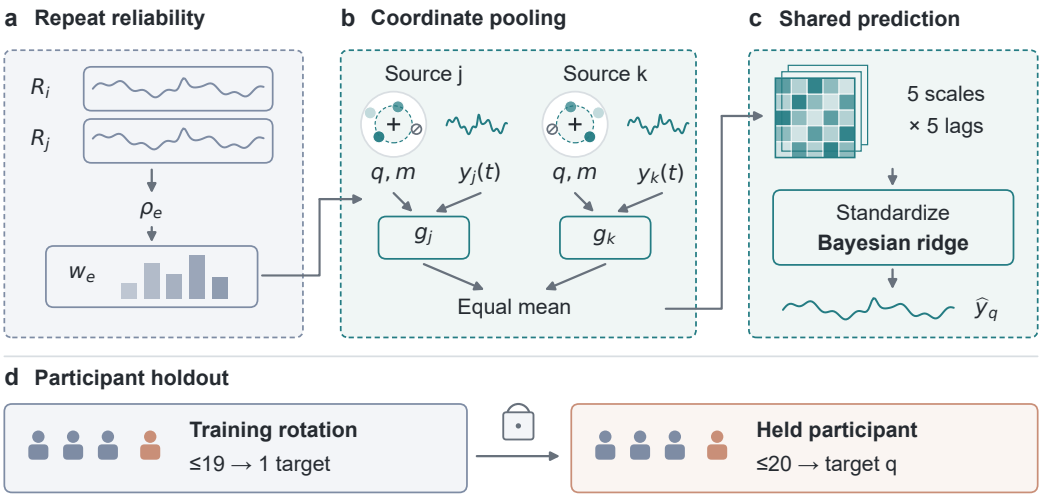

Figure 2: RCNF workflow. (a) Training repeats R_i, R_j yield agreement ρ_e and weight w_e. (b) Source responses $y_j(t)$ enter pooling with query q, mask m and weights; g_j is the source-specific feature. (c) Spatial and temporal features enter a shared readout. (d) The lock marks parameters fixed after fitting. Traces, contact positions and matrices are schematic.

repeats; these folds vary training observations, not the target test set. Source contacts are removed at fractions $0, .3, .5, .7$. A/B integrate performance over these levels, while C/D use the full source pool for their primary endpoint.

The two incomplete music cells are excluded from music training and task B, while their remaining presentations are evaluated in D. Other materials from these participants are retained. This yields five music materials for two participants and six for the other nineteen in B. Task B evaluates 20 participants for S21/S23 and 21 for the other excerpts; task D evaluates all 21 for each excerpt.

4 Repeat-Calibrated Neural Fields

RCNF converts an irregular set of source responses into a fixed-size representation at a target coordinate. It first uses repeated observations to weight contacts, then aggregates within each source participant and averages across participants. Fixed spatial scales and temporal lags expose these summaries to a shared linear readout (Figure 2). This design separates contact calibration from the complexity of the learned predictor.

4.1 Contact calibration and participant-balanced aggregation

For each source contact, training-repeat correlation measures response consistency. In music training, we average these correlations equally across the available complete training excerpts. The resulting $\rho_{j,e}$ defines

$$w_{j,e} = \mathrm{clip}\left((1 + \rho_{j,e})/2,\, 0.05,\, 1\right). \tag{2}$$

The fixed soft mapping retains every eligible contact, assigns weight .5 to zero repeat correlation, and bounds the influence of negative agreement. It encodes relative repeatability rather than an estimate of inverse noise variance.

For query q, scale σ and contact-retention mask m, let $K_\sigma(q, c) = \exp[-\|q - c\|^2/(2\sigma^2)]$, with $K_\infty = 1$. RCNF forms

$$\alpha^{(\sigma)}_{j,e}(q) = \frac{m_{j,e} w_{j,e} K_\sigma(q, c_{j,e})}{\sum_{e'} m_{j,e'} w_{j,e'} K_\sigma(q, c_{j,e'})}, \tag{3}$$

$$\phi_{\sigma,\tau}(q, t) = \frac{1}{|\mathcal{J}_m|} \sum_{j \in \mathcal{J}_m} \sum_{e} \alpha^{(\sigma)}_{j,e}(q) \widetilde{y}_{j,e}(t - \tau), \tag{4}$$

where $\mathcal{J}_m$ contains sources with at least one retained contact. Within-source normalization makes each local feature a weighted average; across-source averaging prevents a densely sampled participant from receiving more total weight solely because it has more contacts. Calibration therefore redistributes each participant's contribution among its contacts.

Five spatial scales (10, 20, 40, 80 mm and global) and five temporal lags (0, 50, 100, 150 and 200 ms) produce a 25-dimensional vector $\Phi(q,t)$. Narrow kernels emphasize nearby observations; broader kernels retain a spatially distributed summary when local coverage is sparse. The global feature removes distance dependence. All scales remain available to the readout. The representation has fixed size across electrode layouts and is invariant to a joint permutation of contact coordinates, responses, weights and retention masks.

Finite-scale kernels use a within-source distance offset before exponentiation, preserving normalized weights and a positive denominator (Appendix D).

4.2 Shared prediction across participants

Within each outer fold, the 20 training participants take turns providing targets; available contacts from the other 19 supply source responses. These examples fit a feature standardizer and Bayesian-ridge readout in scikit-learn (Pedregosa et al., 2011), with uniform target loss:

$$z(q,t) = \mathrm{Standardize}_{\mathrm{train}}(\Phi(q,t)), \qquad \widehat{\widetilde{y}}(q,t) = b + \beta^{\mathsf{T}} z(q,t). \tag{5}$$

We call this model RCNF (FULL). Temporal features are constructed within each trial before training examples are pooled, so lags do not cross stimulus boundaries. At test time, available contacts from the training cohort supply responses for the held participant's query coordinates. Contact weights enter the source representation; they do not reweight target fitting errors. The coordinate basis permits the same readout to operate without a fitted target-specific representation map.

4.3 What repeat calibration changes

A two-contact example illustrates the operation. Suppose two source contacts are equidistant from a query, have weights .9 and .5, and standardized responses 1 and 0 at a given time. Their contributions become $9/14$ and $5/14$, giving local response $9/14$ rather than the spatial-only mean $1/2$. The calibrated feature changes because the more repeatable contact carries a different response.

More generally, for one source, query, scale and mask, let a_e be the normalized spatial-only weights. The calibrated and uniform-reliability features obey

$$g_w - g_0 = \frac{\mathrm{Cov}_a(w,y)}{\mathbb{E}_a[w]}. \tag{6}$$

Calibration changes the local feature through the association between reliability and responses. Its effect vanishes for a single retained contact, equal weights or identical local responses; multiplying all weights of one source by a positive constant also cancels. These properties identify where calibration can act, without guaranteeing a predictive gain after readout fitting. Appendix D derives the identity and the normalization invariances. The matched control below tests whether this redistribution improves prediction when both readouts are refitted.

5 Evaluating cross-subject and cross-domain generalization

5.1 Models and statistical analysis

RCNF (full) denotes the complete model. Its four variants remove reliability weighting, use one training-selected finite scale, retain only zero-lag responses, or replace coordinate-dependent pooling with global pooling. Each refits its standardizer and readout with matched observations, masks and scoring support. SingleScale uses training-only kernel predictions at 100 ms to select its bandwidth, changing both selection and scale count. RCNF (time-shifted) evaluates fixed full-model predictions under 1,000 joint circular shifts at least 2 s from alignment.

Evaluation comprises 21 participant holdouts, three repeat splits and five technical seeds per training domain (630 configurations; 3,150 fitted models). Correlations are averaged within participant, with

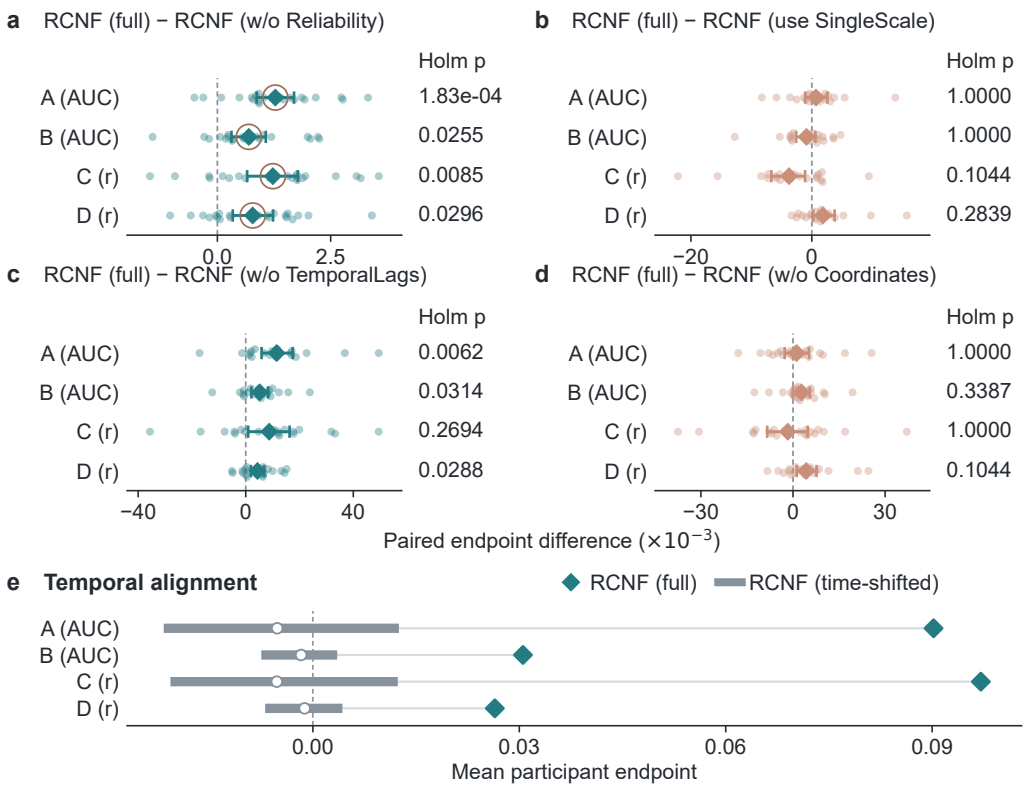

Figure 3: Component and temporal controls. (a–d) RCNF (full)-minus-variant differences: dots show participants, diamonds means and bars unadjusted 95% bootstrap intervals; open rings identify reliability-contrast means. (e) Aligned means (diamonds) and time-shifted means (circles); gray bars span the central 95% of surrogate cohort means.

music materials weighted equally. Primary endpoints are normalized deletion AUC over $[0, .7]$ for A/B and full-source Pearson r for C/D (Table 2).

The 21 participant summaries support unadjusted 95% bootstrap intervals (10,000 resamples) (Efron, 1979), exact two-sided paired sign-flip tests (Nichols & Holmes, 2002), and Holm correction (Holm, 1979) across 16 component contrasts; four temporal contrasts form a separate family. Inference is nominal because this cohort informed development and outer folds share training and test-source participants. Appendix B and Appendix F specify fitting, aggregation and temporal controls.

5.2 Calibration improves all four task endpoints

The matched reliability ablation tests the contribution of repeat consistency beyond spatial pooling. Calibration improves all four endpoints after Holm correction, with small absolute gains (Table 3; Figure 3a). The full model has no corrected advantage over its single-scale variant (Figure 3b). Reliability weighting therefore has more consistent support than the full multiscale representation.

5.3 Repeat splits reveal stable gains and variable absolute performance

Repeat splits test whether calibration depends on a favorable training realization. Gains remain positive in every task–split–seed condition (Figure 4e), while absolute performance varies across folds (Appendix Table S4). The direction of the calibration gain is therefore more stable than the performance estimate from any one split.

Table 3: Mean participant endpoints. Bold and underlined values mark the highest and second-highest means within each task. The time-shifted row averages references from fitted RCNF predictions without refitting.

Variant	A (AUC)	B (AUC)	C (r)	D (r)
RCNF (full)	**.090209**	<u>.030533</u>	.097081	**.026444**
RCNF (w/o Reliability)	.088930	.029840	.095856	<u>.025665</u>
RCNF (use SingleScale)	<u>.089521</u>	**.031422**	**.100857**	.024606
RCNF (w/o TemporalLags)	.078707	.025329	.088353	.022078
RCNF (w/o Coordinates)	.089029	.027796	<u>.098791</u>	.022150
RCNF (time-shifted)	-.005186	-.001715	-.005248	-.001201

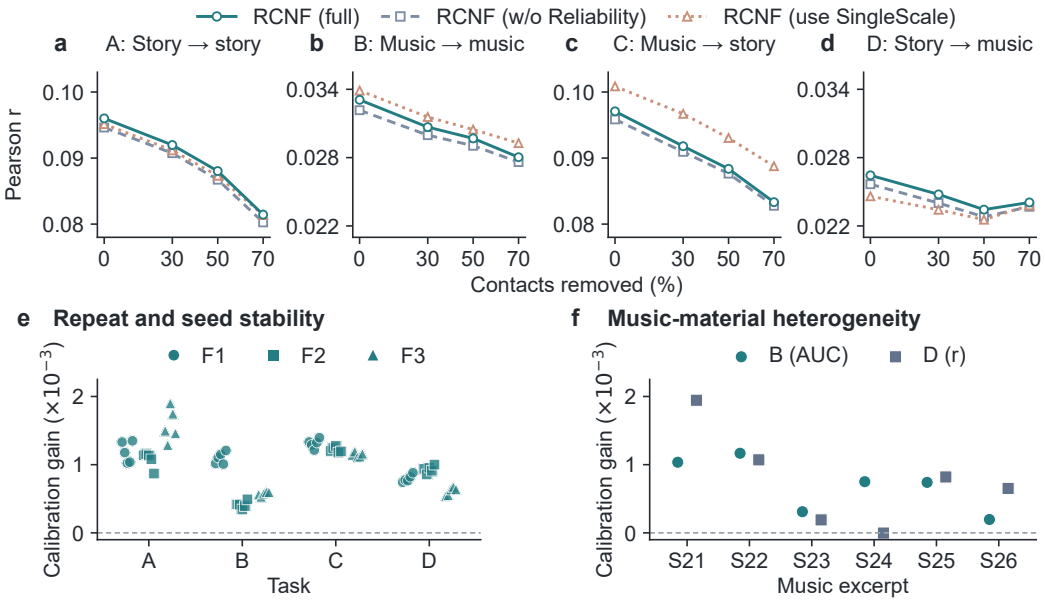

Figure 4: Robustness across coverage, repetitions and materials. (a–d) Mean prediction correlation across contact deletion levels. (e,f) Reliability-calibration gains by repeat-fold/seed and music material, respectively.

5.4 Source coverage and music materials qualify aggregate performance

Contact deletion and material breakdowns test whether aggregate performance masks sensitivity to coverage or content. Removing source contacts reduces overall performance across the displayed variants (Figure 4a–d). With all source contacts retained, absolute music prediction still varies across materials (Appendix Figure S6), while calibration gains are predominantly positive (Figure 4f). Reliability weighting thus provides a refinement without removing coverage and material limitations.

5.5 Component ablations and temporal alignment

Component removals test the contribution of temporal history and query-dependent pooling. Temporal lags benefit A/B/D after correction, whereas coordinate-dependent features show no corrected advantage over global pooling (Figure 3c,d). Temporal history thus has clearer predictive support. All aligned endpoints exceed their temporal references (Figure 3e; Appendix Table S9), supporting prediction of aligned cross-participant structure. An earlier story-fold weight-permutation experiment supports the distinct role of contact-specific reliability assignments (Appendix Figure S9).

5.6 CHANNEL-LEVEL REPEAT CONSISTENCY

Channel-level traces test whether training-repeat consistency persists in a held repetition. High-ranked contacts have higher average held-repeat consistency across participants, but the ordering reverses in the selected example (Figure 5). The observed variability shows why a contact's training-repeat consistency cannot establish that it is uniformly informative or irrelevant.

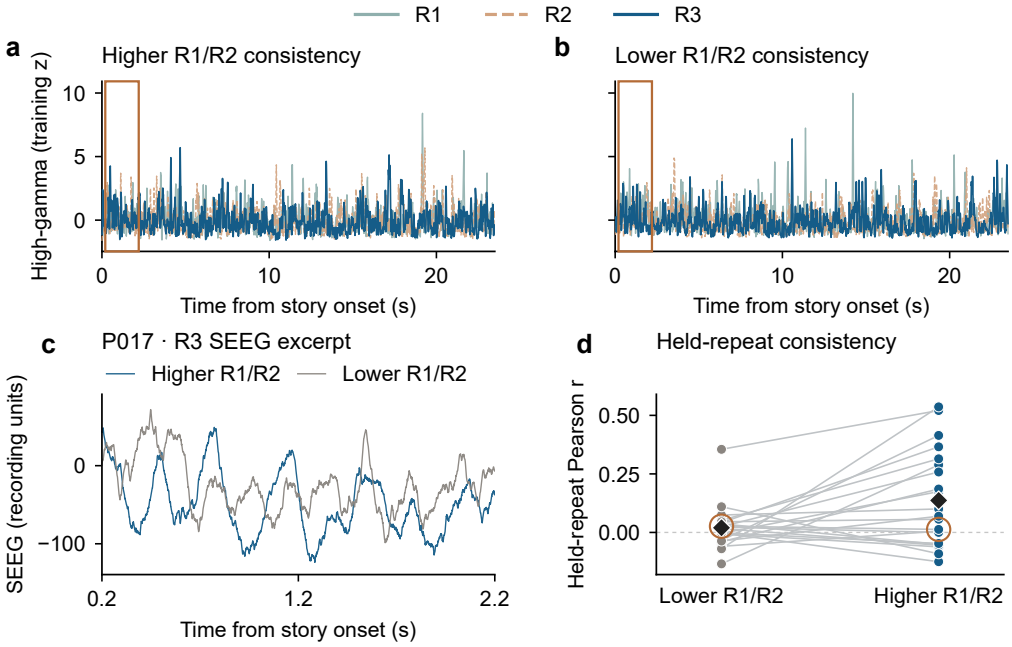

Figure 5: Repeat consistency in recorded SEEG. (a,b) Training-standardized high-gamma responses across three story repetitions at P017 contacts with higher/lower R1/R2 consistency. Boxes mark the training-selected window. (c) Recorded R3 SEEG from that window in stored recording units. (d) R3 correlations with the R1/R2 mean for the selected contacts in all 21 participants; diamonds, means; circles, P017. Selection uses R1/R2 alone (Appendix A).

6 DISCUSSION AND CONCLUSION

RASD provides a controlled setting in which shared auditory content, repeated presentations and contact localization can be examined together. Its value is the ability to distinguish sources of transfer performance within one acquisition design. The short, repeated materials complement resources built around longer naturalistic recordings.

The component results suggest that estimating source consistency is useful even when its effect is modest. A richer spatial representation, however, did not yield a corrected advantage over global pooling. Future models should test whether target-conditioned weighting adds information beyond repeat reliability and temporal history. Matched global-pooling controls remain necessary for interpreting any such gains.

The evidence is limited to a clinically sampled pediatric and adolescent cohort and short auditory materials. Independent cohorts with broader coverage would test whether these findings persist across acquisition settings. Additional excerpts would enable within-domain material holdouts alongside the present repetition and domain splits. More repetitions would permit the calibration budget to vary independently of the readout-training budget. Ethics and data-sharing approvals are in place; the code and dataset will be made available upon acceptance.

REFERENCES

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44937–44956, 2023. doi: 10.52202/075280-1948. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ca113d122584f12a6727341aaf58887-Abstract-Conference.html.

Mehdi Azabou, Krystal Pan, Vinam Arora, Ian Knight, Eva Dyer, and Blake A Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 59654–59677, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/953390c834451505703c9da45de634d8-Abstract-Conference.html.

Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9(1):91, 2022. doi: 10.1038/s41597-022-01173-0. URL https://www.nature.com/articles/s41597-022-01173-0.

Francesco Cardinale, Massimo Cossu, Laura Castana, Giuseppe Casaceli, Marco Paolo Schiariti, Anna Miserocchi, Dalila Fuschillo, Alessio Moscato, Chiara Caborni, Gabriele Arnulfo, and Giorgio Lo Russo. Stereoelectroencephalography: Surgical methodology, safety, and stereotactic application accuracy in 500 procedures. *Neurosurgery*, 72(3):353–366, 2013. doi: 10.1227/NEU.0b013e31827d1161. URL https://pubmed.ncbi.nlm.nih.gov/23168681/.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population Transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations*, pp. 69100–69121, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/acb3e20075b0a2dfa3565f06681578e5-Abstract-Conference.html.

Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James V. Haxby, and Peter J. Ramadge. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, volume 28, pp. 460–468, 2015. URL https://papers.neurips.cc/paper_files/paper/2015/hash/b3967a0e938dc2a6340e258630febd5a-Abstract.html.

Cheol Jun Cho, Edward Chang, and Gopala Anumanchipalli. Neural Latent Aligner: Cross-trial alignment for learning representations of complex, naturalistic neural data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 5661–5676. PMLR, 2023. URL https://proceedings.mlr.press/v202/cho23a.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-Jackknife/10.1214/aos/1176344552.full.

J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6):2919–2934, 2016. doi: 10.1093/cercor/bhw068. URL https://academic.oup.com/cercor/article/26/6/2919/1754308.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, 2010. doi: 10.13026/C2K01R. URL https://physionet.org/content/chbmit/1.0.0/. Version 1.0.0.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A common, high-dimensional

model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. doi: 10.1016/j.neuron.2011.08.026. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3201764/.

Christopher Holdgraf, Stefan Appelhoff, Stephan Bickel, Kristofer Bouchard, Sasha D'Ambrosio, Olivier David, Orrin Devinsky, Benjamin Dichter, Adeen Flinker, Brett L. Foster, Krzysztof J. Gorgolewski, Iris Groen, David Groppe, Aysegul Gunduz, Liberty Hamilton, Christopher J. Honey, Mainak Jas, Robert Knight, Jean-Philippe Lachaux, Jonathan C. Lau, Christopher Lee-Messer, Brian N. Lundstrom, Kai J. Miller, Jeffrey G. Ojemann, Robert Oostenveld, Natalia Petridou, Gio Piantoni, Andrea Pigorini, Nader Pouratian, Nick F. Ramsey, Arjen Stolk, Nicole C. Swann, François Tadel, Bradley Voytek, Brian A. Wandell, Jonathan Winawer, Kirstie Whitaker, Lyuba Zehl, and Dora Hermes. iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Scientific Data*, 6(1):102, 2019. doi: 10.1038/s41597-019-0105-7. URL https://www.nature.com/articles/s41597-019-0105-7.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Kai J. Miller. A library of human electrocorticographic data and analyses. *Nature Human Behaviour*, 3(11):1225–1235, 2019. doi: 10.1038/s41562-019-0678-3. URL https://www.nature.com/articles/s41562-019-0678-3.

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002. doi: 10.1002/hbm.1058. URL https://pubmed.ncbi.nlm.nih.gov/11747097/.

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 Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12 (85):2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6.

Oliver Schoppe, Nicol S. Harper, Ben D. B. Willmore, Andrew J. King, and Jan W. H. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10:10, 2016. doi: 10.3389/fncom.2016.00010. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2016.00010/full.

Vighnesh Subramaniam, Colin Conwell, Christopher Wang, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Revealing vision-language integration in the brain with multimodal networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 46868–46890. PMLR, 2024. URL https://proceedings.mlr.press/v235/subramaniam24a.html.

Lena Trebaul, Pierre Deman, Viateur Tuyisenge, Maciej Jedynak, Etienne Hugues, David Rudrauf, Manik Bhattacharjee, François Tadel, Blandine Chanteloup-Foret, Carole Saubat, Gina Catalina Reyes Mejia, Claude Adam, Anca Nica, Martin Pail, François Dubeau, Sylvain Rheims, Agnès Trébuchon, Haixiang Wang, Sinclair Liu, Thomas Blauwblomme, Mercedes Garcés, Luca De Palma, Antonio Valentin, Eeva-Liisa Metsähonkala, Ana Maria Petrescu, Elizabeth Landré, William Szurhaj, Edouard Hirsch, Luc Valton, Rodrigo Rocamora, Andreas Schulze-Bonhage,

Ioana Mindruta, Stefano Francione, Louis Maillard, Delphine Taussig, Philippe Kahane, and Olivier David. Probabilistic functional tractography of the human cortex revisited. *NeuroImage*, 181:414–429, 2018. doi: 10.1016/j.neuroimage.2018.07.039. URL https://doi.org/10.1016/j.neuroimage.2018.07.039.

Christopher Wang, Adam Yaari, Aaditya K. Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain Treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *Advances in Neural Information Processing Systems*, volume 37, pp. 96505–96540, 2024. doi: 10.52202/079017-3060. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/aefa2385b3f33abf1526ae4e2c208cd9-Abstract-Datasets_and_Benchmarks_Track.html.

Junjie Yu, Wenxiao Ma, Chen Wei, Jianyu Zhang, Haotian Deng, Zihan Deng, and Quanying Liu. Local intrinsic dimension of representations predicts alignment and generalization in AI models and human brain. *arXiv preprint arXiv:2601.22722*, 2026. doi: 10.48550/arXiv.2601.22722. URL https://arxiv.org/abs/2601.22722.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Xueyi Zhang, Ruicong Wang, Jialu Sun, Siqi Cai, and Haizhou Li. Listening to the brain: Multi-band sEEG auditory reconstruction via dynamic spatio-temporal hypergraphs. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-4302. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bbc6396ae4b4dbf78ce328aa8beafd27-Abstract-Datasets_and_Benchmarks_Track.html.