NEURAL EVIDENCE TRANSPORT: REUSING BRAIN DECODERS ACROSS SPEECH TOKENIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech tokenizers define different vocabularies for the same audio, preventing a brain decoder trained on one vocabulary from directly serving another. We introduce **Neural Evidence Transport** to reuse a participant's decoder with a frozen target model. The method divides source probabilities by reference token frequencies and averages the resulting ratios through a soft bridge estimated from ordinary audio. This transfers support relative to baseline, avoiding repeated frequency weighting when updating the target prior. We prove that transport preserves neutral evidence and derive an exact identity for the loss of chi-square contrast. Fusion weights depend on the available target-token context, are selected on development participants, and remain fixed for confirmation. On magnetoencephalography (MEG) data from 11 held-out LibriBrain100 participants, transport reduces target-token negative log-likelihood relative to the prior and single-weight fusion in both HuBERT–Wav2Vec2 directions. It outperforms matched target-space decoders for Wav2Vec2 and retains 97% of Ridge's improvement over the prior for HuBERT. In an independent electroencephalography (EEG) study with 14 held-out participants, transport achieves 60.6% attended-speaker accuracy over 12 seconds (95% confidence interval: 56.4–65.6%), retaining 81% of the target-space Ridge gain over chance. Timing and source-label controls link these gains to aligned neural observations and vocabulary correspondence. These results demonstrate decoder reuse without an additional participant-specific target-space neural fit.

1 INTRODUCTION

Neural speech decoders map brain recordings to continuous audio embeddings (Défossez et al., 2023; Jayalath et al., 2025), language representations (Tang et al., 2023; d'Ascoli et al., 2025), or discrete speech units (Duan et al., 2023; Zheng et al., 2024; Li et al., 2025). A token-based decoder predicts categories defined by a particular codebook. Changing that codebook changes what its scores refer to, preventing direct reuse with a downstream model that uses another vocabulary.

We study reuse of an existing brain decoder when the downstream speech vocabulary changes. One motivating setting is the integration of independently developed neural and speech models with different vocabularies. Retaining the trained neural decoder is particularly useful when the original participant-specific recordings are unavailable for refitting. For example, a participant-specific decoder trained on HuBERT units (Hsu et al., 2021) must supply neural information to a frozen contextual model over Wav2Vec2 units (Baevski et al., 2020). Neural Evidence Transport connects the two through ordinary audio, adapting the output vocabulary while keeping both models fixed. This lets the participant's trained decoder serve the target model without a separate target-space neural fit.

The two vocabularies describe the same audio but partition it differently. Our HuBERT and Wav2Vec2 codebooks each contain 256 states, and their token indices have meaning only within their own codebooks. Time-aligned audio reveals a many-to-many correspondence. We retain this correspondence as a soft bridge that combines support from several source states for each target token.

Fusion introduces a second challenge. A source token can receive a high probability simply because it is common. Propagating source probabilities and multiplying them into the target prior introduces reference frequencies again. We therefore express each source prediction relative to its reference

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.

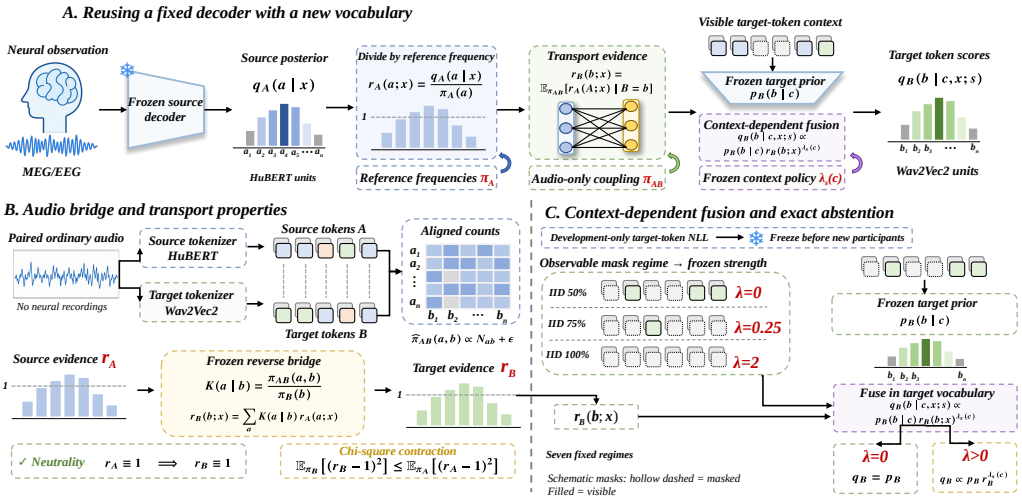


Figure 1: **Neural Evidence Transport across speech vocabularies.** (A) Source posteriors are divided by reference frequencies, transferred through an audio bridge, and combined with a frozen target prior. (B) Time-aligned ordinary audio estimates source-token frequencies conditional on target tokens. Averaging source ratios with these weights preserves neutrality and contracts chi-square contrast. (C) MEG development selects one fusion weight per masking regime; zero leaves the prior unchanged. Three IID examples illustrate the seven-regime policy. Both directions use this procedure. Each new participant supplies a source-vocabulary decoder; no target-space neural refitting is required. The source decoder is frozen during transfer. EEG uses the bridge marginal as its prior (Appendix D.10).

frequency and average these ratios through the audio bridge (Figure 1). A neutral source prediction then leaves the target prior unchanged.

The value of this evidence also depends on context. Neural predictions can fill gaps when target tokens are masked, whereas informative visible tokens may favor a smaller update. We select a fusion weight for each direction and masking regime on development participants, including zero for exact prior fallback, and fix the policy for confirmation participants.

Participant-disjoint studies evaluate target-token negative log-likelihood (NLL) in LibriBrain100 MEG (Mantegna et al., 2026) and attended-speaker accuracy in an independent EEG corpus (Stoll et al., 2025). Matched target-space decoders quantify the performance available with an additional participant-specific fit. Timing, source-label, and operator controls test the sources of the reuse gains.

Our contributions are:

- **Cross-vocabulary evidence transport.** An audio-derived soft bridge for source-model density ratios, with neutrality and normalization guarantees and an exact identity for the loss of chi-square contrast.
- **Context-dependent fusion.** A development-selected weight per direction and masking regime, including zero for exact prior fallback.
- **Decoder-reuse evaluation.** Participant-disjoint MEG and EEG studies with matched target-space training, observation-specific controls, and learned-bridge diagnostics.

2 Related Work

Neural speech decoding. Temporal-response and canonical-correlation methods linked continuous speech to brain activity (Crosse et al., 2016; de Cheveigné et al., 2018). Cortical (Mesgarani & Chang, 2012) and single-trial EEG (O'Sullivan et al., 2015) studies decoded attended speech. Deep approaches extended these mappings to learned representations (Puffay et al., 2023). We study reuse of a trained decoder when its output vocabulary changes.

Speech representations and tokenization. Speech pretraining offers contrastive and masked objectives (Chung et al., 2021), denoising (Chen et al., 2022), and latent-target prediction (Baevski et al., 2022). SUPERB evaluated reusable features through frozen encoders and lightweight heads (Yang et al., 2021). For discrete interfaces, learned vector quantization (van den Oord et al., 2017) underlies RepCodec's representation reconstruction (Huang et al., 2024) and SpeechTokenizer's semantic and acoustic encoding (Zhang et al., 2024). AudioLM used semantic and codec tokens for generation (Borsos et al., 2023). Decoder reuse across these representations additionally requires a correspondence between vocabularies.

Transfer across tokenizers. WECHSEL and FOCUS initialized embeddings for new vocabularies (Minixhofer et al., 2022; Dobler & de Melo, 2023); ZeTT predicted them with a hypernetwork (Minixhofer et al., 2024). Cross-tokenizer scoring evaluated likelihoods under different vocabularies (Phan et al., 2026), while distillation matched likelihoods (Minixhofer et al., 2025) or logit distributions through optimal transport (Boizard et al., 2025). TokenMapper learned speech-token translation from paired audio, taking discrete source sequences as input (Kozakov et al., 2026). We connect a neural decoder's uncertain predictions to a separate target prior. Our bridge averages posterior-to-frequency ratios across source states, preserving soft correspondence and neutral evidence while keeping the source decoder and target prior fixed.

Evidence fusion and prior fallback. Density ratios express support relative to reference distributions (Sugiyama et al., 2007; Kanamori et al., 2009), and products of experts combine multiplicative factors (Hinton, 2002). Density-ratio speech-recognition fusion combined language models within a shared vocabulary (McDermott et al., 2020). Uncertainty-aware methods adjusted fusion weights using predictive confidence (Chen et al., 2024; Cao et al., 2024); our weights are fixed per masking regime from development participants. A zero weight retains the target-prior prediction, unlike prediction rejection in selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

3 Neural Evidence Transport

The method expresses source predictions relative to reference frequencies, transfers them through an audio bridge, and uses the transferred evidence to update a frozen target prior.

3.1 Setting and Reference Distributions

Let $A \in \mathcal{A}$ and $B \in \mathcal{B}$ denote source and target tokens for aligned audio frames. Given a neural observation $X = x$, a participant's source decoder provides $q_A(a \mid x)$. The target prior $p_B(b \mid c)$ predicts a masked target token from visible target-token context $C = c$. Both tokenizers, the source decoder, and the target prior remain fixed during transfer. The vocabularies may differ in size and require no shared token labels.

An independent audio corpus supplies a reference joint distribution π_{AB}, with marginals π_A and π_B. These marginals describe corpus token frequencies; the contextual prior predicts tokens from the available context. Ordinary audio supplies the bridge and trains the target prior. We select fusion weights using target-token outcomes from development participants and apply them to confirmation participants with their own source-vocabulary decoders.

3.2 Extracting Neural Evidence

To measure support relative to baseline frequency, we define the source evidence ratio:

$$r_A(a; x) = \frac{q_A(a \mid x)}{\pi_A(a)}, \qquad q_A(a \mid x) = \pi_A(a) r_A(a; x). \tag{1}$$

Ratios above or below one increase or decrease support relative to baseline. When $q_A = \pi_A$, every ratio equals one, so transport leaves the target prior unchanged.

Each token is represented by an eight-dimensional code vector. We apply principal component analysis (PCA) to its tokenizer's audio-derived cluster centroids and normalize each projected vector to unit length. The source decoder regresses onto these vectors and returns a unit-normalized

prediction. For decoder prediction $\hat{h}(x)$ and source code vector v_a, both unit-normalized, we compute $L_A(a;x) = \exp\{0.5\langle \hat{h}(x), v_a \rangle\}$. The evidence scale is fixed across transfer directions. Normalizing $\pi_A(a)L_A(a;x)$ defines the source distribution $q_A(a \mid x)$, making L_A proportional to q_A/π_A. The shared normalization factor passes through linear transport and cancels in the final target prediction. We therefore transport L_A directly. Appendix A.1 specifies the construction, and Appendix D.12 verifies numerical equivalence.

3.3 Building the Audio Bridge

To estimate vocabulary correspondence, we tokenize the same audio frames with both models. Let N_{ab} count frames assigned to source token a and target token b. Additive smoothing gives

$$\widehat{\pi}_{AB}(a,b) = \frac{N_{ab} + \epsilon}{\sum_{a',b'} (N_{a'b'} + \epsilon)}, \qquad \epsilon = 10^{-4}. \tag{2}$$

The same smoothed joint table supports both directions. Reversing transfer exchanges the source and target roles and changes the conditional normalization. Five-fold segment cross-validation compares bridge prediction with a target unigram and a shuffled-source control (Appendix D.6).

3.4 Transporting Evidence Across Vocabularies

A target token can co-occur with several source tokens. The reverse channel $K(a \mid b) = \pi_{AB}(a,b)/\pi_B(b)$ gives their conditional frequencies. We use these frequencies to average the source evidence:

$$r_B(b;x) = \sum_{a \in \mathcal{A}} K(a \mid b) r_A(a;x) = \mathbb{E}_{\pi_{AB}}[r_A(A;x) \mid B = b]. \tag{3}$$

The weights sum to one for each target token. Appendix A gives the log-space implementation.

Proposition 1 (Neutrality, normalization, and contraction). Let π_{AB} have positive marginals, fix x, and let $q_A = \pi_A r_A$ be normalized. Propagating q_A through the reference channel $\pi(b \mid a) = \pi_{AB}(a,b)/\pi_A(a)$ gives

$$\bar{q}_B(b \mid x) = \sum_a \pi(b \mid a) q_A(a \mid x) = \pi_B(b) r_B(b;x). \tag{4}$$

The transported ratio satisfies

$$r_A \equiv 1 \;\Rightarrow\; r_B \equiv 1, \qquad \mathbb{E}_{\pi_B}[r_B] = 1, \qquad \chi^2(\bar{q}_B \parallel \pi_B) \le \chi^2(q_A \parallel \pi_A), \tag{5}$$

where $\chi^2(q \parallel \pi) = \sum_z (q(z) - \pi(z))^2/\pi(z)$ is Pearson chi-square divergence.

Equation (4) gives an equivalent implementation: propagate q_A through $\pi(b \mid a)$, then divide by π_B. The chi-square contraction follows from data processing through this stochastic channel (Sason & Verdú, 2016; Raginsky, 2016).

Proposition 2 (Exact loss of chi-square contrast). Under the same conditions,

$$\chi^2(q_A \parallel \pi_A) - \chi^2(\bar{q}_B \parallel \pi_B) = \mathbb{E}_{\pi_B}[\mathrm{Var}_{\pi_{AB}}(r_A(A;x) \mid B)]. \tag{6}$$

The bridge averages evidence from source tokens associated with each target token. Differences among their ratios are lost, while a common ratio is preserved. Equation (6) expresses the lost chi-square contrast as the expected within-target variance. Transport is lossless in this measure exactly when the ratio is almost surely constant conditional on the target token. Appendix C gives the proofs, and Appendix A bounds bridge-estimation error.

3.5 CONTEXT-DEPENDENT FUSION

The target prior uses visible tokens to predict the masked token. For transfer direction s, we update this prediction with the transported evidence:

$$q_B(b \mid c, x; s) = \frac{p_B(b \mid c)\, r_B(b; x)^{\lambda_s(c)}}{\sum_{b' \in \mathcal{B}} p_B(b' \mid c)\, r_B(b'; x)^{\lambda_s(c)}}. \tag{7}$$

The nonnegative weight $\lambda_s(c)$ depends on the masking regime; zero returns p_B exactly. Under the sufficient conditions in Appendix C.1, unit-weight fusion gives the Bayesian update of the target prior. We select weights on development participants to set the contribution of the fitted decoder and bridge.

Frequency correction is essential to this update. At unit weight, fusing the propagated posterior $\bar{q}_B = \pi_B r_B$ introduces an extra factor π_B. Even neutral evidence then gives a prediction proportional to $p_B(b \mid c)\pi_B(b)$. Using r_B removes this frequency-driven change.

Proposition 3 (Context risk and abstention). Fix the transfer direction and write $q_B(\cdot \mid c, x; \lambda)$ for Eq. (7) at weight λ. For positive evidence factors, let $\ell_b(x) = \log r_B(b; x)$, and let $R_c(\lambda)$ be the expected NLL conditional on $C = c$. Then

$$R_c'(\lambda) = \mathbb{E}[\mathbb{E}_{B' \sim q_B(\cdot \mid c, X; \lambda)}[\ell_{B'}(X)] - \ell_B(X) \;\mid\; C = c], \tag{8}$$

and

$$R_c''(\lambda) = \mathbb{E}[\mathrm{Var}_{B' \sim q_B(\cdot \mid c, X; \lambda)}[\ell_{B'}(X)] \;\mid\; C = c] \geq 0. \tag{9}$$

The risk is convex in λ. At zero, its derivative compares the prior-weighted log evidence with the log evidence assigned to the observed target. A negative derivative means that a small positive weight reduces risk; a nonnegative derivative makes zero an optimum over $\lambda \geq 0$. Opposite derivative signs in two contexts therefore preclude one positive weight from improving both locally (Appendix C.4).

For each direction and each of seven masking regimes, we select a weight by development NLL from $\{0, 0.25, 0.5, 1, 2, 4, 8\}$. At inference, the policy selects the weight for the transfer direction and masking regime. It is fixed before confirmation, together with the bridge, target prior, and evidence temperature. Each participant's source decoder remains fixed during transfer. Appendix D.4 gives the complete policy and selection procedure.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and participant splits. For MEG, LibriSpeech train-clean-100 provides the non-neural supervision (Panayotov et al., 2015). We independently cluster HuBERT layer 6 and Wav2Vec2 layer 9 into 256 states each. The bridge and priors use 106,962 aligned three-second segments; checkpoint selection uses dev-clean.

LibriBrain100 supplies 20 development and 11 confirmation participants (Mantegna et al., 2026). Each participant's source decoder is fitted on session 11 and evaluated on session 12. The groups are disjoint and hear shared speech. Development session-12 outcomes inform policy finalization, including zero-weight overrides. The resulting policies, comparators, masks, and decision rules remain fixed for confirmation (Appendices D.2 and D.4).

Models and comparators. Frozen target priors are four-layer bidirectional Transformers trained on ordinary audio (Appendix D.1). Source residual MLPs predict eight-dimensional, unit-normalized codes from lagged MEG inputs, with training-only standardization and PCA. Direct target-space MLP and Ridge ($\alpha = 1$) baselines share training examples, preprocessing, priors, masks, and evidence temperature. MLP architecture, optimization, and three-seed ensembles are also matched (Appendix D.3).

Contexts and statistical inference. We evaluate masked target-token completion using visible tokens from the stimulus audio as context. Seven regimes mask 50%, 75%, or 100% independently (IID), or 50% or 75% in a span or suffix. IID and span masks may reveal both sides; suffix predictions share a visible prefix without feeding back suffix tokens. Global fusion uses one development-selected weight. Transport and each direct baseline use regime-specific weights selected from the same grid (Appendix D.4).

The primary MEG endpoint is target-token NLL (Gneiting & Raftery, 2007), averaged equally across regimes within participants and then across participants. Wrong-time controls substitute evidence from the same participant, session, story, and recording at least 30 seconds away, keeping targets and masks fixed. Paired 95% intervals use 100,000 participant-bootstrap resamples.

Confirmation requires gains over the prior, global fusion, and wrong-time evidence in both directions, at least 75% retention of the strongest development-selected direct gain in each direction, and superiority over both direct families in at least one direction. Appendix D.5 specifies context thresholds and all decision rules.

EEG evaluation. OpenNeuro ds006434 (version 1.2.0) supplies 32-channel EEG from 28 listeners attending to one of two simultaneous audiobooks (Stoll et al., 2025). Disjoint groups of 14 develop and confirm the protocol. Development selects Wav2Vec2-to-HuBERT transfer, weight 8, and target-space Ridge before confirmation EEG is opened. The primary endpoint is 12-second attended-speaker accuracy.

For each candidate stream, we add target log frequencies to weighted transported log evidence. EEG uses the bridge marginal as its prior. We sum these scores over adjacent three-second windows and choose the higher-scoring stream. We also evaluate shorter decisions and 100-way retrieval, compare direct and source-native decoders, and apply timing and source-label controls. A paired opposite-attention substitution provides a consistency check (Appendix D.10).

4.2 Context-Dependent Evidence Fusion

A single fusion weight helped under complete masking but hurt when informative context was visible. On 20 development participants, HuBERT-to-Wav2Vec2 fusion at weight 2 reduced NLL by 0.05336 nat/token under complete masking but increased it by 0.00497 and 0.00541 under IID 50% and 75%. Wav2Vec2-to-HuBERT fusion reduced NLL by 0.04096 without visible context but increased it by 0.00679 and 0.00765 under IID 50% and 75% masking.

The final policies assigned zero to IID 50%, 0.25 to IID 75%, and 1 or 2 elsewhere. Development NLL improved over global fusion by 0.00379 nat/token for HuBERT-to-Wav2Vec2 and 0.00210 in reverse (Figure 2a). These policies were fixed for confirmation. The results connect the context-risk analysis to the chosen weights: strong context favors a smaller neural contribution, while missing context permits a larger update.

4.3 Decoder Reuse in MEG

Transport improved target-token prediction on all 11 held-out MEG participants. Mean NLL gains over the target prior were 0.02087 nat/token for HuBERT-to-Wav2Vec2 and 0.01990 in reverse. It also improved over global fusion, with positive paired effects in 9 of 11 participants for HuBERT-to-Wav2Vec2 and all 11 in reverse (Figure 2b).

IID 50% preserved the prior through zero weight. IID 75% produced gains of 0.00038 and 0.00022 nat/token, while complete masking produced the largest gains, 0.06400 and 0.04789. Mean NLL improved in all six nonzero-weight regimes in both directions (Figure 2c).

Wrong-time observations increased NLL in every participant in both directions (Table 1). With participant, session, recording, targets, and masks fixed, these paired effects link the gains to aligned neural observations.

Comparison with direct training. For the Wav2Vec2 target, transport outperformed both matched direct decoders in all 11 participants, with paired intervals excluding zero (Table 1; Figure 2d). Its mean NLL was 3.84129, compared with 3.85012 for MLP and 3.84888 for Ridge. For the HuBERT

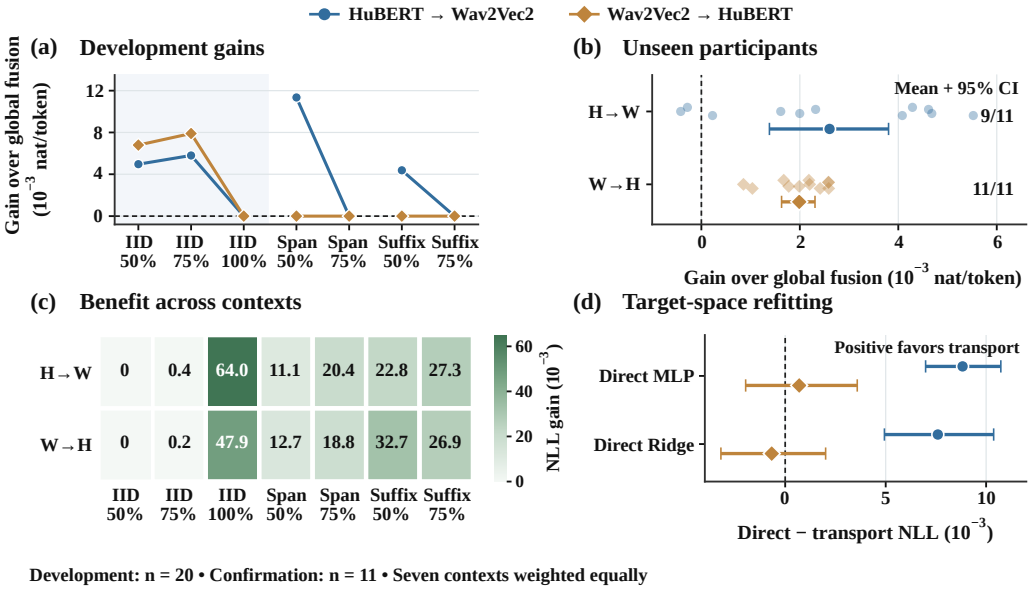


Figure 2: **Context-dependent evidence reuse in MEG.** Policies developed on 20 participants are fixed for 11 held-out participants. (a) Development NLL reductions over global fusion. (b) Confirmation reductions; faint points denote participants and counts denote positive effects. (c) Prior-minus-transport NLL by regime; IID 50% abstains. (d) Direct-minus-transport NLL for matched MLP and Ridge decoders. Positive differences favor transport. Solid points and error bars in (b,d) give means and paired 95% participant-bootstrap intervals. Summaries weight seven regimes equally. Units: 10^{-3} nat/token. H: HuBERT; W: Wav2Vec2.

Table 1: **Paired NLL differences on 11 MEG confirmation participants (nat/token).** Positive values favor context-dependent transport; brackets show paired 95% participant-bootstrap intervals. Transport reuses the source decoder, while direct MLP and Ridge receive target-space neural fits. Appendix D.11 gives absolute NLL.

NLL difference	HuBERT → Wav2Vec2	Wav2Vec2 → HuBERT
Global fusion − transport	0.00260 [0.00138, 0.00380]	0.00199 [0.00163, 0.00231]
Target prior − transport	0.02087 [0.01455, 0.02708]	0.01990 [0.01276, 0.02713]
Wrong-time − aligned evidence	0.04047 [0.03320, 0.04718]	0.04109 [0.03311, 0.04884]
Direct MLP − transport	0.00883 [0.00699, 0.01073]	0.00070 [−0.00197, 0.00358]
Direct Ridge − transport	0.00760 [0.00494, 0.01037]	−0.00067 [−0.00320, 0.00201]

target, transport reached 4.10543, compared with 4.10613 for MLP and 4.10476 for Ridge; both paired intervals included zero.

The HuBERT prior had NLL 4.12533. Transport improved on this prior by 0.01990 nat/token, retaining 97% of Ridge's improvement of 0.02058. Transport obtains this gain from the source-vocabulary decoder; direct MLP and Ridge use separate target-space fits. Appendix D.5 reports the prespecified retention test, and Appendix D.11 lists all absolute losses.

4.4 Attended-Speech Decoding in EEG

On 14 held-out EEG participants, Wav2Vec2-to-HuBERT transport achieved 60.6% attended-speaker accuracy over 12 seconds (95% CI [56.4, 65.6]); every point estimate exceeded 50% chance. Mean accuracy increased from 56.2% at three seconds to 58.3% at six seconds and 60.6% at 12 seconds (Figure 3a).

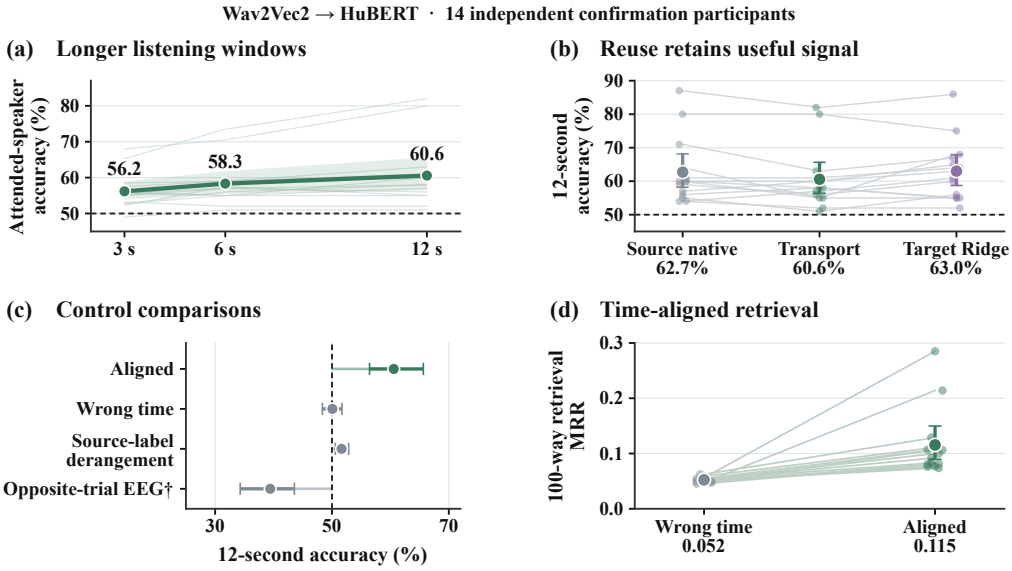


Figure 3: **Independent EEG confirmation in 14 held-out participants.** Wav2Vec2-to-HuBERT transport is evaluated for (a) attended-speaker accuracy across durations, (b) 12-second accuracy against source-native and target-space Ridge decoders, (c) timing/source-label controls and paired opposite-trial EEG, and (d) 100-way retrieval MRR. In (c), † substitutes evidence from the paired opposite-attention trial while retaining the original label; symmetric pairing makes its accuracy complementary to aligned accuracy. Dashed lines mark 50% chance. Faint points and lines represent participants; solid points, error bars, and shaded bands show means and 95% participant-bootstrap intervals.

Target-space Ridge and source-native decoding reached 63.0% and 62.7% accuracy (Figure 3b). Transport retained 81% and 83% of their respective gains over chance (Appendix D.10).

Substituting EEG evidence from the paired opposite-attention presentation yielded 39.4% against the original presentation's label. Symmetric evaluation of both presentations makes this the complement of aligned accuracy, a consistency check. Source-label derangement reduced accuracy by 8.93 percentage points (95% CI [5.43, 13.07]; Figure 3c), linking performance to token correspondence.

In 100-way retrieval, aligned mean reciprocal rank (MRR) was 0.115, versus 0.052 with wrong-time evidence and 0.060 with deranged labels. The paired aligned-over-wrong-time gain was 0.0631 (95% CI [0.0378, 0.0968]; Figure 3d). Wrong-time evidence reduced recall@5 from 0.159 to 0.053.

EEG applies the transferred evidence to offline selection among known speech candidates, using the bridge marginal as its prior. The two-speaker task identifies the attended stream. In 100-way retrieval, the bank contains both streams from every tested chapter and time span; the attended stream at the matching chapter and span is the correct candidate. Aligned, wrong-time, and source-label-deranged evidence score the same bank. The controls test whether candidate scores depend on the matching neural interval and the source tokens' correspondence with the bridge.

4.5 Operator and Bridge Ablations

The audio bridge provides predictive correspondence between the vocabularies. Under five-fold segment cross-validation, its target-token NLL was 2.3805 nat/token for HuBERT-to-Wav2Vec2 and 2.5079 in reverse, compared with target-unigram NLLs of 5.3557 and 5.4831, respectively (Appendix D.6). This evaluates vocabulary correspondence on ordinary audio alone. The neural experiments assess its value when source information comes from a participant's decoded evidence.

Frequency correction and soft correspondence each contribute to target-token prediction. Post-confirmatory diagnostics on the same 11 MEG participants replaced either density ratios with raw

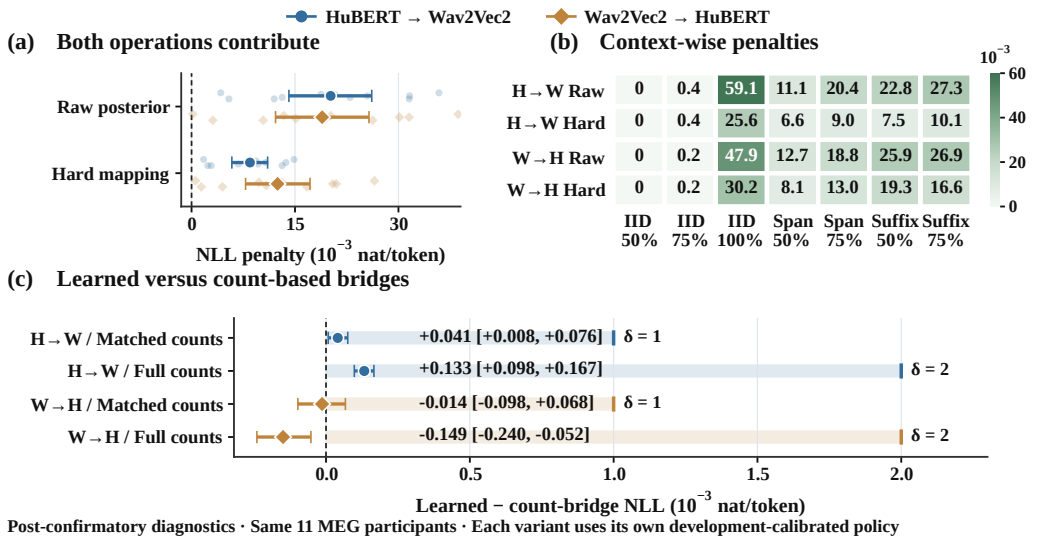


Figure 4: **Frequency correction, soft mapping, and bridge estimation.** Post-confirmatory diagnostics reuse 11 MEG participants with separately development-selected policies. (a) Raw-posterior and hard-mapping NLL penalties relative to the full method; (b) penalties by regime. Positive values favor the full method. (c) Learned-minus-count NLL with matched 4,000-segment and full-corpus count bridges. Upper 95% bounds fall below noninferiority margins $\delta = 1$ and $\delta = 2$, respectively. Units: 10^{-3} nat/token. Error bars give paired 95% participant-bootstrap intervals.

posteriors or the soft bridge with a hard mapping. Each variant received its own development-selected policy from the same weight grid.

Raw-posterior transport increased NLL by 0.02016 and 0.01894 nat/token in the two directions; hard mapping increased it by 0.00847 and 0.01247. All four intervals excluded zero, and every participant favored the full method (Figure 4a–b; Appendix D.7). The raw-posterior variant introduces target frequency a second time during fusion. The hard map discards alternative correspondences. Both penalties persist after retuning fusion strength.

We also tested an audio-only learned bridge while keeping the source decoder, frequency correction, and target prior fixed. A count bridge trained on the same 4,000 audio segments controls the amount of supervision; the full-corpus count bridge provides the main-model comparison. Each estimator received separately development-selected fusion weights.

The learned bridge met noninferiority margins of 0.001 and 0.002 nat/token against the matched-subset and full-corpus count bridges in both directions. The largest upper 95% bound on learned-minus-count NLL was 0.000167 (Figure 4c). It retained gains over the prior and global fusion, and aligned evidence outperformed wrong-time evidence (Appendix D.8). These results support both count-based and learned correspondence within the same transport procedure.

5 Conclusion

Neural Evidence Transport reuses an existing brain decoder with another speech vocabulary through an audio-derived soft bridge. Frequency correction expresses predictions relative to baseline, and context-dependent fusion controls their contribution to the target prior. We establish neutrality and an exact identity for the loss of chi-square contrast. MEG and EEG evaluations demonstrate useful target predictions without an additional participant-specific target-space fit.

The evaluations cover two 256-state vocabularies, with shared speech in MEG and held-out chapters in one English two-speaker EEG corpus. Next steps include unseen MEG stimuli, additional vocabularies, and measurement of practical adaptation costs.

AI USE STATEMENT

Generative AI assisted with writing and polishing, literature retrieval and discovery, and implementing research ideas. The authors are responsible for reviewing AI-assisted content and verifying the final work.

ETHICS STATEMENT

We reanalyze two public, pseudonymized neural-recording datasets. LibriBrain100 reports consent for research data sharing and University of Oxford ethics approval (R90053/RE003; Mantegna et al., 2026). The ds006434 study reports written informed consent and University of Rochester ethics approval (protocol 1227; Stoll et al., 2025). We recruited no participants and introduced no interventions. Neural decoding raises privacy concerns; our group-level results do not establish covert access to individual thoughts.

REPRODUCIBILITY STATEMENT

Appendices A–D provide proofs, preprocessing and optimization settings, frozen policies, and participant-level statistical procedures. The anonymous supplement provides frozen protocols, code, tests, evaluation records, and reconstruction scripts for transport, matched direct baselines, and learned-bridge diagnostics. The separate EEG package provides the frozen protocol, disjoint participant splits, evaluations, and decision records. It documents recalculation of all reported EEG effects from participant-level records in a separate computing environment. The implementation-audit addendum provides original code-vector tables, scoring and candidate reconstruction records, and absolute MEG NLL anchors.

REFERENCES

Alexei Baevski, Henry 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, pp. 12449–12460, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: The universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=buBxXiG09A.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628, 2024. URL https://proceedings.mlr.press/v235/cao24c.html.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Ensiong Chng, and Chao-Han Huck Yang. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0231de0eff264c0639a4c43728b8b55b-Abstract-Conference.html.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian,

Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 244–250, 2021. doi: 10.1109/ASRU51503.2021.9688253.

Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016. doi: 10.3389/fnhum.2016.00604.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, 2025. doi: 10.1038/s41467-025-65499-0.

Alain de Cheveigné, Daniel D. E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund C. Lalor. Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216, 2018. doi: 10.1016/j.neuroimage.2018.01.033.

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

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13440–13454, 2023. doi: 10.18653/v1/2023.emnlp-main.829.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete encoding of EEG waves for EEG-to-text translation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

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. doi: 10.1109/TASLP.2021.3122291.

Zhichao Huang, Chutong Meng, and Tom Ko. RepCodec: A speech representation codec for speech tokenization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5777–5790, 2024. doi: 10.18653/v1/2024.acl-long.314.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27005–27020, 2025.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Tal Kozakov, Tal Rosenwein, and Eliya Nachmani. TokenMapper: A step toward interoperable speech token translation. *arXiv preprint arXiv:2609.12563*, 2026. doi: 10.48550/arXiv.2609.12563.

Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. BrainECHO: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2762–2778, 2025. doi: 10.18653/v1/2025.findings-acl.142.

Francesco Mantegna, Dulhan Jayalath, Gereon Elvers, Tasha Kim, Benjamin Ballyk, Alex Fung, SungJun Cho, Teyun Kwon, Luisa Kurth, Miran Ozdogan, Gilad Landau, Pratik Somaiya, Natalie Voets, Mark Woolrich, and Oiwi Parker Jones. LibriBrain100: One hundred hours of broad and deep MEG data for neural speech decoding at scale. *arXiv preprint arXiv:2608.25204*, 2026. doi: 10.48550/arXiv.2608.25204.

Erik McDermott, Hasim Sak, and Ehsan Variani. A density ratio approach to language model fusion in end-to-end automatic speech recognition, 2020. URL `https://arxiv.org/abs/2002.11268`. Presented at the 2019 IEEE Automatic Speech Recognition and Understanding Workshop.

Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012. doi: 10.1038/nature11020.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3992–4006, 2022. doi: 10.18653/v1/2022.naacl-main.293.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1484.

Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal cross-tokenizer distillation via approximate likelihood matching. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2653.

James A. O'Sullivan, Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C. Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 25(7):1697–1706, 2015. doi: 10.1093/cercor/bht355.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Truong Buu Phan, Ashish Khisti, and Karen Ullrich. Cross-tokenizer likelihood scoring algorithms for language model distillation. In *International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/hash/40d50fe20d5e6922dbb52d7da6de03f1-Abstract-Conference.html`.

Corentin Puffay, Bernd Accou, Lies Bollens, Mohammad Jalilpour Monesi, Jonas Vanthornhout, Hugo Van Hamme, and Tom Francart. Relating EEG to continuous speech using deep neural networks: A review. *Journal of Neural Engineering*, 20(4):041003, 2023. doi: 10.1088/1741-2552/ace73f.

Maxim Raginsky. Strong data processing inequalities and Φ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151.

Thomas J. Stoll, Nathan D. Vandjelovic, Melissa J. Polonenko, Nadja R. S. Li, Adrian K. C. Lee, and Ross K. Maddox. The auditory brainstem response to natural speech is not affected by selective attention. *PLOS Biology*, 23(10):e3003407, 2025. doi: 10.1371/journal.pbio.3003407.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal PERformance benchmark. In *Interspeech 2021*, pp. 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/86d1ab582afb247ccaa84bec4a7e24f7-Abstract-Conference.html.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-IN: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A Implementation and Bridge Stability

A.1 Source Evidence Factor

Code geometry. Each tokenizer has its own 256 audio-derived K-means centroids in $\mathbb{R}^{768}$. We fit unwhitened PCA with full singular-value decomposition to that centroid matrix, center across the 256 equally weighted centroids, retain eight components, and normalize each projected row to unit Euclidean norm. There is no feature standardization or pre-PCA centroid normalization. With centroid μ_a, their unweighted mean $\bar{\mu}$, and loading matrix $W_8 \in \mathbb{R}^{768\times 8}$,

$$\widetilde{v}_a = (\mu_a - \bar{\mu})W_8, \qquad v_a = \frac{\widetilde{v}_a}{\|\widetilde{v}_a\|_2}.$$

Token ID a indexes row a of the resulting table. Each PCA uses only its own tokenizer's centroids, without the other vocabulary or neural data. Eight dimensions were fixed as the common regression interface before held-out participant confirmation; dimensionality was not tuned in this evaluation. The same vectors define neural regression targets and evidence templates.

The encoders are `facebook/hubert-base-ls960` and `facebook/wav2vec2-base-960h`; their exact revisions are listed in the supplementary code-geometry manifest. We use `hidden_states[6]` and `hidden_states[9]`, respectively: index zero denotes the Transformer input and index k the output after k blocks. Each MiniBatchKMeans quantizer uses Euclidean distance on 96,000 unstandardized features, comprising 24 frames sampled without replacement from each of 4,000 three-second LibriSpeech train-clean-100 segments. Segment/frame sampling and K-means use seed 20270902. K-means uses batch size 8192, three initializations, at most 100 iterations, and reassignment ratio 0.01. These 4,000 segments are a subset of the full bridge corpus; only the centroids enter PCA.

Saved decoders use the original 256×8 vector tables, preserved with their construction inputs in the supplementary code-geometry package. The original generator did not save the PCA transformer; labeled reconstructions of its means and loadings reproduce both tables bit for bit.

The source decoder predicts a unit-normalized code-geometry vector $\widehat{h}(x) \in \mathbb{R}^8$. For each source token a, we compare this prediction with its code vector v_a and form the positive evidence factor

$$L_A(a; x) = \exp(\beta\langle\widehat{h}(x), v_a\rangle), \qquad \beta = 0.5. \tag{A.1}$$

Both codebooks use this construction, with source vectors chosen for the transfer direction. The source distribution is $q_A(a \mid x) \propto \pi_A(a)L_A(a; x)$, so its ratio to π_A is obtained by normalizing L_A. Let

$$Z_A(x) = \sum_{a\in\mathcal{A}} \pi_A(a)L_A(a; x), \qquad r_A(a; x) = \frac{L_A(a; x)}{Z_A(x)}. \tag{A.2}$$

The constant $Z_A(x)$ is shared by all source tokens for the same observation. Because transport is linear, the same constant is shared by all target tokens after transport. It therefore cancels when the fused target prediction is normalized. We can compute with L_A directly.

A.2 Log-Space Transport and Fusion

We denote the transported, unnormalized factor by $L_B(b; x)$. Using aligned audio counts N_{ab} and smoothing $\epsilon = 10^{-4}$, we evaluate it in log space:

$$\begin{aligned}\log L_B(b; x) = \mathrm{LSE}_a[\log(N_{ab} + \epsilon) + \beta\langle\widehat{h}(x), v_a\rangle] \\ - \log \sum_a (N_{ab} + \epsilon),\end{aligned} \tag{A.3}$$

where $\mathrm{LSE}_a[z_a] = \log \sum_a \exp(z_a)$.

Equation (A.3) computes the conditional average of L_A under the count-based reverse channel. Dividing by the target token's total smoothed count removes its frequency as a separate multiplicative factor. A frequent target token receives stronger evidence only when its associated source tokens carry stronger evidence.

The normalized ratio from the estimated bridge is $\widehat{r}_B(b;x) = L_B(b;x)/Z_A(x)$. Substituting this expression into the fusion rule gives

$$q_B(b \mid c, x; s) = \mathrm{softmax}_b[\log p_B(b \mid c) + \lambda_s(c) \log L_B(b;x)]. \tag{A.4}$$

The omitted term, $-\lambda_s(c) \log Z_A(x)$, is identical for every target token and cancels in the softmax. Thus, using the unnormalized factors produces the same target prediction as explicitly normalizing the source ratio before transport.

Appendix D.12 reports numerical equivalence and neutrality checks.

A.3 Stability to Bridge Estimation

The bridge is estimated from finite audio data. The following bound describes how an error in its conditional probabilities affects the transported evidence.

Proposition 4 (Stability to an estimated bridge). Let $K(\cdot \mid b)$ and $\widehat{K}(\cdot \mid b)$ be the population and estimated reverse channels. For a fixed observation x, suppose $0 \le r_A(a;x) \le R_x$ for every source token. Define r_B and $\widehat{r}_B$ by averaging the same source ratio under K and $\widehat{K}$, respectively. Then, for every target token b,

$$| \widehat{r}_B(b;x) - r_B(b;x) | \le R_x \| \widehat{K}(\cdot \mid b) - K(\cdot \mid b) \|_1. \tag{A.5}$$

Here, the ℓ_1 term sums the absolute errors in the source-token probabilities conditional on b. The proof appears in Appendix C.3.

The bound separates the magnitude of the source evidence from the accuracy of the audio bridge. More accurate audio estimates reduce the channel-error term.

B Supervision and Frozen Policies

B.1 Supervision by Component

The audio bridge, target prior, and fusion policy use different sources of supervision (Table 2). Ordinary audio trains the bridge and prior. Target-token prediction outcomes from the 20 MEG development participants determine the fusion weights. For each confirmation participant, transport reuses the source-vocabulary neural decoder. The direct MLP and Ridge comparators receive a separate participant-specific fit using target-vocabulary labels.

Table 2: **Supervision by component in the MEG study.** The target-token outcome column refers to outcomes used for neural fitting or policy selection. A source-vocabulary neural fit is required for confirmation transport.

Component	Neural supervision	Target-token outcomes used for fitting or selection	Target-space fit for a new participant
Target prior and audio bridge	None; ordinary audio only	No	No
Context policy: seven weights per direction	20 development participants	Yes	No
Transport at confirmation	Participant's source-vocabulary decoder	No	No
Direct MLP and Ridge comparators	Participant's target-vocabulary labels	Yes	Yes

B.2 Frozen Transport Policies

The policy assigns one development-selected weight to each transfer direction and masking regime. At application, it reads only these two inputs. The zero weights under IID 50% leave the target prior unchanged. Weights, bridge, target prior, and evidence temperature remain fixed across confirmation participants; each participant's source decoder is frozen during transfer. Appendix D.4 gives the selection procedure and the complete transport and direct-comparator policies in Table 4.

C Proofs

C.1 Proof of Proposition 1

Fix a neural observation x. Let π_{AB} have positive marginals, and define the forward and reverse reference channels as

$$\pi(b \mid a) = \frac{\pi_{AB}(a,b)}{\pi_A(a)}, \qquad K(a \mid b) = \frac{\pi_{AB}(a,b)}{\pi_B(b)}.$$

Substituting $q_A(a \mid x) = \pi_A(a) r_A(a;x)$ into the forward-channel prediction gives

$$\begin{aligned} \bar{q}_B(b \mid x) &= \sum_a \pi(b \mid a) q_A(a \mid x) \\ &= \sum_a \pi_{AB}(a,b) r_A(a;x) \\ &= \pi_B(b) \sum_a K(a \mid b) r_A(a;x) \\ &= \pi_B(b)\, \mathbb{E}_{\pi_{AB}}[r_A(A;x) \mid B = b] \\ &= \pi_B(b) r_B(b;x). \end{aligned} \tag{C.1}$$

This establishes the ratio push-forward identity.

If $r_A \equiv 1$, its conditional expectation is also one, so $r_B \equiv 1$. Normalization follows from iterated expectation and the normalization of q_A:

$$\mathbb{E}_{\pi_B}[r_B(B;x)] = \mathbb{E}_{\pi_A}[r_A(A;x)] = \sum_a q_A(a \mid x) = 1. \tag{C.2}$$

To prove contraction, apply conditional Jensen's inequality to the squared deviation from one:

$$\begin{aligned} \chi^2(\bar{q}_B \parallel \pi_B) &= \mathbb{E}_{\pi_B}[(r_B(B;x) - 1)^2] \\ &= \mathbb{E}_{\pi_B}[(\mathbb{E}_{\pi_{AB}}[r_A(A;x) - 1 \mid B])^2] \\ &\le \mathbb{E}_{\pi_{AB}}[(r_A(A;x) - 1)^2] \\ &= \chi^2(q_A \parallel \pi_A). \end{aligned} \tag{C.3}$$

Thus, transport preserves neutrality and normalization and cannot increase chi-square contrast relative to the reference distribution. □

Relation to the target posterior. Let $p(a,b,x,c)$ denote the joint evaluation distribution. Suppose $\pi_{AB} = p_{AB}$, $q_A(a \mid x) = p(a \mid x)$, and $X \perp B \mid A$. Equation (4) then gives

$$\bar{q}_B(b \mid x) = \sum_a p(b \mid a) p(a \mid x) = \sum_a p(b \mid a, x) p(a \mid x) = p(b \mid x).$$

Consequently, $r_B(b;x) = p(b \mid x)/p(b)$. If, in addition, $p_B(b \mid c) = p(b \mid c)$ and $X \perp C \mid B$, unit-weight fusion in Eq. (7) yields $p(b \mid c, x)$. These sufficient conditions connect the transport operator to Bayesian updating.

C.2 Proof of Proposition 2

Under the reference joint π_{AB}, define $Y = r_A(A; x) - 1$. Its second moment is $\chi^2(q_A \parallel \pi_A)$, and its conditional mean is

$$\mathbb{E}_{\pi_{AB}}[Y \mid B] = r_B(B; x) - 1.$$

Decomposing the second moment into the expected conditional variance and the squared conditional mean gives

$$\begin{aligned} \chi^2(q_A \parallel \pi_A) &= \mathbb{E}_{\pi_{AB}}[Y^2] \\ &= \mathbb{E}_{\pi_B}[\mathrm{Var}_{\pi_{AB}}(Y \mid B)] + \mathbb{E}_{\pi_B}[(\mathbb{E}_{\pi_{AB}}[Y \mid B])^2] \\ &= \mathbb{E}_{\pi_B}[\mathrm{Var}_{\pi_{AB}}(r_A(A; x) \mid B)] + \chi^2(\bar{q}_B \parallel \pi_B). \end{aligned} \tag{C.4}$$

Subtracting the target divergence proves Eq. (6). The gap vanishes exactly when the conditional variance is zero almost surely. For the fixed observation x, this means that the source tokens associated with each target token carry the same evidence ratio. □

C.3 Proof of Proposition 4

Fix a target token b and observation x. Let

$$\Delta_a = \widehat{K}(a \mid b) - K(a \mid b),$$

and suppose $0 \le r_A(a; x) \le R_x$ for every source token. The difference between the two transported ratios satisfies

$$\begin{aligned} \mid \widehat{r}_B(b; x) - r_B(b; x) \mid &= \mid \sum_a \Delta_a r_A(a; x) \mid \\ &\le \sum_a \mid \Delta_a \mid r_A(a; x) \\ &\le R_x \sum_a \mid \Delta_a \mid \\ &= R_x \| \widehat{K}(\cdot \mid b) - K(\cdot \mid b) \|_1. \end{aligned} \tag{C.5}$$

The first inequality is the triangle inequality; the second uses the upper bound on the source ratio. This proves Eq. (A.5). □

Both channels sum to one. When the minimum and maximum source ratios are known, a range-based total-variation bound can further tighten the result.

C.4 Proof of Proposition 3

Fix the transfer direction and context c. For the positive evidence factors used in fusion, write $\ell_b(x) = \log r_B(b; x)$. Define the partition function

$$Z_\lambda(c, x) = \sum_{b' \in \mathcal{B}} p_B(b' \mid c) \exp(\lambda \ell_{b'}(x)). \tag{C.6}$$

For an observation x and target label b, the negative log-likelihood under Eq. (7) is

$$\mathcal{L}_\lambda(b, c, x) = -\log p_B(b \mid c) - \lambda \ell_b(x) + \log Z_\lambda(c, x). \tag{C.7}$$

Differentiating with respect to λ gives

$$\frac{\partial \mathcal{L}_\lambda}{\partial \lambda} = -\ell_b(x) + \frac{\sum_{b'} p_B(b' \mid c) e^{\lambda \ell_{b'}(x)} \ell_{b'}(x)}{Z_\lambda(c,x)} = -\ell_b(x) + \mathbb{E}_{B' \sim q_B(\cdot|c,x;\lambda)}[\ell_{B'}(x)]. \tag{C.8}$$

Differentiating again yields

$$\frac{\partial^2 \mathcal{L}_\lambda}{\partial \lambda^2} = \mathbb{E}_{B' \sim q_B(\cdot|c,x;\lambda)}[\ell_{B'}(x)^2] - (\mathbb{E}_{B' \sim q_B(\cdot|c,x;\lambda)}[\ell_{B'}(x)])^2 = \mathrm{Var}_{B' \sim q_B(\cdot|c,x;\lambda)}[\ell_{B'}(x)] \geq 0. \tag{C.9}$$

Taking expectations conditional on $C = c$ gives the risk derivatives in Eqs. (8)–(9). The nonnegative second derivative establishes that $R_c(\lambda)$ is convex. Consequently, $R_c'(0) \geq 0$ makes the left boundary $\lambda = 0$ a risk minimizer over $\lambda \geq 0$. At this weight, the fused prediction equals the target prior exactly.

If two contexts have derivatives of opposite signs at zero, a sufficiently small positive weight decreases risk in one and increases it in the other. A single positive weight therefore cannot improve both contexts locally. □

D Experimental Methods, Evidence, and Reproducibility

Sections D.1–D.5 specify the MEG methods and inference procedure. Sections D.6–D.11 report supporting evaluations and the EEG protocol. Section D.12 describes implementation checks and reproducibility records.

D.1 Target-Prior Training and Evaluation

Training uses 106,962 three-second segments (15,937,338 frames); checkpoint selection uses 5,067 dev-clean segments. Each target prior is a four-layer bidirectional Transformer with width 256 and eight attention heads. It predicts masked tokens within a 129-frame window and is trained on ordinary audio for 12,000 AdamW steps. The selected checkpoints remain fixed during neural evaluation.

Both priors achieve lower NLL than the training-corpus unigram and uniform baselines at all three masking ratios on LibriSpeech dev-clean (Table 3). These comparisons evaluate the audio-trained priors independently of the neural decoders.

Table 3: **Target-prior NLL on LibriSpeech dev-clean.** Mask percentages indicate missing tokens. Values are in nat/token; lower is better.

Target vocabulary	Mask	Target prior	Unigram	Uniform
HuBERT	50%	**1.1771**	5.4938	5.5452
HuBERT	75%	**1.9483**	5.4942	5.5452
HuBERT	100%	**5.3379**	5.4941	5.5452
Wav2Vec2	50%	**0.9845**	5.3647	5.5452
Wav2Vec2	75%	**1.6707**	5.3645	5.5452
Wav2Vec2	100%	**5.2779**	5.3645	5.5452

D.2 Participant Split and MEG Preprocessing

The MEG development population comprises LibriBrain100 participants 12–28 and 30–32. Participant 29 failed the fixed alignment and sensor-geometry checks before development outcomes were analyzed. Participants 1–11 form the confirmation population. All 11 passed checks for file integrity, monotonic timestamps, common recording duration, sensor geometry, and training-only PCA. No participant was excluded or replaced based on prediction outcomes.

Session 11 provides 90 consecutive, non-overlapping three-second windows from the first 270 seconds. The first 70 form the neural-training partition; the final 20 support policy calibration only for development participants. Session 12 provides 90 evaluation windows over the same duration. The final 20 session-11 windows remain unused for confirmation participants.

MEG is aligned to audio event anchors, shifted by a fixed 150-ms neural-response delay, and resampled to 100 Hz. Standardization and 64-component PCA are fitted on the 70-window training partition. Downstream predictions use these stored transformations without refitting.

For MEG, the wrong-time donor is selected deterministically from the same participant, session, and recording, with a separation of at least 30 seconds. Target tokens and masks remain fixed. This changes the alignment of the neural observation while preserving the evaluated stimulus distribution.

D.3 Neural Optimization and Direct Comparators

The MEG decoder receives 320-dimensional inputs formed by concatenating the 64 PCA components at lags of 0, 50, 100, 200, and 300 ms. The residual MLP uses layer normalization, two 256-unit GELU layers, dropout of 0.1, an eight-dimensional readout, and a linear skip connection from the input.

Training minimizes cosine distance with AdamW, using a learning rate of 3×10^{-4}, weight decay of 10^{-4}, and gradient clipping at 1.0. A fixed participant-specific split assigns 56 windows to fitting and 14 to early stopping. Training runs for at most 80 epochs with patience 12. Predictions from seeds 17, 29, and 41 are individually normalized, averaged, and normalized again.

Ridge uses $\alpha = 1$ on the same standardized, lagged inputs. Direct models predict the target codebook's eight-dimensional geometry. Both direct baselines use the same frozen target-vector table for regression and evidence scoring. Reversing direction swaps the source table and source-decoder predictions together. The centroid PCA in Appendix A.1 is distinct from the 64-dimensional PCA of MEG inputs. For transport, the same model family is fitted in the source vocabulary and its output passes through the audio bridge (Appendix A.2). Training examples, preprocessing, masks, target priors, evidence temperature, and participant aggregation are matched across comparisons.

D.4 Complete Frozen Policies

Candidate weights minimize development NLL on the final 20 session-11 windows over the grid $\{0, 0.25, 0.5, 1, 2, 4, 8\}$. We then inspect development session-12 outcomes and finalize the policies before accessing confirmation recordings. For Wav2Vec2-to-HuBERT under IID 50%, calibration selected 0.25, but development session 12 showed a small NLL increase. We set this weight to zero before confirmation. Table 4 gives the final transport policies and the separately selected direct-comparator policies.

Table 4: **Complete frozen fusion policies.** H and W denote HuBERT and Wav2Vec2. I, C, and S denote IID, contiguous-span, and suffix masking; subscripts indicate the percentage of missing tokens. "Global" is the single weight used across all regimes.

Model	I_{50}	I_{75}	I_{100}	C_{50}	C_{75}	S_{50}	S_{75}	**Global**
H $\rightarrow$ W transport	0	0.25	2	1	2	1	2	2
W $\rightarrow$ H transport	0	0.25	2	2	2	2	2	2
Direct W MLP	0	0	2	1	1	1	1	1
Direct W Ridge	0	0	2	1	1	1	1	1
Direct H MLP	0.25	0.25	2	2	2	2	2	2
Direct H Ridge	0.25	0.25	2	2	2	2	2	2

D.5 Prospective Criteria and Statistical Inference

The MEG confirmation decision has five requirements. At least 10 participants who pass structural checks must complete both transfer directions and all four direct baselines. In each direction,

the paired global-minus-context-dependent NLL interval must have a positive lower bound. The prior-minus-transport and wrong-time-minus-aligned intervals must also have positive lower bounds.

Context-level checks require a nonnegative pooled gain across IID 50% and IID 75%. No nonzero-weight regime may lose more than 0.001 nat/token, and at least five regimes must improve. Finally, both directions must retain at least 75% of the strongest development-selected direct-decoder gain, and at least one direction must outperform both direct model families. All 12 instantiated checks pass.

For each participant, we average token losses within each regime and then weight the seven regimes equally. Confidence intervals use 100,000 resamples of the 11 paired participant effects, with the 2.5th and 97.5th percentiles as interval bounds. The participant is the statistical unit. The margins above the 75%-retention threshold are 0.01091 nat/token (95% CI [0.00746, 0.01443]) for HuBERT-to-Wav2Vec2 and 0.00447 (95% CI [0.00222, 0.00680]) for Wav2Vec2-to-HuBERT.

D.6 Held-Out Audio-Bridge Prediction

We evaluate the count bridge by five-fold cross-validation over all 106,962 aligned LibriSpeech segments. Each fold fits the conditional table on four partitions and predicts target tokens in the fifth. Every fitted fold covers all source and target states. We compare with a target unigram and a shuffled-source control. For this control, the fitted table remains fixed while source segments in the held-out partition are permuted relative to the unchanged target segments.

Aligned source tokens reduce target-token NLL relative to both controls (Table 5). Confidence intervals use 20,000 bootstrap resamples over 1,024 deterministic analysis blocks. Both NLL reductions are positive in every block. This evaluation measures the predictive correspondence between the audio tokenizers; the neural experiments measure how much decoded evidence can use that correspondence.

Table 5: **Audio-bridge NLL under five-fold segment cross-validation.** Values are in nat/token; lower is better. "Paired conditional" predicts target tokens from aligned source tokens.

Direction	Paired conditional	Target unigram	Shuffled test source
HuBERT → Wav2Vec2	**2.3805**	5.3557	14.1713
Wav2Vec2 → HuBERT	**2.5079**	5.4831	14.2978

D.7 Operator Controls on Target-Token Prediction

This post-confirmatory diagnostic evaluates two replacements on the same 11 MEG confirmation participants and target-token NLL endpoint. The raw-posterior variant retains the soft bridge but omits frequency correction. The hard-bridge variant retains density ratios but replaces the soft correspondence. Each variant receives its own development-only search over the same seven candidate weights used by the full method. Confirmation losses are not used to select weights or zero-weight overrides.

Both replacements increase NLL in every participant in both directions. All four paired confidence intervals exclude zero (Table 6).

Table 6: **NLL penalties from replacing the transport operator.** Values are in nat/token. Positive differences favor the full density-ratio/soft-bridge method. Brackets give paired 95% participant-bootstrap intervals.

Direction	Raw posterior − full	Hard bridge − full	Participants favoring full
HuBERT → Wav2Vec2	0.02016 [0.01411, 0.02612]	0.00847 [0.00583, 0.01102]	11/11 in both comparisons
Wav2Vec2 → HuBERT	0.01894 [0.01218, 0.02573]	0.01247 [0.00783, 0.01717]	11/11 in both comparisons

D.8 Learned Bridge with Matched Audio Supervision

We fixed this diagnostic after the prospective MEG evaluation and before computing the diagnostic's seven-regime NLL endpoint. The learned estimator predicts a target token from a source-token embedding and the fixed source-code geometry. Training uses 4,000 ordinary-audio segments containing 516,000 frames; model selection uses 1,000 held-out audio segments. A matched count bridge uses the same 4,000 training segments. Both fitting and model selection use audio alone.

Each estimator receives separately selected global and context-dependent weights from the 20 development participants. These weights are applied unchanged to the original 11 confirmation participants. The diagnostic therefore evaluates a change of bridge estimator using the existing confirmation population.

The learned bridge improves over the target prior and global fusion in both directions, and aligned observations outperform wrong-time observations. It also meets the noninferiority margins of 0.001 nat/token against matched counts and 0.002 nat/token against full-data counts (Table 7).

Table 7: **Post-confirmatory learned-bridge results.** NLL differences are in 10^{-3} nat/token, with paired 95% participant-bootstrap intervals. The first three rows compare predictions using the learned bridge; positive values favor context-dependent or aligned transport. In the two learned-minus-count rows, positive values indicate higher NLL for the learned bridge. The final row reports participant counts.

Comparison	**HuBERT → Wav2Vec2**	**Wav2Vec2 → HuBERT**
Global fusion − context-dependent fusion	2.643 [1.421, 3.838]	1.982 [1.615, 2.308]
Target prior − transport	20.738 [14.464, 26.915]	20.053 [12.902, 27.222]
Wrong-time − aligned evidence	40.283 [33.117, 47.033]	41.151 [33.070, 48.890]
Learned bridge − matched count bridge	0.041 [0.008, 0.076]	−0.014 [−0.098, 0.068]
Learned bridge − full-data count bridge	0.133 [0.098, 0.167]	−0.149 [−0.240, −0.052]
Participants benefiting over the prior	11/11	10/11

D.9 Complete Context-Level Confirmation Results

Table 8 reports the mean target-prior-minus-transport NLL for every masking regime on the 11 MEG confirmation participants. Both directions improve mean NLL in all six nonzero-weight regimes. IID 50% uses zero weight; its small numerical residual reflects floating-point evaluation of the unchanged prior.

Table 8: **Context-level gains over the target prior.** Values are in 10^{-3} nat/token. Positive values favor transport. Mask abbreviations follow Table 4.

Direction	I_{50}	I_{75}	I_{100}	C_{50}	C_{75}	S_{50}	S_{75}
HuBERT → Wav2Vec2	−0.00003	0.379	64.001	11.141	20.444	22.794	27.342
Wav2Vec2 → HuBERT	−0.00003	0.220	47.888	12.737	18.850	32.711	26.918

D.10 Independent Attended-Speech EEG Confirmation

Participants and data split. The EEG study uses OpenNeuro ds006434, version 1.2.0. Its protocol was fixed before any confirmation EEG sample was opened. Participants sub-diotic01–sub-diotic14 form the development population; sub-diotic16–sub-diotic29 form the confirmation population. All 14 confirmation participants completed the pipeline, exceeding the prespecified minimum of 12. No replacements were made.

For each participant, chapters 0–35 fit the source decoder and chapters 36–39 select its stopping epoch. Chapters 40–49 support policy selection in development participants. Chapters 50–59 are evaluated once.

Preprocessing and models. We apply the released participant-specific ICA before resampling to 100 Hz. Channel centering and scaling use only source-fitting chapters. Inputs contain all 32 channels at lags of 0, 50, 100, 200, and 300 ms. The residual MLP uses width 256, dropout of 0.1, AdamW, and seeds 17, 29, and 41. Ridge receives the same lagged input.

Policy and endpoints. Development selection fixes Wav2Vec2-to-HuBERT transfer, weight 8, and target-space Ridge. The primary endpoint is 12-second attended-speaker accuracy; additional endpoints are three-/six-second accuracy and 100-way retrieval. Wrong-time and source-label-derangement controls perturb the transferred information. The paired opposite-attention quantity is a dependent consistency check, as detailed below.

EEG audio bridge. For the selected Wav2Vec2-to-HuBERT direction, the frozen bridge uses clean audio from chapters 0–39: two speakers, twenty windows per chapter, and all 149 tokenizer frames per window (238,400 paired frames). This in-domain bridge and the existing LibriSpeech bridge were compared on clean audio from chapters 40–49 (59,600 frames), using source-to-target conditional NLL and smoothing $\epsilon = 10^{-4}$. The in-domain bridge was selected at 2.70453 versus 2.76832 nat/token. Construction and this selection use audio alone, without EEG or attention labels. Neural scoring subsequently retains 129 frames per window, as specified below.

Candidate scoring. Let $\mathcal{W}$ contain the elementary windows in a decision, $F = 129$, and $b_{j,w,t}$ be the token for candidate speaker j at frame t of window w. The evaluator computes

$$S_j(\mathcal{W}) = \sum_{w \in \mathcal{W}} \frac{1}{F} \sum_{t=0}^{F-1} \left[\log \pi_B(b_{j,w,t}) + \lambda e_{w,t}(b_{j,w,t}) \right], \qquad \widehat{j} = \underset{j \in \{0,1\}}{\arg\max}\, S_j(\mathcal{W}), \tag{D.1}$$

where $e_{w,t}(b) = \log \sum_a K(a \mid b) \exp\{0.5 \langle \widehat{h}_{w,t}, v_a \rangle\}$ and $\lambda = 8$. The prior π_B is the target marginal of the frozen smoothed audio bridge. EEG uses no candidate-specific contextual model. The normalization of $\pi_B(b) \exp\{\lambda e_{w,t}(b)\}$ is common to candidates at each frame and cancels in ranking. Weight 8 therefore scales neural evidence relative to the nonuniform frequency term. Source-native and direct decoders use the same scoring rule with their respective vocabularies and $e_{w,t}(k) = 0.5 \langle \widehat{h}_{w,t}, v_k \rangle$. The attended-speaker label is used only to evaluate the completed candidate comparison.

Decision windows and temporal aggregation. Each 64-second trial contributes twenty nonoverlapping three-second windows starting at 2, 5, …, 59 seconds, covering seconds 2–62. We sum elementary-window scores over successive pairs or groups of four for six-/twelve-second decisions, with corresponding strides of six/twelve seconds. Groups start at the first elementary window, never cross trial boundaries, and are complete. Each test participant has 20 trials (ten chapters under both attention conditions), yielding 400, 200, and 100 decisions. An exact score tie earns half credit; there is no majority voting.

A three-second audio crop yields 149 tokenizer frames, of which the first 129 are retained. At frame index $t = 0, \ldots, 128$, the EEG design concatenates samples $2t + \ell$ for $\ell \in \{0, 5, 10, 20, 30\}$ across the 32 channels. This implements a 20-ms frame-index correspondence with 0–300-ms EEG lags. All indices lie within the 300-sample EEG crop (maximum 286); the last 20 tokenizer frames are excluded. This discrete alignment does not equate EEG sample timestamps with the encoder's receptive-field centers.

Retrieval candidates and ranking. The bank is the complete Cartesian product of chapters 50–59, five nonoverlapping four-window spans, and two clean speech streams: 100 identity-distinct candidates, ordered by chapter, span, then speaker. The attended stream at the matching chapter/span is correct. The bank also includes the competing stream, other spans, and other chapters; distractors are not randomly sampled. Candidate scores average Eq. (D.1) over four windows, differing from decision scores only by a common factor of four. Token sequences are not deduplicated. Exact ties receive midrank $1 + n_> + (n_= - 1)/2$, where $n_>$ and $n_=$ count scores greater than and equal to the correct candidate's score. MRR and recall@5 average 100 queries per participant, then weight participants equally. Aligned, wrong-time, and source-label-derangement evaluations use the same bank and ordering.

Paired opposite-attention substitution. For each participant, chapter, and elementary segment, the branch substitutes decoded evidence from the other recorded presentation of the same mixture, identified by the published `attend` field, while retaining the original query's attention label. The

EEG observation and prediction thus come from a distinct trial. Candidate tokens, frequencies, and weights remain unchanged. This pairing is an involution, and both presentations enter with equal weight. It permutes candidate scores while evaluating them against the opposite labels. With half credit for ties, participant accuracy satisfies $a_i^{\mathrm{opp}} = 1 - a_i$ and $a_i - a_i^{\mathrm{opp}} = 2(a_i - 0.5)$. The aggregate quantity is a dependent consistency check, not an independent effect of changing attention.

Evaluation criteria. The 17 prespecified checks combine predictive comparisons, the dependent paired-opposite check, and an independent reconstruction check. The supplement preserves the full protocol and decision record. Table 9 reports the paired effects. A half-gain margin subtracts half of a reference decoder's gain over chance from transport's gain over chance; a positive lower bound meets the prespecified retention requirement.

Table 9: **Paired EEG effects.** Intervals use 100,000 participant-bootstrap resamples. Accuracy differences and half-gain margins are expressed as proportions.

Fixed comparison	**Mean effect**	**95% interval**
Transport accuracy − chance, 12 s	0.1057	[0.0643, 0.1564]
Aligned accuracy − paired opposite-trial accuracy†	0.2114	[0.1286, 0.3143]
Aligned − wrong-time MRR	0.0631	[0.0378, 0.0968]
Transport − source-label-deranged accuracy	0.0893	[0.0543, 0.1307]
Transport − source-label-deranged MRR	0.0552	[0.0311, 0.0867]
Half-gain margin versus target-space Ridge	0.0407	[0.0146, 0.0711]
Half-gain margin versus source-native decoder	0.0421	[0.0186, 0.0689]

† Evidence is exchanged between recorded attention conditions. Symmetric pairing makes this difference twice the aligned accuracy above chance; it is a dependent consistency check.

D.11 Absolute NLL of the MEG Comparators

Table 10 uses the same models and equal weighting of masking regimes as Table 1. Frozen prior checkpoints, evaluation tokens, and masks recover prior NLL; subtracting archived paired improvements gives absolute model losses without refitting.

Table 10: **Absolute MEG confirmation NLL (nat/token).** Lower is better. The target vocabulary determines the column; transport uses the other vocabulary as its source. Values average seven regimes equally, then 11 participants.

Model	Wav2Vec2 target	HuBERT target
Frozen target prior	3.86216	4.12533
Global-weight transport	3.84389	4.10742
Context-dependent transport	3.84129	4.10543
Direct target MLP	3.85012	4.10613
Direct target Ridge	3.84888	4.10476

For HuBERT, transport and Ridge improve on the prior by 0.0199033 and 0.0205759 nat/token. Dividing transport's gain by Ridge's yields 0.967313 (97% when rounded).

D.12 Implementation Checks and Reproducibility

Numerical checks. Finite-state tests verify normalization, equivalence of explicit probability push-forward and log-space transport, and neutrality (maximum float64 error below 6×10^{-17}). On saved predictions, ratio normalization and direct factor transport yield fused probabilities differing by less than 10^{-14} in both directions.

EEG reconstruction. Recalculation in a separate computing environment reproduced all EEG effects and scientific checks from 14 participant records. Saved-prediction replay exactly matched 1,120 accuracy, retrieval, and weight fields. The supplement contains code, tests, frozen protocols, geometry inputs and manifests, and decision, retrieval, and paired-trial records.