MERIT: MAKING EVERY MODALITY EARN ITS CONTRIBUTION IN MULTIMODAL ATTENTION DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

More than 1.5 billion people live with hearing loss, and when several people talk at once, a hearing aid amplifies every voice alike. Auditory attention decoding (AAD) aims to solve this: it reads a listener's brain signals, finds the talker they are attending to, and lets the device amplify that talker. A real device will not see the brain alone. Glasses and earbuds can also track the eyes, the head and the scene, and these signals help, because people tend to look at the person they listen to. But they show where a listener is looking, not whom they are listening to. A decoder that leans on them works when the two agree and fails when they differ — when a listener cannot move their eyes or head, or listens to someone they are not looking at — which is exactly when a brain-based device is needed. A multimodal decoder must therefore use these signals without letting them replace the electroencephalogram (EEG), and accuracy alone cannot show whether it does. We present MERIT, a decoder and an evaluation method built around one question: how much of the accuracy does each input earn? MERIT shuffles each input stream across test examples, measures how much accuracy drops, and splits the result exactly across the streams with Shapley values. It also blocks two shortcuts by design: the candidate audio is checked before training so that it cannot reveal the answer, and the scoring function cannot rise above chance if it ignores the brain signal. On a public dataset with EEG, eye tracking, head motion and scene video, MERIT improves on the published benchmark, reaching 0.82 four-way accuracy within listeners and 0.74 on unseen listeners, where chance is 0.25. The measurement also shows what accuracy hides: adding the behavioural streams raises accuracy but almost halves the brain's share of it. On two other public datasets, the same tests find a label imbalance in a widely used benchmark and flag results that accuracy alone would report as brain decoding.

1 INTRODUCTION

Why attention decoding matters. About 1.5 billion people live with some hearing loss, and the number is expected to reach 2.5 billion by 2050 (World Health Organization, 2021). The hardest situation for them is the one Cherry (1953) called the cocktail party problem: several people talking at once. A hearing aid can make sound louder, but it cannot tell which voice matters, so it amplifies all of them, including the ones the listener is trying to ignore. *Auditory attention decoding* (AAD) addresses this (O'Sullivan et al., 2015; Geirnaert et al., 2021). The brain's response follows the speech a listener attends to (Ding & Simon, 2012; Lalor & Foxe, 2010), so a decoder can read that response from the electroencephalogram (EEG), find the attended talker, and steer a hearing device toward them (Pan et al., 2024).

Why several streams matter. Such a device will not see the brain alone. Glasses and earbuds already carry motion sensors, and eye trackers and cameras are being added to them. These signals are useful. In everyday listening, people turn their head and eyes toward the talker they follow (Hendrikse et al., 2019; Best et al., 2023; Kidd Jr. et al., 2005), and a hearing aid steered by an eye tracker has already been built (Kidd Jr. et al., 2013). A realistic decoder should therefore combine these streams with the EEG, and doing so raises accuracy (Section 6.2).

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.

Why the EEG must still carry the decision. Gaze and head motion show where a listener is *facing*, not whom they are *listening to*, and the two often differ. A driver listens to a passenger while watching the road. At a dinner table, people listen to one person while looking at another. Some listeners cannot move their eyes or head freely at all. A decoder that leans on the behavioural streams does well when facing and listening agree, and fails in exactly these cases. Those are the cases where a brain-based device is needed: when facing and listening agree, an eye tracker is enough. Whatever the EEG adds, it adds through the brain signal. A multimodal decoder should therefore use the other streams as helpers, but it must not let them replace the EEG. Accuracy cannot tell us whether this holds. A multimodal model is reported with one number, and that number does not say which input produced it. A decoder that silently decides from gaze and head motion and a decoder that truly reads attention from the EEG can therefore report the same accuracy, even though only the second keeps working when facing and listening disagree. The risk is real: Rotaru et al. (2024) found that decoders believed to read spatial attention from EEG owed most of their accuracy to eye movements captured in the recording.

The usual formulation is a matching task. The model sees a window of the recording and K candidate speech signals, one of which is the attended talker, and it must pick that one; accuracy is compared with chance, $1/K$ (Accou et al., 2021; Jalilpour Monesi et al., 2020; Puffay et al., 2023). We study it on MAESTRO (Hassan et al., 2026), a public dataset in which sixteen listeners each attend one of four talkers speaking at once while EEG, eye tracking, head motion and scene video are recorded together (Section 5). We show that this task can be solved without using the brain signal at all, in three ways (Section 3). **(i) The audio gives the answer away.** Scaling each candidate to the same loudness does not make the candidates look alike: statistics such as skew and sparsity do not change with scale. On this dataset they differ between attended and ignored talkers, and a simple classifier that sees only the audio reaches 0.5581 four-way accuracy, against a chance of 0.2500. **(ii) The standard scoring function gives the model a path to the leak.** Problem (i) says the answer is *present* in the audio; this one says the standard architecture can *reach* it without the brain. If the brain encoder outputs the same vector at every time step — the collapse an encoder falls into when its input stops helping — the usual score reduces to a classifier on the audio alone, and training often finds this solution because it is easier than the intended one. **(iii) Easy streams crowd out the brain.** When several streams are available, training leans on the easiest one (Wang et al., 2020; Huang et al., 2022; Wu et al., 2022; Peng et al., 2022). Here the easy streams measure where the listener is *looking*, not what they are *listening to*, so accuracy rises while the evidence for the claim shrinks.

One test catches all three problems. We shuffle a recorded stream across the test windows, so each window gets another window's recording but keeps its own candidates and its own label, and we measure how much accuracy drops. We call the drop the **contribution** of that stream. A decoder that does not use a stream loses nothing when it is shuffled. This paper turns that test from an occasional check into the centre of the method.

MAESTRO was released with a benchmark, and the benchmark already reports a permutation test: it shuffles the EEG, gaze, head motion and video *together*, and its contributions are positive. That shows the listener's signals are used. It cannot show *which* signal is used, and problem (iii) sits exactly there, because a joint test gives its largest value when a strong behavioural stream has pushed the brain aside. MERIT shuffles each stream *separately* and splits the total exactly across streams. It also decodes better than the benchmark: 0.82 four-way accuracy within listeners and 0.74 on unseen listeners, against the benchmark's 0.698 and 0.674 (Section 6.3).

Contributions. **(1) A per-stream measure of credit** (Section 4). We compute each stream's Shapley value over all 2^M subsets of the M recorded streams ($M = 5$ here), keeping the streams in a subset intact and shuffling the rest. The values add up exactly to the total contribution, and they are cheap enough to report for every run. **(2) A decoder in which the shortcuts are blocked by design** (Section 4). A time-centred correlation score gives every candidate the same score when the brain encoder is constant, so that shortcut lands exactly at chance. The classifier heads that read gaze and head motion receive no audio, so they cannot use audio cues. The candidate audio is tested before training. **(3) Choosing checkpoints by contribution, not accuracy**, which trades a little accuracy for a larger share of credit on the brain. **(4) Findings.** Adding behavioural streams cuts the brain's share of the credit almost in half; at two-second windows the EEG earns almost none of it. On two other public datasets, the same tests find a label imbalance in a widely used benchmark and show that

high accuracy on both cannot be confirmed as brain activity. We also tested adding the permutation test to the training loss (Section 4); it does not increase the brain's share, and we report this in full.

2 CANDIDATE SETS AND SPLITS

Candidate sets. In the matching task the decoder is shown K candidate speech signals and must pick the one the listener attended. The candidates are inputs, not labels: the decoder sees them, so anything in them that marks the correct one is a shortcut that solves the task without the recording. How the candidates are built therefore decides what the task measures, and we name each set by how it is built. UNMATCHED: the four talkers of the trial, each standardised to zero mean and unit variance. This is the usual choice in the literature. Standardising removes loudness, but the four candidates are still four different recordings, and any property beyond the mean and the variance — skew, sparsity, the balance between frequency bands — can still tell them apart; on this dataset it does (Section 3). MATCHED: the same four talkers after *quantile matching*. We sort each candidate's values, average the sorted values across the four candidates, and give each candidate these averages back in its own time order. All four then contain exactly the same set of values, so no statistic of the values — loudness, skew, sparsity or any other — can tell them apart. Each keeps its own timing, and timing is what a brain response follows. This is the most we can remove from the audio without removing the signal the task is about. SAME-TALKER: the wrong candidates are the *attended talker's own* speech, taken from other, non-overlapping parts of the same trial. All candidates then come from one voice and one loudspeaker, so neither the audio nor the listener's orientation can say which is correct, and chance is exact by construction (Accou et al., 2021). Trials are long enough for $K=3$ and $K=2$. SAME-TALKER MATCHED applies both steps. We use MATCHED for the four-way task and SAME-TALKER MATCHED as the strictest check that the EEG follows the speech.

Splits. *Within-listener*, five folds: trials are split by their audio content, so no stimulus heard in training is heard again at test; every listener appears on both sides. *Leave-one-listener-out* (LOSO), sixteen folds: one listener is held out, and 20% of the audio content is also held out, so the test listener is new in both identity and content. A validation set is always taken from the training part, and each test set is scored once.

3 THREE WAYS TO BE RIGHT FOR THE WRONG REASON

(i) The audio gives the answer away. The decoder sees the candidate audio, so anything in the audio that marks the attended talker solves the task without the brain. Holding out listeners or stimuli does not help, because the mark belongs to the *role* of attended talker, so it is present for every listener and every stimulus. Making all candidates equally loud does not help either. Loudness is removed exactly by standardising each candidate, but standardising only fixes the mean and the variance. Every statistic that does not change with scale survives: skew, kurtosis, sparsity and the balance between frequency bands. On this dataset the attended and ignored talkers differ in exactly these. The peak-to-average ratio of the envelope alone separates them with an effect size of $d = -1.22$. The dataset's authors trace this to the loudness adjustment, which clipped the attended talker's peaks (Hassan et al., 2026).

We therefore test each candidate set *before* training, with an **acceptance probe**: a logistic regression on eight statistics that do not depend on scale, evaluated on held-out content. A clean candidate set must score $1/K$. The UNMATCHED set scores +0.3081 above chance. MATCHED brings this down to 0.2603 against 0.25, and same-talker candidates to 0.5134 against 0.50 (Table 3).

(ii) The standard scoring function has a shortcut. The standard score normalises the brain and audio embeddings at each time step, multiplies them, averages over time and applies a linear layer (Accou et al., 2021; Jalilpour Monesi et al., 2020). Suppose the brain encoder outputs the same unit vector $\hat{c}$ at every step. This is not a hypothetical state: once the audio already carries the answer, the easiest way for gradient descent to lower the loss is to let the encoder ignore its noisy input and emit a constant, and we observe exactly this collapse when the standard model is trained on this data (Section 6.1). With a constant $\hat{c}$ the time average splits apart and the score becomes $w^\top(\hat{c} \odot \frac{1}{T}\sum_t \hat{a}_{k,t}) + b$, where $\hat{a}_{k,t}$ is the audio embedding of candidate k at time step t, T is the number of time steps in the window, $\odot$ is the elementwise product, and w and b are the weights and

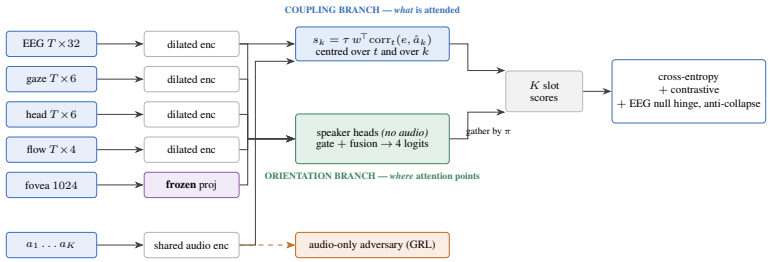


Figure 1: **MERIT.** The coupling branch compares the EEG with each candidate using a correlation over time with the mean removed. A brain encoder that outputs the same vector at every step therefore gives every candidate a score of zero, which is chance. The orientation branch gets *no audio*, so it cannot use audio cues. The model encodes once and scores separately, so a trained model can be re-scored under any shuffle of any stream without encoding again.

bias of the final linear layer. This is a linear classifier on the time-averaged audio, with exactly the power needed to read the audio mark of (i). So the architecture contains an audio-only decoder, which it reaches by letting the brain encoder go constant. Problem (i) makes the audio shortcut *enough*; problem (ii) makes it *reachable*.

(iii) Easy streams crowd out the brain. When several streams are trained together, the model relies most on the easiest one (Wang et al., 2020; Huang et al., 2022; Wu et al., 2022; Peng et al., 2022). Here this changes what the result means. Gaze, head motion and scene video show where the listener is *facing*, which in a room of loudspeakers predicts the attended talker well (Kidd Jr. et al., 2005; Rotaru et al., 2024). Only the EEG follows what the talker is *saying*. A fused decoder that quietly relies on gaze reports a higher number while supporting a weaker claim.

As Section 1 argues, it also fails exactly for the listeners a brain-based device is for: those who cannot, or choose not to, face the talker they listen to.

One test catches all three. Shuffling a recorded stream across test windows, while each window keeps its own candidates and label, costs accuracy only if the decoder uses that stream. We shuffle the recording, not the labels: shuffling the labels would also break the link between the audio and the answer, and would hide the very shortcut we are testing for (Appendix A.4). Let $v(T)$ be the accuracy when the streams in the set T are left intact and the others are shuffled. Then $v(\emptyset)$ is what *this* model can get from the audio alone, measured through its own audio pathway. The probe bounds what a simple reader can take from the audio; the shuffled accuracy measures what this decoder actually takes. Appendix A gives all three problems in more detail.

4 MERIT

The name states the goal: every input should earn its share of the result. MERIT has two branches that answer two different questions. The **coupling branch** asks whether the recording follows a candidate's speech over time. Only EEG enters it, because only EEG has a timed relationship with a speech signal. The **orientation branch** asks which loudspeaker the listener was facing. It receives no audio, so it cannot use audio cues. Figure 1 shows the whole decoder.

Encoders sized to the brain response. Each stream goes through five dilated temporal convolutions (Yu & Koltun, 2016) with dilations $1, 2, 4, 8, 16$ and GroupNorm (Wu & He, 2018) after each layer. This gives a window of 63 samples (0.98 s at 64 Hz), about the time over which the cortical response to speech unfolds (Lalor & Foxe, 2010). There is no activation on the last layer, because a rectified output has no negative values and squeezes all the correlations the score depends on toward one. Padding is symmetric, because the brain responds 100–300 ms *after* the sound, so an encoder that looks only backward in time cannot see the response. Appendix D details both choices. One-sided versions restrict the model to a chosen range of delays (Section 6.4).

A score that cannot use a constant brain signal. For each of the $D = 16$ embedding dimensions we compute the correlation over time between the EEG embedding e and the embedding $\hat{a}_k$ that the shared audio encoder computes from the k-th candidate envelope. We combine the D values with a weight vector that has no bias, scale by a learned temperature, and subtract the mean over candidates:

$$\tilde{e} = \frac{e - \overline{e}}{\|e - \overline{e}\|_2}, \qquad c_d(k) = \textstyle\sum_t \tilde{e}_{t,d}\, \tilde{a}_{k,t,d}, \qquad s_k = \tau\, w^\top c(k), \qquad \mathrm{score}_k = s_k - \tfrac{1}{K} \textstyle\sum_j s_j. \tag{1}$$

Removing the time average is the key step. If the encoder outputs the same vector at every time step — the state it falls into once it learns to ignore its input — then $e - \overline{e}$ is *exactly zero*. Every $c_d(k)$ is then zero, all K scores are equal, and accuracy is exactly chance. The shortcut of Section 3(ii) is not discouraged by a penalty; it is impossible. Two more properties follow. Correlation ignores scale, so a candidate cannot win by being louder. Subtracting the mean over candidates removes anything added equally to all of them, which is why w has no bias.

Orientation heads with no audio, and an audio-only adversary. Gaze, head motion, scene motion and the foveal stream — a frozen V-JEPA 2 embedding of the patch of scene around the gaze point (Section 5) — show where the listener is facing. They have no timed relationship with speech, so asking them to match a speech envelope makes no sense — and doing so made them collapse in our first attempt. Instead, each one feeds a classifier head that predicts the loudspeaker index from the mean and standard deviation of its embedding over time. These classifier heads get **no audio input**, so shuffling them drops accuracy to chance, and anything above chance is earned by the recording. The EEG also gets such a classifier head, because attention shifts brain activity toward one side. A learned, non-negative weight per stream and a small fusion layer combine the loudspeaker scores, which are then added to the coupling scores. Separately, a small head tries to find the correct candidate from the audio alone, behind a gradient-reversal layer (Ganin & Lempitsky, 2015). Any audio cue it can read is thereby trained out of the audio encoder. This matters because the probe of Section 3(i) is linear and can miss cues that a network finds.

Training objective. Let a batch hold B windows from one listener. For window i, $\ell_i \in \mathbb{R}^K$ are the final candidate scores, y_i is the correct candidate, r_i is the attended loudspeaker, e_i is the EEG embedding and $a_i^+ = \hat{a}_{i,y_i}$ is the embedding of the correct candidate. We write $s(e, a) = \tau\, w^\top c(e, a)$ for the coupling score of Eq. 1 before centring over candidates, and ρ for a shift of the batch by one window. MERIT minimises

$$\mathcal{L} = \mathcal{L}_{\text{task}} + 0.3\,\mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{con}} + 0.5\,\mathcal{L}_{\text{null}} + 0.1\,\mathcal{L}_{\text{col}} + 0.3\,\mathcal{L}_{\text{adv}}, \tag{2}$$

with the six terms

$$\begin{aligned}
\mathcal{L}_{\text{task}} &= \tfrac{1}{B} \textstyle\sum_i \mathrm{CE}_{0.1}\big(\ell_i,\, y_i\big), \qquad \mathcal{L}_{\text{aux}} = \tfrac{1}{B|\mathcal{O}|} \textstyle\sum_i \sum_{m \in \mathcal{O}} \mathrm{CE}\big(u_i^m,\, r_i\big), \\
\mathcal{L}_{\text{con}} &= -\tfrac{1}{2B} \textstyle\sum_i \Big[\log \frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}} + \log \frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}} \Big], \qquad S_{ij} = s\big(e_i, a_j^+\big), \\
\mathcal{L}_{\text{null}} &= \tfrac{1}{B} \textstyle\sum_i \Big[\mathrm{sp}\big(s(e_{\rho(i)}, a_i^+) - s(e_i, a_i^+) + \mu\big) + \mathrm{sp}\big(s(\mathbf{0}, a_i^+) - s(e_i, a_i^+) + \mu\big) \Big], \\
\mathcal{L}_{\text{col}} &= \tfrac{1}{BD} \textstyle\sum_{i,d} \max\big(0,\ \gamma - \mathrm{std}_t\, e_{i,t,d}\big) + \tfrac{0.04}{D} \textstyle\sum_{d \neq d'} C_{dd'}^2, \\
\mathcal{L}_{\text{adv}} &= \tfrac{1}{B} \textstyle\sum_i \mathrm{CE}\big(\big[g(R(f_{i,k}))\big]_{k=1}^K,\, y_i\big), \qquad f_{i,k} = \big[\, \mathrm{mean}_t\, \hat{a}_{i,k}\, ;\, \mathrm{std}_t\, \hat{a}_{i,k} \big].
\end{aligned}$$

Here $\mathrm{CE}_{0.1}$ is cross-entropy with label smoothing 0.1 and sp is softplus. $\mathcal{L}_{\text{task}}$ is the decision loss: the cross-entropy between the K final candidate scores ℓ_i and the index y_i of the attended candidate, so it trains everything that feeds those scores. $\mathcal{L}_{\text{aux}}$ gives each orientation head $m \in \mathcal{O}$ its own loudspeaker prediction u_i^m, so the strongest stream cannot take all the gradient. $\mathcal{L}_{\text{con}}$ is a two-way contrastive loss (van den Oord et al., 2018; Radford et al., 2021) that asks which window's EEG goes with which window's speech; the batch comes from one listener, so the EEG cannot be matched by listener identity, and a constant encoder cannot lower this loss. $\mathcal{L}_{\text{null}}$ requires the real EEG to give the correct candidate a higher coupling score than another window's EEG or a zeroed EEG does, by a margin $\mu = 0.5$. $\mathcal{L}_{\text{col}}$ keeps the spread over time of each embedding dimension above $\gamma = 0.5$ — the quantity that falls to zero when an encoder stops using its input — and discourages dimensions from copying each other, where C is the covariance of the time-centred embeddings over all windows and time steps (Bardes et al., 2022). $\mathcal{L}_{\text{adv}}$ trains a small network g to find the correct candidate from

summary statistics of the audio embeddings alone; R is a gradient-reversal layer (Ganin & Lempitsky, 2015), which leaves the forward pass unchanged and flips the sign of the gradient that reaches the audio encoder, so any audio cue g can read is trained out of it.

Tested but not used: a contribution hinge. The EEG hinge above acts on the coupling score only. A natural next step is to apply the same test to the *final* decision, for every stream. For each stream m we tried

$$\mathcal{L}_m^{\text{contrib}} = \text{softplus}\big(\sigma_y(z_{m\leftarrow\rho}) - \sigma_y(z) + \delta\big) + \text{softplus}\big(\sigma_y(z_{m\leftarrow 0}) - \sigma_y(z) + \delta\big), \tag{3}$$

where $\sigma_y(\ell) = \ell_y - \frac{1}{K}\sum_k \ell_k$ is the score of the correct candidate minus the mean score, $z_{m\leftarrow\rho}$ replaces stream m with another window's copy, $z_{m\leftarrow 0}$ replaces it with zeros, and $\delta = 0.3$ is the margin. The hinge arms add $0.5\sum_m \lambda_m \mathcal{L}_m^{\text{contrib}}$ to Eq. 2. After each epoch the weight of stream m is set from its validation contribution c_m in the fused model and the contribution c_m^{solo} it reaches through its own head alone:

$$\lambda_m = \text{clip}\Big(\frac{\max(0,\ c_m^{\text{solo}} - c_m)}{c_m^{\text{solo}}},\ 0,\ 1\Big), \tag{4}$$

so a stream that can do well alone but is ignored by the fused model is pushed hardest. The hinge does not raise the brain's share of credit (Section 6.2), so MERIT trains without it; we report it because the reason it fails is informative.

Choosing checkpoints by contribution. If a shortcut exists, picking the epoch with the best validation accuracy picks the epoch that uses the shortcut best. We keep instead the epoch that maximises

$$\big(v(M) - v(\emptyset)\big) + \beta \min_m \phi_m, \qquad \beta = 1, \tag{5}$$

the validation contribution plus the credit of the *weakest* stream. An epoch that is accurate because it rides one stream loses to a slightly less accurate epoch that uses all of them. We compute both criteria at every epoch and report both.

Credit: a split that adds up exactly. The contributions of single streams do not add up to the total, because streams overlap — gaze and the foveal embedding both describe where the listener looks — and interact. The Shapley value (Shapley, 1953) over subsets of *intact* streams is the only split that does add up:

$$\phi_m = \sum_{T\subseteq M\setminus\{m\}} \frac{|T|!\,(|M| - |T| - 1)!}{|M|!}\Big[v(T\cup\{m\}) - v(T)\Big], \qquad \sum_m \phi_m = v(M) - v(\emptyset). \tag{6}$$

This is permutation feature importance (Breiman, 2001; Fisher et al., 2019; Covert et al., 2020) applied to whole streams. With $M = 5$ streams there are only 32 subsets, so we compute it exactly, and no subset needs the inputs to be encoded again. The same random shuffles are used for every subset, so differences between subsets are not shuffle noise.

5 EXPERIMENTAL SETUP

Data and task. MAESTRO (Hassan et al., 2026) has sixteen listeners and one hundred trials each. In every trial four talkers speak at once from loudspeakers at fixed positions, and the listener is told beforehand which one to attend to. Trials last about 30 s. Each talker recording is used only once, so no audio is shared between training and test. The attended talker is about 15 dB louder than the others in every trial. We use five streams, all resampled to 64 Hz and standardised per channel: EEG (32 channels), gaze (6), head motion (6), optical flow from the scene camera (4), and one new stream, a *frozen* V-JEPA 2 ViT-L embedding (Assran et al., 2025) of the part of the scene the listener is looking at (one 1024-d vector per window, never trained). Signals are cut into windows of W seconds with hop H ($W = 10$, $H = 5$ unless stated). Each example holds the streams, K candidate speech envelopes of which one is correct, the index of the correct one, the loudspeaker each candidate came from, and the listener's identity. The candidates are placed in random order in every window, so their position says nothing. At test time the order is fixed by the window's index, so every model we compare sees exactly the same inputs (Appendix S).

Every result is defined by four choices: the candidate set, the split, the decision window and the set of streams. Every accuracy comes with its own shuffled baseline; an accuracy without one is not a result. Besides accuracy, the baseline and the contribution, we report a permutation p-value $(\#\{v_{\text{null}} \geq v\} + 1)/(n + 1)$, whose smallest possible value with twenty shuffles is 0.048; the Shapley credit ϕ_m; the *flip rate*, the fraction of windows whose chosen candidate changes when the stream is shuffled; accuracy with the stream replaced by zeros; and the *collapse*, the mean cosine similarity between the time-centred embeddings of different windows, where a value near one means every window produces the same pattern.

Training uses AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3}, batches of 32 windows from one listener, and at most 50 epochs. Checkpoints are chosen on validation data by the rule of Section 4, with three shuffles, and the chosen checkpoint is scored once on the test set with twenty. **We did no hyper-parameter search**: every setting was fixed once in advance (Appendix P), so these numbers are not tuned. To compare architectures, the first row of our ablation runs the benchmark's model (Hassan et al., 2026) — its encoder, its cosine-then-linear score and plain cross-entropy — inside our training loop and splits, so that differences between rows come from the model and not from the protocol. We run it on both candidate sets. Neither run is meant to reproduce the benchmark's published numbers, which use its own training loop, a longer schedule and all streams; when we cite those numbers we take them from the paper.

Why this dataset. Crowding out between streams can only be studied when the streams are recorded *and* the listener is free to produce them; Section 7 reviews why the other public corpora rule this out by design. As far as we know, MAESTRO is the only public AAD dataset that records eye tracking, head motion and scene video together with EEG from a listener who is free to move in a room with several loudspeakers (Appendix G). The parts of MERIT that do not need these streams — the audio probe and the coupling score — work on any matching dataset, and we test them on two (Section 6.4).

6 RESULTS

6.1 WHAT THE MATCHING TASK ALLOWS

We change one thing at a time, starting from the benchmark's model run in our training loop. The benchmark itself uses matched candidates and reports positive contributions. Here we first pair its model with UNMATCHED candidates, the usual choice in the wider literature. This shows what the task allows; it is not a reproduction of the benchmark's result.

With UNMATCHED candidates the benchmark's model reaches 0.4991 four-way accuracy, but its contribution is +0.0000, its decisions never change when the EEG is shuffled (flip rate 0.0000), and its embeddings are nearly identical across windows (collapse 0.9138). The accuracy is real, but none of it comes from the EEG: this is the shortcut of Section 3(ii). A better encoder raises accuracy to 0.5800 and leaves the contribution at +0.0000 — it simply reads the audio better. A change that raises accuracy is not, by itself, evidence; only the shuffle test shows this. The time-centred score gives the first non-zero contribution. With matched candidates the same benchmark model scores 0.2833, which is chance.

6.2 MORE STREAMS, LESS CREDIT FOR THE BRAIN

The brain's share falls by almost half. Adding the behavioural streams raises accuracy from 0.6915 to 0.8173. But the EEG's credit falls from +0.3971, when it is the only stream, to +0.2217 in the fused model. More streams have not taught the model more about the brain; they have let it rely on the brain less. A joint shuffle cannot see this, because the joint contribution *rises*, from +0.3971 to +0.5439. This is why the per-stream split is the measurement, and the total is not. In the fused model the foveal embedding earns +0.2866, more than the EEG: much of the accuracy is a statement about where the listener was looking.

Putting the test in the loss does not fix it. The contribution hinge never raised the brain's credit above the matching model without it. With modality dropout it lowered the EEG's credit from +0.2229 to +0.1600; without dropout it made no real difference (+0.2169 against +0.2217). The

Table 1: **More streams give more accuracy, and less credit to the brain.** Within-listener, MATCHED, chance 0.25. The four fused models share streams, data, folds, architecture and listener adaptation; they differ only in the contribution hinge and in modality dropout. ϕ_m is each stream's Shapley credit and the five add up to the contribution. The EEG-only model shows how much credit the brain earns on its own. Appendix J gives the same four models on unseen listeners.

				Shapley credit ϕ_m				
Model	Accuracy	Null	Contrib.	EEG	gaze	head	flow	fovea
EEG alone	0.6915	0.2944	+0.3971	+0.3971	—	—	—	—
MERIT (no hinge, no dropout)	**0.8173**	0.2733	**+0.5439**	+0.2217	+0.0194	+0.0039	+0.0123	+0.2866
No hinge, with dropout	0.8091	0.2793	+0.5298	+0.2229	+0.0206	+0.0079	+0.0128	+0.2656
With hinge, no dropout	0.8242	0.2685	+0.5557	+0.2169	+0.0340	+0.0074	+0.0295	+0.2679
With hinge, with dropout	0.7884	0.2774	+0.5111	+0.1600	+0.0362	+0.0144	+0.0247	+0.2758

Table 2: **Improvement ladder**, EEG only, within-listener, MATCHED, chance 0.25. Each row adds one change to the row above. The last row is a control, not a step: it limits the coupling to delays *before* the sound, where no brain response to that sound can be.

	Accuracy	Null	Contribution
Random crops, one decision per window	0.4956	0.2982	+0.1974
+ one decision per trial	0.6676	0.3810	+0.2866
+ only delays after the sound	0.6502	0.3231	+0.3271
MERIT, EEG only (+ per-listener correction)	**0.6915**	0.2944	**+0.3971**
control: only delays before the sound	0.6145	0.4017	+0.2127

reason is the interaction. Dropout removes streams at random during training, so the validation contribution that sets the hinge's weights is measured on a model that keeps being asked to work without them, and the hinge then pushes whichever stream looks unused — rarely the EEG. On unseen listeners the models without the hinge keep more credit on the brain in both pairs (Appendix J). A measurement that cannot be pushed up by training against it is one worth trusting. The one training choice that did raise the brain's credit was choosing checkpoints by contribution: in an earlier version of this experiment, at $W = 10$ s, it moved the EEG's credit from +0.1039 to +0.1182 (Table 14).

Which model we call MERIT. MERIT is the model without the hinge and without dropout. It is not the most accurate of the four in either protocol. It keeps the most credit on the brain on unseen listeners (+0.1762) and ties for the most within listeners — and credit is what we select for. Choosing it costs less than 0.02 in accuracy in each protocol.

6.3 THE FULL SYSTEM, AND THE BENCHMARK

The results above make one decision per window. But the listener attends to one talker for the whole trial, so Table 2 adds, one at a time: one decision per trial, made by summing the window scores; training on random crops of the whole trial; limiting the coupling to delays after the sound; a small per-listener correction of the EEG channels; and a second audio channel that marks speech onsets. The candidates stay matched and the splits stay the same.

What each change is worth. One decision per trial is by far the largest gain: the contribution rises from +0.1974 to +0.2866 on the same windows. Limiting the coupling to delays after the sound adds a little, and the control confirms the direction: delays before the sound reach only +0.2127. The per-listener correction raises the contribution from +0.3271 to +0.3971, and the onset channel raises accuracy from 0.6709 to 0.6915.

Against the benchmark. We compare with the benchmark at its longest window, 30 s, which — like our per-trial decision — gives one decision per trial (Hassan et al., 2026). There it reports 0.591 for EEG alone within listeners and 0.619 on unseen listeners, and 0.698 and 0.674 for its best

multimodal models; its contributions are $+0.320$ and $+0.363$ for EEG alone, and $+0.413$ for all streams on unseen listeners. MERIT reaches 0.6915 for EEG alone within listeners (contribution +0.3971), 0.8173 with all streams (+0.5439), and 0.7438 with all streams on unseen listeners (+0.4737, positive in 16/16 folds). Three of the four comparisons improve on both accuracy and contribution.

The fourth, EEG alone on unseen listeners, does not: 0.6000 against 0.619, with contribution +0.3132. The per-listener correction behind our largest within-listener gain cannot, by design, help a listener it has never seen. Our gains on the *brain* signal are therefore shown within listeners.

6.4 IS THE EEG SIGNAL REAL?

Four checks, with details in Appendix L. **Same-talker candidates.** When all candidates come from the attended talker's own speech, the audio alone is at chance and no orienting cue can help; EEG alone still contributes +0.2156 at $K = 3$ and +0.1986 at $K = 2$. **Direction in time.** Limited to delays of $+125$ to $+1109$ ms after the sound, where a brain response must be, the contribution is +0.2240; the mirror image before the sound gives +0.0967. This rules out an electrical echo of the sound at zero delay, and slow drift that is the same in both directions. **Window length.** The contribution grows with the window, as it should for this task (Geirnaert et al., 2021) (Figure 3). At short windows the behavioural streams take over: at 2 s the EEG alone contributes +0.0263 while the foveal embedding alone contributes +0.4570 (Appendix L.1). **Other datasets.** On two public two-talker datasets the audio probe is at chance on DTU but +0.1947 above chance on KU Leuven, because one of its two talkers is the attended one in 256 of 320 trials (Appendix H). On KU Leuven and DTU the coupling branch reaches 0.8149 and 0.8170 even with same-talker candidates, but the direction-in-time check does not pass: delays before and after the sound score the same. We therefore do not count these as brain decoding (Appendix I). At the 5 s window used by recent work, the coupling branch is more accurate on unseen listeners of both datasets than NeuroToken (Alavi & Williamson, 2026), the strongest prior result without data leakage (Table 8). A plain accuracy report would have done so.

7 RELATED WORK

Decoding auditory attention. Brain activity follows the speech a listener attends to (Ding & Simon, 2012; Lalor & Foxe, 2010). The oldest decoders are linear: they rebuild the speech envelope from the EEG and compare the reconstruction with each talker (O'Sullivan et al., 2015). They are simple and well understood, but they need windows of tens of seconds to decide reliably (Geirnaert et al., 2021), too slow for a hearing device. Deep models shorten the window in two ways. *Match-mismatch* models learn directly whether a window of EEG goes with a window of speech (Accou et al., 2021; Jalilpour Monesi et al., 2020; Puffay et al., 2023); our decoder belongs to this family. *Direction* decoders drop the audio and classify which side the listener attends (Vandecappelle et al., 2021; Yan et al., 2024a; Fan et al., 2025; Ni et al., 2024; Xu et al., 2024; Zhang et al., 2025). Both lines share a weakness: the reported accuracy often measures something other than attention. Rotaru et al. (2024) showed that direction decoding on several public datasets relies largely on eye movements captured in the EEG rather than on attention, and when within-trial leakage between training and test is removed, recent direction decoders lose 17 to 45 points and several fall to chance (Alavi & Williamson, 2026) (Table 8). We bring the concern of Rotaru et al. (2024) to the matching task and to several recorded streams at once, and make the check part of training and model choice rather than an after-the-fact test.

Datasets. The public AAD corpora record EEG and little else, and their designs deliberately hold the listener still, which is good for isolating the brain response and rules out the multimodal question: DTU uses earphones that simulate a room (Fuglsang et al., 2018), KU Leuven a quiet booth (Biesmans et al., 2017; Das et al., 2019), and a recent four-speaker dataset a chin rest (Yan et al., 2024b). A stream that was never recorded, from a listener who was not free to move, cannot compete with the EEG for credit. KU Leuven has a further limitation we measure in Appendix H: one of its two talkers is attended in 256 of 320 trials, so the audio alone carries part of the answer. AV-GC-AAD (Geirnaert et al., 2024) records gaze and decouples it from attention on purpose — the right tool for asking *whether* a decoder reads gaze instead of the brain, but its other streams are designed to carry no information, so it cannot say how credit is shared among streams that do. MAESTRO (Hassan et al., 2026) records EEG, eye tracking, head motion and scene video from a listener who moves freely in a

room of loudspeakers, which is why we work on it; its own limitation, an audio confound in the usual candidate construction, is the subject of Section 3.

Shortcut learning and multimodal balance. The problem is a case of shortcut learning (Geirhos et al., 2020; Torralba & Efros, 2011; Gururangan et al., 2018). The usual fixes rebalance the data (Goyal et al., 2017) or train against a shortcut-only branch (Cadene et al., 2019); we add a scoring function in which the shortcut is impossible. Work on balancing modalities (Wang et al., 2020; Huang et al., 2022; Wu et al., 2022; Peng et al., 2022) evens out how fast each stream learns during training; we measure each stream's effect on the decision at test time instead. The measurement is permutation importance (Breiman, 2001; Fisher et al., 2019) and the Shapley value (Shapley, 1953; Lundberg & Lee, 2017; Covert et al., 2020) applied to whole streams; what is new is using it to choose models and to report every result. Appendix B gives a longer discussion.

8 Conclusion

A decoder can become more accurate because it has stopped using the input the claim is about. We showed three ways this happens and built a decoder in which two of them are impossible. For the third, we made each stream's credit a number that is measured, used to choose models, and reported for every result. It shows that adding behavioural streams almost halves the brain's share of the credit, and that on other public datasets, high accuracy can hide an imbalance in the labels or a signal that cannot be confirmed as neural. Training against the measurement did not raise the brain's share. That the measurement resists this is the best evidence we have that it is measuring something real.

AI use statement

Generative AI tools were used substantively in this work and their use is disclosed here in full. An AI coding assistant was used to write the decoder, data and evaluation code in `src/` and to produce the first draft of this manuscript's text, working from the authors' prior analyses, design decisions and experimental results. It was not used to generate research ideas, to select which hypotheses to test, or to produce any experimental result: every number reported here comes from a training run whose configuration, execution and outputs are recorded on disk, and the manuscript's numbers are inserted mechanically from those outputs rather than written by hand, so that an unsupported cell is visible rather than silent (Appendix O). The authors have read and verified the code, re-derived the coupling-score and Shapley claims of Section 4 independently, and checked the cited literature against the primary sources. The authors take full responsibility for the final content of this work, including all text, claims and artifacts.

Ethics statement

The dataset was collected under institutional review with informed consent, and we use it as released; no new data from people was collected for this work. Attention decoding is meant for hearing devices. We state clearly that the behavioural streams measure where a listener faces, not what they attend to (Section 1), because a device that confused the two would fail exactly for the listeners who cannot turn their head or eyes.

Reproducibility statement

The decoder, the data code and all tests form one package with one configuration file, and every experiment in the paper is one command (Appendix O). All settings are listed in Appendix P, and none were tuned. At test time the order of the candidates is fixed by each window's index, so every model we compare sees exactly the same inputs, and the models inside one run share one dataset and one set of folds. Each test set is scored once, and the shuffled baseline behind every reported number is never used for training or for choosing a checkpoint.

REFERENCES

Bernd Accou, Mohammad Jalilpour Monesi, Jair Montoya, Hugo Van hamme, and Tom Francart. Predicting speech intelligibility from EEG in a non-linear classification paradigm. *Journal of Neural Engineering*, 18(6):066008, 2021.

Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Scientific Reports*, 13:812, 2023.

Ali Alavi and Donald S. Williamson. NeuroToken: Joint source and directional AAD with envelope decoding via conditional flow matching. arXiv preprint, 2026.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Virginia Best, Alex D. Boyd, and Kamal Sen. An effect of gaze direction in cocktail party listening. *Trends in Hearing*, 27:23312165231152356, 2023.

Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5): 402–412, 2017.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019.

E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *NeurIPS*, 2020.

Neetha Das, Tom Francart, and Alexander Bertrand. Auditory attention detection dataset KULeuven, 2019.

Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

Cunhang Fan, Xiaoke Yang, Hongyu Zhang, Ying Chen, Lu Li, Jian Zhou, and Zhao Lv. ListenNet: A lightweight spatio-temporal enhancement nested network for auditory attention detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Søren A. Fuglsang, Daniel D. E. Wong, and Jens Hjortkjær. EEG and audio dataset for auditory attention decoding, 2018.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Simon Geirnaert, Servaas Vandecappelle, Emina Alickovic, Alain de Cheveigné, Edmund Lalor, Bernd T. Meyer, Sina Miran, Tom Francart, and Alexander Bertrand. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4):89–102, 2021.

Simon Geirnaert, Iustina Rotaru, Tom Francart, and Alexander Bertrand. Linear stimulus reconstruction works on the KU Leuven audiovisual, gaze-controlled auditory attention decoding dataset, 2024. Dataset: `https://doi.org/10.5281/zenodo.11058711`.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL*, 2018.

Md Rakibul Hassan, Ali Alavi, and Donald S. Williamson. MAESTRO: A multimodal dataset for auditory attention decoding with simultaneous EEG, eye tracking, head motion and egocentric video. *arXiv preprint*, 2026. Benchmark code: `https://github.com/ASPIRE-OSU/MAESTRO`.

Maartje M. E. Hendrikse, Gerard Llorach, Volker Hohmann, and Giso Grimm. Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. *Trends in Hearing*, 23:2331216519872362, 2019.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*, 2020.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? In *ICML*, 2022.

Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martinez, Tom Francart, and Hugo Van hamme. An LSTM based architecture to relate speech stimulus to EEG. In *ICASSP*, 2020.

Gerald Kidd Jr., Tanya L. Arbogast, Christine R. Mason, and Frederick J. Gallun. The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118(6):3804–3815, 2005.

Gerald Kidd Jr., Sylvain Favrot, Joseph G. Desloge, Timothy M. Streeter, and Christine R. Mason. Design and preliminary testing of a visually guided hearing aid. *The Journal of the Acoustical Society of America*, 133(3):EL202–EL207, 2013. doi: 10.1121/1.4791710.

Edmund C. Lalor and John J. Foxe. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *European Journal of Neuroscience*, 31(1):189–193, 2010.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Qinke Ni, Hongyu Zhang, Cunhang Fan, Shengbing Pei, Chang Zhou, and Zhao Lv. DBPNet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5108–5116, 2024.

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.

Zexu Pan, Marvin Borsdorf, Siqi Cai, Tanja Schultz, and Haizhou Li. NeuroHeed: Neuro-steered speaker extraction using EEG signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4456–4470, 2024. doi: 10.1109/TASLP.2024.3463498.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Iustina Rotaru, Simon Geirnaert, Nicolas Heintz, Iris Van de Ryck, Alexander Bertrand, and Tom Francart. What are we really decoding? unveiling biases in EEG-based decoding of the spatial focus of auditory attention. *Journal of Neural Engineering*, 21(1):016017, 2024. doi: 10.1088/1741-2552/ad2214.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Servaas Vandecappelle, Lucas Deckers, Neetha Das, Amir Hossein Ansari, Alexander Bertrand, and Tom Francart. EEG-based detection of the locus of auditory attention with convolutional neural networks. *eLife*, 10:e56481, 2021.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

World Health Organization. World report on hearing. Technical report, World Health Organization, Geneva, 2021.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, 2022.

Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

Xiran Xu, Bo Wang, Yujie Yan, Xihong Wu, and Jing Chen. A DenseNet-based method for decoding auditory spatial attention with EEG. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1946–1950, 2024.

Sheng Yan, Cunhang Fan, Hongyu Zhang, Xiaoke Yang, Jianhua Tao, and Zhao Lv. DARNet: Dual attention refinement network with spatiotemporal construction for auditory attention detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024a.

Yujie Yan, Xiran Xu, Haolin Zhu, Xihong Wu, and Jing Chen. Four-speaker EEG dataset, 2024b.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

Ziyang Zhang, Andrew Thwaites, Alexandra Woolgar, Brian Moore, and Chao Zhang. SWIM: Short-window CNN integrated with Mamba for EEG-based auditory spatial attention decoding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

A MORE DETAIL ON THE THREE PROBLEMS

This section gives the steps behind Section 3 that the main text states briefly.

A.1 WHY MAKING THE CANDIDATES EQUALLY LOUD IS NOT ENOUGH

Every step of the envelope pipeline is linear, so scaling a recording by α scales its envelope by α: $\mathrm{env}(\alpha x) = \alpha\,\mathrm{env}(x)$. Standardising each candidate then subtracts its mean and divides by its standard deviation, which removes any change of the form $x \mapsto \alpha x + \beta$. Playing a recording 15 dB louder or quieter therefore gives exactly the same standardised envelope. A pure difference in loudness cannot survive, so it cannot be what gives the answer away.

What does survive is every property that scaling and shifting leave unchanged. Standardising fixes only two numbers per candidate, the mean and the variance. Skew, kurtosis, sparsity, ratios between percentiles and the balance between frequency bands all pass through untouched. If the attended and ignored recordings differ in any of these, no rescaling can help — and on this dataset they do: the envelope's peak-to-average ratio alone separates attended and ignored talkers with an effect size of $d = -1.22$. A fix must therefore either make the candidates' whole distributions identical (MATCHED) or remove the difference between attended and ignored recordings altogether (same-talker candidates). Reporting chance does not help either, since the mark lifts accuracy well above it.

A.2 HOW THE STANDARD SCORE REDUCES TO AN AUDIO CLASSIFIER

The standard score has one convolution stack for the recording and one, shared across candidates, for the audio. At each time step t it normalises the brain embedding c_t and the candidate embedding $a_{k,t}$ to unit length, multiplies them element by element, averages over time and applies a linear layer: $\mathrm{score}_k = w^\top\big(\frac{1}{T}\sum_t \hat{c}_t \odot \hat{a}_{k,t}\big) + b$. Now let the brain encoder output the same unit vector at every step, $\hat{c}_t = \hat{c}$. The brain term no longer depends on t, so it moves out of the average:

$$\mathrm{score}_k \;=\; w^\top\Big(\hat{c} \odot \tfrac{1}{T}\textstyle\sum_t \hat{a}_{k,t}\Big) + b \;=\; \tilde{w}^\top \bar{a}_k + b, \qquad \tilde{w} = w \odot \hat{c}. \tag{7}$$

This is a linear classifier on the time-averaged audio embedding $\bar{a}_k$ of each candidate, and it can read any audio mark the audio encoder learns to represent. The model reaches it simply by letting the brain encoder go constant, which is easier than learning to follow the speech. The time-centred score of Eq. 1 closes this route: with a constant brain embedding, $e - \bar{e} = 0$, every correlation is zero, and every candidate gets the same score.

A.3 WHY A JOINT SHUFFLE CANNOT SEE CROWDING OUT

Shuffling all recorded streams together measures how much the model uses the listener's signals as a whole. Suppose EEG alone earns a contribution of 0.40. Now add a gaze stream that is easier to read, and suppose the fused model earns 0.55 in total, but only 0.20 of it from the EEG. The joint contribution went *up*, from 0.40 to 0.55, while the brain's share fell by half. A joint shuffle reports the first change and hides the second. Only a split per stream, such as the Shapley values of Eq. 6, shows both. The numbers of Table 1 follow exactly this pattern.

A.4 WHY WE SHUFFLE THE RECORDING, NOT THE LABELS

A common check shuffles the labels and retrains or re-scores. That is the wrong test here. Shuffling the labels breaks the link between the audio and the answer as well as the link between the brain and the answer, so a decoder that reads only the audio also drops to chance, and the shortcut we are testing for is hidden. We instead shuffle the *recording* across test windows and keep each window's own candidates and label. A decoder that reads the audio keeps its accuracy; only a decoder that uses the recording loses any. The accuracy that remains, $v(\emptyset)$, is what *this* model gets from the audio alone, measured through its own audio pathway. The linear probe bounds what a simple reader can take from the audio; $v(\emptyset)$ measures what this decoder actually takes.

B RELATED WORK IN MORE DETAIL

Auditory attention decoding. Brain activity follows the envelope of the speech a listener attends to (Ding & Simon, 2012; Lalor & Foxe, 2010). Standard decoders either rebuild the envelope from

EEG and correlate it with each candidate (O'Sullivan et al., 2015), or learn the match directly (Accou et al., 2021; Jalilpour Monesi et al., 2020; Accou et al., 2023; Puffay et al., 2023). A separate line decodes the *direction* of attention without any audio (Vandecappelle et al., 2021). The work closest to ours is Rotaru et al. (2024), who showed that direction decoding on several public datasets is driven largely by eye movements and by regularities in the trial design, not by attention. We bring the same concern to the matching task and to several recorded streams at once, and we make the check part of how models are chosen and reported rather than a test run afterwards.

Shortcuts in datasets. Models use whatever solves the task (Geirhos et al., 2020), and benchmarks often contain cues that make the intended input unnecessary (Torralba & Efros, 2011; Gururangan et al., 2018; Goyal et al., 2017). The usual fixes are to rebalance the data (Goyal et al., 2017) or to train against a branch that sees only the shortcut (Cadene et al., 2019). MERIT does both — it checks the candidate sets and it trains an audio-only adversary (Ganin & Lempitsky, 2015) — and adds one thing specific to the matching task: a scoring function in which the shortcut is impossible rather than discouraged (Section 4).

Balancing modalities. When modalities are trained together, the easiest one dominates, and the fused network can end up worse than its best single part (Wang et al., 2020). This has been called modality competition (Huang et al., 2022) and a greedy learning process (Wu et al., 2022), and it is usually treated by evening out *training* — adjusting gradients as training runs (Peng et al., 2022), blending losses (Wang et al., 2020), or dropping modalities at random. MERIT looks at something different. Those methods balance how fast each modality learns; a scientific claim instead rests on how much each modality affects the decision *at test time*, which is what Eq. 6 reports. The two are not the same: in Table 1, turning modality dropout on or off changes the EEG's credit very little, while the EEG's credit is still far below what it earns alone.

Attribution. Permutation importance (Breiman, 2001; Fisher et al., 2019) measures how much a model depends on an input by breaking that input's link to the rest of the example. Shapley values (Shapley, 1953; Lundberg & Lee, 2017) give the only way to split credit that adds up exactly, and SAGE (Covert et al., 2020) uses them for importance measured by performance over a whole dataset. We apply the same idea to whole streams, where there are few enough subsets (2^M) to compute it exactly; Hessel & Lee (2020) make a related point for multimodal models. What is new is not the measurement but how it is used: to choose checkpoints, and to report the brain's share alongside every accuracy. We also tried adding it to the training loss, and report that this does not help (Section 6.2).

Components. The contrastive term is InfoNCE (van den Oord et al., 2018) in the two-way arrangement of CLIP (Radford et al., 2021), with negatives from the same listener; the anti-collapse term follows VICReg (Bardes et al., 2022); and the foveal stream is a frozen V-JEPA 2 encoder (Assran et al., 2025).

C CHECKING THE CANDIDATE SETS

Table 3 gives the audio probe for every candidate set. The probe is a logistic regression on eight statistics of each standardised envelope that do not change when the signal is rescaled: kurtosis, skew, Gini sparsity, the range between the 5th and 95th percentiles, the fraction of near-silent samples, and three frequency-band ratios. It is trained and tested on different audio content, and picks the candidate it rates most likely to be attended. A candidate set that does not give the answer away must score $1/K$.

D ENCODER DESIGN

Each stream goes through five dilated temporal convolutions (Yu & Koltun, 2016) with dilations $1, 2, 4, 8, 16$ and GroupNorm (Wu & He, 2018) after each layer. The EEG first passes through a 1×1 convolution that maps the electrodes to 8 virtual sensors. Three choices matter. **The window matches the brain response.** The encoder sees 63 samples ($0.98\,\mathrm{s}$), about the time over which the cortical response to speech unfolds (Lalor & Foxe, 2010). A much larger window would make the deepest layers mostly see padding, which is the same in every window, so their output would stop depending

Table 3: **Audio-only acceptance probe**: what a linear reader takes from the candidates alone, with no access to any recording. A shortcut-free construction must sit at chance.

Construction		K	Probe	Excess over chance
Competing talkers	UNMATCHED	4	0.5581	+0.3081
+ distribution matching	MATCHED	4	0.2603	+0.0103
Same-talker temporal negatives	SAME-TALKER	3	0.3758	+0.0425
+ distribution matching	SAME-TALKER MATCHED	3	0.3564	+0.0231
Same-talker temporal negatives	SAME-TALKER	2	0.5131	+0.0131
+ distribution matching	SAME-TALKER MATCHED	2	0.5134	+0.0134

on the input. **There is no activation on the last layer.** A rectified output has no negative values, so the cosine between any two embeddings is squeezed into $[0, 1]$, and the correlations the score depends on would all sit close to one. **Padding is symmetric.** The brain responds to a sound 100 to 300 ms after it, so an encoder that looks only backward in time cannot see the response. One-sided versions are used on purpose to restrict the model to delays after, or before, the sound (Section 6.4).

E FURTHER LIMITATIONS

Part of the four-way contribution can come from the trial. The attended talker is the same in every window of a trial, so a decoder that recognises *which trial* it is looking at already knows part of the answer. The same-talker results do not have this issue, because all candidates there come from the same trial and the same voice.

One room, sixteen listeners, no tuning. The data come from one room, one language and sixteen listeners, and the frozen video stream in particular reads *this* room. We did not tune any setting. The numbers are therefore not tuned in our favour, but we also cannot claim they are the best this model can reach.

The cause of the audio mark is not tested directly. Every recording is used in only one role, so the role and the recording cannot be separated. The dataset's authors explain the mark by clipping during the loudness adjustment (Hassan et al., 2026), but the released data cannot test this directly. No result depends on it: the probe measures how readable the mark is, whatever its cause, and both fixes work either way. A dataset in which each recording is the attended talker in some trials and an ignored one in others would remove the mark at the source.

F PER-STREAM CREDIT, DRAWN

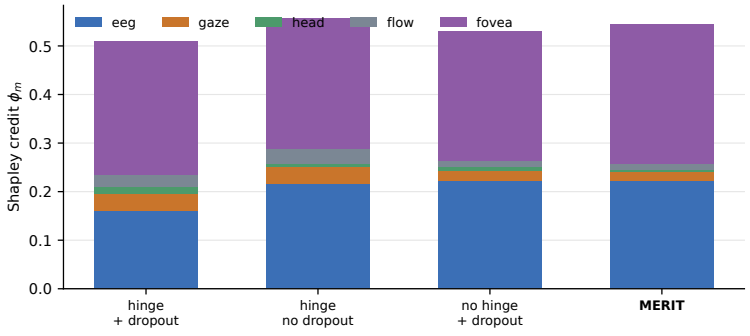


Figure 2: **Shapley credit per stream**, stacked, for the four fused rows of Table 1. Each bar's total height is that model's contribution $v(M) - v(\emptyset)$, decomposed exactly. The question the figure answers is not how tall the bars are but how much of each one is blue.

G WHY THIS DATASET, AND NOT ANOTHER

Table 4 compares the public AAD datasets we know of on the two things that matter here: which streams are recorded next to the EEG, and whether the listener is free to produce them.

Table 4: **Public AAD corpora.** "EOG" denotes a few electro-oculography electrodes recorded for artifact rejection, which is not calibrated gaze and carries no head or scene information. The last column is the one that decides whether modality freeloading is observable at all.

Corpus	Gaze	Head	Scene video	Presentation	Listener
DTU (Fuglsang et al., 2018)	6 EOG	—	—	simulated rooms, earphones	soundproof booth
Four-speaker (Yan et al., 2024b)	—	—	—	4 loudspeakers, anechoic	head fixed by chinrest
AV-GC-AAD (Geirnaert et al., 2024)	4 EOG	—	—	earphones, simulated HRTFs	gaze controlled by design
MAESTRO (Hassan et al., 2026)	**eye tracker**	**IMU**	**egocentric**	**6 loudspeakers, real room**	**unconstrained**

This pattern is not a flaw in those datasets. It is good design for the question each one was built to answer. Holding the head and eyes still removes a source of confusion, and AV-GC-AAD goes further and controls gaze on purpose, so that a decoder that reads gaze can be caught. That makes it the best available test of *whether* a decoder uses gaze instead of the brain. But that is a yes-or-no question about one stream. Our question is about amounts: when a listener produces five streams at once, how does a fused decoder divide its evidence among them, and how much stays with the brain? A dataset in which four of the five streams are held fixed cannot answer it, because the effect has been designed out. Testing MERIT there would not be a harder test; it would be a test in which the effect cannot appear.

Two parts of MERIT do not need these streams and work on any matching dataset: the audio probe, which only looks at the candidates, and the coupling score, whose guarantee is a matter of arithmetic. We test both on two public datasets in Appendices H and I.

H THE AUDIO PROBE ON TWO OTHER DATASETS

A check that only flags problems in the authors' own dataset is not yet a reliable tool. We therefore ran the probe of Section 3(i), unchanged, on two public two-talker datasets collected by other groups: KU Leuven (Biesmans et al., 2017; Das et al., 2019) and DTU (Fuglsang et al., 2018), with 16 and 18 listeners. Neither records gaze, head motion or video, so only the audio part of MERIT applies — which is the point, since the probe depends only on the candidates. We use one envelope method for all three (power-law compression with exponent 0.6, a 20 Hz low-pass filter, and resampling to 64 Hz), standardise each candidate, put the candidates in random order, and hold out whole stimulus pairs for testing. Chance is 0.5.

Table 5: **The probe transfers, and it is calibrated.** Audio-only two-way accuracy on 10 s windows, content-disjoint folds, no EEG. The last column is the corpus's prior on which talker is attended; the probe tracks it.

Corpus	Listeners	Windows	Probe	Excess	Attended-role prior
KU Leuven	16	9,564	**0.6947**	**+0.1947**	0.800 track 1
DTU	18	9,720	0.5008	+0.0008	0.494 male

One dataset is clean, and that is what gives the other its meaning. Across four window lengths (Table 6) the probe stays within 0.005 of chance on DTU. The probe is not a warning that goes off everywhere. Given two talkers who are attended about equally often, a classifier on eight scale-free statistics cannot tell which one was attended, as it should not be able to.

On KU Leuven, the audio alone gives +0.1947 above chance. On the most widely used AAD benchmark, the same classifier finds the attended story with accuracy 0.6947 on 10 s windows, against a chance of 0.5, and it is above chance at every window we tried. The cause is not the audio mark of Section 3(i) but an imbalance in the labels: track 1 is the attended one in 256 of 320 trials, because

Table 6: **The same probe at four decision windows**, excess over chance. On DTU, where the attended-role prior is balanced, no cell exceeds 0.005. The KU Leuven excess is present at every window; its 30 s cell rests on 4,608 windows drawn from eight stimulus pairs and is the least powered of the four.

Corpus	2 s	5 s	10 s	30 s
KU Leuven, as released	+0.0945	+0.1384	**+0.1947**	+0.0729
distribution-matched	+0.0321	+0.0401	+0.0974	+0.0208
DTU, as released	+0.0010	+0.0001	+0.0008	-0.0019

the third part of the experiment repeats the first, and in the first the listener always attends track 1. With two fixed talkers and one of them attended far more often, the talker's identity predicts the label, and voice identity is exactly what these statistics carry. The dataset's own notes warn about two other problems, trial-level overfitting and gaze; this is a third, separate one. No choice of split can remove it, because it is part of the design, not of the split.

Matching the distributions only halves it. Quantile matching lowers the excess from +0.1947 to +0.0974, and no further. This tells us something. Matching makes every candidate's set of values the same, which removes the shape statistics. It keeps each candidate's order in time, so the balance between frequency bands survives, and with two fixed talkers that balance still reveals who is speaking. When the problem is a label imbalance between two voices, the fix is to balance the labels, not to match the distributions. This is a limit of the candidate set we recommend, found by running our own tool on someone else's data — and it is why we report the probe for each dataset instead of calling one candidate set safe everywhere.

I THE COUPLING BRANCH ON TWO OTHER DATASETS

The probe of Appendix H needs no model. Here we run the other part of MERIT that needs no behavioural stream — the EEG coupling branch, its shuffled baseline and the direction-in-time check — on the same two datasets. It runs through exactly the same code as every MAESTRO result: the dataset is handed to the runner through the same data interface, and nothing after that is changed. Candidates are matched, $K = 2$, chance is 0.5, and $W = 10$ s. *Same talker* replaces the competing talker with the attended talker's own speech from another part of the same trial, so neither the talker's identity nor their position can tell the candidates apart.

Table 7: **What the full battery says about two public datasets.** Accuracy and contribution for the unconstrained coupling branch, then contribution with the branch confined to the causal ($+125 \cdots +1109$ ms) and mirrored acausal ($-1109 \cdots -125$ ms) lag bands. *Folds+* is for the unconstrained branch. The last row is MAESTRO's same-talker $K = 2$ result, the reference for what a directional response looks like.

	Accuracy	Contribution	Causal	Acausal	Folds+
KU Leuven, within, competing	0.7252	+0.2141	+0.2104	+0.2074	5/5
KU Leuven, within, same talker	0.8149	+0.3150	+0.3115	+0.3080	5/5
KU Leuven, LOSO, competing	0.7138	+0.1766	+0.1764	+0.1687	16/16
DTU, within, competing	0.7590	+0.2584	+0.2560	+0.2499	5/5
DTU, within, same talker	0.8170	+0.3175	+0.3164	+0.3145	5/5
DTU, LOSO, competing	0.7269	+0.2264	+0.2006	+0.2060	18/18
MAESTRO, within, same talker	0.6896	+0.1905	**+0.2959**	**+0.1478**	5/5

The coupling branch transfers, and passes the hardest test we have. On KU Leuven and DTU the coupling branch reaches 0.7252 and 0.7590 within listeners, positive in every fold, and holds up

on unseen listeners (0.7138 and 0.7269, positive in 16/16 and 18/18 folds). The same-talker test is the one that matters. It removes talker identity, and with it a route that needs no timing at all: the EEG could reveal *which side* was attended — through brain activity that leans to one side, or through small eye movements, which are known to affect datasets of this kind (Rotaru et al., 2024) — and the audio could reveal which talker is which. Accuracy does not fall under this test; it rises, to 0.8149 and 0.8170. Whatever the branch reads, it reads it from the timing.

But the direction-in-time check does not confirm it is the brain. On MAESTRO, limiting the coupling to delays after the sound gives about twice the contribution of the mirror image before the sound (+0.2959 against +0.1478), as a response to the sound must. On KU Leuven and DTU the two are the same in every row of Table 7. A match that depends on timing but not on its direction is what a signal at *zero* delay would give, spread equally into both directions by the slow changes of the speech envelope itself. Both datasets play the audio through in-ear phones close to the electrodes, and an electrical leak of the sound into the recording is the classic zero-delay artifact; MAESTRO plays it through loudspeakers across a room, and there the check passes. We think this is the most likely explanation, but we have not shown it, and we do not count the KU Leuven or DTU accuracies as brain decoding.

Against published results, at the same window. Accuracy on these two datasets depends strongly on how the data are split. Alavi & Williamson (2026) re-ran five recent decoders on KU Leuven at 5 s, once with each paper's own split and once with a split that holds out whole trials, allows no overlap between training and test windows, and fits all preprocessing on training trials only. Table 8 gives the result next to NeuroToken's own numbers and ours, all at 5 s. Chance is 0.5.

Table 8: **Two-talker accuracy at 5 s: with each paper's own split, and without leakage.** The first five rows decode the *direction* of attention from EEG alone, and are NeuroToken's re-runs on KU Leuven listeners 1–4 (Alavi & Williamson, 2026); the headline numbers in those papers are higher still, 95–97% at 1 s. NeuroToken and MERIT decode the attended *talker* from EEG and the candidate speech. NeuroToken holds out whole trials; MERIT holds out whole stimuli and uses matched candidates. "Unseen" means listeners not seen in training.

		KU Leuven		DTU
Method	Decodes	own split	no leakage	no leakage
DARNet (Yan et al., 2024a)	direction	0.895	0.552	—
ListenNet (Fan et al., 2025)	direction	0.930	0.484	—
DBPNet (Ni et al., 2024)	direction	0.884	0.483	—
DenseNet-3D (Xu et al., 2024)	direction	0.864	0.698	—
SWIM (Zhang et al., 2025)	direction	0.644	0.353	—
NeuroToken, within listeners	talker	—	0.864	0.837
NeuroToken, unseen listeners	talker	—	0.634	0.592
MERIT, within listeners	talker	—	0.6763	0.6909
MERIT, unseen listeners	talker	—	0.6664	0.6635

Reading the table. With each paper's own split, the direction decoders reach 0.64 to 0.93. Once leakage is removed they fall by 17 to 45 points, to between 0.35 and 0.70, and several land near or below chance: the published numbers mostly measured the split, not the decoder. NeuroToken is the strongest prior result that survives a split without leakage. **On unseen listeners, MERIT is more accurate than NeuroToken on both datasets**: 0.6664 on KU Leuven and 0.6635 on DTU (Table 8), and positive for every one of the 16 and 18 held-out listeners. Within listeners it is not: NeuroToken is more accurate than our 0.6763 and 0.6909 on both, and DenseNet-3D is also higher on the four KU Leuven listeners it was run on. The five direction decoders were re-run within listeners only, so they have no unseen-listener number to compare. Two differences from our setting remain. First, NeuroToken holds out whole trials, while we hold out whole stimuli. KU Leuven repeats stories across trials — its third part replays the first — so a split by trial can still put the same story in training and test, and a split by stimulus cannot. Second, our candidates are matched, while on KU Leuven the candidates as released let the audio alone reach +0.1384 above chance at 5 s (Table 6). For our

own rows, the caution of this section still applies: on these two datasets the direction-in-time check does not confirm that the coupling branch reads a brain response. We have not run that check on the other methods.

This is the argument of the paper, on other people's data. A standard evaluation would report 0.82 on DTU and stop there. The audio probe clears the candidates; the same-talker test rules out identity and position; and the direction check then declines to confirm what is left. Each step is cheap, and without the last one the number would have been reported as brain decoding.

J THE HINGE EXPERIMENT ON UNSEEN LISTENERS

Table 9: **The 2×2 of Table 1, both protocols.** ϕ_{EEG} is the EEG's Shapley credit; *Folds+* counts leave-one-listener-out folds with a positive contribution. No arm is best in every column.

	Within-subject		Leave-one-listener-out		
Arm	Accuracy	ϕ_{EEG}	Accuracy	ϕ_{EEG}	Folds+
Hinge, modality dropout	0.7884	+0.1600	0.7469	+0.1149	16/16
No hinge, modality dropout	0.8091	+0.2229	0.7625	+0.1560	16/16
MERIT (no hinge, no dropout)	0.8173	+0.2217	0.7438	+0.1762	16/16
Hinge, no dropout	0.8242	+0.2169	0.7219	+0.1479	16/16

The order of the four models changes between the two splits. The model with the hinge and no dropout is the most accurate within listeners and the least accurate on unseen ones. In both pairs, the model without the hinge keeps more credit on the brain on unseen listeners. MERIT is the model without the hinge and without dropout: it has the most EEG credit on unseen listeners and ties for the most within listeners.

K EVERY STREAM SET, BOTH SPLITS

Table 10 gives every stream set within listeners, and Table 11 gives the same models on unseen listeners. Both use per-window decisions at $W = 10$ s.

L CHECKS ON THE EEG SIGNAL

Table 12 gives the same-talker results and the direction-in-time check. Figure 3 shows how the contribution changes with the decision window.

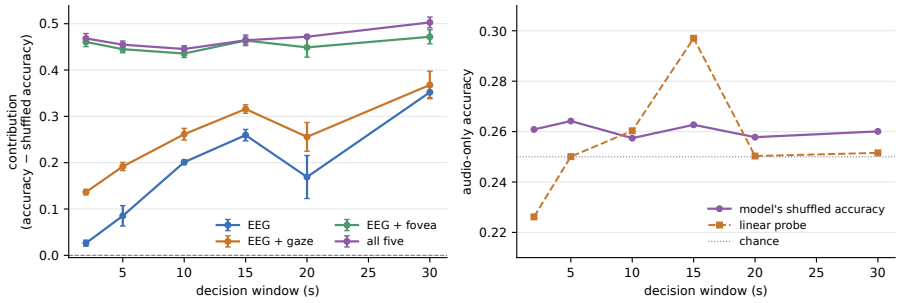


Figure 3: **Contribution against decision-window length**, within-listener, MATCHED. Left: contribution for the four stream sets measured at every window; the eleven stream sets at 2 and 5 s are in Table 13. Right: the full model's shuffled accuracy, which is what the contribution subtracts, and the linear audio probe, both measured at each window. The model's shuffled accuracy stays within 0.015 of chance throughout; the probe varies more, between 0.23 and 0.30.

Table 10: **Accuracy, null and contribution by stream set**, within-subject (five content-disjoint folds), MATCHED, $W = 10$ s, chance 0.25. *Folds+* counts folds with a positive contribution. The leave-one-listener-out counterpart is Table 11.

Streams	Accuracy	Null	Contribution	p	Folds+
EEG	0.4788	0.2882	+0.1906	0.0476	5/5
EEG + gaze	0.5295	0.2610	+0.2685	0.0476	5/5
EEG + head	0.4995	0.2589	+0.2405	0.0476	5/5
EEG + flow	0.5325	0.2854	+0.2471	0.0476	5/5
EEG + fovea	0.6909	0.2550	+0.4358	0.0476	5/5
Behaviour only	0.6552	0.2541	+0.4010	0.0476	5/5
All five	0.7101	0.2597	+0.4504	0.0476	5/5

Table 11: **Leave-one-listener-out**, sixteen folds, each with a held-out listener on held-out content. MATCHED, $W = 10$ s, chance 0.25.

Streams	Accuracy	Null	Contribution	p	Folds+
EEG	0.4831	0.3130	+0.1701	0.0863	15/16
EEG + gaze	0.4844	0.2760	+0.2084	0.0506	16/16
EEG + head	0.4944	0.2782	+0.2162	0.0536	16/16
EEG + flow	0.5013	0.2767	+0.2245	0.0506	16/16
EEG + fovea	0.6700	0.2649	+0.4051	0.0476	16/16
Behaviour only	0.6300	0.2677	+0.3623	0.0476	16/16
All five	0.6706	0.2668	+0.4038	0.0476	16/16

L.1 SHORT DECISION WINDOWS, BY STREAM SET

Figure 3 changes the window for a few stream sets. Table 13 does the opposite at the two shortest windows: every stream set at $W = 2$ s and $W = 5$ s, within listeners, one decision per window, with half-window hops. The rows extend Figure 3 downward and can be compared directly with a benchmark that decides one window at a time. Short windows are where the problem this paper is about is strongest. The behavioural streams show *where* the listener faces, which a two-second look already settles. The coupling branch needs time for a brain response to build up, and its own window is already 0.98 s. The table shows this clearly. At $W = 2$ s the EEG alone contributes +0.0263, while the frozen foveal embedding alone reaches 0.7130 with a contribution of +0.4570, and the full five-stream model reaches 0.7291. At that window a fused accuracy is almost entirely about where the listener was looking. It should not be read as a brain result, and without the per-stream split nothing in the headline number would say so.

M THE HINGE EXPERIMENT AT TEN SECONDS

Table 14 repeats the hinge experiment at the earlier setting ($W = 10$ s, one decision per window). It also changes how checkpoints are chosen, which Table 1 does not. Choosing by contribution instead of by accuracy is the one training choice that raises the EEG's credit.

N THE TRAINING OBJECTIVE, EXPLAINED

Equation 2 gives the loss. This section explains two choices behind it. **Why the contrastive term uses one listener at a time.** Raw EEG identifies the listener with accuracy 0.90 in a sixteen-way test, against a chance of 0.0625. In a batch that mixes listeners, the contrastive term could match EEG to speech by recognising the listener, which says nothing about attention. Every batch is therefore drawn from a single listener. **Why the contrastive term and the anti-collapse term cannot be satisfied by collapse.** If the EEG encoder outputs a vector that does not change over time, the time-centred score gives $s(e, a) = 0$ for every pair, so every entry of S in $\mathcal{L}_{\text{con}}$ is zero and the loss is exactly $\log B$, its value at chance. And the first part of $\mathcal{L}_{\text{col}}$ measures exactly the spread over time that such an encoder

Table 12: **Orienting-free decoding and neural directionality.** EEG only unless stated, within-subject, $W = 10\,\text{s}$. Same-talker candidates are the attended talker at different times, so slot index carries no spatial meaning and no orienting cue can help. The lag band is set by the encoder's direction plus an eight-sample shift; a contribution as large in the acausal band would indicate a shared artifact rather than envelope tracking.

Candidates	K	Streams	EEG visible at	Chance	Accuracy	Null	Contribution
SAME-TALKER MATCHED	3	EEG	centred	0.333	0.5474	0.3318	**+0.2156**
SAME-TALKER MATCHED	3	all five	centred	0.333	0.5578	0.3321	**+0.2257**
SAME-TALKER MATCHED	2	EEG	centred	0.500	0.6981	0.4995	**+0.1986**
SAME-TALKER MATCHED	2	all five	centred	0.500	0.6857	0.4988	**+0.1869**
MATCHED	4	EEG	$+125 \cdots + 1109\,\text{ms}$	0.250	0.4990	0.2750	**+0.2240**
MATCHED	4	EEG	$-1109 \cdots - 125\,\text{ms}$	0.250	0.4171	0.3204	**+0.0967**
SAME-TALKER MATCHED	2	EEG	$+125 \cdots + 1109\,\text{ms}$	0.500	0.8085	0.5126	**+0.2959**
SAME-TALKER MATCHED	2	EEG	$-1109 \cdots - 125\,\text{ms}$	0.500	0.6486	0.5008	**+0.1478**

Table 13: **Every stream set at the two shortest decision windows**, within-subject (content-disjoint, five folds), MATCHED, $K = 4$, chance 0.25, hop $= W/2$. *Contribution* is accuracy minus the model's own permutation null; p is the permutation p-value, floored at 0.048 for twenty permutations. The first block is single streams, the second EEG paired with each behavioural stream.

	$W = 2\,\text{s}$				$W = 5\,\text{s}$			
Streams	Acc.	Null	Contrib.	p	Acc.	Null	Contrib.	p
EEG	0.3281	0.3018	+0.0263	0.0647	0.3954	0.3101	+0.0853	0.0418
gaze	0.3448	0.2304	+0.1144	0.0050	0.3349	0.2236	+0.1113	0.0050
head motion	0.3541	0.2389	+0.1151	0.0050	0.3800	0.2456	+0.1345	0.0050
optical flow	0.4109	0.2477	+0.1632	0.0050	0.4126	0.2458	+0.1668	0.0050
fovea (V-JEPA 2)	0.7130	0.2560	+0.4570	0.0050	0.7096	0.2566	+0.4530	0.0050
behaviour (all four)	0.7174	0.2575	+0.4599	0.0050	0.6857	0.2575	+0.4282	0.0050
EEG + gaze	0.4167	0.2803	+0.1365	0.0050	0.4845	0.2927	+0.1918	0.0050
EEG + head	0.4134	0.2952	+0.1182	0.0050	0.4914	0.3198	+0.1716	0.0050
EEG + flow	0.4509	0.2939	+0.1569	0.0050	0.4994	0.3003	+0.1992	0.0050
EEG + fovea	0.7251	0.2645	+0.4606	0.0050	0.7198	0.2748	+0.4450	0.0050
all five	0.7291	0.2608	+0.4682	0.0050	0.7189	0.2642	+0.4547	0.0050

has lost. MERIT trains without modality dropout; the dropout models of Table 1 drop each stream with probability 0.3, never all of them at once, and fill a dropped stream with zeros so the fusion layer always runs.

O REPRODUCIBILITY

The decoder, the data code and all tests are one package with one configuration file, and every experiment is one command. The main results come from:

```
# Table 1 and Appendix J: the four fused models, both splits
python -m src.main data=merit model=merit
runner=merit_hinge2 $CROP \
data.euclidean_align=true data.audio_feats=env_onset
python -m src.main data=merit model=merit
runner=merit_hinge2loso0 $CROP ...
# Table 2:  the EEG-only ladder
python -m src.main data=merit model=merit
runner=merit_ladder2 $CROP \
data.audio_feats=env_onset
# Appendices:  stream sets, short windows, other datasets
```

Table 14: **Hinge, selection rule and modality dropout at $W = 10\,\mathbf{s}$**, per-window decisions — the earlier operating point, kept because it varies the selection rule, which Table 1 does not. Identical streams, data and folds; within-subject, MATCHED, $W = 10\,\mathrm{s}$, chance 0.25. ϕ_m is Shapley credit and sums to the contribution by construction. The single-stream EEG model is shown for reference.

				Shapley credit ϕ_m				
Model	Accuracy	Null	Contrib.	EEG	gaze	head	flow	fovea
EEG alone (reference)	0.4788	0.2882	+0.1906	+0.1906	—	—	—	—
No hinge, accuracy-selected	0.7154	0.2600	+0.4554	+0.1039	+0.0296	+0.0087	+0.0178	+0.2954
No hinge, credit-selected	0.7250	0.2639	+0.4611	+0.1182	+0.0307	+0.0109	+0.0197	+0.2816
Hinge, accuracy-selected	0.7087	0.2602	+0.4486	+0.0882	+0.0442	+0.0156	+0.0270	+0.2736
Hinge, credit-selected	0.7088	0.2614	+0.4473	**+0.0931**	+0.0446	+0.0186	+0.0258	+0.2653
Hinge, uniform weights	0.7283	0.2594	+0.4688	+0.0785	+0.0362	+0.0124	+0.0282	+0.3135
Hinge, no modality dropout	0.7007	0.2549	+0.4458	+0.1236	+0.0393	+0.0156	+0.0297	+0.2376
No hinge, no modality dropout	0.7226	0.2542	+0.4684	+0.1207	+0.0254	+0.0064	+0.0161	+0.2999

```
python -m src.main data=merit model=merit
runner=merit_streams
python -m src.main data=merit model=merit
runner=merit_winmod0 data.window_sec=2 data.hop_sec=1
python -m src.main data=external model=merit
runner=merit_external data.corpus=dtu
python -m src.main mode=certify data=merit model=merit
runner=merit
```

where `$CROP` is `data.window_sec=30 data.hop_sec=15 data.crop_len=640 data.n_train_crops=8 data.n_eval_crops=5`. The job scripts that ran each experiment are released with the code.

The models inside one run share one dataset and one set of folds, so a difference between two rows comes from the models alone. At test time the order of the candidates is fixed by each window's index, so every model sees exactly the same inputs. Every number in the paper is written from the run outputs by a script and inserted automatically, never typed by hand, and a result is only used once every fold of its run has finished.

P HYPER-PARAMETERS

Q WHY THE SHAPLEY CREDITS ADD UP

Equation 6 guarantees $\sum_m \phi_m = v(M) - v(\emptyset)$ exactly. The code reports both sides for every run as a check, and they agree to the printed precision in every result in this paper. All $2^5 = 32$ subsets are computed rather than sampled, and the same shuffles are reused across subsets, so a difference between two subsets is not shuffle noise. The cost is mostly in re-scoring, not re-encoding, because the stream embeddings and audio embeddings are computed once per evaluation.

Two cautions. First, ϕ_m is the credit *for this trained model on this test set*, not a fixed property of the stream: a stream may carry information the model did not use. Second, when two streams overlap — gaze and the foveal embedding both describe where the listener looks — Shapley splits their shared credit between them. That is the right answer to "how much would I lose by removing this one, averaged over the order of removal," but it does not mean that each stream holds half of the information.

R THREE PER-STREAM MEASURES

The runs report three per-stream numbers, which answer different questions. The *contribution* $v(M) - v(M \setminus \{m\})$ is the accuracy lost by shuffling stream m while all others stay intact; it is

Table 15: Settings. None were tuned; each was set once from first principles.

Item	Value
Sampling rate / window / hop	64 Hz / 10 s / 5 s for per-window results (sweep: $2, 5, 10, 15, 20, 30$ s at 50% hop)
Per-trial decisions (MERIT)	10 s crops of 30 s trials: 8 random crops per trial when training, 5 fixed crops at test; window scores summe
Stream channels	EEG 32 (64 on KU Leuven and DTU), gaze 6, head 6, flow 4, fovea 1024 (frozen)
Descriptor width D / layers	16 / 5 (dilations $1, 2, 4, 8, 16$; RF 63 samples $= 0.98$ s)
Spatial filters (EEG)	8
Normalisation / activation	GroupNorm(4 groups) after every conv; no activation on the last
Dropout / modality dropout	0.1 / none in MERIT; 0.3 in the dropout models of Table 1 (never all streams; dropped streams filled with ze
Delays (MERIT)	EEG encoder sees only delays after the sound, offset 8 samples ($+125$ to $+1109$ ms)
Listener correction (MERIT)	rank-2 correction of the EEG channels per listener; in the fused model the EEG is also whitened per listener
Audio channels (MERIT)	envelope and its rectified derivative (speech onsets)
Coupling temperature	learned, initialised at $1/0.07$
Optimiser / lr / weight decay	AdamW / 10^{-3} / 10^{-4}; grad-norm clip 1.0
Batch	32, drawn from a single listener
Schedule	lr $\times 0.5$ after 5 epochs without improvement; ≤ 50 epochs; patience 12
Label smoothing	0.1
Loss weights (Eq. 2)	task 1.0 (label smoothing 0.1); auxiliary 0.3; contrastive 1.0; EEG null hinge 0.5 (margin $\mu = 0.5$); anti-co
Contribution hinge (Eq. 3)	none in MERIT; weight 0.5, margin $\delta = 0.3$ in the hinge models
Hinge weights	Eq. 4 in Table 1; $\lambda_m = \mathrm{clip}((\kappa - c_m)/\kappa, 0, 1)$, $\kappa = 0.05$, in Table 14
Checkpoint choice	Eq. 5, $\beta = 1$ (accuracy recorded as well)
Permutations	3 on inner validation; 20 at test; 5 draws per Shapley coalition
Splits	within: 5 content-disjoint folds, 20% inner val (content-disjoint)
	LOSO: 16 folds, 20% global content holdout, inner val listener-disjoint

small for a stream whose information is also in another stream. The *Shapley credit* ϕ_m is the same quantity averaged over all subsets of the other streams, and it is the one that adds up to the total. *Leave-one-out* is the accuracy lost by *zeroing* stream m; it differs from shuffling when the model uses a stream's overall level rather than its window-by-window values. All three are in the released result tables.

S DATASET

Design. MAESTRO (Hassan et al., 2026) has sixteen listeners with one hundred trials each. In every trial four talkers speak at once from loudspeakers at fixed, known positions. The listener is told before the trial which talker to attend to and answers a question about it afterwards. Trials last about 30 s, and each loudspeaker is the attended one equally often. The talkers of a trial are different recordings and no recording is used twice, so no audio is shared between training and test. The attended talker is played about 15 dB louder than the others in every trial; this is why the difference between loudness and shape in Section 3(i) matters.

Streams. Next to the EEG, the dataset records gaze, head motion and a head-mounted scene video. All signals are resampled to 64 Hz and standardised per channel over each trial. This gives four time-series streams — EEG (32 channels), gaze (6), head motion (6) and optical-flow summaries of the scene video (4) — and, for each talker, the speech envelope. We add a fifth stream that the released benchmark does not use: a *frozen* V-JEPA 2 ViT-L (Assran et al., 2025) embedding of the part of the scene video around the gaze point, one 1024-d vector per window, never trained.

Examples. Signals are cut into windows of W seconds with hop H; unless stated, $W = 10$ s and $H = 5$ s, which gives $T = 640$ samples and 7,990 windows. One example holds the streams $x^m \in \mathbb{R}^{T \times C_m}$; K candidate envelopes, exactly one of them correct; the index y of the correct one; the attended loudspeaker; the permutation π that records which loudspeaker is in which candidate slot; and the listener's identity. The candidates are put in a random order in every window, so their position says nothing. At test time that order is fixed by the window's index, so it is the same for every model we compare.

Two details of the candidate sets that matter. First, same-talker negatives must be drawn *at random* from windows of the same trial that do not overlap the target window. Taking the window furthest from the target seems natural, but it pushes the negatives toward the start and end of the trial, and the audio probe then learns the position: it reached 0.60 on a task with a chance of 0.50. Second, a negative must never overlap the target window; when a trial is too short to provide enough non-overlapping windows, we lower K rather than allow overlap. For per-trial decisions, the training crops are drawn at random positions and the candidates are matched separately in each crop, and the same-talker negatives are random non-overlapping crops of the same trial.