THE ALIGNMENT BOUNDARY IN CROSS-SUBJECT IMAGINED-SPEECH EEG: PROVABLE LIMITS OF PRE-ALIGNMENT AND IN-TRAINING ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Imagined-speech brain–computer interfaces (BCIs) must generalize across people: a decoder trained on some subjects should work on an unseen subject without recalibration. Two families of remedies dominate—*pre-alignment* (fixed transforms before the decoder, e.g. Euclidean Alignment and its class-conditional refinements) and *in-training alignment* (objectives during training, e.g. adversarial or correlation-based domain adaptation, and episodic meta-learning). We prove both families are fundamentally bounded and characterize the boundary precisely. On the pre-alignment side, class-conditional Euclidean alignment (CC-EA) is not an improvement over marginal EA but a *catastrophe*: it whitens each class covariance to the identity, erasing the between-class contrast that Common Spatial Patterns (CSP) decoders read, and accuracy collapses to chance ($72.3\% \to 49.3\%$ on BCI IV-2a; replicated on Cho2017, 49 subjects, $p < 10^{-11}$). A gauge-invariance argument proves the CSP decoder is unchanged by *any* class-agnostic linear pre-alignment, making marginal EA optimal within the linear family up to a residual rotation we rule out empirically. On the in-training side, we prove minibatch objectives such as DANN and CORAL carry *zero* expected cross-subject $\mathcal{H}$-divergence signal, while episodic prototype alignment is a principled $\mathcal{H}$-divergence surrogate; the predicted gradient alignment is confirmed across $29/30$ proxy–dataset pairs ($p < 0.001$). Finally, the two bounds *combine* on imagined speech: on a new multilingual benchmark (Telugu, Kannada, English; consumer-grade 8-channel EEG), cross-subject decoding is near chance for every method—not a failure but the quantitative prediction of the $\mathcal{H}$-divergence bound at the available signal-to-subject ratio. Motor imagery serves throughout as a signal-present control. Our analysis converts assumed "better alignment" heuristics into precisely characterized boundaries and tells the field where cross-subject gains cannot come from.

1 INTRODUCTION

Cross-subject alignment in EEG has a fixed *transfer budget*. The Ben-David bound writes a decoder's target error as source error plus a cross-subject divergence term plus an irreducible joint risk; reducing the divergence is the only lever alignment methods pull. This paper's organizing claim is that this budget is spent in provably separated places, and that once its shares are accounted for, the space of possible gains is fully mapped. *Pre-alignment* (fixed transforms applied before the decoder) can reduce only the class-agnostic component of the divergence, and marginal Euclidean alignment already drives that component to its infimum—so no pre-alignment, however elaborate, can do better, and the one refinement that acts differently does so by destroying the signal itself. *In-training alignment* (objectives applied during learning) can in principle spend the residual, but only under episodic cross-subject structure; minibatch objectives such as DANN and CORAL provably carry zero expected cross-subject signal. What remains after both shares are spent is an irreducible floor set by the data itself. On imagined speech from consumer-grade EEG, that floor is at chance: the budget is exhausted before any method acts, so cross-subject decoding is near chance *by prediction, not by failure*. We prove each of these claims and confirm them on motor imagery (where signal is present

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.

and the mechanism is visible) and on three imagined-speech datasets (where the budget is exhausted). The result is a sharp, unified map of where cross-subject gains can and cannot come from.

Imagined speech—the internal generation of speech without overt articulation—offers a communication pathway for individuals with severe motor impairment. Consumer-grade EEG makes the paradigm deployable outside the lab, but deployment demands *cross-subject* generalization: a decoder trained on subjects A, B, C must work on an unseen subject D without recalibration. EEG is strongly subject-specific, and this cross-subject domain gap is the central obstacle. On our multilingual benchmark it averages 19.7 percentage points (pp) between within-subject and cross-subject accuracy.

Two families of remedies dominate. **Pre-alignment** applies a fixed transform to each subject's data before any decoder: Euclidean Alignment (EA) (He & Wu, 2019) whitens each subject by a reference covariance so subjects become comparable. **In-training alignment** instead shapes the representation during learning: adversarial (DANN, Ganin et al., 2016), correlation-based (CORAL, Sun & Saenko, 2016), or episodic meta-learning (MLDG, Li et al., 2018) objectives try to make features subject-invariant. Both are applied to imagined speech with limited success, and a basic question has gone unanswered: *why*—at a mechanistic, provable level—do these remedies fail or saturate, and is there a boundary that no member of either family can cross?

This paper draws that boundary on both axes and shows the two bounds combine precisely on imagined speech.

Contributions.

- **Pre-alignment boundary.** We characterize CC-EA's collapse exactly (between-class contrast to zero, accuracy to chance), prove a gauge-invariance lemma and ceiling theorem establishing marginal EA as optimal within the linear pre-alignment family, an equalization criterion (Proposition 2, Corollary 1) confirmed across three operators (CC-EA, conditional CORAL, conditional MMD), and a causal signal-placement control (Sections 4–7).
- **In-training boundary.** We prove episodic sampling is necessary for prototype gradients to carry $\mathcal{H}$-divergence signal, that classification and minibatch objectives (DANN, CORAL) carry none, and validate the gradient alignment on imagined speech across $29/30$ proxy–dataset pairs (Section 8).
- **Combined boundary and benchmark.** We introduce the first multilingual cross-subject imagined-speech benchmark (Telugu, Kannada, English; consumer-grade EEG) and show cross-subject decoding is near chance for all methods, matching the $\mathcal{H}$-divergence prediction (Section 9).
- **Unified negative boundary.** We connect the two axes through a transfer-budget decomposition (Proposition 3): pre-alignment saturates its share of the bound, and the residual is reachable only in training or at test time—giving the field a sharp map of where gains cannot come from.

2 Related Work

Cross-subject alignment. Euclidean Alignment (He & Wu, 2019) and Riemannian approaches (Zanini et al., 2017; Barachant et al., 2011) whiten or align covariances to make subjects comparable; these are class-agnostic pre-alignments, and our gauge-invariance result (Theorem 1) shows the whole family is equivalent for a refit CSP decoder while the class-conditional refinement (CC-EA) destroys the signal.

Domain adaptation and generalization. Adversarial (Ganin et al., 2016), correlation-based (Sun & Saenko, 2016), MMD-based (Gretton et al., 2012), and meta-learning (Li et al., 2018) objectives minimize a distributional discrepancy for domain generalization (Wang et al., 2022); we analyze them on the Ben-David bound (Ben-David et al., 2010) and show pooled-minibatch objectives carry no cross-subject signal (Corollary 3).

Imagined-speech decoding. Cross-subject inner-speech decoding remains hard (Panachakel & Ramakrishnan, 2021) on small public datasets (KaraOne (Zhao & Rudzicz, 2015), FEIS (Wellington &

Clayton, 2019)); our motor-imagery control uses FBCSP (Ang et al., 2008) on BCI IV-2a (Tangermann et al., 2012) and Cho2017 (Cho et al., 2017), and we compare against EEGNet (Lawhern et al., 2018).

3 Background and Setup

Notation. A trial is $X \in \mathbb{R}^{n_c \times T}$ (n_c channels, T samples), with spatial covariance $C = \frac{1}{T} X X^\top$; for a set of trials we write the mean covariance $\bar{C}$. Whitening by a symmetric positive-definite (SPD) matrix W maps $X \mapsto WX$, so $C \mapsto WCW^\top$.

Euclidean Alignment. *Marginal* EA whitens every trial of a subject by the pooled reference: $W_{\text{marg}} = \bar{C}^{-1/2}$. After marginal EA the pooled covariance is the identity while the two class covariances move together, preserving their relative geometry. *Class-conditional* EA (CC-EA) whitens each class by its own reference, $W_c = (\bar{C}_c)^{-1/2}$, driving each class covariance to the identity.

Decoders. For motor imagery we use the standard Filter-Bank Common Spatial Patterns (FBCSP) pipeline: band-pass into five sub-bands (8–30 Hz), per-band CSP with Ledoit–Wolf regularization, log-variance features, and LDA. CSP reads the between-class covariance difference. For imagined speech we use Riemannian tangent-space features (Barachant et al., 2011) with a shared MLP encoder (Section 8).

$\mathcal{H}$-divergence and LOSO. For a hypothesis class $\mathcal{H}$, the $\mathcal{H}$-divergence is $d_{\mathcal{H}}(P_S, P_Q) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{P_S}[h{=}1] - \Pr_{P_Q}[h{=}1] \right|$; smaller values indicate more subject-invariant features. We evaluate by leave-one-subject-out (LOSO): train on all-but-one subject, test on the held-out one, and report mean balanced accuracy (chance $= 50\%$).

Datasets. *Motor imagery (signal-present control):* BCI Competition IV-2a (BNCI2014-001; 9 subjects, 22 channels, binary left/right hand) and Cho2017 (49 subjects after quality screening, 64 channels, 128 Hz). *Imagined speech (target application):* a new multilingual benchmark (Table 5) recorded on consumer-grade 8-channel devices (Enobio, UNICORN) across three languages and three speech unit types, and the public FEIS dataset (Wellington & Clayton, 2019) (21 subjects, 14-channel, binary voiced/unvoiced) used for gradient-alignment validation.

4 Pre-Alignment I: A Between-Class Contrast Instrument

To reason about why pre-alignment helps or hurts we measure the quantity CSP consumes: how different the two classes' covariances are.

Definition 1 (Between-class contrast)**.** *For class covariances C_0, C_1, let* $\text{riem}(C_0, C_1) = \| \log(C_0^{-1/2} C_1 C_0^{-1/2}) \|_F$*, the affine-invariant (AIRM) Riemannian distance, and let* gev *be the standard deviation of the log-odds of the generalized eigenvalues of the pencil $(C_1, C_0{+}C_1)$—the spread CSP directly exploits. Both vanish iff $C_0 = C_1$.*

The mechanism, exactly. Full CC-EA whitens class c by $(\bar{C}_c)^{-1/2}$, so the transformed class covariance is $(\bar{C}_c)^{-1/2} \bar{C}_c (\bar{C}_c)^{-1/2} = I$ for *both* classes. Hence $C_0 = C_1 = I$ and the contrast is identically zero—by construction.

Table 4 confirms both halves on real data: marginal EA leaves the contrast identical to the raw value—for every subject individually—while full CC-EA drives it to exactly 0.0000. The instrument cleanly separates "safe" alignment (marginal, contrast-preserving) from "destructive" alignment (class-conditional, contrast-erasing).

5 Pre-Alignment II: Class-Conditional EA Collapses to Chance

Table 7 reports the deployable comparison. Marginal EA lifts accuracy from 66.8% to 72.3%; CC-EA collapses to 49.3%, statistically indistinguishable from chance. Its low standard deviation (5.66) reflects that every subject is stuck near chance once the between-class signal is erased.

The failure is not a labeling artifact. Improving the target alignment—pseudo-labels (58.5%), confidence-gated pseudo-labels (62.1%), and an oracle with *true* target labels (63.4%)—monotonically raises CC-EA but plateaus 9 pp below marginal EA (Table 8); perfect labels cannot close the gap, so the residual deficit is the intrinsic cost of class-conditional whitening, not label noise.

Replication with statistical power. On Cho2017 (49 subjects) the collapse escapes the small-sample floor of BCI IV-2a: CC-EA drives the contrast to zero and accuracy to chance (50.1%, $p = 0.92$; the 52.9% under Ledoit–Wolf shrinkage is a regularization residual). Marginal EA dominates CC-EA across all 49 subjects (Wilcoxon $W = 1190$, $p < 10^{-11}$).

6 Pre-Alignment III: Causal Isolation and the Two-Mode Principle

Design. We generate synthetic multi-subject data matched to the real dimensions (9 subjects, 22 channels); each trial is Gaussian with a class-conditional mean and covariance we set by hand, plus a mild per-subject random linear mixing as cross-subject nuisance. In the **COV** regime classes share a mean and differ only in covariance (MI-like); in **MEAN** they share a covariance and differ only in mean. We run the identical raw/marginal/CC-EA ladder on each (a mean-based feature for MEAN, since CSP is blind to mean shifts).

Table 9 is the causal centerpiece: the only variable that changed between the two rows is *where the signal lives*. In motor imagery the class signal *is* covariance structure, so the nuisance CC-EA removes and the signal it must preserve are the same object—removing the former per class destroys the latter. We now formalize this.

Proposition 1 (Second-order class-conditional alignment: fatal iff the signal is second-order)**.** *Model class $c \in \{0, 1\}$ as $\mathcal{N}(\mu_c, \Sigma_c)$ and let second-order class-conditional alignment be $T_c : x \mapsto \Sigma_c^{-1/2} x$.* (i) *If $\mu_0 = \mu_1 = 0$ and $\Sigma_0 \neq \Sigma_1$, then T_c sends each class to $\mathcal{N}(0, I)$; the class laws coincide and the Bayes error is $1/2$.* (ii) *If $\Sigma_0 = \Sigma_1 = \Sigma$ and $\mu_0 \neq \mu_1$, then $T_0 = T_1 = \Sigma^{-1/2}$ is a single shared map; the separation $\Sigma^{-1/2}(\mu_0 - \mu_1) \neq 0$ and the Bayes error are unchanged.*

Proof in Appendix A.

Proposition 2 (The equalization criterion for class-conditional alignment)**.** *Let the two classes be zero-mean with distinct spatial covariances $C_0 \neq C_1 \in \mathrm{SPD}(n_c)$. A* class-conditional second-order alignment *maps class-c trials by an invertible $B_c \in \mathrm{GL}(n_c)$ chosen from class statistics, sending $C_c \mapsto C_c' = B_c C_c B_c^\top$. Call the alignment* equalizing *if $C_0' = C_1'$ and* distinctness-preserving *otherwise. Then a class-conditional second-order alignment drives the covariance signal to chance* if and only if *it is equalizing:* (i) *If $C_0' = C_1' =: M$ (equalizing), the between-class contrast (Definition 1) is exactly zero, the aligned class-conditional laws coincide ($\mathcal{N}(0, M)$ for both), and the Bayes error of any covariance-based decoder is $1/2$.* (ii) *If $C_0' \neq C_1'$ (distinctness-preserving), the contrast is strictly positive and the Bayes error is strictly below $1/2$; the signal survives. Survival does* not *imply optimality: a per-class map with $B_0 \neq B_1$ breaks the class-agnostic gauge equivalence of Theorem 1, so such an alignment may match or underperform marginal EA but cannot exceed it in general.*

Proof in Appendix B.

Corollary 1 (Three operator instances)**.** *The criterion of Proposition 2 classifies existing conditional aligners by whether they equalize the class covariances:* (a) *CC-EA is equalizing (collapse);* (b) *conditional CORAL and* (c) *conditional MMD in its distinctness-preserving form are non-equalizing (survival); conditional MMD in its equalizing form collapses like CC-EA.*

Proof in Appendix C.

Table 3 (Appendix L) confirms all four: equalizing operators drive the source contrast to exactly zero and accuracy to chance (49.3, 51.0), while distinctness-preserving operators retain positive contrast and stay above chance (64.2, 69.8) without reaching marginal EA.

7 PRE-ALIGNMENT IV: THE CEILING IS INTRINSIC

7.1 A GAUGE-INVARIANCE ACCOUNT

Write the log-variance feature $\phi_w(X) = \log(w^\top C_X w)$ with $C_X = \frac{1}{T} X X^\top$, and let the CSP+LDA class with $2p$ filters be $\mathcal{H} = \{X \mapsto \text{sign}(\sum_k \beta_k \phi_{w_k}(X) + \beta_0)\}$. The group $\text{GL}(n_c)$ acts by $X \mapsto AX$, i.e. $C_X \mapsto A C_X A^\top$.

Lemma 1 (CSP gauge invariance)**.** *Fix $A \in \text{GL}(n_c)$.* (a) *$C_c \mapsto A C_c A^\top$ leaves the generalized eigenvalues of $(C_1, C_0{+}C_1)$ unchanged and sends $w \mapsto A^{-\top} w$.* (b) *$\mathcal{H}$ is closed under $h \mapsto h \circ A$ ($w_k \mapsto A^\top w_k$, β fixed), a bijection of $\mathcal{H}$.* (c) *If the same A is applied to all training and test trials, the trained CSP+LDA decision is unchanged.*

Proof in Appendix D.

Remark 1. *The invariance is a property of the pipeline's covariance-congruence structure, not of CSP alone: it holds equally for Riemannian MDM and for tangent-space features read by a* linear *classifier, but fails for front-ends that are not covariance-congruence-invariant—per-channel band power, or a nonlinear (e.g. MLP) head on tangent-space features. We state it for CSP+LDA as the decoder used in our motor-imagery analysis; the imagined-speech boundary is the $\mathcal{H}$-divergence one (Theorems 2–3), not the gauge one.*

Corollary 2 (The $\mathcal{H}\Delta\mathcal{H}$-divergence is gauge-invariant)**.** *For any $A \in \text{GL}(n_c)$ applied to both domains, $d_{\mathcal{H}\Delta\mathcal{H}}(A_\# D_S, A_\# D_T) = d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$.*

Proof in Appendix E.

Theorem 1 (Gauge reduction of class-agnostic linear alignment)**.** *Let $\mathcal{A}_{\text{lin}} = \{A_s \in \text{GL}(n_c)\}_s$ be the per-subject, class-agnostic invertible linear alignments; marginal EA is $A_s = \bar{C}_s^{-1/2}$.* (i) *For every subject the generalized-eigenvalue spectrum of its class pencil is invariant to the choice of A_s.* (ii) *Two schemes related by a single global A yield identical LOSO accuracy.*

Proof in Appendix F.

Remark 2. *A residual per-subject orthogonal freedom survives, $A_s = M^{1/2} O_s \bar{C}_s^{-1/2}$. Each O_s leaves subject s's own class-pencil spectrum invariant by (i), but since LOSO pools subjects an inconsistent choice $\{O_s\}$ alters the cross-subject geometry the pooled decoder reads. Marginal EA fixes $O_s = I$ (canonical registration); a non-canonical registration (Procrustes) is not globally gauge-equivalent and may hurt—its 57.99% in Table 3. That EA's canonical registration is not beaten empirically is the content of this section.*

Nothing in the linear or nonlinear family clears marginal EA. Against marginal EA's 72.3%, Procrustes rotation (58.0%), a shared discriminative basis (72.3%), and discriminative reweighting (72.8%) all fail to clear it, as Lemma 1 predicts; a nonlinear Riemannian tangent-space pipeline (71.5%) and tuned EEGNet variants with contrast-preservation and/or subject-adversarial losses (74.0–75.0%) likewise do not beat it within the ± 11 across-subject std, and on Cho2017 the adversarial+contrast objective (71.3%) does not beat the EEGNet baseline (71.8%) despite $6\times$ more subjects (full results in Table 3, Appendix L).

Bridge: where the residual budget lives. The following proposition locates, in the standard transfer bound, where the remaining budget can be spent—directly at in-training alignment (Section 8).

Proposition 3 (Pre-alignment exhausts its share of the transfer budget)**.** *Consider the bound $\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \lambda$ with the CSP+LDA class $\mathcal{H}$. Marginal EA achieves the minimum divergence term attainable by any class-agnostic linear pre-alignment — a value that is in fact constant across the family (Theorem 1, Corollary 2) — and no class-conditional linear map lowers it further without driving the source error to the chance floor. Consequently the residual target-error budget can be reduced only* (a) *in training, by enlarging the source support so that λ and the divergence fall together, or* (b) *at test time, by adapting h on unlabeled target data.*

Proof in Appendix G.

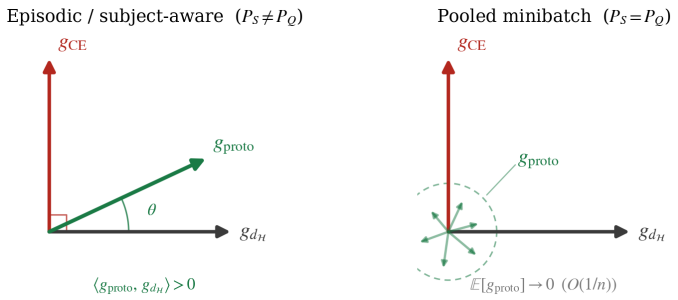


Figure 1: Gradient-alignment geometry. g_{CE} is orthogonal in expectation to the signed domain-classifier gradient (Theorem 3(ii)) and empirically nearly orthogonal to $g_{d_{\mathcal{H}}}$ (Table 1). When $P_S \neq P_Q$ (left), g_{proto} tilts toward $g_{d_{\mathcal{H}}}$; under pooled minibatching ($P_S = P_Q$, right) $\mathbb{E}[g_{\mathrm{proto}}]$ reduces to an $O(1/n)$ within-class term with no preferred direction (Theorem 2).

8 In-Training Alignment Is Bounded

Proposition 3 routes the residual budget to training. We now show the dominant in-training remedies are themselves bounded, at the gradient level. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a differentiable encoder, and let $p_\theta(y \mid x) = \mathrm{softmax}(W f_\theta(x))_y$ be the class-probabilities from a linear head on its features. For source P_S and query P_Q with class-conditional distributions P_c^S, P_c^Q, define population centroids $\mu_c^S(\theta) = \mathbb{E}_{x \sim P_c^S}[f_\theta(x)]$ and $\mu_c^Q(\theta)$ analogously. The prototype and classification losses are

$$\mathcal{L}_{\mathrm{proto}} = \sum_c \|\mu_c^S(\theta) - \mu_c^Q(\theta)\|^2, \qquad \mathcal{L}_{\mathrm{CE}} = -\mathbb{E}_{(x,y)\sim P_Q}[\log p_\theta(y \mid x)],$$

with gradients $g_{\mathrm{proto}}, g_{\mathrm{CE}}$ and $g_{d_{\mathcal{H}}}$ the gradient of an $\mathcal{H}$-divergence proxy. Episodic sampling designates one training subject as query and the rest as support, directly simulating LOSO ($P_S \neq P_Q$); minibatch sampling draws both from the pooled distribution ($P_S = P_Q$).

Lemma 2 ($\mathcal{L}_{\mathrm{proto}}$ is class-conditional MMD)**.** *Under the linear kernel $k(x, y) = x^\top y$, $\mathrm{MMD}_c^2(P_c^S, P_c^Q) = \|\mu_c^S - \mu_c^Q\|^2$, hence $\mathcal{L}_{\mathrm{proto}} = \sum_c \mathrm{MMD}_c^2(P_c^S, P_c^Q)$.*

Proof in Appendix H.

Theorem 2 (Episodic necessity)**.** *Under minibatch sampling ($P_S = P_Q = P$), $\mathbb{E}[\mathcal{L}_{\mathrm{proto}}] \to 0$ as batch size grows, and $\mathbb{E}[g_{\mathrm{proto}}] \to 0$: the prototype gradient carries no $\mathcal{H}$-divergence signal. Under episodic sampling ($P_S \neq P_Q$), $\mathbb{E}[g_{\mathrm{proto}}]$ is non-zero and, by Lemma 2, is the gradient of class-conditional MMD, an $\mathcal{H}$-divergence surrogate under (B) (Theorem 3(i)).*

Proof sketch. Under minibatch sampling the mean difference is exactly zero, $\mathbb{E}[\mu_c^S - \mu_c^Q] = 0$ by linearity, while the squared norm is strictly positive at finite batch size, $\mathbb{E}[\|\mu_c^S - \mu_c^Q\|^2] = \left(\frac{1}{n_S^c} + \frac{1}{n_Q^c}\right)\mathrm{tr}\,\Sigma_c$ — an $O(1/n)$ finite-sample variance bias that vanishes as the batch grows (law of large numbers); the chain rule gives the corresponding $\mathbb{E}[g_{\mathrm{proto}}] = O(1/n) \to 0$. Under episodic sampling $P_Q = P_v \neq P_S$, so $\|\mu_c^{P_S} - \mu_c^{P_Q}\|^2 > 0$ for some class, making g_{proto} non-zero; Lemma 2 identifies it as the class-conditional MMD gradient. The full computation, including the exact $O(1/n)$ minibatch bias $\mathbb{E}[g_{\mathrm{proto}}] = \sum_c (1/n_S^c + 1/n_Q^c)\nabla_\theta \mathrm{tr}\,\Sigma_c$, is given in Appendix I. □

Theorem 3 (Gradient selectivity)**.** *Let $z = f_\theta(x)$ be the features of a differentiable encoder, consider balanced binary classification, and let $\mathcal{H} = \{z \mapsto \mathbf{1}[w^\top z + b > 0]\}$ be the linear hypothesis class.* (i) *If class-conditional features are Gaussian with a domain-independent covariance, $\mathcal{N}(\mu_c^D, \Sigma_c)$ with $\Sigma_c \succ 0$ (assumption (B)), then $d_{\mathcal{H}}(P_S, P_Q) \leq C\sqrt{\mathcal{L}_{\mathrm{proto}}}$ with $C = (\pi\lambda_\star)^{-1/2}$, $\lambda_\star = \min_c \lambda_{\min}(\Sigma_c)$.* (ii) *Fix the current parameters θ_0 and the query law P_Q, and let L_{DC} be the signed domain-classifier objective evaluated with the optimal linear witness at θ_0, held fixed (so $d_{\mathcal{H}} = 2|L_{\mathrm{DC}}|$ at θ_0). Under a symmetric prior over source laws, assumptions (S0)–(S3) and (R1)–(R2) of Appendix J,*

$$\mathbb{E}_{P_S \sim \pi}\left[\langle g_{\mathrm{CE}}, -\nabla_\theta L_{\mathrm{DC}} \rangle\right]_{\theta=\theta_0} = 0.$$

Table 1: Prototype gradient alignment gap (episodic − minibatch) on imagined speech. $\Delta > 0$ confirms Theorem 2. Proto/CE ratio $= |\cos(g_{\text{proto}}, g_{d_{\mathcal{H}}})|/|\cos(g_{\text{CE}}, g_{d_{\mathcal{H}}})|$ (KMMD) is consistent with Theorem 3(ii), measured against a divergence proxy rather than the signed objective of the theorem. † fewer than 4 usable subjects; ratio not computed. ‡ near-perfect g_{CE} orthogonality ($|\cos(g_{\text{CE}}, g_{d_{\mathcal{H}}})| < 0.003$). *** $p < 0.001$; ns = not significant.

Dataset	NeuralDC Δ	KMMD Δ	Energy Δ	Proto/CE
FEIS (21 subj.)	$+0.133$***	$+0.316$***	$+0.249$***	$27.4\times$
Telugu Words†	$+0.099$***	$+0.258$***	$+0.208$***	—
Telugu Phonemes†	$+0.154$***	$+0.347$***	$+0.289$***	—
Telugu Vowels	$+0.048$***	$+0.292$***	$+0.247$***	$> 100\times$‡
Kannada Vowels	-0.024ns	$+0.345$***	$+0.277$***	$17.2\times$
Kannada Phonemes	$+0.083$***	$+0.305$***	$+0.250$***	$> 100\times$‡
Kannada Words	$+0.079$***	$+0.298$***	$+0.218$***	$77.0\times$
English Vowels	$+0.089$***	$+0.312$***	$+0.227$***	$24.7\times$
English Phonemes†	$+0.204$***	$+0.367$***	$+0.306$***	—
English Words	$+0.185$***	$+0.335$***	$+0.269$***	$18.9\times$
Confirmed	9/10	10/10	10/10	7/7

Proof sketch. (i) Under (B), a linear classifier's acceptance rate is a Gaussian CDF of its projected class means; the CDF is $(2\pi)^{-1/2}$-Lipschitz, and Cauchy–Schwarz bounds each projected mean gap by $\|\mu_c^S - \mu_c^Q\|/\sqrt{\lambda_\star}$. This bounds $d_{\mathcal{H}}$ directly; the Ben-David theorem is not used. Assumption (B) is necessary: with matched means but unequal covariances, $\mathcal{L}_{\text{proto}} = 0$ while $d_{\mathcal{H}} > 0$. The bound fixes the divergence *magnitude*, not the gradient *direction*; the positive alignment $\langle g_{\text{proto}}, g_{d_{\mathcal{H}}} \rangle > 0$ is established empirically (Section 9.1). (ii) A prior-preserving mirror that negates L_{DC} on a neighbourhood of θ_0 makes the integrand odd, so its expectation vanishes; the claim is for the signed objective only (Appendix J). □

Corollary 3 (Plug-in impossibility)**.** *DANN and CORAL, applied as* standard pooled-minibatch *in-training objectives—drawing each minibatch from the pooled training distribution, without a held-out-subject query split—cannot systematically reduce cross-subject $\mathcal{H}$-divergence, regardless of loss weight or architecture. An episodic, subject-aware variant that induces $P_S \neq P_Q$ escapes this corollary—precisely the structure Theorem 2 shows to be necessary.*

Proof in Appendix K.

Corollary 3 unifies the two axes on CORAL specifically: as a *pre-alignment* operator, conditional CORAL preserves distinctness and cannot beat marginal EA (Proposition 2(ii)); as a *pooled in-training* objective, minibatch CORAL carries no cross-subject signal at all. The same method is bounded on both axes, for related structural reasons.

9 Empirical Validation on Imagined Speech

9.1 Gradient alignment

We measure $\text{align}(g, g') = g \cdot g'/(\|g\|\|g'\|)$ over 200 episodes (episodic vs. minibatch) using three $\mathcal{H}$-divergence proxies (Neural DC, KMMD with σ=1, Energy Distance); Δ = episodic − minibatch.

Table 1 confirms both theorems on imagined speech: the episodic−minibatch prototype-alignment gap is positive across $29/30$ proxy–dataset pairs ($p < 0.001$), confirming Theorem 2, while the prototype-to-classification alignment ratio ($17\times$ to $> 100\times$, KMMD) is consistent with Theorem 3(ii). This is the paper's strongest imagined-speech result—the mechanism is real and replicates across three languages and three speech-unit types.

The signal is the distribution gap, not LOSO machinery. To separate the effect of episodic LOSO structure from the mere presence of distinct subjects, we add a *subject-aware minibatch* scheme: support and query are drawn from two distinct random subjects ($P_S \neq P_Q$) with no held-out-subject query. Its prototype-alignment gap over pooled minibatching is positive on all ten conditions and

Table 2: Cross-subject balanced accuracy (%) on the multilingual benchmark (mean of 5 seeds). No method—adaptive or not—clears chance by a meaningful margin. This is the quantitative prediction of the $\mathcal{H}$-divergence bound at the available signal-to-subject ratio, not a method failure. DANN and CORAL fall *below* the episodic MLDG baseline, confirming Corollary 3. ([†]English Phonemes: 4 subjects; interpret with caution.)

Method	Tel.W	Tel.Ph	Tel.V	Kan.W	Kan.Ph	Kan.V	Eng.W	Eng.Ph[†]	Eng.V
Normal	51.6	51.9	52.8	55.2	51.3	48.3	54.5	50.7	46.2
DANN	50.9	52.7	52.3	54.8	51.9	49.8	54.6	51.1	46.8
CORAL	48.8	53.7	52.6	56.0	51.5	49.8	54.3	50.9	47.2
MLDG	48.2	53.3	55.5	56.0	51.4	49.0	54.4	52.1	46.1

significant on all four with ≥ 6 subjects (mean $+0.29$ KMMD, $p \leq 0.016$, Wilcoxon; Table 6), matching the episodic gap ($+0.27$). The episodic$-$subject-aware difference is negligible and never significant (-0.02 mean), so the LOSO held-out structure adds no signal beyond the raw cross-subject distribution gap. This isolates the mechanism exactly as Theorem 2 predicts—prototype gradients carry $\mathcal{H}$-divergence signal whenever $P_S \neq P_Q$, however that gap is induced—and it directly confirms the escape clause of Corollary 3: a subject-aware variant inducing $P_S \neq P_Q$ recovers the signal that pooled minibatching cannot.

9.2 Cross-subject classification: the boundary binds

Table 2 reports cross-subject balanced accuracy: every method sits near chance ($\sim$ 51–52% mean), and DANN and CORAL fall below the episodic MLDG baseline, exactly as Corollary 3 predicts (both operate without episodic subject structure). With measured $\mathcal{H}$-divergence 0.80–1.00 and only 4–14 subjects per condition, the Ben-David bound predicts near-chance performance for *any* procedure; near-chance classification is thus a prediction of our framework, not an anomaly—a coherent picture of a real mechanism (replicating at $p < 0.001$) operating under an exhausted signal budget.

9.3 Public confirmation: the boundary on KaraOne and FEIS

We replicate the boundary diagnostics on two public datasets. On **KaraOne** (Zhao & Rudzicz, 2015) (14 subjects, 62–64 channels, vowels-vs-nasals, marginal EA, same FBCSP pipeline) cross-subject decoding is at chance—LOSO balanced accuracy $49.6 \pm 1.6\%$ ($p = 0.34$, $n = 14$), tightly concentrated at the floor—while the between-class contrast is large but subject-idiosyncratic: mean AIRM contrast 3.05 ± 2.31 (range 1.26–8.19), far above motor imagery (Table 4: 0.93 on IV-2a, 1.23 on Cho2017), with a standard deviation nearly equal to its mean. Each subject shows substantial class-covariance difference, but in directions that do not align across subjects, so the pooled decoder does not transfer. On **FEIS** (Wellington & Clayton, 2019) (21 subjects, 14-channel, voiced/unvoiced), the most favourable signal-to-subject ratio in our study, cross-subject LOSO accuracy is $50.0 \pm 0.4\%$ ($p = 0.72$, $n = 21$) while the contrast is again large (3.28 ± 1.18). Across three imagined-speech datasets spanning 8–64 channels and consumer- to lab-grade hardware, cross-subject decoding does not exceed chance despite substantial between-class covariance structure, indicating an intrinsic property of cross-subject imagined speech rather than a dataset artifact. FEIS instantiates both halves of our claim at once: the gradient-alignment mechanism replicates strongly (Table 1, $+0.316$ KMMD, $p < 0.001$) while classification sits at chance—the mechanism is real, yet the cross-subject signal budget is exhausted.

10 Discussion

Where gains can and cannot come from. Both alignment axes are bounded: pre-alignment cannot exceed marginal EA (Lemma 1, Theorem 1), and minibatch in-training objectives carry no cross-subject signal (Corollary 3); CORAL is bounded on *both* axes, a unification the two analyses make precise. Improvements over marginal EA therefore cannot come from covariance-geometry manipulation or minibatch invariance objectives; the remaining levers lie outside both families—episodic training that genuinely simulates the LOSO shift, cross-dataset transfer that expands the effective support, or test-time target adaptation (Proposition 3). On imagined speech the two bounds

combine to near-chance decoding for every method. We read this as a positive result: the mechanism (gradient alignment) is confirmed at $p < 0.001$ while the classification budget is provably exhausted at the available signal-to-subject ratio, so what must change is not the objective but the data—more subjects and higher signal quality, which the $\mathcal{H}$-divergence bound quantifies.

Limitations. *(1)* The pre-alignment ceiling is established on motor imagery; on imagined speech the operative boundary is the in-training/$\mathcal{H}$-divergence one, and the mean-carried (non-covariance) regime is characterized only synthetically (Proposition 1(ii)). *(2)* Theorem 3(ii) assumes a symmetric source prior whose mirror negates the signed domain objective on a neighbourhood of the current parameters, and a classification gradient invariant under that mirror (automatic when the CE loss is computed on the query subject, but not for source-side CE); both are idealized at 4–14 subjects, so we rely on the measured proto/CE ratios ($\geq 17\times$), not the exact identity. Theorem 3(i) further assumes Gaussian features with shared covariance across domains, an idealization for an MLP encoder; outside this regime $\mathcal{L}_{\text{proto}}$ remains a first-moment discrepancy but no longer provably bounds $d_{\mathcal{H}}$. *(3)* Classification gains are small by design, so the contribution is the boundary and its mechanism, not a decoding advance. *(4)* The synthetic control is Gaussian and deliberately idealized. *(5)* Weak features cannot explain the null result: a broken pipeline would not produce the theory-predicted gradient-alignment signal that replicates on the same datasets ($29/30$ pairs, $p < 0.001$); a larger cohort could raise the floor, as the framework predicts.

11 Conclusion

We established a two-axis boundary for cross-subject EEG alignment. On the pre-alignment axis, CC-EA collapses to chance by erasing the between-class contrast, and a gauge-invariance argument proves marginal EA optimal within the linear family (motor imagery, 9 and 49 subjects, $p < 10^{-11}$). On the in-training axis, minibatch objectives (DANN, CORAL) provably carry no cross-subject $\mathcal{H}$-divergence signal while episodic prototype alignment is a principled surrogate ($29/30$ pairs, $p < 0.001$). On a new multilingual benchmark, cross-subject decoding is near chance for every method—the $\mathcal{H}$-divergence prediction. The result is a sharp map of where cross-subject gains cannot come from, and why consumer-grade imagined-speech decoding remains hard.

Reproducibility Statement

All theoretical claims are stated with explicit assumptions and proved in full. Propositions 1, 2, and 3, Lemmas 1 and 2, Theorems 1, 2, and 3, and Corollaries 2, 1, and 3 are proved in full in Appendices A–K, where the operative assumptions (finite feature moments, the $\nabla_\theta/\mathbb{E}$ exchange, the shared-covariance class-conditional regime, the symmetric source prior with its local antisymmetry condition, and a mirror-invariant classification gradient) are made explicit. The motor-imagery experiments use the public BCI IV-2a (BNCI2014-001) and Cho2017 datasets, and the public KaraOne and FEIS datasets, with the preprocessing, FBCSP configuration, and LOSO protocol specified in Sections 3–7; the synthetic causal control is fully specified in Section 6. Data-processing steps for the multilingual imagined-speech benchmark are described in Section 3 and Table 5. Anonymized code reproducing every experiment, table, and the benchmark-processing pipeline will be released upon acceptance.

Ethics Statement

This work includes EEG newly recorded from human participants (the multilingual imagined-speech benchmark). Data collection was carried out under approval from the responsible institutional ethics review board, with written informed consent obtained from every participant, who could withdraw at any time. Recordings were pseudonymized at source and contain no personally identifying information. The other datasets (BCI IV-2a, Cho2017, KaraOne, FEIS) are publicly available and used in accordance with their licenses. Imagined-speech BCIs are intended as an assistive communication channel for people with severe motor impairment; our findings are deliberately conservative about present-day cross-subject decoding and do not enable covert or non-consensual decoding of mental content. We are aware of no other ethical concerns. To preserve double-blind review, institution and ethics-approval identifiers are omitted here and will be restored in the camera-ready version.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, we disclose that generative AI tools were used to improve the readability of the writing, to assist with LaTeX and reference formatting, and to help check the mathematical proofs for internal consistency. All claims, proofs, experimental design, implementation, and analysis were formulated and verified by the authors; any AI-suggested change to a proof was independently checked for correctness by at least one author. We have reviewed all AI-assisted content and take responsibility for the final content of this work.

REFERENCES

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 2390–2397. IEEE, 2008.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. Eeg datasets for motor imagery brain–computer interface. *GigaScience*, 6(7):gix034, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. Decoding covert speech from eeg-a comprehensive review. *Frontiers in neuroscience*, 15:642251, 2021.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

S Wellington and J Clayton. Fourteen-channel eeg with imagined speech (feis) dataset. *University of Edinburgh*, 2019.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2017.

Shunan Zhao and Frank Rudzicz. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 992–996. IEEE, 2015.

A Proof of Proposition 1

Proof. T_c sends $\mathcal{N}(\mu_c, \Sigma_c)$ to $\mathcal{N}(\Sigma_c^{-1/2}\mu_c, I)$. (i) With $\mu_c = 0$ both classes become $\mathcal{N}(0, I)$; identical laws give Bayes error $1/2$. (ii) With $\Sigma_c = \Sigma$ the map is common and invertible; an invertible map preserves Bayes error, and the means remain distinct. □

B Proof of Proposition 2

Proof. Since $\mu_c = 0$, class c after alignment is $\mathcal{N}(0, C'_c)$. *(i)* If $C'_0 = C'_1 = M$, the two class-conditional laws are the identical Gaussian $\mathcal{N}(0, M)$, so no measurable function of X separates them and the Bayes error is $1/2$; equivalently $\mathrm{riem}(C'_0, C'_1) = 0$ and every generalized eigenvalue of $(C'_1, C'_0{+}C'_1)$ equals $\frac{1}{2}$, so the contrast of Definition 1 vanishes. *(ii)* If $C'_0 \neq C'_1$, then $\mathrm{riem}(C'_0, C'_1) > 0$ (Definition 1 vanishes iff the covariances coincide), so the two zero-mean Gaussian densities differ on a positive-measure set and the likelihood-ratio test attains error strictly below $1/2$; CSP reads a nonzero generalized-eigenvalue spread. *Exhaustiveness:* the contrast is zero iff $C'_0 = C'_1$, so collapse holds iff the alignment equalizes. □

C Proof of Corollary 1

Proof. Each operator applies a per-class map $C_c \mapsto C'_c = B_c C_c B_c^\top$; by Proposition 2 it collapses the covariance signal iff $C'_0 = C'_1$. *(a) CC-EA* uses $B_c = (\bar{C}_c)^{-1/2}$, so $C'_c = (\bar{C}_c)^{-1/2}\bar{C}_c(\bar{C}_c)^{-1/2} = I$ for both classes; thus $C'_0 = C'_1 = I$ is equalizing, hence collapse (Proposition 2(i)). *(b) Conditional CORAL* recolors class c onto the class-c target covariance, $C'_c = M_c$; because the target classes themselves carry between-class contrast, $M_0 \neq M_1$, so the map is distinctness-preserving, hence survival (Proposition 2(ii)). *(c) Conditional MMD* in its distinctness-preserving form likewise sends the two classes to distinct targets $M_0 \neq M_1$ (survival); in its equalizing form it transports both classes to a single common target $M_0 = M_1 = M$, which is equalizing, hence collapse. In all four cases the outcome is fixed by the equalization test, not by the choice of operator. □

D Proof of Lemma 1

Proof. (a) From $C_1 w = \lambda(C_0{+}C_1)w$, left-multiplying by A and inserting $A^\top A^{-\top} = I$ gives $AC_1A^\top(A^{-\top}w) = \lambda A(C_0{+}C_1)A^\top(A^{-\top}w)$; invertibility makes this a bijection of eigenpairs, so the spectrum is preserved, as is the normalization $w^\top(C_0{+}C_1)w = 1$. (b) $\phi_w(AX) = \log((A^\top w)^\top C_X(A^\top w)) = \phi_{A^\top w}(X)$, so $h \circ A \in \mathcal{H}$; $w_k \mapsto A^\top w_k$ is a bijection. (c) On A-transformed data CSP returns filters $A^{-\top}w_k$ by (a), and the feature $\phi_{A^{-\top}w_k}(AX) = \phi_{w_k}(X)$ is identical, so LDA fits the same β and the decision is unchanged. □

E Proof of Corollary 2

Proof. $\Pr_{A_\# D}[h \neq h'] = \Pr_D[h \circ A \neq h' \circ A]$; by Lemma 1(b) $h \mapsto h \circ A$ is a bijection of $\mathcal{H}$, so the supremum defining the divergence is unchanged. □

F Proof of Theorem 1

Proof. (i) A_s acts on both class covariances by the same congruence; apply Lemma 1(a). (ii) The datasets differ by one global A on every trial; apply Lemma 1(c). □

G Proof of Proposition 3

Proof. By Corollary 2, $d_{\mathcal{H}\Delta\mathcal{H}}$ is invariant under any global $A \in \mathrm{GL}(n_c)$ applied to both domains, and by Theorem 1(i) the within-subject class-pencil spectra — the only statistics CSP reads — are invariant to the choice of class-agnostic per-subject alignment A_s. Hence every class-agnostic linear pre-alignment yields the same divergence term as marginal EA, up to the residual per-subject orthogonal freedom of Remark 2, which leaves the spectra fixed. Hence the divergence term is constant across the class-agnostic linear family, and marginal EA already attains that constant value, so no further class-agnostic linear map reduces it. The only linear maps that alter the class spectra are class-conditional ($B_0 \neq B_1$); in the covariance-signal regime, by Propositions 1 and 2, any such map that reduces the divergence equalizes the two class covariances and sends the between-class contrast — hence the attainable source accuracy — to the chance floor, raising ϵ_S (and λ) by at least what it removes from the divergence. Thus no linear pre-alignment lowers the bound. The remaining terms can move only outside this family: *(a)* enlarging the source support at training time, which shrinks the source–target gap and so lowers λ and $d_{\mathcal{H}\Delta\mathcal{H}}$ jointly, or *(b)* transductive adaptation of h on unlabeled target data at test time. □

H Proof of Lemma 2

Restatement. Under the linear kernel $k(a,b) = a^\top b$ on encoder features, $\mathrm{MMD}_c^2(P_c^S, P_c^Q) = \|\mu_c^S - \mu_c^Q\|^2$ and hence $\mathcal{L}_{\mathrm{proto}} = \sum_c \mathrm{MMD}_c^2$.

Proof. Let $z = f_\theta(x) \in \mathbb{R}^d$ and, for domain $D \in \{S, Q\}$ and class c, write P_c^D also for the pushforward law of z under $x \sim P_c^D$. For a kernel k with RKHS $\mathcal{H}_k$ and feature map ψ ($k(a,b) = \langle \psi(a), \psi(b) \rangle_{\mathcal{H}_k}$), the population MMD is

$$\begin{aligned}\mathrm{MMD}_c^2 &= \left\| \mathbb{E}_{z\sim P_c^S}[\psi(z)] - \mathbb{E}_{z\sim P_c^Q}[\psi(z)] \right\|_{\mathcal{H}_k}^2 \\ &= \mathbb{E}_{z,z'\sim P_c^S} k(z,z') - 2\,\mathbb{E}_{\substack{z\sim P_c^S \\ z'\sim P_c^Q}} k(z,z') + \mathbb{E}_{z,z'\sim P_c^Q} k(z,z'),\end{aligned}$$

with z, z' drawn independently inside each expectation. For the linear kernel the feature map is the identity, $\psi(z) = z$, so the mean embedding is the ordinary mean $\mu_c^D = \mathbb{E}_{z\sim P_c^D}[z]$. Independence factorizes each expectation, $\mathbb{E}[z^\top z'] = \mathbb{E}[z]^\top \mathbb{E}[z']$, giving

$$\mathrm{MMD}_c^2 = \|\mu_c^S\|^2 - 2(\mu_c^S)^\top \mu_c^Q + \|\mu_c^Q\|^2 = \|\mu_c^S - \mu_c^Q\|^2.$$

Summing over the (balanced) classes gives $\mathcal{L}_{\mathrm{proto}} = \sum_c \|\mu_c^S - \mu_c^Q\|^2 = \sum_c \mathrm{MMD}_c^2$. □

Remark 3. *The linear kernel makes* MMD *a pseudometric sensitive only to first moments, so* $\mathcal{L}_{\mathrm{proto}}$ *is* exactly *the squared Euclidean distance between class-mean embeddings. This is what licenses reading the prototype loss as a class-conditional MMD; a characteristic kernel would make* MMD *a full metric but would not reduce it to the prototype term.*

I Full Proof of Theorem 2

Assumptions. **(A1)** finite feature second moments, $\mathrm{tr}\,\Sigma_c^D := \mathbb{E}_{x\sim P_c^D}\|f_\theta(x) - \mu_c^D\|^2 < \infty$; **(A2)** f_θ is differentiable in θ with Jacobian $\mathrm{J}_\theta f_\theta(x) \in \mathbb{R}^{d\times p}$ ($p = \dim\theta$) under a domination condition permitting exchange of ∇_θ and $\mathbb{E}$ (so $\mathrm{J}_\theta \mu_c^D = \mathbb{E}[\mathrm{J}_\theta f_\theta(x)]$); **(A3)** support and query samples are mutually independent, with n_S^c, n_Q^c samples per class. We use the estimators $\hat{\mu}_c^D = \frac{1}{n_D^c}\sum_i f_\theta(x_i)$, the empirical loss $\hat{\mathcal{L}}_{\mathrm{proto}} = \sum_c \|\hat{\mu}_c^S - \hat{\mu}_c^Q\|^2$, and $g_{\mathrm{proto}} = \nabla_\theta \hat{\mathcal{L}}_{\mathrm{proto}}$. Two identities are used repeatedly: $\mathbb{E}\|u\|^2 = \|\mathbb{E}u\|^2 + \mathrm{tr}\,\mathrm{Cov}(u)$ and $\nabla_\theta \|u(\theta)\|^2 = 2(\mathrm{J}_\theta u)^\top u$.

Proof. **Minibatch case ($P_S = P_Q = P$).** Both class-c samples are i.i.d. from P_c with common mean μ_c and covariance Σ_c. By (A3) $\hat{\mu}_c^S, \hat{\mu}_c^Q$ are independent and unbiased, so $\mathbb{E}[\hat{\mu}_c^S - \hat{\mu}_c^Q] = 0$ and $\mathrm{Cov}(\hat{\mu}_c^S - \hat{\mu}_c^Q) = \Sigma_c/n_S^c + \Sigma_c/n_Q^c$. Hence

$$\mathbb{E}[\hat{\mathcal{L}}_{\mathrm{proto}}] = \sum_c \left(\tfrac{1}{n_S^c} + \tfrac{1}{n_Q^c}\right)\mathrm{tr}\,\Sigma_c \xrightarrow[n\to\infty]{} 0.$$

For the gradient, with $\mathrm{J}_c^D = \frac{1}{n_D^c}\sum_i \mathrm{J}_\theta f_\theta(x_i^D)$,

$$g_{\mathrm{proto}} = \sum_c 2(\mathrm{J}_c^S - \mathrm{J}_c^Q)^\top(\hat{\mu}_c^S - \hat{\mu}_c^Q).$$

Write $\bar{\mathrm{J}}_c = \mathbb{E}[\mathrm{J}_\theta f_\theta(x)]$ and $\rho_c = \mathbb{E}[(\mathrm{J}_\theta f_\theta(x))^\top f_\theta(x)]$. Expanding a same-set term $\mathbb{E}[(\mathrm{J}_c^S)^\top \hat{\mu}_c^S] = \frac{1}{(n_S^c)^2}\sum_{i,j}\mathbb{E}[(\mathrm{J}_\theta f_i)^\top f_j]$ into the $i \neq j$ pairs (independent, value $\bar{\mathrm{J}}_c^\top \mu_c$) and the $i = j$ pairs (value ρ_c) gives $\mathbb{E}[(\mathrm{J}_c^S)^\top \hat{\mu}_c^S] = \frac{n_S^c - 1}{n_S^c}\bar{\mathrm{J}}_c^\top \mu_c + \frac{1}{n_S^c}\rho_c$; the cross (support–query) terms are exactly $\bar{\mathrm{J}}_c^\top \mu_c$ by independence. Collecting the four terms,

$$\mathbb{E}\big[(\mathrm{J}_c^S - \mathrm{J}_c^Q)^\top(\hat{\mu}_c^S - \hat{\mu}_c^Q)\big] = \Big(\tfrac{1}{n_S^c} + \tfrac{1}{n_Q^c}\Big)\big(\rho_c - \bar{\mathrm{J}}_c^\top \mu_c\big).$$

Because $\nabla_\theta \mathrm{tr}\,\Sigma_c = \nabla_\theta\big(\mathbb{E}\|f_\theta\|^2 - \|\mu_c\|^2\big) = 2(\rho_c - \bar{\mathrm{J}}_c^\top \mu_c)$, this yields the closed form

$$\mathbb{E}[g_{\mathrm{proto}}] = \sum_c \Big(\tfrac{1}{n_S^c} + \tfrac{1}{n_Q^c}\Big)\nabla_\theta \mathrm{tr}\,\Sigma_c = O(1/n) \xrightarrow[n\to\infty]{} 0.$$

Thus the *only* expected minibatch prototype gradient is an $O(1/n)$ within-class variance-shrinkage term; it carries no $\mathcal{H}$-divergence (cross-subject) content and vanishes as the batch grows.

Episodic case ($P_S \neq P_Q$). Now the class means differ across domains, $\mu_c^S \neq \mu_c^Q$ generically. Unbiasedness gives $\mathbb{E}[\hat{\mu}_c^S - \hat{\mu}_c^Q] = \mu_c^S - \mu_c^Q$, so $\mathbb{E}[\hat{\mathcal{L}}_{\mathrm{proto}}] = \sum_c \|\mu_c^S - \mu_c^Q\|^2 + O(1/n)$, which is $O(1)$ and nonzero. Repeating the four-term expansion with distinct source/query laws, the $O(1)$ part is $(\bar{\mathrm{J}}_c^S)^\top\mu_c^S + (\bar{\mathrm{J}}_c^Q)^\top\mu_c^Q - (\bar{\mathrm{J}}_c^S)^\top\mu_c^Q - (\bar{\mathrm{J}}_c^Q)^\top\mu_c^S = (\bar{\mathrm{J}}_c^S - \bar{\mathrm{J}}_c^Q)^\top(\mu_c^S - \mu_c^Q)$, hence

$$\mathbb{E}[g_{\mathrm{proto}}] = \sum_c 2(\bar{\mathrm{J}}_c^S - \bar{\mathrm{J}}_c^Q)^\top(\mu_c^S - \mu_c^Q) + O(1/n) = \nabla_\theta \sum_c \|\mu_c^S - \mu_c^Q\|^2 + O(1/n),$$

the population gradient of the class-conditional MMD (Lemma 2). It is nonzero whenever $\mu_c^S \neq \mu_c^Q$ for some class — precisely the cross-subject shift episodic sampling induces. □

J Full Proof of Theorem 3

Common setup. Features are $z = f_\theta(x) \in \mathbb{R}^d$, with f_θ differentiable in θ. There are two classes $c \in \{0, 1\}$ with prior $\frac{1}{2}$ in every domain. The hypothesis class is linear, $\mathcal{H} = \{h_{w,b} : z \mapsto \mathbf{1}[w^\top z + b > 0]\}$. For a law P over trials, write $q_P^h(\theta) = \Pr_{x\sim P}\big[h(f_\theta(x)) = 1\big]$ for the acceptance rate of h, and $d_{\mathcal{H}}(P_S, P_Q) = 2\sup_{h\in\mathcal{H}}|q_{P_S}^h - q_{P_Q}^h|$.

Part (i): $d_{\mathcal{H}}(P_S, P_Q) \le C\sqrt{\mathcal{L}_{\mathrm{proto}}}$

Assumption (B). At the θ under consideration, for each domain $D \in \{S, Q\}$ and class c, the feature law is $z \mid (c, D) \sim \mathcal{N}(\mu_c^D, \Sigma_c)$. The covariance Σ_c is the same in both domains, and $\Sigma_c \succ 0$, so $\lambda_\star := \min_c \lambda_{\min}(\Sigma_c) > 0$.

Proof. If $w = 0$, then $h_{w,b}$ is constant and contributes zero to the supremum, so take $w \neq 0$. Then $w^\top z \mid (c, D) \sim \mathcal{N}(w^\top\mu_c^D,\, w^\top\Sigma_c w)$ with $w^\top\Sigma_c w > 0$, and with equal class priors

$$q_{P_D}^h = \tfrac{1}{2}\sum_c \Phi(a_c^D), \qquad a_c^D = \frac{w^\top\mu_c^D + b}{\sqrt{w^\top\Sigma_c w}}.$$

Since $\sup\Phi' = 1/\sqrt{2\pi}$, the mean value theorem gives

$$\big|q_{P_S}^h - q_{P_Q}^h\big| \le \frac{1}{2\sqrt{2\pi}}\sum_c |a_c^S - a_c^Q|.$$

The covariance is domain-independent, so the offset b cancels in $a_c^S - a_c^Q$. Cauchy–Schwarz and $w^\top\Sigma_c w \ge \lambda_{\min}(\Sigma_c)\|w\|^2$ give, uniformly in (w, b),

$$|a_c^S - a_c^Q| = \frac{|w^\top(\mu_c^S - \mu_c^Q)|}{\sqrt{w^\top\Sigma_c w}} \le \frac{\|\mu_c^S - \mu_c^Q\|}{\sqrt{\lambda_\star}}.$$

Cauchy–Schwarz over the two classes gives $\sum_c \|\mu_c^S - \mu_c^Q\| \le \sqrt{2}\,\mathcal{L}_{\mathrm{proto}}^{1/2}$. Taking the supremum over h and multiplying by 2,

$$d_{\mathcal{H}}(P_S, P_Q) \le 2 \cdot \frac{1}{2\sqrt{2\pi}} \cdot \frac{\sqrt{2}}{\sqrt{\lambda_\star}} \sqrt{\mathcal{L}_{\mathrm{proto}}} = \frac{1}{\sqrt{\pi\lambda_\star}} \sqrt{\mathcal{L}_{\mathrm{proto}}}.$$

Hence $C = (\pi\lambda_\star)^{-1/2}$. □

The linear kernel enters only through Lemma 2, which identifies $\mathcal{L}_{\mathrm{proto}}$ with class-conditional MMD. The bound controls the divergence magnitude, not the gradient direction.

PART (II): ORTHOGONALITY IN EXPECTATION

Objects. Fix the current parameters θ_0 and the query law P_Q. For $h \in \mathcal{H}$ define the signed objective

$$L^h(\theta; P_S) = q^h_{P_S}(\theta) - q^h_{P_Q}(\theta).$$

For each source law P_S, a *witness* $h_{P_S} \in \arg\max_{h\in\mathcal{H}} |L^h(\theta_0; P_S)|$ is chosen once, at θ_0, and then *held fixed* as θ varies. Set

$$L_{\mathrm{DC}}(\theta; P_S) := L^{h_{P_S}}(\theta; P_S), \qquad G(P_S) := \nabla_\theta L_{\mathrm{DC}}(\theta_0; P_S),$$

so that $d_{\mathcal{H}}(P_S, P_Q) = 2|L_{\mathrm{DC}}(\theta_0; P_S)|$ at θ_0. Write $g_{\mathrm{CE}}(P_S)$ for the classification gradient at θ_0, allowing it to depend on P_S.

Assumptions.

- **(S0)** *Prior.* π is a probability measure on a measurable set $\mathcal{P}$ of source laws; the query P_Q is fixed.
- **(S1)** *Measure-preserving mirror.* There is a measurable involution $\iota : \mathcal{P} \to \mathcal{P}$, $\iota \circ \iota = \mathrm{id}$, with $\iota_\# \pi = \pi$. Write $\tilde{P}_S = \iota(P_S)$.
- **(S2′)** *Local antisymmetry.* There is an open neighbourhood $U \ni \theta_0$ such that, for π-a.e. P_S, $h_{\tilde{P}_S} = h_{P_S}$ and
$$L^{h_{P_S}}(\theta; \tilde{P}_S) = -L^{h_{P_S}}(\theta; P_S) \qquad \text{for all } \theta \in U.$$
- **(S3)** *Mirror-invariant classification gradient.* For π-a.e. P_S, $g_{\mathrm{CE}}(\tilde{P}_S) = g_{\mathrm{CE}}(P_S)$. This holds automatically when $\mathcal{L}_{\mathrm{CE}} = -\mathbb{E}_{(x,y)\sim P_Q}[\log p_\theta(y \mid x)]$ is computed on the query, since then g_{CE} does not depend on P_S.
- **(R1)** *Differentiability.* For π-a.e. P_S and $P \in \{P_S, \tilde{P}_S, P_Q\}$, $\theta \mapsto q_P^{h_{P_S}}(\theta)$ is differentiable at θ_0. This requires that $w^\top f_\theta(x) + b$ place no probability mass on the decision boundary near θ_0.
- **(R2)** *Integrability.* $P_S \mapsto \langle g_{\mathrm{CE}}(P_S), G(P_S)\rangle$ is π-integrable.

(S2′) at $\theta = \theta_0$ recovers the pointwise reflection $q^h_{\tilde{P}_S}(\theta_0) = 2q^h_{P_Q}(\theta_0) - q^h_{P_S}(\theta_0)$. The requirement that $2q_Q - q_S \in [0,1]$ is then automatic, since $\tilde{P}_S$ is a genuine law. The pointwise reflection alone is *not* sufficient (see the remark below).

Proof. Step 1 (gradient antisymmetry). Fix P_S outside the null sets in (S2′) and (R1), and let $h = h_{P_S}$. By (S2′) the functions $\theta \mapsto L^h(\theta; \tilde{P}_S)$ and $\theta \mapsto -L^h(\theta; P_S)$ coincide on the open set $U \ni \theta_0$. By (R1) both are differentiable at θ_0, and functions that agree on an open neighbourhood of a point have equal gradients there. Since $h_{\tilde{P}_S} = h$, the left-hand function is $L_{\mathrm{DC}}(\cdot\,; \tilde{P}_S)$. Hence

$$G(\tilde{P}_S) = -G(P_S).$$

Step 2 (odd integrand). Let $F(P_S) := \langle g_{\mathrm{CE}}(P_S), -G(P_S)\rangle$. By (S3) and Step 1,

$$F(\tilde{P}_S) = \langle g_{\mathrm{CE}}(P_S),\, G(P_S)\rangle = -F(P_S) \qquad \text{for } \pi\text{-a.e. } P_S.$$

Step 3 (integrate). F is integrable by (R2). Since $\iota_\#\pi = \pi$ by (S1), the change-of-variables formula gives

$$\mathbb{E}_\pi[F] = \int F\, d\pi = \int F\, d(\iota_\#\pi) = \int F \circ \iota\, d\pi = -\int F\, d\pi = -\mathbb{E}_\pi[F].$$

Hence $\mathbb{E}_{P_S\sim\pi}\big[\langle g_{\mathrm{CE}}, -\nabla_\theta L_{\mathrm{DC}}\rangle\big]_{\theta=\theta_0} = 0$. □

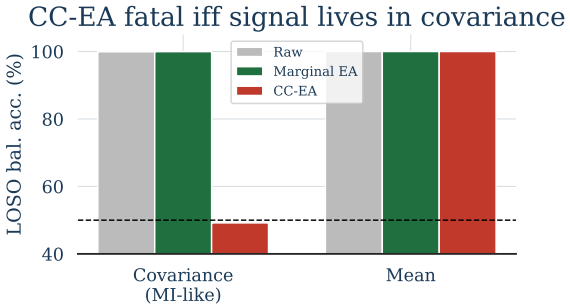


Figure 2: Causal two-mode principle (Proposition 1). CC-EA collapses accuracy to chance only when the class signal lives in the covariance (left); when it lives in the mean (right), CC-EA is a single shared whitening and leaves accuracy untouched. Same operation, opposite outcomes, flipped only by where the signal lives.

Special case. If g_{CE} does not depend on P_S (query-side CE) and G is π-integrable, the same argument applied to G gives the stronger statement $\mathbb{E}_\pi[\nabla_\theta L_{\mathrm{DC}}(\theta_0;\cdot)] = 0$, and the theorem follows by pulling g_{CE} out of the expectation.

Remark 4 (Why pointwise reflection is not enough)**.** *If the mirror negates L_{DC} only at θ_0, the gradients need not be negatives of each other. For example, $\ell_1(\theta) = \theta$ and $\ell_2(\theta) = \theta - 2$ satisfy $\ell_2(1) = -\ell_1(1)$, yet $\ell_1'(1) = \ell_2'(1) = 1$. Step 1 therefore requires antisymmetry on a neighbourhood of θ_0, which is (S2′).*

Remark 5 (Fixed witness and the unsigned divergence)**.** *The witness h_{P_S} is held fixed as θ varies, and the result concerns the* signed *objective. It does not extend to $\nabla_\theta d_{\mathcal{H}}$. Under the mirror, L^h changes sign, so the maximiser of the signed objective for $\tilde{P}_S$ is the complementary classifier. Consequently $d_{\mathcal{H}} = 2|L_{\mathrm{DC}}|$, and its gradient, are* invariant *rather than antisymmetric under ι, and no cancellation occurs.*

Remark 6 (Status of the assumptions)**.** *(S1)–(S2′) together constitute the symmetric-prior assumption. They are postulated, not derived: exchangeability of LOSO subjects makes them plausible, but at* 4–14 *subjects they hold at best approximately. (S3) holds exactly for query-side CE but generally fails for source-side CE, since a mirrored source pool has different data. These are idealizations, so we treat exact orthogonality as a structural prediction rather than an identity: it holds in expectation over π, not per instance, and is validated empirically through the proto/CE ratios of Table 1.*

K Proof of Corollary 3

Proof. By Theorem 2, episodic sampling ($P_S \neq P_Q$) is necessary for the alignment gradient to carry $\mathcal{H}$-divergence signal: under minibatch sampling ($P_S = P_Q$) the expected gradient's $O(1)$ component vanishes and only an $O(1/n)$ within-class variance term survives. DANN and CORAL, in their standard form, draw each minibatch from the pooled training-subject distribution without a held-out-subject query split, so they operate entirely in the $P_S = P_Q$ regime; loss weight and encoder architecture do not alter the sampling regime. Moreover, under the symmetric-prior conditions of Theorem 3(ii), the signed domain-classifier gradient is orthogonal in expectation to any classification gradient that is invariant under the source mirror (assumption (S3)), so the label objective supplies no systematic divergence-reducing direction either. Hence neither method, so implemented, induces the episodic structure that Theorem 2 shows to be necessary, and neither systematically reduces cross-subject $\mathcal{H}$-divergence. □

L Additional Tables and Figures

Table 3: The pre-alignment ceiling across method families (LOSO balanced accuracy, %). No linear alignment, nonlinear Riemannian feature, or deep invariance objective clears marginal EA meaningfully. ([†]Cho2017, 49-subject LOSO; all others BCI IV-2a.) [‡]Conditional-CORAL oracle uses target labels; a shuffled-label control retains 50.5%, i.e. the oracle gain is entirely label leakage (Section 6).

Family	Method	Balanced acc. (%)
None	Raw	66.78
Linear (marg.)	Marginal EA	**72.34**
Linear (CC)	CC-EA (honest / oracle)	49.34 / 63.43
Linear (align)	Procrustes rotation	57.99
Linear (align)	Shared discriminative basis	72.34
Linear (align)	Discriminative reweighting (best)	72.76
Linear (cond.)	Conditional CORAL (honest / oracle)	64.16 / 82.14[‡]
Linear (cond.)	Conditional MMD (equalizing)	50.96
Linear (cond.)	Conditional MMD (distinct)	69.83
Nonlinear	Riemannian tangent space	71.53
Deep	EEGNet (tuned)	74.15
Deep	+ contrast loss	74.50
Deep	+ adversarial loss	73.96
Deep	+ both	74.96
Deep[†]	EEGNet baseline (49 subj.)	71.76
Deep[†]	+ contrast + adversarial	71.27

Table 4: Between-class contrast under three alignments, mean over all 9 subjects of BCI IV-2a. Marginal EA preserves the contrast to four decimals; full CC-EA drives it to exactly zero for every subject.

	Raw	Marginal EA	CC-EA (full)
riem (mean)	0.9342	0.9342	0.0000
gev (mean)	0.1903	0.1903	0.0000

Table 5: Multilingual cross-subject imagined-speech benchmark. Binary classification; 36-dimensional Riemannian tangent-space features from 8-channel EEG. Subject counts are pre-screening totals.

Language	Unit	Subjects
Telugu	Words	6
Telugu	Phonemes	6
Telugu	Vowels	8
Kannada	Words	14
Kannada	Phonemes	11
Kannada	Vowels	14
English	Words	6
English	Phonemes	4
English	Vowels	7

Table 6: Subject-aware ablation on imagined speech: separating the $P_S \neq P_Q$ distribution gap from LOSO episodic structure (KMMD prototype-alignment gap, mean over the four conditions with ≥ 6 usable subjects). A subject-aware minibatch (support and query drawn from two distinct subjects, no held-out query) recovers the full episodic signal; the episodic$-$subject-aware difference is negligible and never significant. One KANNADA-Words subject (single-class, degenerate covariances) is excluded as in the LOSO classification protocol.

Scheme comparison	Mean KMMD gap	Significant ($p \leq 0.05$)
Episodic $-$ pooled	$+0.270$	$4/4$
Subject-aware $-$ pooled	$+0.291$	$4/4$
Episodic $-$ subject-aware	-0.020	$0/4$

Table 7: LOSO balanced accuracy (%) on BCI IV-2a. Marginal EA improves over raw by $+5.6$ pp; CC-EA, applied realistically, collapses to chance (49.3%). The alignment machinery is identical—the only change is grouping by class before whitening.

Subject	Raw	Marginal EA	CC-EA
S1	81.60	84.38	46.18
S2	52.43	52.78	51.39
S3	77.08	81.94	54.17
S4	61.81	77.08	42.71
S5	60.42	64.58	46.88
S6	63.89	68.75	43.06
S7	59.72	68.75	45.83
S8	81.60	81.94	57.64
S9	62.50	70.83	56.25
Mean	**66.78**	**72.34**	**49.34**
Std	10.56	10.14	5.66

Table 8: Removing every excuse for CC-EA on BCI IV-2a (LOSO, %). Even an oracle with true target labels plateaus 9 pp below marginal EA.

Method	Balanced acc. (%)
Marginal EA (reference)	**72.34**
CC-EA, honest (target fallback)	49.34
CC-EA, pseudo-labeled target	58.49
CC-EA, confidence-gated pseudo-labels	62.08
CC-EA, oracle (true target labels; upper bound)	63.43

Table 9: Causal control (LOSO balanced accuracy, %; 30 dataset draws). CC-EA is catastrophic exactly when the signal lives in the covariance it whitens, and completely harmless when the signal lives in the mean. Same operation, opposite outcomes, flipped only by signal placement (interaction $p \approx 4 \times 10^{-53}$).

Signal placement	Raw	Marginal EA	CC-EA (honest)	CC-EA (oracle)
Covariance (MI-like)	99.92	99.96	**49.19**	**49.85**
Mean	100.00	100.00	**100.00**	**100.00**

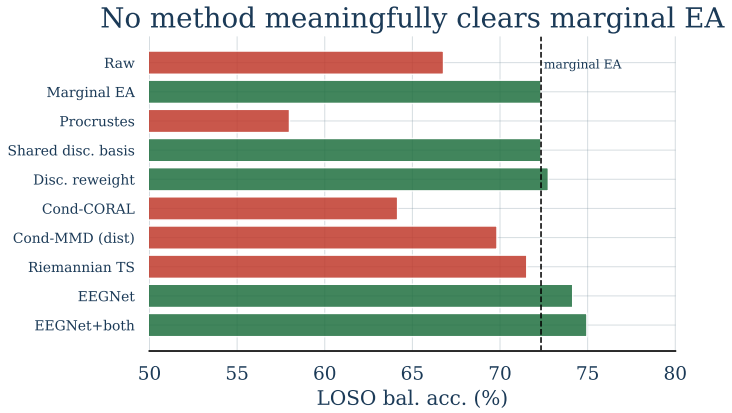


Figure 3: The pre-alignment ceiling. No linear alignment, nonlinear Riemannian feature, or deep invariance objective meaningfully clears marginal EA (dashed line), as Lemma 1 and Theorem 1 predict.

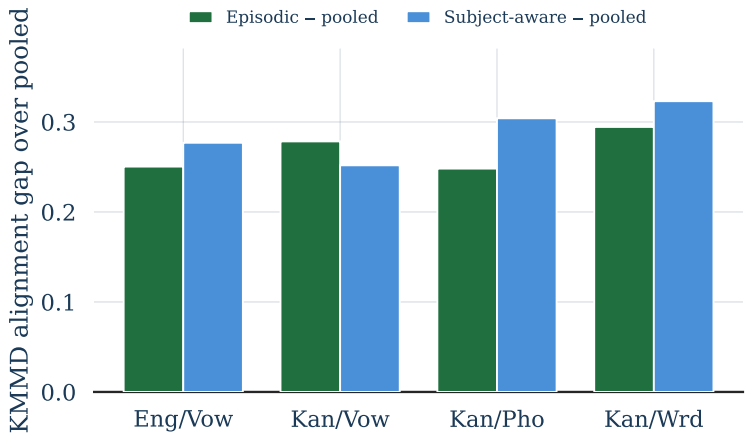


Figure 4: Subject-aware ablation (KMMD, four conditions with ≥ 6 subjects). Episodic$-$pooled and subject-aware$-$pooled gaps are both large and indistinguishable, confirming the signal is the $P_S \neq P_Q$ distribution gap rather than LOSO-specific structure (Table 6, Theorem 2).