# MULTI-LEVEL TRANSPORT ALIGNMENT FOR TIME-SERIES DOMAIN GENERALIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Time-series domain generalization is difficult because class semantics are often entangled with domain-specific signal variation: the same class may appear differently across subjects, sensors, datasets, and recording conditions. Existing pairwise or contrastive alignment methods typically define relations locally, either by matching selected sample pairs or by normalizing relations separately for each sample, which can miss the global structure induced by class semantics and class-conditional domain variation. We propose Multi-Level Transport Alignment (MULTA), a balanced optimal transport objective for time-series domain generalization. MULTA learns a globally organized transport structure in the unit representation space, encouraging class-discriminative representations while suppressing spurious source-domain shortcuts. We further introduce an adaptive alignment-gap reweighting mechanism that focuses optimization on samples whose learned alignment structure deviates most from the target relation. We show that globally balanced optimal transport yields a tighter bound on class-conditional source variation than row-wise local matching for fixed representations, and derive a generalization bound linking Multa's transport objective to unseen-domain risk. Across five leave-one-domain-out cross-dataset sleep staging tasks and 25 cross-person human activity recognition tasks, MULTA consistently improves over existing baselines under both dataset-level and subject-level domain shifts.

## 1 INTRODUCTION

Time-series signals exhibit substantial variability across subjects, devices, and recording conditions (Lara & Labrador, 2012; Chen et al., 2021), which introduces major shifts in how the same underlying class is expressed (Wang et al., 2024b). In biosignal settings, this variability appears as shifts in amplitude, temporal dynamics, frequency content (Zhang et al., 2022), and noise characteristics across different people (Wu et al., 2020; Wang et al., 2024b) and sensors (Kunze & Lukowicz, 2014). Nonstationary temporal and phase-related variation can further change the structure of time-series representations across domains (Liu et al., 2022; Mohapatra et al., 2025). Such structured, multi-scale transformations degrade generalization, even when models perform well within-domain (Gagnon-Audet et al.; Fan et al., 2023).

Here, we study target-free zero-shot domain generalization (DG). In this setting, the model is trained only on labeled source domains, selected using only source-domain validation data, and evaluated on unseen target domains without any access to target-domain (Zhou et al., 2022; Gulrajani & Lopez-Paz, 2020). In time-series, this problem is particularly challenging because domain shifts act directly on the signal, reshaping the temporal patterns that define each class (Lu et al., 2022c). As a result, shifts across datasets exhibit a natural hierarchy: samples cluster by class, but within each class they are further organized by domain-specific factors such as subject identity (Qian et al., 2021), sensor configuration or placement (Lu et al., 2022b), and recording protocol or dataset (Xiong et al., 2025). Generalization therefore, requires preserving class structure while also accounting for the structured, domain-dependent variation.

Most existing approaches address this problem through pairwise or local alignment (Mahajan et al., 2021), or contrastive learning (Khosla et al., 2020; Dave et al., 2022). Time-series contrastive methods typically construct positive relations through temporal views (Tonekaboni et al., 2021) or hierarchical contexts (Yue et al., 2022). Masked time-series modeling also relies on neighborhood aggregation to



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.

reconstruct masked signals (Dong et al., 2023). Related representation objectives also disentangle factors such as seasonal and trend components (Woo et al., 2022). However, these approaches do not explicitly control how samples are organized across domains at a global level (Wang & Isola, 2020). We argue that global organization is important for time-series generalization because local similarity can be unreliable (Chuang et al., 2020). Different classes may share short-term patterns (Ye & Keogh, 2009; Grabocka et al., 2014), while same-class samples can differ substantially across domains (Lu et al., 2022c). Moreover, models may rely on domain-specific shortcuts (Geirhos et al., 2020), such as amplitude scaling (Kim et al., 2021) or spectral biases (Rahaman et al., 2019), further limiting transfer. Together, these limitations reveal a narrower gap in current methods, which often specify which pairs should be close without determining how alignment should be globally organized across domains (Shi et al., 2023; Chen et al., 2024).

In this work, we address this gap by formulating time-series DG as a problem of structured, multi-level distribution alignment. We use an optimal transport (OT) framework to define a *multi-level target transport plan* that encodes how alignment mass should be distributed across samples across classes and domains (Fig 1). The idea is to learn embedding that allow for this multi-level transport to be achieved with minimal transport cost. Further, we define a notion of hardness in transport which we use to guide learning through a *alignment-gap weighted transport objective*. This encourages representations that preserve class structure while remaining robust to domain-specific distortions.

Section 4 analyzes the main design choices of MULTA. We first bound class-conditional source variation by the transport-matching loss (Proposition 1), and use this analysis to derive an unseen-domain risk bound (Theorem 1). We further show how global balancing tightens this source-variation bound by quantifying its exact reduction in unweighted matching loss over independent row normalization, at fixed affinities and a common balanced target (Proposition 2). Finally, to justify the proposed reweighting, we derive a lower bound on row-level KL divergence in terms of the alignment gap (Proposition 3), linking the weighting score to mismatch between the learned and target relations.

We conduct a comprehensive evaluation on two challenging DG benchmarks, HAROOD (Lu et al., 2025) and SleepDG (Wang et al., 2024a), covering cross-dataset and cross-subject distribution shifts. Across 30 DG tasks, including five cross-dataset sleep staging tasks and 25 cross-person HAR tasks, our method consistently outperforms existing DG approaches, demonstrating the benefits of multi-level alignment. Ablation studies and sensitivity analyses further confirm the flexibility and robustness of our method.

**Contributions**

- We identify a key improvement opportunity in time-series DG: existing methods enforce pairwise alignment but do not control the global organization of samples across domains.
- We propose *Multi-Level Transport Alignment* (MULTA) for timeseries, a distribution-level formulation of class-aware alignment using optimal transport, enabling explicit control over cross-domain relational structure globally.
- We introduce a *transport alignment-gap weighted objective* that prioritizes samples with large multi-level alignment mismatches, and provide a theoretical analysis that connects the matching loss to class-conditional source variation and yields an unseen-domain risk bound.
- We demonstrate consistent improvements on SleepDG cross-dataset sleep staging and cross-person human activity recognition scenarios from HAROOD benchmarks.

## 2 BACKGROUND AND RELATED WORK

**Time-series DG and class-domain relations.** Domain generalization aims to learn representations that transfer to unseen domains without target domain samples during training (Wang et al., 2022). While vision DG often focuses on shifts in appearance, texture, background or style, domains in multivariate sensor and biosignal time series commonly arise from subjects, sensors, devices, sessions, recording environment, or datasets. Such factors can substantially alter amplitude, phase, channel morphology,noise and etc. even within the same class (Dissanayake et al., 2020; Ballas & Diou, 2024). Related sleep-staging DG work further highlights the importance of handling cross-dataset shifts through hierarchical alignment of epoch and sequence-level representations (Wang et al., 2024a).

Therefore, same-class cross-domain pairs are not just additional positives, but informative supports for modeling how class-shared semantics interact with domain-conditioned signal variation.

**Class-aware alignment.** Class-aware DG and supervised contrastive methods have been widely used in the CV community, which use labels to align same-class samples and separate different-class samples across domains (Motiian et al., 2017; Khosla et al., 2020). More specialized DG methods infer cross-domain matches (Mahajan et al., 2021), construct cross-domain class-contrastive positives (Dave et al., 2022), or use proxies to reduce unreliable sample-level matching under large domain gaps (Yao et al., 2022). These methods motivate the use of class and domain labels, but standard alignment objectives do not explicitly coordinate how mass is distributed across identity pairs, same-class cross-domain pairs, same-class same-domain pairs, and different-class pairs.

**Row-wise contrastive matching versus global transport.** Contrastive objectives commonly define anchor-wise matching distributions by row-normalizing pairwise similarities. This local normalization is useful for semantic matching, but it lets different anchors allocate relation mass independently. Optimal transport provides a complementary view by computing a globally constrained coupling under marginal constraints (Cuturi, 2013; Gabriel & Marco, 2019). Recent contrastive-transport analyses connect contrastive learning with entropic distribution alignment and target-plan matching (Shi et al., 2023; Chen et al., 2024). MULTA belongs to this optimal transport category not for augmentation matching, but to organize labeled source samples according to a class-domain target relation.

**OT for time-series adaptation versus target-free DG.** For time series, transport-based methods have mostly been developed for adaptation or test-time settings where target/test-domain samples are observed. Match-and-Deform combines OT with temporal alignment for time-series domain adaptation (Painblanc et al., 2023). Monge-based methods normalize sleep or multivariate signals through convolutional or spatio-temporal alignment (Gnassounou et al., 2023; 2024), while related test-time normalization adapts sleep-staging representations at inference time (Gnassounou et al.). Deep OT on SPD manifolds addresses domain adaptation for covariance-structured representations (Ju & Guan, 2025). In contrast, MULTA uses transport to organize source-source class-domain relations before any target-domain samples are available. Our setting is different: target samples are unavailable during training. MULTA therefore uses OT as a source-source relation organizer for target-free DG, not a source-target adaptation mechanism.

**Positioning of MULTA.** Taken together, prior work suggests that class-aware relations are useful for DG and that transport can globally organize matching mass. However, the source-side relation-structure problem in target-free time-series DG remains less directly addressed: how should labeled source samples be organized when class semantics are shared across domains but their signal expressions vary by domain? MULTA addresses this question by constructing a multi-level target relation from class labels and source-domain identities, computing a globally balanced transport relation over source samples, and upweighting anchors whose learned relation under-allocates same-class cross-domain mass or assigns mass to different-class pairs.

# 3 METHOD

## 3.1 PROBLEM SETUP

We consider supervised time-series domain generalization with a source batch $\mathcal{B} = \{(x_i, y_i, d_i)\}_{i=1}^{B}$, where $x_i$ is a time-series window, $y_i$ its class label, and $d_i$ its source domain. Target-domain samples are inaccessible during training. We train an encoder and classifier using source cross-entropy $\mathcal{L}_{\mathrm{CE}}^{\mathrm{src}}$. The classifier uses $\mu_i = f_\theta(x_i)$, while alignment uses the normalized feature $z_i = \mu_i / \|\mu_i\|_2$.

## 3.2 METHOD OVERVIEW

Source classification encourages class separation but does not explicitly align same-class samples across domains. Given $\mathcal{B}$, MULTA constructs a multi-level target transport plan $T$ and learns a balanced Sinkhorn transport plan $P_\theta$ from latent affinities. This transport-plan matching view is related to prior representation-learning work (Chen et al., 2024), while our target plan and alignment-gap weighting are designed for supervised time-series DG. Here an *anchor* is the row sample $x_i$, whose transport mass is allocated over candidate samples $x_j$. MULTA computes an *alignment gap* for each anchor, measuring missing mass on cross-source positives and leakage to different-label samples, and uses this gap to reweight the transport-matching loss. The right panels of Figure 1

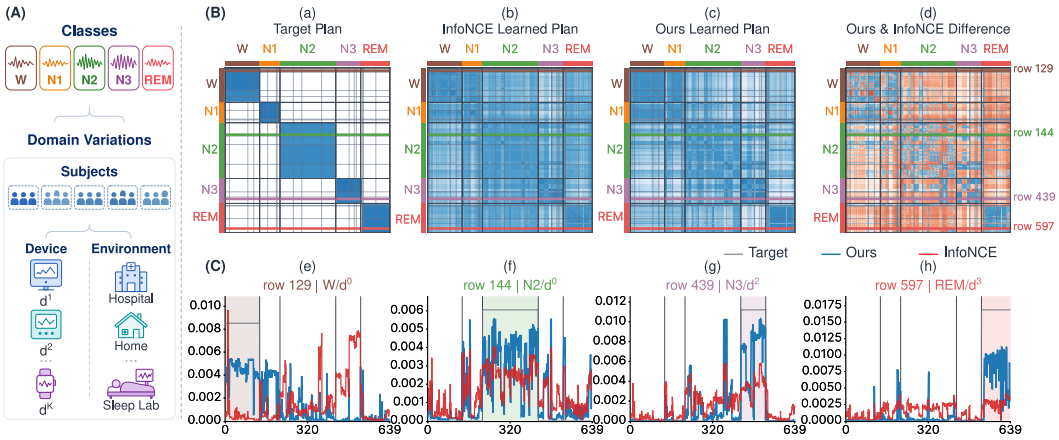


Figure 1: **MULTA exploits multi-level domain structure and learns globally organized affinity. (A):** Sleep stages are shared across domains, while subjects, devices, and environments create domain-specific variation. This multi-level structure motivates class-aware alignment rather than flat domain alignment. **(B):** affinity matrices on a sampled source set ordered by sleep-stage labels and four source domains (darker blue means higher value, sorted by class then domain). From left to right, the target plan, the InfoNCE plan, and MULTA transport plan, and the difference of the transport plan learned by MULTA minus the InfoNCE affinity where the red color indicates negative values) on the off-diagonal (across-classes). **(C)** Below, we show the learned row-wise coupling relation for MULTA (blue) and InfoNCE (red) for samples from the four representative classes. Shaded regions indicate the target support for each anchor, gray denotes the target plan value.

qualitatively illustrate the target plan and learned Sinkhorn transport plan on a sampled source set. Compared with a non-global matching baseline such as InfoNCE, MULTA places more coupling mass on same-label cross-source supports and less mass on different-label regions.

### 3.3 MULTI-LEVEL TRANSPORT ALIGNMENT

We construct $T$ from the labels and domains in the source batch. For each anchor, we give a score to its self-pair, its same-class partners in other domains, and its same-class partners in the same domain. Different-class pairs receive zero score. The three relations have construction weights $\alpha_{\mathrm{id}} > 0$ and $\alpha_{\mathrm{cd}}, \alpha_{\mathrm{sd}} \geq 0$. For each relation, its weight is divided among the anchor's eligible partners.

We first balance the resulting scores to form $T$. Every sample then sends and receives total mass $1/B$. Let $u = B^{-1}\mathbf{1}$, and let $\mathcal{U}(u,u) = \{A \geq 0 \mid A\mathbf{1} = u,\ A^\top \mathbf{1} = u\}$ denote the plans with these row and column sums. Thus $T \in \mathcal{U}(u,u)$. Balancing adjusts the scores assigned to individual pairs while keeping zero-score pairs at zero. The $\alpha$ values set the relative scores before balancing, so they need not equal the final proportions of mass in $T$. Further details are given in App. A.1. Figure 1, panel (a), shows the resulting same-class cross-domain blocks.

For the learned plan from data, let $K_\theta$ be the feature similarity matrix with entries $K_{\theta,ij} = \exp(z_i^\top z_j/\tau)$, where $\tau > 0$ is the temperature. The learned plan is $P_\theta = \mathrm{Sinkhorn}(K_\theta; u, u) \in \mathcal{U}(u,u)$. Sinkhorn alternates row and column scaling to meet these marginals (Cuturi, 2013; Gabriel & Marco, 2019). The target $T$ depends on the batch labels and domains, while $P_\theta$ changes with the encoder features.

### 3.4 ALIGNMENT-GAP WEIGHTING AND OBJECTIVE

For anchor $i$, let $t_{ij} = BT_{ij}$ and $q_{ij} = BP_{\theta,ij}$ be the target and learned probabilities assigned to candidate $j$. We match the two row distributions with $\ell_i = -\sum_j t_{ij} \log q_{ij}$. To measure discrepancies in the relations of interest, define $M^+_{ij} = \mathbf{1}\{i \neq j,\ y_i = y_j,\ d_i \neq d_j\}$ and $M^-_{ij} = \mathbf{1}\{y_i \neq y_j\}$. The first is one when $j$ is a same-class partner from another domain, and the second is one when $j$ has a different class. Their rows $M^+_i$ and $M^-_i$ select these two groups of candidates for anchor $i$. We

define the corresponding gaps and their combined score by

$$\Delta_i^+ = [\langle t_i - q_i, M_i^+ \rangle]_+, \qquad \Delta_i^- = [\langle q_i - t_i, M_i^- \rangle]_+, \qquad \zeta_i = \frac{\beta_+ \Delta_i^+ + \beta_- \Delta_i^-}{\beta_+ + \beta_-}. \tag{1}$$

Here $\langle a, b \rangle = \sum_j a_j b_j$ adds the differences over the selected candidates, and $[a]_+ = \max(a, 0)$ discards negative totals. The first gap measures the shortfall of learned mass on same-class cross-domain partners relative to $T$. Because $T$ gives no mass to different-class partners, the second gap is precisely the learned mass assigned to them. The coefficients $\beta_+, \beta_- \geq 0$ satisfy $\beta_+ + \beta_- > 0$ and set the gaps' relative contributions. Thus $\zeta_i$ is the alignment-gap score for anchor $i$.

Let $\zeta_{\min} = \min_k \zeta_k$ and $\zeta_{\max} = \max_k \zeta_k$. With a numerical floor $\delta > 0$, we rescale the scores within each batch and set the anchor weights as $\hat{h}_i = \frac{\zeta_i - \zeta_{\min}}{\max\{\zeta_{\max} - \zeta_{\min}, \delta\}}$, $w_i = 1 + \lambda_h \hat{h}_i$ (2). The floor makes this expression defined when every gap in the batch is equal. The parameter $\lambda_h \geq 0$ sets the maximum additional weight, so $1 \leq w_i \leq 1 + \lambda_h$. In other words, $\lambda_h \geq 0$ controls the strength of alignment-gap reweighting. The training objective is $L_w = \frac{\sum_i w_i \ell_i}{\sum_i w_i}$, $\mathcal{L} = \mathcal{L}_{\mathrm{CE}}^{\mathrm{src}} + \lambda_{\mathrm{tr}} L_w$ (3), where $\lambda_{\mathrm{tr}} \geq 0$ controls the contribution of transport-plan matching relative to source cross-entropy. Anchors with larger alignment gaps contribute more to the matching loss. During each update, $T$ and the computed weights are held fixed, while gradients through $P_\theta$ update the encoder. Per-batch pseudocode is given in App. P.

## 4 Analysis

We derive an upper bound on unseen-domain risk to explain the alignment objective in Sec. 3. The bound includes source risk, an explicit target-to-source discrepancy, and a source-variation term controlled in part by transport matching. We also show that global balancing can tighten the unweighted source-variation bound at fixed features, and that the alignment gap certifies a lower bound on an anchor's row-level matching error.

We use the quantities defined in Sec. 3 throughout and write $z(x)$ for the normalized alignment feature. To compare with the weighted training loss, define the unweighted matching loss

$$L_{\mathrm{uw}} = \frac{1}{B} \sum_i \ell_i, \qquad 1 \leq w_i \leq 1 + \lambda_h. \tag{4}$$

The results concern a fixed encoder and exact transport plans with the prescribed marginals, including self-pairs. Expectations are over source batches. Conditional on batch labels and domains, each sample follows its corresponding class–domain distribution. Full assumptions are given in App. A and App. F.

**What does target matching control?** Source classification can separate classes while leaving same-class features in different domains far apart. Let $\nu_{d,y}$ be the distribution of normalized features $z$ for class $y$ in source domain $d$. To measure this variation, define

$$S_z = \sum_{y, d \neq d'} \lambda_{y,d,d'} W_2^2(\nu_{d,y}, \nu_{d',y}), \tag{5}$$

where $W_2$ is the 2-Wasserstein distance with Euclidean ground metric. Here $\lambda_{y,d,d'}$ is the *expected target-plan mass* on ordered pairs from class $y$ in domains $d$ and $d'$. It measures coverage of that pair of domains under $T$ and batch sampling. The sum covers distinct source domains containing $y$.

**Proposition 1** (Control of class-conditional source variation)**.** *For the target described in Sec. 3.3, under the sampling assumptions of App. A,*

$$S_z \leq 2\tau \mathbb{E} L_{\mathrm{uw}} \leq 2\tau(1 + \lambda_h) \mathbb{E} L_w. \tag{6}$$

The target mass determines which same-class cross-domain distribution pairs contribute to $S_z$. The first inequality bounds their variation by the transport-matching loss. The second expresses this bound using the weighted loss $L_w$ optimized during training, since the anchor weights lie in $[1, 1 + \lambda_h]$. The proof and a comparison with source classification are in App. B.

**Why global transport?** Row normalization makes each anchor distribute mass independently. Global balancing also limits the total mass received by each candidate. For the same similarity matrix $K_\theta$, define the row-only plan $Q_{ij} = u_i K_{\theta,ij} / \sum_k K_{\theta,ik}$ and its unweighted matching loss $L_{\mathrm{row}} = -\sum_{ij} T_{ij} \log(BQ_{ij})$. This uses the same target $T$ as $L_{\mathrm{uw}}$, but matches it with $Q$ instead of $P_\theta$.

**Proposition 2** (Global target matching)**.** *For a fixed affinity matrix and target $T$ with both marginals $u$,*

$$L_{\mathrm{row}} - L_{\mathrm{uw}} = \mathrm{KL}(P_\theta \| Q) \geq \mathrm{KL}(u \| Q^\top \mathbf{1}) \geq 0. \tag{7}$$

*The inequality is strict if the column sums $Q^\top \mathbf{1}$ differ from $u$.*

Thus, at fixed features and target, global balancing gives an *unweighted* matching loss no larger than row normalization. It lowers the loss whenever the row-normalized plan has unequal column sums, and can therefore tighten the bound in Proposition 1. This comparison does not order the weighted losses or the results of separately trained models (App. C).

**Can source alignment bound unseen-domain risk?** The preceding results bound variation across source domains. To extend the analysis to an unseen domain, we include the distance between each target class's feature distribution and mixtures of the corresponding source-class feature distributions (Albuquerque et al., 2019; Dayal et al., 2023). This discrepancy appears explicitly in the risk bound. For each source sample $i$, let $\mu_i = f_\theta(x_i)$ be its classifier feature and assume $0 < \|\mu_i\|_2 \leq R$. The term $A_r$ measures the expected $T$-weighted squared difference between $\|\mu_i\|_2$ and $\|\mu_j\|_2$ on same-class cross-domain pairs. It is needed because alignment uses $z_i = \mu_i / \|\mu_i\|_2$, which does not control feature norms.

**Theorem 1** (Unseen-domain risk)**.** *Suppose source feature norms are nonzero and bounded, the classifier is linear in $\mu$ with bounded weights, and every pair of source domains containing the same class receives positive expected target-plan mass. Under the remaining assumptions of App. A,*

$$R_t^{\mathrm{bal,CE}} \leq R_{s,\gamma}^{\mathrm{bal,CE}} + \Delta_{\mathrm{unseen}} + \kappa_{\mathrm{cov}} \sqrt{A_r + 2\tau R^2 (1 + \lambda_h) \mathbb{E} L_w}. \tag{8}$$

*Here $R_t^{\mathrm{bal,CE}}$ is target cross-entropy averaged equally over classes, and $R_{s,\gamma}^{\mathrm{bal,CE}}$ is the analogous risk for a fixed, class-specific mixture $\gamma$ of source domains. $\Delta_{\mathrm{unseen}}$ is proportional to the sum of classwise $W_1$ distances from target feature distributions to their closest source mixtures. The coefficient $\kappa_{\mathrm{cov}}$ depends on the classifier weight bound and the smallest positive cross-domain target mass for each class. Definitions and proof are in App. D.*

The theorem gives an upper bound on target risk in terms of source risk, target discrepancy, and the source-alignment term. The target discrepancy is unknown and may be large. Training uses pooled source CE $\mathcal{L}_{\mathrm{CE}}^{\mathrm{src}}$, while the theorem uses class-balanced source risk. Relating the two may introduce a class-imbalance factor (App. D). When the other terms are fixed, a smaller matching loss tightens this bound.

**What does the alignment gap measure?** The source-variation and risk bounds require only the range of $w_i$ in Eq. (4). A separate result explains what the method's gap score detects within each row.

**Proposition 3** (Alignment gaps and row-level mismatch)**.** *For $\zeta_i$ in Eq.* (1)*,*

$$\mathrm{KL}(t_i \| q_i) \geq 2\zeta_i^2, \qquad \ell_i = H(t_i) + \mathrm{KL}(t_i \| q_i), \tag{9}$$

*where $H(t_i)$ is the entropy of the target row distribution.*

The target entropy $H(t_i)$ is fixed with respect to the encoder, while the KL term measures row-level matching error. A large gap therefore certifies a minimum mismatch of $2\zeta_i^2$. When $\lambda_h > 0$, anchors with larger gaps receive larger relative weights in the matching update. The proof is in App. E.

## 5 Experiments

This section presents a comprehensive evaluation of our approach. We first compare our method against standard benchmark baselines, as discussed in Section 5.2. We then perform ablation studies to analyze the impact of different model components, as described in Section 5.2.1.

### 5.1 SETUP

**Datasets.** We evaluate MULTA on two complementary zero-shot time-series DG benchmarks. SleepDG cross-dataset sleep staging (Wang et al., 2024a) tests dataset-level shifts in clinical PSG signals, where domains are distinct datasets with shared sleep-stage labels but different cohorts, sites, protocols, devices, and signal statistics. Following SleepDG, we use SleepEDFx (Kemp et al., 2000; Goldberger et al., 2000), HMC (Alvarez-Estevez & Rijsman, 2021), ISRUC (Khalighi et al., 2016), SHHS (Zhang et al., 2018; Quan et al., 1997), and P2018 (Ghassemi et al., 2018) under leave-one-dataset-out evaluation. HAROOD (Lu et al., 2025) complements this setting by testing subject-level shifts in wearable HAR, where held-out subject groups induce variation in motion patterns, body dynamics, sensor responses, and noise. We use the benchmark-preprocessed tasks on DSADS (Barshan & Yüksek, 2014), USC-HAD (Zhang & Sawchuk, 2012), UCI-HAR (Reyes-Ortiz et al., 2016), PAMAP2 (Reiss & Stricker, 2012), EMG (Lobov et al., 2018), and WESAD (Schmidt et al., 2018) with the official leave-one-subject-group-out protocol. Together, these benchmarks cover dataset- and subject-level shifts across clinical and wearable time series (App. G).

**Baselines.** We compare our OT-based method against the full set of domain generalization baselines included in the HAROOD benchmark, covering a broad spectrum of representative DG approaches under a standardized evaluation protocol with the same backbone: ERM, ERM++ (Teterwak et al., 2025), Mixup (Yan et al., 2020), CORAL (Sun & Saenko, 2016), MMD (Long et al., 2015), DANN (Ganin et al., 2016), MLDG (Li et al., 2018), GroupDRO (Sagawa et al., 2019), RSC (Huang et al., 2020), ANDMask (Parascandolo et al., 2020), VREx (Krueger et al., 2021), URM (Krishnamachari et al., 2024), Fish (Shi et al., 2021), Fishr (Rame et al., 2022), LAG (Lu et al., 2022a), DDLearn (Qin et al., 2023), and further include SupCon (Khosla et al., 2020) and PCL (Yao et al., 2022). For sleep staging, we also include baseline SleepDG(Wang et al., 2024a).

**Evaluation.** We evaluate under zero-shot DG following the original SleepDG and HAROOD train/validation/test protocols. In SleepDG, each dataset is held out once as the target domain; in HAROOD, the official cross-person subject-group splits cover 6 datasets and define 25 target tasks. We report balanced accuracy as the main table metric, with macro-F1, kappa, accuracy, and weighted-F1 in the appendix, and summarize results as mean $\pm$ standard deviation over five seeds $\{0, 42, 96, 123, 3407\}$. To assess statistical significance beyond average performance, we also report a significance-aware average rank: for each held-out target task, methods are compared across the same seeds using paired two-sided tests on balanced accuracy at $\alpha = 0.05$. A method's rank is 1 plus the number of methods significantly better than it, with statistically indistinguishable methods tied where lower is better. HAROOD ranks are averaged first over target tasks within each dataset and then across datasets. Further details are in App. G, I.2, and H.

**Implementation.** All methods use the benchmark encoder family for fair comparison: the official HAROOD CNN encoder for HAROOD tasks and the SleepDG CNN encoder for sleep staging. For sleep staging, each model encodes 30-second epochs, and a one-layer Transformer encodes sequences of 20 consecutive epochs, following AASM scoring practice and prior sleep staging work (Berry, 2014; Phan & Mikkelsen, 2022; Wang et al., 2024a). For HAROOD, we keep the official benchmark data loaders, subject-group domains, and CNN backbone. All methods are trained with AdamW, checkpoint selection and early stopping use only the source-domain validation split, never the held-out target domain. Full implementation, hyperparameter, and training details are provided in App. H.

### 5.2 MAIN RESULTS

Tables 1 and 2 report the main zero-shot DG results on SleepDG cross-dataset sleep staging (Wang et al., 2024a) and HAROOD cross-person human activity recognition (Lu et al., 2025).

**Cross-dataset sleep staging.** Table 1 shows that MULTA achieves the best average balanced accuracy and the best average rank under leave-one-dataset-out evaluation. MULTA has the highest mean balanced accuracy on four of the five held-out sleep datasets and obtains an average balanced accuracy of $72.22\%$ with average rank $1.00$, outperforming SleepDG (Wang et al., 2024a) ($71.05\%$, rank $1.60$), SupCon (Khosla et al., 2020) ($70.87\%$, rank $2.00$), and PCL (Yao et al., 2022) ($69.85\%$, rank $3.00$). The largest gain is on P2018, where MULTA improves over the strongest competing method, PCL, by $1.75$ points and it also improves the best competing result on Sleep-EDFx, SHHS, and HMC. Overall, the sleep results indicate that MULTA is effective under dataset-level shifts caused by differences in cohorts, recording protocols, and dataset-specific signal statistics.

Table 1: **Leave-one-dataset-out performance for sleep staging datasets.** Balanced accuracy is reported as mean ± sample standard deviation across 5 seeds. The Avg column first averages balanced accuracy across target datasets within each seed and then reports mean ± sample standard deviation across the five seed-level averages. Lower Avg. Rank is better. Best and second-best performance/rank entries are bolded and underlined, respectively.

| Method | P2018 | ISRUC | Sleep-EDFx | SHHS | HMC | Avg | Avg. Rank ↓ |
|---|---|---|---|---|---|---|---|
| ERM | 70.47 ± 1.76 | 69.50 ± 1.86 | 68.92 ± 1.19 | 65.77 ± 3.02 | 69.87 ± 1.82 | 68.91 ± 1.62 | 3.40 |
| ERM++ | 72.32 ± 1.10 | **73.26 ± 0.66** | 67.31 ± 2.04 | 66.06 ± 1.42 | 72.64 ± 0.59 | 70.32 ± 0.88 | 2.20 |
| Mixup | 71.24 ± 1.34 | 68.73 ± 1.82 | 68.18 ± 2.32 | 66.75 ± 2.96 | 68.90 ± 1.26 | 68.76 ± 1.36 | 3.80 |
| CORAL | 69.82 ± 1.38 | 69.95 ± 1.08 | 69.66 ± 1.35 | 68.38 ± 1.64 | 68.71 ± 2.38 | 69.31 ± 0.97 | 4.00 |
| MMD | 71.18 ± 0.93 | 69.45 ± 0.84 | 70.67 ± 0.87 | 68.07 ± 2.20 | 68.04 ± 1.26 | 69.48 ± 0.83 | 4.00 |
| DANN | 70.73 ± 0.81 | 68.36 ± 2.26 | 67.37 ± 2.70 | 65.13 ± 1.54 | 69.33 ± 0.76 | 68.18 ± 1.07 | 5.40 |
| MLDG | 20.89 ± 2.02 | 20.81 ± 1.25 | 31.34 ± 6.55 | 30.22 ± 9.46 | 23.55 ± 6.81 | 25.36 ± 2.57 | 19.00 |
| GroupDRO | 68.74 ± 1.95 | 68.10 ± 1.55 | 64.45 ± 4.27 | 66.77 ± 0.37 | 67.25 ± 1.44 | 67.06 ± 1.74 | 6.20 |
| RSC | 72.37 ± 1.53 | 69.32 ± 1.23 | 66.83 ± 1.81 | 65.33 ± 1.72 | 69.93 ± 1.89 | 68.76 ± 1.19 | 4.00 |
| ANDMask | 69.50 ± 1.66 | 69.07 ± 1.07 | 68.38 ± 1.20 | 65.18 ± 2.21 | 69.13 ± 1.64 | 68.25 ± 0.66 | 4.80 |
| VREx | 71.82 ± 1.83 | 69.03 ± 2.02 | 67.14 ± 2.01 | 65.14 ± 1.68 | 68.82 ± 1.23 | 68.39 ± 1.17 | 4.60 |
| URM | 68.93 ± 3.93 | 66.51 ± 3.40 | 65.63 ± 3.70 | 63.10 ± 2.58 | 65.43 ± 2.03 | 65.92 ± 2.13 | 7.60 |
| Fishr | 70.24 ± 2.39 | 70.51 ± 1.84 | 66.37 ± 3.25 | 66.15 ± 0.96 | 69.85 ± 1.76 | 68.63 ± 1.59 | 3.60 |
| Fish | 64.45 ± 1.82 | 65.28 ± 1.86 | 62.68 ± 2.73 | 65.36 ± 0.75 | 62.95 ± 2.27 | 64.14 ± 0.53 | 12.60 |
| LAG | 71.30 ± 1.73 | 69.45 ± 1.17 | 69.43 ± 3.31 | 67.19 ± 2.34 | 68.12 ± 2.38 | 69.09 ± 1.31 | 2.80 |
| DDLearn | 20.97 ± 0.32 | 20.61 ± 0.40 | 20.21 ± 0.34 | 20.19 ± 0.34 | 21.52 ± 0.95 | 20.70 ± 0.26 | 19.20 |
| SupCon | 73.21 ± 0.41 | 72.37 ± 0.36 | 69.90 ± 0.37 | 67.75 ± 1.50 | 71.09 ± 0.83 | 70.87 ± 0.53 | 2.00 |
| PCL | 73.24 ± 1.16 | 69.35 ± 1.92 | 69.13 ± 0.62 | 66.70 ± 1.67 | 70.82 ± 0.56 | 69.85 ± 0.61 | 3.00 |
| SleepDG | 72.96 ± 0.62 | 73.10 ± 0.42 | 71.31 ± 0.44 | 66.93 ± 1.21 | 70.96 ± 0.87 | 71.05 ± 0.34 | 1.60 |
| Ours | **75.00 ± 0.75** | 72.56 ± 0.63 | **72.16 ± 0.55** | **68.63 ± 0.96** | **72.77 ± 0.87** | **72.22 ± 0.33** | **1.00** |

Table 2: **Leave-one-subject-group-out balanced accuracy results for HAROOD human activity benchmark.** Balanced accuracy (%) is reported as mean ± sample standard deviation across five seeds. For each dataset column, we first average the target tasks within each seed and then compute the mean and standard deviation across seed-level dataset averages.The Avg column averages all dataset–task targets within each seed. Avg. Rank is computed using paired seed-level comparisons on each dataset-specific task; lower is better. Best and second-best performance/rank entries are bolded and underlined, respectively.

| Method | DSADS | USC | HAR | EMG | WESAD | PAMAP | Avg | Avg. Rank ↓ |
|---|---|---|---|---|---|---|---|---|
| ERM | 81.86 ± 1.40 | 70.01 ± 0.38 | 92.12 ± 0.36 | 74.19 ± 1.04 | 38.72 ± 4.55 | 65.97 ± 1.06 | 71.35 ± 0.83 | 3.66 |
| ERM++ | 79.56 ± 2.20 | 68.95 ± 1.15 | 92.48 ± 0.36 | 76.19 ± 0.73 | 38.36 ± 1.70 | 65.30 ± 1.17 | 71.03 ± 0.72 | 4.54 |
| Mixup | 85.36 ± 0.87 | 70.67 ± 0.65 | 93.21 ± 0.49 | 72.73 ± 1.20 | 39.20 ± 3.82 | 68.90 ± 0.86 | 72.54 ± 0.64 | 2.35 |
| CORAL | 82.32 ± 1.18 | 70.87 ± 0.74 | 92.14 ± 0.25 | 76.90 ± 0.65 | 33.43 ± 2.88 | 65.97 ± 0.90 | 71.15 ± 0.44 | 3.48 |
| MMD | 81.12 ± 1.96 | 69.54 ± 1.01 | 92.21 ± 0.62 | 76.70 ± 0.81 | 34.70 ± 2.60 | 65.62 ± 1.16 | 70.87 ± 0.83 | 3.88 |
| DANN | 81.87 ± 1.82 | 69.11 ± 1.23 | 91.85 ± 0.32 | 63.79 ± 6.69 | 32.56 ± 2.92 | 65.80 ± 1.23 | 68.47 ± 1.42 | 4.76 |
| MLDG | 30.00 ± 29.85 | 23.05 ± 4.43 | 41.86 ± 29.10 | 17.45 ± 1.15 | 25.04 ± 0.09 | 23.05 ± 1.23 | 27.35 ± 10.51 | 18.29 |
| GroupDRO | 81.73 ± 2.27 | 70.15 ± 0.45 | 92.10 ± 0.37 | 77.03 ± 0.81 | 31.67 ± 0.92 | 65.89 ± 0.81 | 70.65 ± 0.55 | 4.16 |
| RSC | 83.83 ± 0.87 | 73.65 ± 0.81 | 92.79 ± 0.38 | 77.20 ± 1.94 | 35.37 ± 3.14 | 69.28 ± 0.70 | 72.85 ± 0.58 | 1.94 |
| ANDMask | 82.74 ± 2.39 | 69.81 ± 0.60 | 91.81 ± 0.25 | 75.00 ± 0.97 | 37.03 ± 4.39 | 64.98 ± 2.15 | 71.09 ± 0.81 | 3.96 |
| VREx | 81.86 ± 1.48 | 69.94 ± 0.66 | 92.07 ± 0.42 | 75.48 ± 0.92 | 39.47 ± 2.93 | 66.34 ± 1.17 | 71.71 ± 0.65 | 3.57 |
| URM | 83.27 ± 1.22 | 71.49 ± 0.33 | 91.16 ± 1.21 | 76.59 ± 0.91 | 36.50 ± 3.08 | 68.77 ± 1.72 | 72.09 ± 0.63 | 2.72 |
| Fishr | 81.97 ± 1.33 | 70.20 ± 0.51 | 92.12 ± 0.37 | 74.09 ± 1.13 | 40.14 ± 4.78 | 66.02 ± 1.35 | 71.61 ± 0.93 | 3.43 |
| Fish | 83.03 ± 0.98 | 71.79 ± 0.80 | 92.20 ± 0.24 | 67.50 ± 8.97 | 28.79 ± 2.07 | 67.22 ± 1.70 | 69.37 ± 1.48 | 3.68 |
| LAG | 83.51 ± 1.27 | 68.62 ± 0.99 | 92.35 ± 0.31 | 71.49 ± 2.23 | 39.96 ± 1.53 | 63.71 ± 2.10 | 70.84 ± 0.50 | 5.34 |
| DDLearn | 61.35 ± 2.59 | 46.98 ± 2.47 | 80.69 ± 2.16 | 50.52 ± 2.44 | 34.24 ± 0.46 | 47.06 ± 3.48 | 54.56 ± 1.03 | 16.00 |
| SupCon | 86.61 ± 1.10 | 70.89 ± 0.65 | 93.42 ± 0.31 | 76.88 ± 0.72 | 46.10 ± 1.65 | 67.87 ± 0.97 | 74.42 ± 0.35 | 1.89 |
| PCL | 81.37 ± 0.69 | 69.80 ± 0.46 | 92.83 ± 0.38 | 74.47 ± 1.37 | 39.72 ± 1.41 | 66.19 ± 1.18 | 71.61 ± 0.25 | 3.86 |
| Ours | **88.35 ± 0.84** | **74.62 ± 0.46** | **94.48 ± 0.27** | **80.46 ± 0.21** | **50.82 ± 2.68** | **71.56 ± 1.05** | **77.43 ± 0.26** | **1.00** |

**Across-person human activity recognition.** Table 2 shows that MULTA obtains the best balanced accuracy on all six HAROOD datasets and the best overall rank. MULTA reaches an average balanced accuracy of $77.43\%$ with average rank $1.00$, improving over the strongest overall baseline, SupCon (Khosla et al., 2020) ($74.42\%$, rank $1.89$), by $3.01$ points. The gains are consistent across sensing benchmarks: MULTA improves the strongest non-MULTA result by $1.73$ points on DSADS, $0.97$ on USC-HAD, $1.06$ on UCI-HAR, $3.26$ on EMG, $4.72$ on WESAD, and $2.28$ on PAMAP2. These improvements show that the proposed global relation alignment is not specific to sleep staging, but also benefits subject-level generalization in wearable sensing.

Taken together, Tables 1 and 2 support the main empirical claim: organizing same-class cross-domain relations globally improves robustness under both dataset-level and subject-level zero-shot domain shifts. Detailed sleep-staging per-target results and additional metrics are provided in App. J. HAROOD per-task results and task-level rankings are provided in App. K, and the significance-aware

ranking protocol is described in App. I. We also provide a variance decomposition of HAROOD results across target tasks and random seeds in Table 15.

### 5.2.1 Ablations

We ablate MULTA on EMG and PAMAP across eight cross-subject group tasks, using source-only checkpoint selection. We report balanced accuracy as mean $\pm$ std over seeds $\{0, 42, 96\}$, after averaging tasks within each dataset and seed.

Table 3: **Ablations.** BA (%, mean $\pm$ std).

| Variant | EMG | PAMAP |
|---|---|---|
| MULTA (full) | **80.15 $\pm$ 0.27** | **71.21 $\pm$ 0.50** |
| *(a) Component ablations* | | |
| w/o Reweight | 78.90 $\pm$ 0.48 | 69.47 $\pm$ 1.36 |
| High-$\lambda_{\text{tr}}$ | 79.45 $\pm$ 0.80 | 69.09 $\pm$ 0.68 |
| w/o Align ($\lambda_{\text{tr}} = 0$) | 74.22 $\pm$ 0.29 | 65.32 $\pm$ 0.96 |
| InfoNCE+Rwt | 77.32 $\pm$ 0.88 | 69.55 $\pm$ 1.21 |
| InfoNCE | 76.17 $\pm$ 0.21 | 66.64 $\pm$ 1.21 |
| *(b) Target-plan ablations* | | |
| ID only | 72.63 $\pm$ 0.24 | 64.72 $\pm$ 1.47 |
| ID + SD | 73.12 $\pm$ 2.72 | 64.70 $\pm$ 1.54 |
| ID + CD | 79.76 $\pm$ 0.50 | 70.36 $\pm$ 3.03 |

**Alignment mechanism.** Table 3(a) shows that removing alignment causes the largest degradation, while a larger $\lambda_{\text{tr}}$ yields no further benefit, possibly due to overconstraining within-class geometry. With the same class-domain target structure, global matching outperforms row-normalized matching both with and without gap reweighting, isolating the benefit of joint row-and-column constraints. Reweighting improves both schemes, supporting complementary roles: global balancing coordinates assignments across anchors, while reweighting prioritizes poorly aligned anchors. Task-level results shows similar trend in Tables 22 and 23.

**Target relation design.** Table 3(b) separates identity (ID), same-class cross-domain (CD), and same-class same-domain (SD) relations. ID is retained throughout to avoid empty target rows. Adding SD to ID yields no consistent gain, whereas adding CD substantially improves performance, suggesting that explicit cross-domain connections provide the main transferable alignment signal. Adding SD to ID+CD gives smaller improvements in mean performance, supporting a complementary role.

### 5.2.2 Sensitivity Analysis

We further evaluate the sensitivity of MULTA on HAROOD to the main hyperparameters controlling transport alignment and alignment-gap reweighting, including the transport weight $\lambda_{\text{tr}}$ and the alignment-gap reweighting strength $\lambda_h$ As shown in Figure 2, performance remains stable across a broad range of settings on both EMG and PAMAP, indicating that the method does not rely on a fragile configuration. The results suggest that moderate transport alignment and alignment-gap reweighting consistently outperform either disabling or overemphasizing these components. Additional dataset- and task-level analyses in App. M, Tables 25 and 26. show the same overall trend, supporting the robustness of MULTA across hyperparameter choices.

## 6 Discussion

In this work, we introduced MULTA, a framework for time-series domain generalization based on structured, multi-level transport alignment. Rather than enforcing only local pairwise similarity constraints, MULTA models how alignment should be globally organized across classes and domains. By combining global transport matching with adaptive reweighting, the method encourages representations that align semantically related samples while remaining robust to domain-specific distortions and shortcuts.Our analysis connects this alignment objective to class-conditional source variation and an upper bound on unseen-domain risk, while evaluation across 30 tasks provides broad empirical support. Together, these findings underscore the value of multi-level, globally coordinated alignment for generalization to unseen domains.

**Limitations and future work.** Despite strong empirical performance, MULTA may introduce additional computational overhead due to optimal transport estimation. The method also assumes that meaningful cross-domain structure can be captured through transport geometry, which may become more challenging in highly heterogeneous or weakly related domains. Future work could explore scalable approximations for large-scale settings, extensions to multimodal and adaptive transport priors that evolve during training. More broadly, integrating transport-based objectives with pretrained sequence foundation models may provide a promising direction for improving robustness and transfer in real-world biosignal applications.

### AI USE STATEMENT

We used generative AI tools to polish the writing and identify related work. We also used them to review mathematical proofs and identify potentially missing assumptions, and to write and debug baseline reproduction scripts, cluster job scripts, and scripts for aggregating results across seeds and computing statistical tests. AI tools generated the small illustrative icons in panel A of Figure 1. We did not use generative AI to design MULTA or to draft original paragraphs or sections from scratch. Generative AI was not used to create the latent-space visualizations, target- or learned-plan heatmaps, or line plots from scratch but to read the results file and write matplotlib node for visualizaiton. The authors reviewed all AI-assisted work and take responsibility for the final text, proofs, code, figures, and reported results.

### REPRODUCIBILITY STATEMENT

The method and training objective are described in Sec. 3. The assumptions and proofs for the theoretical results are provided in App. A. Dataset protocols and splits are documented in App. G, and all datasets are public. Training, source-only model selection, hyperparameters, and compute resources are described in App. H. Evaluation and aggregation procedures are detailed in App. I. And we also upload the anonymous code as supplementary material.

## REFERENCES

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Diego Alvarez-Estevez and RM Rijsman. Haaglanden medisch centrum sleep staging database (version 1.0. 1). *PhysioNet*, 2021.

Aristotelis Ballas and Christos Diou. Towards domain generalization for ecg and eeg classification: Algorithms and benchmarks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):44–54, 2024. doi: 10.1109/TETCI.2023.3306253.

Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11):1649–1667, 2014.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Richard B Berry. The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. version 2.1. *Darien Illinois: American Academy of Sleep Medicine*, 2014.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54 (4):1–40, 2021.

Zihao Chen, Chi-Heng Lin, Ran Liu, Jingyun Xiao, and Eva L Dyer. Your contrastive learning problem is secretly a distribution alignment problem. *Advances in Neural Information Processing Systems*, 37:91597–91617, 2024.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Saransh Dave, Ritam Basu, and Vineet Gandhi. Cross-domain class-contrastive learning: finding lower dimensional representations for improved domain generalization. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–8, 2022.

Aveen Dayal, Vimal KB, Linga Reddy Cenkeramaddi, C Mohan, Abhinav Kumar, and Vineeth N Balasubramanian. Madg: margin-based adversarial learning for domain generalization. *Advances in Neural Information Processing Systems*, 36:58938–58952, 2023.

Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Houman Ghaemmaghami, Sridha Sridharan, and Clinton Fookes. Domain generalization in biosignal classification. *IEEE Transactions on Biomedical Engineering*, 68(6):1978–1989, 2020.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7522–7529, 2023.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Guillaume Dumas, and Irina Rish. Woods: Benchmarks for out-of-distribution generalization in time series tasks. arxiv 2022. *arXiv preprint arXiv:2203.09978*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

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.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Theo Gnassounou, Antoine Collas, Rémi Flamary, and Alexandre Gramfort. Psdnorm: Temporal normalization for deep learning in sleep staging. In *The Fourteenth International Conference on Learning Representations*.

Theo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution monge mapping normalization for learning on sleep data. *Advances in Neural Information Processing Systems*, 36:10457–10476, 2023.

Théo Gnassounou, Antoine Collas, Rémi Flamary, Karim Lounici, and Alexandre Gramfort. Multi-source and test-time domain adaptation on multivariate signals using spatio-temporal monge alignment. *arXiv preprint arXiv:2407.14303*, 2024.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 392–401, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European conference on computer vision*, pp. 124–140. Springer, 2020.

Ce Ju and Cuntai Guan. Deep optimal transport for domain adaptation on spd manifolds. *Artificial Intelligence*, 345:104347, 2025.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Kiran Krishnamachari, See-Kiong Ng, and Chuan-Sheng Foo. Uniformly distributed feature representations for fair and robust learning. *Transactions on Machine Learning Research*, 2024.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Kai Kunze and Paul Lukowicz. Sensor placement variations in wearable activity recognition. *IEEE Pervasive Computing*, 13(4):32–41, 2014.

Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

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.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35:9881–9893, 2022.

Sergey Lobov, Nadia Krilova, Innokentiy Kastalskiy, Victor Kazantsev, and Valeri A Makarov. Latent factors limiting the performance of semg-interfaces. *Sensors*, 18(4):1122, 2018.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Wang Lu, Jindong Wang, and Yiqiang Chen. Local and global alignments for generalizable sensor-based human activity recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3833–3837. IEEE, 2022a.

Wang Lu, Jindong Wang, Yiqiang Chen, Sinno Jialin Pan, Chunyu Hu, and Xin Qin. Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–19, 2022b.

Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. *arXiv preprint arXiv:2209.07027*, 2022c.

Wang Lu, Yao Zhu, and Jindong Wang. Harood: A benchmark for out-of-distribution generalization in sensor-based human activity recognition. *arXiv preprint arXiv:2512.10807*, 2025.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pp. 7313–7324. PMLR, 2021.

Payal Mohapatra, Lixu Wang, and Qi Zhu. Phase-driven generalizable representation learning for nonstationary time series classification. *Transactions on Machine Learning Research*, 2025.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

François Painblanc, Laetitia Chapel, Nicolas Courty, Chloé Friguet, Charlotte Pelletier, and Romain Tavenard. Match-and-deform: Time series domain adaptation through optimal transport and temporal alignment. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 341–356. Springer, 2023.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological measurement*, 43(4):04TR01, 2022.

Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11921–11929, 2021.

Xin Qin, Jindong Wang, Shuo Ma, Wang Lu, Yongchun Zhu, Xing Xie, and Yiqiang Chen. Generalizable low-resource activity recognition with diverse and discriminative representation learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1943–1953, 2023.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pp. 108–109. IEEE, 2012.

Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171:754–767, 2016.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning*, pp. 31408–31421. PMLR, 2023.

Yuge Shi, Jeffrey Seely, Philip HS Torr, Narayanaswamy Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve. *Machine Learning*, 112(7):2685–2721, 2023.

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

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan Plummer. Erm++: An improved baseline for domain generalization. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8514–8524, 2025.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

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.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024a.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yiming Wang, Bin Zhang, and Yujiao Tang. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 628–636, 2024b.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*, 2022.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2020.

Di Xiong, Shuoyuan Wang, Lei Zhang, Wenbo Huang, and Chaolei Han. Generalizable sensor-based activity recognition via categorical concept invariant learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 923–931, 2025.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7097–7107, 2022.

Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947–956, 2009.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Mi Zhang and Alexander A Sawchuk. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 1036–1043, 2012.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

# A DEFINITIONS, NORMALIZATION, AND ASSUMPTIONS

Let $B$ be the batch size, $u = B^{-1}\mathbf{1}$, and $\mathcal{Y}$ the shared label set, with $C = |\mathcal{Y}| \geq 2$. Throughout, logarithms are natural and $0 \log 0 = 0$. Sample $i$ is the anchor for row $i$, and $j$ indexes its candidate partners.

## A.1 TARGET PLAN

For a labeled source batch, the pair indicators are

$$I_{ij} = \mathbf{1}\{i = j\}, \tag{10}$$

$$M^{+}_{ij} = \mathbf{1}\{i \neq j,\ y_i = y_j,\ d_i \neq d_j\}, \tag{11}$$

$$M^{\mathrm{sd}}_{ij} = \mathbf{1}\{i \neq j,\ y_i = y_j,\ d_i = d_j\}, \tag{12}$$

$$M^{-}_{ij} = \mathbf{1}\{y_i \neq y_j\}. \tag{13}$$

As in Sec. 3.4, $M^{+}$ selects same-class cross-domain partners and $M^{-}$ selects different-class partners. The indicators $I$ and $M^{\mathrm{sd}}$ select self-pairs and same-class partners from the same domain. Let $n^{+}_i = \sum_j M^{+}_{ij}$ and $n^{\mathrm{sd}}_i = \sum_j M^{\mathrm{sd}}_{ij}$ count anchor $i$'s eligible partners for the two same-class relations. For each relation type, we distribute its construction weight uniformly over these partners. When no eligible partners are available, that relation contributes zero. Together with the identity weight, the resulting target score is

$$G_{ij} = \alpha_{\mathrm{id}} I_{ij} + \alpha_{\mathrm{cd}} \frac{M^{+}_{ij}}{\max(n^{+}_i, 1)} + \alpha_{\mathrm{sd}} \frac{M^{\mathrm{sd}}_{ij}}{\max(n^{\mathrm{sd}}_i, 1)}. \tag{14}$$

Here $\alpha_{\mathrm{id}} > 0$ and $\alpha_{\mathrm{cd}}, \alpha_{\mathrm{sd}} \geq 0$. We obtain the target by balancing these scores to the prescribed row and column marginals.

$$T = \mathrm{Sinkhorn}(G; u, u). \tag{15}$$

Each relation contributes its construction weight to a row of $G$ whenever eligible partners are present, independently of their number. The coefficients therefore specify relation weights before balancing, rather than fixed mass fractions in the final target rows. The resulting $T$ depends only on source labels and domain identities, has row and column marginals $u$, and preserves zero different-class entries. It is held fixed with respect to the encoder. Finite-iteration conventions are given in App. P and F.

Set $\Omega_G = \{(i, j) \mid G_{ij} > 0\}$ and

$$\mathcal{U}(u, u) = \{A \geq 0 \mid A\mathbf{1} = u,\ A^{\top}\mathbf{1} = u\}. \tag{16}$$

An exact doubly stochastic scaling with unchanged support exists when $G$ has total support, meaning that every positive entry belongs to a positive permutation diagonal. Equation (14) satisfies this condition because its support is symmetric and includes the diagonal. Any positive entry $(i, j)$ extends to such a permutation by swapping $i, j$ and using self entries elsewhere.

The exact target is the unique matrix

$$T = \operatorname*{arg\,min}_{\substack{A \in \mathcal{U}(u,u) \\ A_{ij} = 0 \text{ if } (i,j) \notin \Omega_G}} \sum_{(i,j) \in \Omega_G} A_{ij} \log \frac{A_{ij}}{G_{ij}}. \tag{17}$$

The target may be sparse, whereas the learned kernel $K$ and plan $P_\theta$ are positive. The proofs use the balanced marginals of $T$ and do not require a symmetry assumption. Disconnected support components can make the scaling factors nonunique, but the scaled matrix remains unique.

## A.2 LEARNED PLAN AND MATCHING LOSSES

For the same normalized features on both axes and temperature $\tau > 0$, define

$$K_{ij} = e^{z_i^{\top} z_j / \tau}, \qquad P_\theta = \operatorname*{arg\,min}_{A \in \mathcal{U}(u,u)} \sum_{ij} A_{ij} \log \frac{A_{ij}}{K_{ij}}. \tag{18}$$

The plans $T, P_\theta$ have total mass one, while $t = BT$ and $q = BP_\theta$ have unit row and column sums. Their rows $t_i, q_i$ are the target and learned matching distributions for anchor $i$. Self-pairs are retained. Finite iterations and the subsequent row normalization are treated in App. F.

For comparison, row normalization alone gives the plan and loss

$$Q_{ij} = u_i \frac{K_{ij}}{\sum_k K_{ik}}, \qquad L_{\mathrm{row}} = -\sum_{ij} T_{ij} \log(BQ_{ij}). \tag{19}$$

The unweighted loss in Eq. (4) uses the row distributions. Its relation to cross-entropy between the full plans is

$$L_{\mathrm{uw}} = -\sum_{ij} T_{ij} \log(BP_{\theta,ij}), \qquad -\sum_{ij} T_{ij} \log P_{\theta,ij} = L_{\mathrm{uw}} + \log B. \tag{20}$$

Both $L_{\mathrm{row}}$ and $L_{\mathrm{uw}}$ average anchors equally. The loss $L_w$ weights anchors when matching $P_\theta$ to $T$. The weights satisfy $1 \le w_i \le 1 + \lambda_h$, where $\lambda_h \ge 0$ controls the maximum additional weight assigned to an anchor.

Let $H(t_i) = -\sum_j t_{ij} \log t_{ij}$. Then

$$L_{\mathrm{uw}} = \frac{1}{B} \sum_i H(t_i) + \mathrm{KL}(T \| P_\theta). \tag{21}$$

The source-variation bound uses this full matching loss, including target entropy. Replacing it by KL alone is not generally valid.

### A.3 Features and batch sampling

The classifier uses features $\mu(x)$ and predicts $\widehat{y} = \arg\max_k (w_k^\top \mu(x) + b_k)$. Alignment uses $z(x) = \mu(x)/\|\mu(x)\|_2$. Writing $r(x) = \|\mu(x)\|_2$ gives $\mu(x) = r(x)z(x)$. We assume $0 < r(x) \le R$ almost surely in each source distribution and exact feature normalization. Target features need only have finite first moments.

The encoder is a fixed samplewise map, with fixed batch-normalization statistics when applicable. For contextual encoders, $x$ denotes the full input context to which the sampling assumption below applies.

Let $\mathcal{D}_y$ be the nonempty set of source domains containing class $y$. Every target class occurs in at least one source. We take expectations over fresh batches drawn from the source distributions. Let $\mathcal{M} = ((y_i, d_i))_{i=1}^B$ denote all batch labels and domain identities. Conditional on $\mathcal{M}$, each sample $X_i$ retains its class-domain distribution, and $T$ and its stopping rule depend only on $\mathcal{M}$. Independent draws within class-domain groups satisfy this condition. More generally, each cross-domain pair only needs to define a coupling of the corresponding source distributions. Temporally dependent windows are covered when these conditional marginals remain valid.

Every batch contains valid labeled source samples, and every anchor enters the loss. In MULTA, the score used to assign weights lies in $[0, 1]$ and the weights are detached. If weighting strength varies during training, $\lambda_h$ is its bound at the stage being analyzed.

### A.4 Source variation and class-balanced risks

Let

$$\nu_{d,y} = \mathcal{L}(z(X) \mid d, y), \qquad \eta_{d,y} = \mathcal{L}(\mu(X) \mid d, y). \tag{22}$$

All Wasserstein distances use Euclidean ground metric. For ordered distinct-domain pairs, define

$$\lambda_{y,d,d'} = \mathbb{E}\left[ \sum_{\substack{i: y_i = y, d_i = d \\ j: y_j = y, d_j = d'}} T_{ij} \right]. \tag{23}$$

These expected block masses include the $1/B$ plan normalization. Define the source discrepancies and feature-norm term by

$$\begin{aligned} S_z &= \sum_{y,d\neq d'} \lambda_{y,d,d'} W_2^2(\nu_{d,y}, \nu_{d',y}), \\ S_\mu &= \sum_{y,d\neq d'} \lambda_{y,d,d'} W_2^2(\eta_{d,y}, \eta_{d',y}), \\ A_r &= \mathbb{E} \sum_{(i,j)\in\Omega_{\mathrm{cd}}} T_{ij}(r_i - r_j)^2, \qquad \Omega_{\mathrm{cd}} = \{(i,j) : y_i = y_j,\ d_i \neq d_j\}. \end{aligned} \tag{24}$$

Here $S_z$ compares feature directions, $S_\mu$ compares classifier features, and $A_r$ accounts for norm differences within the pairs weighted by $T$. Write $D_{\mu,y} = \sum_{d\neq d'} \lambda_{y,d,d'} W_2^2(\eta_{d,y}, \eta_{d',y})$, so $S_\mu = \sum_y D_{\mu,y}$, with $D_{\mu,y} = 0$ for classes present in only one source. For $\mathcal{Y}_2 = \{y : |\mathcal{D}_y| \geq 2\}$, assume

$$\lambda_{y,\min} = \min_{\substack{d,d'\in\mathcal{D}_y \\ d\neq d'}} \lambda_{y,d,d'} > 0. \tag{25}$$

This sufficient condition requires positive expected target mass between every pair of source domains containing the same class.

For target class distribution $\eta_{t,y}$, define the set of source mixtures and the target's distance to it by

$$\mathcal{H}_y = \left\{ \sum_{d\in\mathcal{D}_y} \alpha_d \eta_{d,y} : \alpha \in \Delta(\mathcal{D}_y) \right\}, \qquad \rho_{t,y} = \inf_{\eta\in\mathcal{H}_y} W_1(\eta_{t,y}, \eta). \tag{26}$$

Here $\Delta(\mathcal{D}_y)$ is the simplex of nonnegative weights summing to one. Fix a reference source mixture $\eta_{s,\gamma,y} = \sum_d \gamma_{y,d}\eta_{d,y}$ with $\gamma_y \in \Delta(\mathcal{D}_y)$. For a loss $\ell$, the risks below average its expectation within each class equally across classes.

$$R_t^{\mathrm{bal}} = \frac{1}{C}\sum_y \mathbb{E}_{\eta_{t,y}}\ell(\mu, y), \qquad R_{s,\gamma}^{\mathrm{bal}} = \frac{1}{C}\sum_y \mathbb{E}_{\eta_{s,\gamma,y}}\ell(\mu, y). \tag{27}$$

Every class in $\mathcal{Y}$ has positive target probability. The corresponding balanced accuracy is

$$\mathrm{BA}_t = C^{-1}\sum_y \Pr(\widehat{Y} = Y \mid Y = y).$$

# B Source alignment and matching loss

We bound the full target-weighted feature cost by the matching loss, then restrict to same-class cross-domain pairs and average over source batches. The target determines which source relations enter this bound.

## B.1 A transport cost identity

Set $C_{ij}^z = \|z_i - z_j\|_2^2$ and $K_{ij}^0 = \exp(-C_{ij}^z/(2\tau))$. Unit normalization gives $K = e^{1/\tau}K^0$. The global factor is absorbed by diagonal scaling. For any positive scaling $q = \mathrm{diag}(a)K^0\,\mathrm{diag}(b)$, set $h_i = \log a_i - \log b_i$. Since $K_{ii}^0 = 1$,

$$C_{ij}^z = -2\tau\log q_{ij} + \tau\log q_{ii} + \tau\log q_{jj} + \tau(h_i - h_j). \tag{28}$$

For this calculation, write $L_{\mathrm{uw}} = -\sum_{ij} T_{ij}\log q_{ij}$. This is the loss in Eq. (4) when $q = BP_\theta$. Summing against a target $T$ with row marginal $u$ and column marginal $c_T$ gives

$$\langle T, C^z\rangle = 2\tau L_{\mathrm{uw}} + \tau\sum_i (u_i + c_{T,i})\log q_{ii} + \tau\langle u - c_T, h\rangle. \tag{29}$$

If $q$ is row stochastic then $q_{ii} \leq 1$. For a balanced target, $c_T = u$, the last term cancels, giving

$$\langle T, C^z\rangle \leq 2\tau L_{\mathrm{uw}}. \tag{30}$$

For a balanced target, positive finite diagonal scaling followed by row normalization suffices. The learned column marginal need not have converged. At exact convergence, symmetry of $q$ also gives $q_{ij} = \sqrt{q_{ii}q_{jj}}e^{-C^z_{ij}/(2\tau)}$.

We apply this cancellation to the full target, since an isolated cross-domain block may have unequal marginals. We can then restrict the nonnegative cost to the cross-domain pairs of interest.

### B.2 Proof of Proposition 1

Conditional on $\mathcal{M}$, a pair $(z_i, z_j)$ in block $(y, d, d')$ has marginals $\nu_{d,y}$ and $\nu_{d',y}$. It is therefore a coupling of these distributions, so $\mathbb{E}[C^z_{ij} \mid \mathcal{M}] \geq W_2^2(\nu_{d,y}, \nu_{d',y})$. Since $T$ is fixed by $\mathcal{M}$, multiplying by $T_{ij}$, summing within the block, and averaging over batches gives

$$\mathbb{E} \sum_{\substack{i:y_i=y,d_i=d \\ j:y_j=y,d_j=d'}} T_{ij}C^z_{ij} \geq \lambda_{y,d,d'} W_2^2(\nu_{d,y}, \nu_{d',y}). \tag{31}$$

Summing over the ordered distinct-domain blocks and using nonnegative remaining costs proves

$$S_z \leq \mathbb{E} \sum_{(i,j)\in\Omega_{\mathrm{cd}}} T_{ij}C^z_{ij} \leq \mathbb{E}\langle T, C^z\rangle \leq 2\tau\mathbb{E}L_{\mathrm{uw}}. \tag{32}$$

Since $q_i$ is a probability distribution, $\ell_i \geq 0$ and the weight bounds give

$$L_w = \frac{\sum_i w_i\ell_i}{\sum_i w_i} \geq \frac{\sum_i \ell_i}{B(1+\lambda_h)} = \frac{L_{\mathrm{uw}}}{1+\lambda_h}. \tag{33}$$

This proves Proposition 1. The weight inequality requires only bounded weights and nonnegative losses. It does not depend on how the weights are computed.

### B.3 Extension to classifier features

The same coupling argument for classifier features gives $S_\mu \leq \mathbb{E}\sum_{\Omega_{\mathrm{cd}}} T_{ij}\|\mu_i - \mu_j\|^2$.

For $\mu_i = r_i z_i$ and unit $z_i$, an expansion shows

$$\|\mu_i - \mu_j\|^2 = (r_i - r_j)^2 + r_i r_j\|z_i - z_j\|^2 \leq (r_i - r_j)^2 + R^2 C^z_{ij}. \tag{34}$$

Consequently

$$S_\mu \leq A_r + R^2\mathbb{E}\sum_{\Omega_{\mathrm{cd}}} T_{ij}C^z_{ij} \leq A_r + 2\tau R^2\mathbb{E}L_{\mathrm{uw}}. \tag{35}$$

Combining this with the weight inequality gives the source-variation term used in Theorem 1. The term $A_r$ remains explicit because normalized alignment does not control feature norms. No independence between norm and direction is assumed.

### B.4 Comparison with the row-normalized source bound

For $q^{\mathrm{row}}_{ij} = K^0_{ij}/\sum_k K^0_{ik}$, self entries give $\sum_k K^0_{ik} \geq 1$, and hence $C^z_{ij} \leq -2\tau\log q^{\mathrm{row}}_{ij}$. Thus both $L_{\mathrm{uw}}$ and $L_{\mathrm{row}}$ give valid upper bounds on source variation. For $S_z$, these are $2\tau\mathbb{E}L_{\mathrm{uw}}$ and $2\tau\mathbb{E}L_{\mathrm{row}}$. Proposition 2 gives their difference as $2\tau\mathbb{E}\,\mathrm{KL}(P_\theta\|Q)$. For classifier features, the corresponding bounds on $S_\mu$ are

$$\mathcal{B}_{\mathrm{global}} = A_r + 2\tau R^2\mathbb{E}L_{\mathrm{uw}}, \qquad \mathcal{B}_{\mathrm{row}} = A_r + 2\tau R^2\mathbb{E}L_{\mathrm{row}}, \tag{36}$$

and Proposition 2 gives

$$\mathcal{B}_{\mathrm{row}} - \mathcal{B}_{\mathrm{global}} = 2\tau R^2\mathbb{E}\,\mathrm{KL}(P_\theta\|Q) \geq 0. \tag{37}$$

Global balancing gives an unweighted source-variation bound no larger than the bound from row normalization. The bound is strictly tighter when $Q^\top\mathbf{1} \neq u$ with positive probability over source batches. The features and hence $S_z, S_\mu$ are fixed in this comparison. It does not establish an ordering between separately trained models.

### B.5 Source feature geometry and classification

For class/domain distributions $\nu, \nu'$ with means $m, m'$, every coupling satisfies $\mathbb{E}\|Z - Z'\|^2 \geq \|\mathbb{E}Z - \mathbb{E}Z'\|^2$ by Jensen's inequality. Taking the infimum gives

$$\|m - m'\|^2 \leq W_2^2(\nu, \nu'). \tag{38}$$

The class-conditional Wasserstein distances therefore bound source class centroid separation.

For independent class/domain features $Z, Z'$ with means $m, m'$ and covariances $\Sigma, \Sigma'$,

$$\mathbb{E}\|Z - Z'\|^2 = \|m - m'\|^2 + \operatorname{tr}\Sigma + \operatorname{tr}\Sigma'. \tag{39}$$

Centering and independence eliminate cross-covariance. Same-class cross-domain distances reflect both centroid shifts and variation within each class. Reducing them does not selectively remove domain variation. Similarly, independent feature norms may differ even when identically distributed, so $A_r$ can remain positive and is not controlled by normalized alignment.

Source CE alone leaves this geometry undetermined. Let $y, d \in \{-1, +1\}$, $a, b, M > 0$, and $a^2 + b^2 = 1$. Under logits $\phi_k(z) = Mke_1^\top z$, the normalized representations

$$z_{\mathrm{inv}}(y, d) = (ay, b), \qquad z_{\mathrm{var}}(y, d) = (ay, bd) \tag{40}$$

have identical CE $\log(1 + e^{-2Ma})$, but same-class cross-domain Wasserstein distances $0$ and $2b$, respectively. This motivates the explicit cross-domain relations in $T$.

## C Global transport matching

We compare plans obtained from the same affinity matrix and evaluated against the same target $T$. Row normalization gives $Q$. Balancing both rows and columns gives $P_\theta$.

### C.1 Proof of Proposition 2

Since $P_\theta, Q$ are positive diagonal scalings of $K$, write $\log(P_{\theta,ij}/Q_{ij}) = a_i + b_j$. The averages of these scaling terms under $T$ and $P_\theta$ agree because both plans have marginals $u$.

$$L_{\mathrm{row}} - L_{\mathrm{uw}} = \sum_{ij} T_{ij} \log \frac{P_{\theta,ij}}{Q_{ij}} = \langle u, a\rangle + \langle u, b\rangle \tag{41}$$

$$= \sum_{ij} P_{\theta,ij} \log \frac{P_{\theta,ij}}{Q_{ij}} = \mathrm{KL}(P_\theta \| Q). \tag{42}$$

Zeros in $T$ cause no difficulty because $P_\theta, Q > 0$. Applying the log-sum inequality within each column gives

$$\mathrm{KL}(P_\theta \| Q) \geq \sum_j u_j \log \frac{u_j}{\sum_i Q_{ij}} = \mathrm{KL}(u \| Q^\top \mathbf{1}). \tag{43}$$

For probability plans, KL is nonnegative and vanishes exactly when $P_\theta = Q$. By uniqueness of the scaling with prescribed marginals, this holds if and only if $Q$ already has column marginal $u$.

**Scope.** The identity follows the KL projection structure of entropic transport and its use in contrastive relation matching Benamou et al. (2015); Chen et al. (2024). It compares unweighted losses at fixed features and target. Anchor reweighting generally changes the target marginals in the loss, so the identity need not hold for $L_w$.

## D Unseen-domain risk

Following source-mixture analyses of DG (Albuquerque et al., 2019, Lemma 1 and Theorem 1) and related source-mixture bounds (Dayal et al., 2023), we first compare each target class with a source mixture, then compare that mixture with the fixed source reference. These two steps give the target discrepancy and source-variation terms, respectively. We bound the latter using Eq. (35).

### D.1 Full statement and constants

Let $C = |\mathcal{Y}| \geq 2$ and let the classifier logits be $w_k^\top \mu + b_k$, with $\max_k \|w_k\|_2 \leq \Lambda$. Under the distribution, sampling, and norm assumptions in App. A, set

$$\Delta_{\text{unseen}} = \frac{2\Lambda}{C} \sum_y \rho_{t,y}, \qquad \kappa_{\text{cov}} = \frac{2\Lambda}{C} \left( \sum_{y \in \mathcal{Y}_2} \lambda_{y,\min}^{-1} \right)^{1/2}. \tag{44}$$

Smaller classwise minimum masses $\lambda_{y,\min}$ increase $\kappa_{\text{cov}}$. With $\rho_{t,y}$ from Eq. (26), Theorem 1 becomes

$$\begin{aligned} R_t^{\text{bal,CE}} \leq R_{s,\gamma}^{\text{bal,CE}} &+ \frac{2\Lambda}{C} \sum_y \rho_{t,y} \\ &+ \frac{2\Lambda}{C} \left( \sum_{y \in \mathcal{Y}_2} \lambda_{y,\min}^{-1} \right)^{1/2} \sqrt{A_r + 2\tau R^2 (1 + \lambda_h) \mathbb{E} L_w}. \end{aligned} \tag{45}$$

A class present in only one source contributes to the classification and target-distance terms, but not to cross-domain source variation. We also obtain $1 - \text{BA}_t \leq R_t^{\text{bal,CE}} / \log 2$ below.

### D.2 Source-mixture proof of Theorem 1

For distributions $\eta, \eta'$ with finite first moments, any $L$-Lipschitz function $f$, and any coupling $(X, X')$ of these distributions, $|\mathbb{E}_\eta f - \mathbb{E}_{\eta'} f| \leq L \mathbb{E} \|X - X'\|$. Taking the coupling infimum gives

$$|\mathbb{E}_\eta f - \mathbb{E}_{\eta'} f| \leq L W_1(\eta, \eta'). \tag{46}$$

This standard $W_1$ inequality also follows from Kantorovich–Rubinstein duality (Gabriel & Marco, 2019, Section 6.1), with related use in adaptation bounds (Redko et al., 2017). Here it applies classwise, using the Euclidean Lipschitz constant of CE derived below.

Fix $y \in \mathcal{Y}_2$ and mixtures $\eta_\alpha = \sum_d \alpha_d \eta_{d,y}$, $\eta_\gamma = \sum_d \gamma_d \eta_{d,y}$. Mix $W_1$-optimal couplings between distinct source domains and identity couplings within each domain with weights $\alpha_d \gamma_{d'}$. This couples the mixtures. Using $W_1 \leq W_2$ gives

$$W_1(\eta_\alpha, \eta_\gamma) \leq \sum_{d \neq d'} \alpha_d \gamma_{d'} W_2(\eta_{d,y}, \eta_{d',y}) \leq \sqrt{D_{\mu,y} / \lambda_{y,\min}}. \tag{47}$$

The last inequality uses $\lambda_{y,\min} W_2^2(\eta_{d,y}, \eta_{d',y}) \leq D_{\mu,y}$ and $\sum_{d \neq d'} \alpha_d \gamma_{d'} \leq 1$. For a class present in only one source domain, the two mixtures coincide.

Assume that $\mu \mapsto \ell(\mu, y)$ is $L$-Lipschitz for every class, with a common constant $L$. For each class, choose a source mixture within $\epsilon$ of the infimum in Eq. (26). Apply Eq. (46) first from target to mixture, then from mixture to the fixed reference. Averaging over classes and letting $\epsilon \downarrow 0$ gives

$$R_t^{\text{bal}} \leq R_{s,\gamma}^{\text{bal}} + \frac{L}{C} \sum_y \rho_{t,y} + \frac{L}{C} \sum_{y \in \mathcal{Y}_2} \sqrt{D_{\mu,y} / \lambda_{y,\min}}. \tag{48}$$

Cauchy–Schwarz bounds the last sum by $\sqrt{\sum_{y \in \mathcal{Y}_2} \lambda_{y,\min}^{-1}} \sqrt{S_\mu}$. Equation (35) and $L_{\text{uw}} \leq (1 + \lambda_h) L_w$ bound the second factor. The argument uses shared class semantics and class-conditional mixtures, avoiding reliance on unconditional invariance under label shift Zhao et al. (2019).

### D.3 Cross-entropy risk and balanced accuracy

For logits $w_k^\top \mu + b_k$, the CE loss is

$$\ell_{\text{CE}}(\mu, y) = \log \sum_k e^{w_k^\top \mu + b_k} - w_y^\top \mu - b_y. \tag{49}$$

Its gradient is $\sum_k p_k(\mu) w_k - w_y$, whose norm is at most $2\Lambda$. Therefore Eq. (48) applies with $L = 2\Lambda$ to the classifier on $\mu$, proving Eq. (8). Finite biases do not affect this Lipschitz constant.

An incorrect argmax prediction implies another class has probability at least $p_y$, hence $p_y \le 1/2$ and

$$\mathbf{1}\{\widehat{y} \neq y\} \le \frac{-\log p_y}{\log 2}. \tag{50}$$

Averaging within each target class and then across classes gives the balanced-accuracy bound for any fixed tie-breaking convention.

**Source-risk convention.** The source risk in the theorem averages equally over classes. To relate it to pooled source CE, let $\pi$ be a probability vector over source domains and $p_d(y) = \Pr_d(Y = y)$. Define $R_s^{\mathrm{CE}} = \sum_d \pi_d \mathbb{E}_d \ell_{\mathrm{CE}}(\mu(X), Y)$, $p_s(y) = \sum_d \pi_d p_d(y)$, and choose $\gamma_{y,d} = \pi_d p_d(y)/p_s(y)$. For this reference mixture, nonnegative CE and $p_{s,\min} = \min_y p_s(y) > 0$ imply

$$R_{s,\gamma}^{\mathrm{bal,CE}} \le R_s^{\mathrm{CE}}/(C p_{s,\min}). \tag{51}$$

Keeping the balanced source term explicit avoids this potentially large class-imbalance factor.

**Scope.** The theorem concerns expected risks at a fixed encoder and classifier, with matching losses averaged over fresh source batches. Applied to resampling a fixed dataset, the argument instead concerns its empirical class-domain distributions. A population guarantee from training averages requires a separate estimation argument. A smaller matching loss tightens the bound when source risk, norm variation, target discrepancy, and the coefficients are fixed. These terms may change together during training, as also discussed in prior DG analysis (Sicilia et al., 2023, Section 3.3).

# E ALIGNMENT-GAP WEIGHTING

The alignment-gap score measures deviations from the relations specified by $T$. The result below relates this score to the KL term in each anchor's matching loss. It provides a reason to give larger-gap anchors more weight without assuming that reweighting always improves training.

## E.1 PROOF OF PROPOSITION 3

Let $t_i, q_i$ be row distributions with $q_{ij} > 0$ whenever $t_{ij} > 0$, including finite plans when both final rows sum to one and before log clipping. We use the pair indicators defined in App. A.1. For completeness, the gap scores from Eq. (1) are

$$\Delta_i^+ = [\langle t_i - q_i, M_i^+ \rangle]_+, \qquad \Delta_i^- = [\langle q_i - t_i, M_i^- \rangle]_+, \qquad \zeta_i = \frac{\beta_+ \Delta_i^+ + \beta_- \Delta_i^-}{\beta_+ + \beta_-}. \tag{52}$$

Here $[a]_+ = \max(a, 0)$, $\beta_\pm \ge 0$, and $\beta_+ + \beta_- > 0$. The first gap compares target and learned totals on same-class cross-domain partners. The second measures learned mass on different-class partners, since Eq. (15) gives these pairs zero target mass.

Writing $D_i = \mathrm{KL}(t_i \| q_i)$, we prove

$$D_i \ge 2(\Delta_i^+)^2, \qquad D_i \ge 2(\Delta_i^-)^2, \qquad D_i \ge 2\zeta_i^2, \tag{53}$$

together with $\ell_i = H(t_i) + D_i$ for unclipped conditional CE.

*Proof.* For an index set $A$, write $t_A = \sum_{j \in A} t_{ij}$ and $q_A = \sum_{j \in A} q_{ij}$. The log-sum inequality applied separately inside and outside $A$ yields

$$D_i \ge t_A \log(t_A/q_A) + (1 - t_A) \log((1 - t_A)/(1 - q_A)) \ge 2(t_A - q_A)^2. \tag{54}$$

For the second inequality, subtract $2(p - q)^2$ from Bernoulli KL. The difference has second derivative $1/p + 1/(1 - p) - 4 \ge 0$ in $p$, with value and first derivative zero at $p = q$ for $0 < q < 1$. Endpoint cases follow by limits, with infinite KL when the required support condition fails. Applying the bound to the candidate groups selected by $M_i^+$ and $M_i^-$ and taking positive parts gives the first two inequalities. The third follows because $\zeta_i$ is a convex combination of nonnegative gaps and is at most their maximum. Adding and subtracting $\sum_j t_{ij} \log t_{ij}$ gives the entropy decomposition. □

**From alignment gaps to anchor weights.** Equation (2) rescales the gap scores in each batch. Its floor $\delta > 0$ keeps the denominator positive when the scores coincide. The computed $w_i = 1 + \lambda_h \hat{h}_i$ is detached during differentiation. Every anchor retains its base weight of one. When all scores coincide, all weights equal one. The KL bound applies to $\zeta_i$ before this normalization. Both gaps and the matching loss use the same row distributions $t_i$ and $q_i$.

Since $H(t_i)$ is fixed with respect to the encoder, the KL term is the part of the loss due to matching errors. Larger gaps give larger lower bounds on that term, motivating additional weight on these anchors. The score compares group totals, so it need not rank anchors by their full KL divergence or guarantee an optimization improvement. Log clipping modifies the entropy identity as described in App. F.

# F Finite Sinkhorn iterations and numerical corrections

The main results assume exact balancing. Here we examine the column-marginal errors left by finite iterations and the effects of clipping inside the logarithm. Hats denote quantities computed from the finite plans.

## F.1 Finite plans and row normalization

Let $A_T, A_P$ denote the target and learned matrices from finite Sinkhorn iterations after scaling by $B$, before final row normalization. Their intended row and column sums are one, giving an intended total mass of $B$. The floor $\delta$ in Eq. (2) applies only to batch gap normalization. The separate floor $\varepsilon$ below applies to row denominators or log arguments. Denominator regularization in Sinkhorn preserves diagonal scaling but need not meet the marginals. The loss uses $\widehat{t} = \mathrm{RowNorm}(A_T)$ and $\widehat{q} = \mathrm{RowNorm}(A_P)$. Set $\widehat{T} = \widehat{t}/B$, $\widehat{P}_\theta = \widehat{q}/B$. Here $\mathrm{RowNorm}$ divides each nonzero row by its sum and preserves zero rows.

We initially assume that the final row normalization divides by the actual row sums, so $\widehat{t}$ and $\widehat{q}$ have unit row sums. If the denominator is instead $\max(s_i, \varepsilon)$ for row sum $s_i$, this requires $s_i \geq \varepsilon$. A target accepted with maximum row or column error $\epsilon_T < 1 - \varepsilon$ meets this condition, because $s_i \geq 1 - \epsilon_T > \varepsilon$. The learned rows must be checked separately.

For the feature-cost bound only, the learned denominator floor may be active. In that case, replace $\widehat{q}_{ij}$ by $(A_P)_{ij}/\max(\sum_k (A_P)_{ik}, \varepsilon)$ throughout that bound. This remains a positive diagonal scaling with row sums at most one, so $\widehat{q}_{ii} \leq 1$. The target must still have unit row sums. The probability and projection results require unit sums for both matrices.

For the unclipped learned rows there are positive factors $a, b$ with $\widehat{q} = \mathrm{diag}(a) K^0 \mathrm{diag}(b)$. Let

$$h_i = \log a_i - \log b_i, \quad c_T = \widehat{T}^\top \mathbf{1}, \quad c_P = \widehat{P}_\theta^\top \mathbf{1}. \tag{55}$$

Since $h_i - h_j = \log \widehat{q}_{ij} - \log \widehat{q}_{ji}$, its range is computable from the unclipped matrix. The identity remains valid under uniform logit shifts.

If the code evaluates $\log \max(\widehat{q}_{ij}, \varepsilon)$, define

$$q_{ij}^{\log} = \max(\widehat{q}_{ij}, \varepsilon), \quad \widehat{L}_{\mathrm{uw}} = -\sum_{ij} \widehat{T}_{ij} \log q_{ij}^{\log}, \quad \mathcal{C}_{\mathrm{clip}} = \sum_{ij} \widehat{T}_{ij} \log \frac{q_{ij}^{\log}}{\widehat{q}_{ij}} \geq 0. \tag{56}$$

For $0 < \varepsilon \leq 1$, clipped row losses are nonnegative. Clipped entries are only logarithm arguments and need not sum to one. The unclipped loss is $\widehat{L}_{\mathrm{uw}} + \mathcal{C}_{\mathrm{clip}}$. When both rows are probability distributions, the rowwise identity is $\widehat{\ell}_i^{\mathrm{clip}} = H(\widehat{t}_i) + \mathrm{KL}(\widehat{t}_i \| \widehat{q}_i) - c_{\mathrm{clip},i}$, where $c_{\mathrm{clip},i} = \sum_j \widehat{t}_{ij} \log(q_{ij}^{\log}/\widehat{q}_{ij})$ and $\mathcal{C}_{\mathrm{clip}} = B^{-1} \sum_i c_{\mathrm{clip},i}$.

## F.2 Bounding feature distances with finite iterations

Applying Eq. (29) to the actual finite target and unclipped scaled learned rows gives

$$\langle \widehat{T}, C^z \rangle \leq 2\tau \widehat{L}_{\mathrm{uw}} + 2\tau \mathcal{C}_{\mathrm{clip}} + \tau \langle u - c_T, h \rangle. \tag{57}$$

Because $u - c_T$ has zero sum,

$$|\langle u - c_T, h\rangle| \le \tfrac{1}{2}\|u - c_T\|_1 \operatorname{osc}(h), \quad \operatorname{osc}(h) = \max_i h_i - \min_i h_i. \tag{58}$$

Define the nonnegative, computable correction

$$\mathcal{E}_{\mathrm{num}} = 2\tau\mathcal{C}_{\mathrm{clip}} + \frac{\tau}{2}\|u - c_T\|_1 \operatorname{osc}(h). \tag{59}$$

This correction vanishes for a balanced target without log clipping, even before the learned plan converges. Learned column convergence is required only for the exact projection comparison, not this cost bound.

Define $\widehat{\lambda}$, $\widehat{S}_z$, $\widehat{S}_\mu$, and $\widehat{A}_r$ as in Eqs. (24) and (23), but using the actual metadata-only finite target $\widehat{T}$. The preceding population proof now gives

$$\widehat{S}_z \le 2\tau(1 + \lambda_h)\mathbb{E}\widehat{L}_w + \mathbb{E}\mathcal{E}_{\mathrm{num}}, \tag{60}$$

$$\widehat{S}_\mu \le \widehat{A}_r + R^2\left[2\tau(1 + \lambda_h)\mathbb{E}\widehat{L}_w + \mathbb{E}\mathcal{E}_{\mathrm{num}}\right]. \tag{61}$$

Here $\widehat{L}_w$ is the weighted average of the clipped row losses, with the same weight bounds as in Eq. (4). Provided the expected finite-target masses also satisfy $\widehat{\lambda}_{y,\min} > 0$ for $y \in \mathcal{Y}_2$, the proof of Theorem 1 applies with these masses in $\kappa_{\mathrm{cov}}$ and the right-hand side of the $\widehat{S}_\mu$ bound under the square root. All masses and norm differences in this extension use the actual metadata-only target $\widehat{T}$. Substituting exact-target quantities would require an additional approximation bound.

### F.3 Finite projection comparison and measured marginal errors

Here both finite plans have row marginal $u$. Let $Q$ be the exact row-normalized plan. For any positive plan $A$, define

$$L_A(\widehat{T}) = -\sum_{ij} \widehat{T}_{ij} \log(BA_{ij}), \qquad L_{\mathrm{row}}(\widehat{T}) = L_Q(\widehat{T}).$$

Positive diagonal scaling of $K$ gives $\log(\widehat{P}_{\theta,ij}/Q_{ij}) = \xi_i + \psi_j$. Because $\widehat{T}, \widehat{P}_\theta, Q$ have the same row marginal, the unclipped losses satisfy

$$L_{\mathrm{row}}(\widehat{T}) - L_{\widehat{P}_\theta}(\widehat{T}) = \mathrm{KL}(\widehat{P}_\theta \| Q) + \langle c_T - c_P, \psi\rangle. \tag{62}$$

Indeed, the loss difference is $\langle u, \xi\rangle + \langle c_T, \psi\rangle$, while $\mathrm{KL}(\widehat{P}_\theta\|Q) = \langle u, \xi\rangle + \langle c_P, \psi\rangle$. The remaining term satisfies

$$|\langle c_T - c_P, \psi\rangle| \le \tfrac{1}{2}\|c_T - c_P\|_1 \operatorname{osc}(\psi) \le \tfrac{1}{2}(\|c_T - u\|_1 + \|c_P - u\|_1)\operatorname{osc}(\psi). \tag{63}$$

The correction terms do not depend on the nonuniqueness of scaling factors for a fixed plan. If only the learned loss is clipped, its value decreases by $\mathcal{C}_{\mathrm{clip}}$, so the right-hand side of Eq. (62) gains $\mathcal{C}_{\mathrm{clip}}$.

For a finite matrix $A$ with intended unit row and column sums, define the mean absolute errors before normalization

$$e_r(A) = B^{-1}\sum_i \left|\sum_j A_{ij} - 1\right|, \qquad e_c(A) = B^{-1}\sum_j \left|\sum_i A_{ij} - 1\right|. \tag{64}$$

Assume every row sum is positive and the final normalization uses that sum. Let $H = \mathrm{RowNorm}(A)$ and $c = H^\top \mathbf{1}/B$. Then

$$\|c - u\|_1 \le e_r(A) + e_c(A). \tag{65}$$

To prove this, let $s_i = \sum_j A_{ij} > 0$. By the triangle inequality,

$$\|H^\top\mathbf{1} - \mathbf{1}\|_1 \le \|H^\top\mathbf{1} - A^\top\mathbf{1}\|_1 + \|A^\top\mathbf{1} - \mathbf{1}\|_1 \tag{66}$$

$$\le \sum_{ij} A_{ij}|s_i^{-1} - 1| + \sum_j \left|\sum_i A_{ij} - 1\right| \tag{67}$$

$$= \sum_i |1 - s_i| + \sum_j \left|\sum_i A_{ij} - 1\right|. \tag{68}$$

Dividing by $B$ bounds the column error after row normalization. A small projection remainder also requires control of the logarithmic scaling range in Eq. (63).

**Numerical conditions.** The learned kernel must remain numerically positive and symmetric, using the same features for rows and columns. Denominator regularization preserves diagonal scaling. Kernel underflow, entrywise modifications before scaling, or different features on the two axes require separate analysis. Log clipping preserves the unclipped identity and changes the observed loss by Eq. (56). Marginal errors, clipping corrections, and normalization floors determine which bounds apply.

# G Datasets and Established Evaluation Protocols

To evaluate out-of-distribution generalization under two complementary and literature-established distribution shifts, we adopt (i) the cross-dataset sleep-staging protocol introduced in SleepDG Wang et al. (2024a) and (ii) the cross-person human activity recognition protocol defined in HAROOD Lu et al. (2025). The SleepDG setting measures robustness to inter-dataset heterogeneity caused by differences in subject populations, channel configurations, acquisition hardware, and clinical environments. The HAROOD setting measures robustness to person-specific variability in wearable sensing data. Both are benchmark-style evaluation protocols rather than ad hoc data splits, and therefore provide a rigorous basis for assessing generalization.

**Table 4:** SleepDG cross-dataset benchmark used in this work. Each dataset is treated as one domain; evaluation follows a five-fold leave-one-dataset-out protocol, where one dataset is held out as the unseen target domain and the remaining four are used as source domains.

| Dataset | Orig. freq. (Hz) | Recordings (all) | Recordings (used) | EEG channel | EOG channel | Classes | Input dim. |
|---|---|---|---|---|---|---|---|
| SleepEDFx Kemp et al. (2000); Goldberger et al. (2000) | 100 | 197 | 197 | Fpz–Cz | horizontal | 5 | $2 \times 3000$ |
| HMC Alvarez-Estevez & Rijsman (2021) | 256 | 151 | 151 | F4–M1 | E1–M2 | 5 | $2 \times 3000$ |
| ISRUC Khalighi et al. (2016) | 200 | 126 | 126 | F4–M1 | E1–M2 | 5 | $2 \times 3000$ |
| SHHS Quan et al. (1997); Zhang et al. (2018) | 125 | 5,793 | 150 | C4–M1 | ROC–LOC | 5 | $2 \times 3000$ |
| P2018 Ghassemi et al. (2018); Goldberger et al. (2000) | 200 | 995 | 150 | C3–M2 | E1–M2 | 5 | $2 \times 3000$ |

*Notes.* The five sleep stages are $\{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$. For SleepEDFx and SHHS, the original R&K annotations are harmonized by merging N3 and N4 into a single N3 stage. SleepEDFx is additionally trimmed to the interval from 30 minutes before the first non-wake epoch to 30 minutes after the last non-wake epoch. Following the benchmark preprocessing, one EEG channel and one EOG channel are retained for each dataset. In our implementation, recordings are first band-pass filtered to 0.3–35 Hz and then resampled to 100 Hz (i.e., *filter first, resample second, this is very important in avoiding issues with aliasing and overlooked in SleepDG.*); therefore each 30-second epoch contains $2 \times 3000$ samples. During training, SleepDG further uses sequences of length $L = 20$ epochs. We cite the original dataset or repository papers in the Dataset column and follow the original access terms, licenses, or data-use agreements for each dataset.

**Table 5:** HAROOD cross-person benchmark used in this work. The table combines the dataset statistics and the published cross-person subject splits into a single paper-ready summary.

| Dataset | Classes | Domains | Subjects | Sampling rate | Approx. instances |
|---|---|---|---|---|---|
| DSADS Barshan & Yüksek (2014) | 19 | 4 | 8 | 25 Hz | 1,140,000 |
| USC-HAD Zhang & Sawchuk (2012) | 12 | 4 | 14 | 100 Hz | 5,441,000 |
| UCI-HAR Reyes-Ortiz et al. (2016) | 6 | 5 | 30 | 50 Hz | 1,310,000 |
| PAMAP2 Reiss & Stricker (2012) | 12 | 4 | 9 | 100 Hz | 3,850,505 |
| EMG Lobov et al. (2018) | 6 | 4 | 36 | 1000 Hz | 33,903,472 |
| WESAD Schmidt et al. (2018) | 4 | 4 | 15 | chest: 700 Hz; wrist: 32–700 Hz | 63,000,000 |

| Dataset | Subject splits (domains) | Input dim. | Channels / sensors used |
|---|---|---|---|
| DSADS | [[0,1],[2,3], [4,5],[6,7]] | $(45, 1, 125)$ | 45 channels: 5 body positions $\times$ {acc, gyro, mag} $\times$ 3 axes |
| USC-HAD | [[1,11,2,0],[6,3,9,5], [7,13,8,10],[4,12]] | $(6, 1, 200)$ | 6 channels: hip accelerometer + gyroscope (3 axes each) |
| UCI-HAR | [[0–5],[6–11],[12–17], [18–23],[24,29]] | $(6, 1, 128)$ | 6 channels: smartphone accelerometer + gyroscope (3 axes each) |
| PAMAP2 | [[3,2,8],[1,5], [0,7],[4,6]] | $(27, 1, 200)$ | 27 channels: 3 IMUs (wrist/chest/ankle) $\times$ {acc, gyro, mag} $\times$ 3 axes |
| EMG | [[0–8],[9,17], [18,26],[27,35]] | $(8, 1, 200)$ | 8-channel sEMG |
| WESAD | [[0,1,2,3],[4,5,6,7], [8,9,10,11],[12,13,14]] | $(8, 1, 200)$ | 8 chest-worn channels: ECG, EDA, EMG, respiration, body temperature, and 3-axis accel. |

*Dataset credit and licensing.* We cite the original dataset or repository papers in the Dataset column. We use these datasets through the corresponding benchmark preprocessing and evaluation protocols, do not redistribute raw data, and follow the original access terms, licenses, or data-use agreements for each dataset.

# H MODEL, TRAINING, MODEL SELECTION, HYPERPARAMETER SWEEPS, AND COMPUTE RESOURCES

Within each benchmark, the methods use the same evaluation protocol and benchmark encoder family. MULTA changes the training objective; checkpoint and hyperparameter selection use source-domain validation only. Method-specific hyperparameters apply only to the relevant method.

**Model setup.** The shared interface represents inputs as $(B, L, C, T)$, where $B$ is batch size, $L$ is the number of epochs or windows per example, $C$ is the channel count, and $T$ is the number of samples per epoch or window. For SleepDG, $L = 20$, $C = 2$ (one EEG and one EOG channel), and $T = 3000$ in the base setting: each item contains 20 consecutive 30-second sleep epochs. The benchmark CNN encodes each epoch, a one-layer Transformer with eight attention heads and 512-dimensional hidden features incorporates cross-epoch context, and a linear head predicts a label per epoch. For HAROOD, $L = 1$ and each item is a preprocessed activity window. The adapter reshapes it to the benchmark input layout $(B, C, 1, T)$ and uses the official HAROOD CNN, which has two convolution–batch-normalization–ReLU–max-pooling blocks (pool size $(1, 2)$ and stride 2), followed by the activity classifier. We follow the benchmark preprocessing and leave-one-domain-out splits in App. G.

**Batch construction.** SleepDG draws mixed-domain training batches from pooled source datasets. HAROOD uses one infinite training loader per source environment and 100 update steps per epoch; the HAROOD batch sizes in Table 6 are per source environment. Validation and test loaders are evaluated separately.

**Training and model selection.** For each zero-shot task, the held-out domain is used only for final testing. In HAROOD, each source environment is partitioned into 80% training and 20% validation data, and source validation splits are pooled for model selection. Hyperparameters, checkpoints, and early-stopping decisions use source-domain validation only. Unless otherwise specified by a benchmark baseline, models use AdamW with weight decay $10^{-4}$ and no learning-rate scheduler. Checkpoints are selected by the source-validation loss domain macro validation loss criterion. Early stopping is used to prevent possible overfitting for better generlizaton.

**Hyperparameter grid.** Table 6 records the MULTA candidates used for source-validation model selection. The candidates are the same across seeds and target tasks within each benchmark, with benchmark-specific batch sizes as described above. Baseline-specific parameters are selected under the same source-only validation rule. Ablation- and sensitivity-only settings, such as no-alignment controls, high alignment weights, and alternative retained-row fractions, are outside the final model-selection grid.

**Table 6:** MULTA hyperparameter candidates for source-validation model selection. The held-out target domain is not used to choose any entry.

| Hyperparameter | Candidate values |
|---|---|
| Optimizer | AdamW |
| Learning rate | $\{10^{-2}, 10^{-3}, 5 \times 10^{-4}\}$ |
| Weight decay | $10^{-4}$ |
| Batch size, SleepDG | 160 per domain per batch, 80 per domain per batch, 40 per domain per batch |
| Batch size, HAROOD | $\{32, 64, 128\}$ per source environment |
| MULTA alignment weight $\lambda_{\mathrm{tr}}$ | $\{0.10, 0.15, 0.25\}$ |
| Feature-similarity temperature $\tau$ | $\{0.05, 0.07, 0.10\}$ |
| Diagonal/self-consistency weight | $\{0.00, 0.05, 0.10\}$ |
| Gap-weight scale $\lambda_h$ | $\{0.50, 1.00, 1.50\}$ |
| Gap-score weights $(\beta_+, \beta_-)$ | $\{(0.50, 0.25), (1.00, 0.50), (2.00, 0.50), (2.00, 1.00)\}$ |
| Sinkhorn iterations | 10 |
| Checkpoint selection | Best source-validation checkpoint |
| Early stopping | Source-validation early stopping |

**Compute resources.** Experiments were run as independent single-GPU jobs on a shared high-performance computing cluster and workstation GPUs. The compute pool included NVIDIA B200 and A40 GPUs, and additional runs were executed on NVIDIA GeForce RTX 5090 GPUs when available. B200 jobs used a 45 GB GPU-memory partition rather than the full physical GPU memory; A40 and RTX 5090 jobs used 48 GB and 32 GB GPUs, respectively. SleepDG runs typically required 1–2 hours per held-out dataset and seed. HAROOD runs typically required 20–40 minutes per dataset and seed under the HAROOD protocol.

# I Evaluation Aggregation and Uncertainty Reporting

Our HAROOD evaluation has three nested levels: dataset, target task, and random seed. Let $\mathcal{D}$ denote the set of datasets. For each dataset $d \in \mathcal{D}$, let $\mathcal{T}_d$ be the set of cross-person target tasks in that dataset, and let $\mathcal{R} = \{0, 42, 96, 123, 3407\}$ be the set of random seeds. For method $m$, dataset $d$, task $t$, and seed $r$, we denote the test metric by

$$y_{m,d,t,r}. \tag{69}$$

In the main HAROOD tables, $y$ is balanced accuracy. The same aggregation rules are used for other metrics in the appendix, with balanced accuracy and accuracy reported in percent and Cohen's $\kappa$ reported on its original scale.

## I.1 Traditional Mean–Standard-Deviation Reporting

The traditional performance table reports the mean and sample standard deviation across random seeds after aggregating tasks within each seed. For a fixed method $m$ and dataset $d$, we first compute a seed-level dataset score by averaging over all target tasks in that dataset:

$$a_{m,d,r} = \frac{1}{|\mathcal{T}_d|} \sum_{t \in \mathcal{T}_d} y_{m,d,t,r}. \tag{70}$$

The dataset-level mean is then

$$\mu_{m,d} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} a_{m,d,r}, \tag{71}$$

and the reported standard deviation is the sample standard deviation over the seed-level dataset scores:

$$s_{m,d} = \sqrt{\frac{1}{|\mathcal{R}| - 1} \sum_{r \in \mathcal{R}} \left(a_{m,d,r} - \mu_{m,d}\right)^2}. \tag{72}$$

Thus, each dataset entry in the traditional table is reported as

$$\mu_{m,d} \pm s_{m,d}. \tag{73}$$

This standard deviation measures how much the dataset-level aggregate changes when the random seed is changed. It is not the standard deviation across target tasks.

For the overall Avg column, we first average over all dataset–task targets within each seed:

$$a^{\text{all}}_{m,r} = \frac{1}{\sum_{d \in \mathcal{D}} |\mathcal{T}_d|} \sum_{d \in \mathcal{D}} \sum_{t \in \mathcal{T}_d} y_{m,d,t,r}. \tag{74}$$

We then report the mean and sample standard deviation of $\{a^{\text{all}}_{m,r} : r \in \mathcal{R}\}$ across seeds. This follows the same convention as the per-dataset columns, but uses all dataset–task targets before computing the seed-level mean and standard deviation.

For per-task appendix tables, the dataset and task are fixed, so the only aggregation dimension is the random seed. For each method, dataset, and task, we report

$$\mu_{m,d,t} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} y_{m,d,t,r}, \qquad s_{m,d,t} = \sqrt{\frac{1}{|\mathcal{R}| - 1} \sum_{r \in \mathcal{R}} \left(y_{m,d,t,r} - \mu_{m,d,t}\right)^2}. \tag{75}$$

Each per-task entry is written as $\mu_{m,d,t} \pm s_{m,d,t}$.

## I.2 Task-Wise Significance-Aware Ranking

In addition to reporting absolute performance, we compute a task-wise significance-aware ranking of methods. The comparison unit is a dataset-specific target task $(d, t)$; we do not aggregate across heterogeneous tasks before testing. For each task, we first compute the seed mean for every method,

$$\bar{y}_{m,d,t} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} y_{m,d,t,r}. \tag{76}$$

For any two methods $m_i$ and $m_j$, we compare them on the same task using paired seed-level differences,

$$\Delta_r^{(i,j,d,t)} = y_{m_i,d,t,r} - y_{m_j,d,t,r}, \qquad r \in \mathcal{R}. \tag{77}$$

We perform a two-sided paired $t$-test on these five paired differences at significance level $\alpha = 0.05$. If the difference is significant and the mean difference is positive, method $m_i$ is considered significantly better than method $m_j$ on task $(d, t)$; if the difference is significant and negative, $m_j$ is considered significantly better than $m_i$. If the paired test is not significant, the two methods are treated as tied for that task.

The task-level rank of method $m$ is computed by counting how many other methods are significantly better than it on that task:

$$\mathrm{rank}_{m,d,t} = 1 + \left|\{m' : m' \text{ is significantly better than } m \text{ on task } (d,t)\}\right|. \tag{78}$$

This rule assigns tied ranks to methods that are not statistically distinguishable on a task. A smaller rank is better. To summarize ranks within a dataset, we average task ranks:

$$\mathrm{AvgRank}_{m,d} = \frac{1}{|\mathcal{T}_d|} \sum_{t \in \mathcal{T}_d} \mathrm{rank}_{m,d,t}. \tag{79}$$

The overall average rank is dataset-balanced:

$$\mathrm{AvgRank}_m = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathrm{AvgRank}_{m,d}. \tag{80}$$

This prevents datasets with more target tasks from dominating the final ranking.

### I.3 Variance Summary in the Appendix

The ranking analysis separates method comparison from variability reporting. In the appendix, we therefore report three complementary quantities. Results can be seen at Table 15. First, for each method and dataset, we report the mean of task means,

$$\mu_{m,d}^{\mathrm{task}} = \frac{1}{|\mathcal{T}_d|} \sum_{t \in \mathcal{T}_d} \bar{y}_{m,d,t}. \tag{81}$$

Second, we report variability across target tasks,

$$s_{m,d}^{\mathrm{task}} = \sqrt{\frac{1}{|\mathcal{T}_d| - 1} \sum_{t \in \mathcal{T}_d} \left(\bar{y}_{m,d,t} - \mu_{m,d}^{\mathrm{task}}\right)^2}. \tag{82}$$

This quantity captures heterogeneity across cross-person target tasks and should not be interpreted as optimization uncertainty.

Third, we report the pooled within-task seed variance. For each task, the seed variance is

$$v_{m,d,t} = \frac{1}{|\mathcal{R}| - 1} \sum_{r \in \mathcal{R}} \left(y_{m,d,t,r} - \bar{y}_{m,d,t}\right)^2. \tag{83}$$

We pool these variances across tasks using their degrees of freedom:

$$\sigma_{\mathrm{seed},m,d}^2 = \frac{\sum_{t \in \mathcal{T}_d} (|\mathcal{R}| - 1) v_{m,d,t}}{\sum_{t \in \mathcal{T}_d} (|\mathcal{R}| - 1)}. \tag{84}$$

When every task uses the same number of seeds, this is equivalent to averaging the task-level seed variances. We also report $\sigma_{\mathrm{seed},m,d} = \sqrt{\sigma_{\mathrm{seed},m,d}^2}$ when a standard-deviation scale is easier to interpret. This seed variability quantifies variation due to random initialization and optimization within a fixed target task, whereas $s_{m,d}^{\mathrm{task}}$ quantifies variation across tasks.

# J Detailed SleepDG Cross-Dataset Results

This section provides the full cross-dataset sleep-staging result tables, including overall averages and leave-one-dataset-out target-specific results.

## J.1 Significance-aware ranks

**Table 7:** Significance-aware ranks for Sleep LODO with SleepDG included. For each held-out target dataset, paired two-sided t-tests are run across the same five seeds. A method's rank is 1 plus the number of methods significantly better than it on that target dataset; methods without significant differences can share the same rank. Lower is better.

| Method | P2018 | ISRUC | Sleep-EDFx | SHHS | HMC | Avg. Rank |
|---|---|---|---|---|---|---|
| ERM | 4 | 5 | 4 | 1 | 3 | 3.40 |
| ERM++ | 2 | 1 | 4 | 3 | 1 | 2.20 |
| Mixup | 3 | 5 | 4 | 1 | 6 | 3.80 |
| CORAL | 8 | 5 | 3 | 1 | 3 | 4.00 |
| MMD | 6 | 5 | 1 | 1 | 7 | 4.00 |
| DANN | 7 | 5 | 5 | 4 | 6 | 5.40 |
| MLDG | 19 | 19 | 19 | 19 | 19 | 19.00 |
| GroupDRO | 8 | 9 | 5 | 2 | 7 | 6.20 |
| RSC | 2 | 5 | 6 | 4 | 3 | 4.00 |
| ANDMask | 9 | 5 | 4 | 3 | 3 | 4.80 |
| VREx | 2 | 5 | 5 | 5 | 6 | 4.60 |
| URM | 2 | 9 | 6 | 7 | 14 | 7.60 |
| Fishr | 3 | 3 | 5 | 4 | 3 | 3.60 |
| Fish | 17 | 12 | 11 | 5 | 18 | 12.60 |
| LAG | 3 | 5 | 1 | 1 | 4 | 2.80 |
| DDLearn | 19 | 19 | 20 | 19 | 19 | 19.20 |
| SupCon | 2 | 1 | 3 | 1 | 3 | 2.00 |
| PCL | 2 | 5 | 4 | 1 | 3 | 3.00 |
| SleepDG | 2 | 1 | 1 | 2 | 2 | <u>1.60</u> |
| Ours | 1 | 1 | 1 | 1 | 1 | **1.00** |

## J.2 Per-target metrics

**Table 8:** Sleep leave-one-dataset-out selected-checkpoint performance with **P2018** as the held-out target dataset. Values are mean ± sample standard deviation across seeds $\{0, 42, 96, 123, 3407\}$. Bal Acc / Acc / F1 metrics are percentages; $\kappa$ is reported on its raw scale. Best is bold and second best is underlined for each metric column.

| Method | Bal. Acc. | Acc. | F1-Macro | F1-Weighted | $\kappa$ |
|---|---|---|---|---|---|
| ERM | 70.47 ± 1.76 | 71.56 ± 0.66 | 67.63 ± 1.07 | 70.52 ± 0.85 | 0.6192 ± 0.0113 |
| ERM++ | 72.32 ± 1.10 | 71.22 ± 1.40 | 68.26 ± 1.33 | 70.68 ± 1.33 | 0.6213 ± 0.0172 |
| Mixup | 71.24 ± 1.34 | 72.25 ± 1.83 | 67.90 ± 2.43 | 71.06 ± 2.14 | 0.6303 ± 0.0206 |
| CORAL | 69.82 ± 1.38 | 72.69 ± 1.09 | 67.17 ± 1.19 | 71.22 ± 0.98 | 0.6315 ± 0.0137 |
| MMD | 71.18 ± 0.93 | 72.99 ± 1.58 | 68.45 ± 1.57 | 71.82 ± 1.49 | 0.6372 ± 0.0174 |
| DANN | 70.73 ± 0.81 | 73.15 ± 1.10 | 68.10 ± 1.00 | 71.69 ± 1.15 | 0.6381 ± 0.0134 |
| MLDG | 20.89 ± 2.02 | 19.61 ± 4.92 | 8.60 ± 5.80 | 9.32 ± 8.87 | 0.0095 ± 0.0211 |
| GroupDRO | 68.74 ± 1.95 | 70.55 ± 1.82 | 65.76 ± 2.60 | 69.48 ± 1.97 | 0.6065 ± 0.0220 |
| RSC | 72.37 ± 1.53 | 73.44 ± 1.84 | 69.52 ± 2.18 | 72.45 ± 1.94 | 0.6452 ± 0.0222 |
| ANDMask | 69.50 ± 1.66 | 71.18 ± 1.88 | 66.61 ± 2.73 | 69.86 ± 2.15 | 0.6134 ± 0.0221 |
| VREx | 71.82 ± 1.83 | 72.12 ± 1.75 | 68.14 ± 2.42 | 71.05 ± 1.95 | 0.6302 ± 0.0209 |
| URM | 68.93 ± 3.93 | 70.81 ± 1.56 | 66.34 ± 3.19 | 69.68 ± 2.31 | 0.6075 ± 0.0276 |
| Fishr | 70.24 ± 2.39 | 71.35 ± 2.64 | 67.21 ± 3.67 | 70.16 ± 3.18 | 0.6170 ± 0.0315 |
| Fish | 64.45 ± 1.82 | 69.21 ± 0.97 | 62.08 ± 2.83 | 66.86 ± 2.05 | 0.5783 ± 0.0145 |
| LAG | 71.30 ± 1.73 | 73.68 ± 0.39 | 69.24 ± 1.60 | 72.66 ± 0.84 | 0.6448 ± 0.0065 |
| DDLearn | 20.97 ± 0.32 | 43.21 ± 0.23 | 13.94 ± 0.62 | 27.16 ± 0.57 | 0.0186 ± 0.0063 |
| SupCon | 73.21 ± 0.41 | 72.69 ± 1.03 | 69.65 ± 0.57 | 72.04 ± 0.43 | 0.6390 ± 0.0113 |
| PCL | <u>73.24 ± 1.16</u> | **74.79 ± 1.07** | <u>71.09 ± 1.18</u> | <u>73.92 ± 1.04</u> | **0.6609 ± 0.0130** |
| SleepDG | 72.96 ± 0.62 | 73.20 ± 0.79 | 70.28 ± 1.12 | 72.80 ± 0.91 | 0.6439 ± 0.0083 |
| Ours | **75.00 ± 0.75** | <u>74.05 ± 1.39</u> | **72.08 ± 1.18** | **74.03 ± 1.29** | <u>0.6579 ± 0.0157</u> |

**Table 9:** Sleep leave-one-dataset-out selected-checkpoint performance with **ISRUC** as the held-out target dataset. Values are mean ± sample standard deviation across seeds $\{0, 42, 96, 123, 3407\}$. Bal Acc / Acc / F1 metrics are percentages; $\kappa$ is reported on its raw scale. Best is bold and second best is underlined for each metric column.

| Method | Bal. Acc. | Acc. | F1-Macro | F1-Weighted | $\kappa$ |
|---|---|---|---|---|---|
| ERM | 69.50 ± 1.86 | 75.04 ± 1.21 | 70.59 ± 1.86 | 73.84 ± 1.53 | 0.6706 ± 0.0174 |
| ERM++ | **73.26 ± 0.66** | 76.60 ± 0.65 | 73.10 ± 0.92 | 75.86 ± 0.84 | 0.6957 ± 0.0084 |
| Mixup | 68.73 ± 1.82 | 74.81 ± 0.98 | 69.39 ± 2.38 | 73.06 ± 1.74 | 0.6666 ± 0.0145 |
| CORAL | 69.95 ± 1.08 | 75.82 ± 0.51 | 70.84 ± 1.24 | 74.25 ± 0.79 | 0.6803 ± 0.0075 |
| MMD | 69.45 ± 0.84 | 74.90 ± 0.98 | 70.16 ± 0.96 | 73.57 ± 0.89 | 0.6689 ± 0.0128 |
| DANN | 68.36 ± 2.26 | 74.76 ± 1.82 | 68.51 ± 2.33 | 72.49 ± 2.00 | 0.6652 ± 0.0256 |
| MLDG | 20.81 ± 1.25 | 25.29 ± 3.06 | 10.39 ± 3.96 | 13.23 ± 6.23 | 0.0149 ± 0.0231 |
| GroupDRO | 68.10 ± 1.55 | 73.64 ± 1.23 | 68.63 ± 1.59 | 72.15 ± 1.37 | 0.6525 ± 0.0166 |
| RSC | 69.32 ± 1.23 | 75.48 ± 0.61 | 69.39 ± 1.40 | 73.27 ± 1.03 | 0.6756 ± 0.0094 |
| ANDMask | 69.07 ± 1.07 | 74.89 ± 0.85 | 69.76 ± 1.33 | 73.27 ± 1.09 | 0.6681 ± 0.0115 |
| VREx | 69.03 ± 2.02 | 74.88 ± 1.37 | 70.29 ± 1.61 | 73.63 ± 1.39 | 0.6677 ± 0.0198 |
| URM | 66.51 ± 3.40 | 73.39 ± 2.62 | 67.36 ± 3.31 | 71.41 ± 2.83 | 0.6466 ± 0.0357 |
| Fishr | 70.51 ± 1.84 | 75.99 ± 1.23 | 71.08 ± 1.92 | 74.48 ± 1.59 | 0.6835 ± 0.0173 |
| Fish | 65.28 ± 1.86 | 72.28 ± 1.19 | 64.97 ± 1.70 | 69.45 ± 1.37 | 0.6311 ± 0.0181 |
| LAG | 69.45 ± 1.17 | 74.67 ± 1.41 | 69.90 ± 1.54 | 73.23 ± 1.35 | 0.6666 ± 0.0174 |
| DDLearn | 20.61 ± 0.40 | 32.10 ± 0.45 | 10.85 ± 0.82 | 16.53 ± 0.97 | 0.0114 ± 0.0076 |
| SupCon | 72.37 ± 0.36 | 77.09 ± 0.48 | <u>73.23 ± 0.77</u> | <u>76.18 ± 0.59</u> | 0.6993 ± 0.0060 |
| PCL | 69.35 ± 1.92 | 75.29 ± 1.43 | 70.27 ± 1.51 | 73.72 ± 1.37 | 0.6730 ± 0.0202 |
| SleepDG | <u>73.10 ± 0.42</u> | **77.82 ± 0.25** | **74.27 ± 0.41** | **76.99 ± 0.33** | **0.7082 ± 0.0036** |
| Ours | 72.56 ± 0.63 | <u>77.26 ± 0.45</u> | 72.99 ± 0.68 | 76.03 ± 0.56 | <u>0.7020 ± 0.0057</u> |

**Table 10:** Sleep leave-one-dataset-out selected-checkpoint performance with **Sleep-EDFx** as the held-out target dataset. Values are mean ± sample standard deviation across seeds $\{0, 42, 96, 123, 3407\}$. Bal Acc / Acc / F1 metrics are percentages; $\kappa$ is reported on its raw scale. Best is bold and second best is underlined for each metric column.

| Method | Bal. Acc. | Acc. | F1-Macro | F1-Weighted | $\kappa$ |
|---|---|---|---|---|---|
| ERM | 68.92 ± 1.19 | 69.61 ± 2.46 | 64.02 ± 1.64 | 69.91 ± 1.93 | 0.5977 ± 0.0293 |
| ERM++ | 67.31 ± 2.04 | 64.94 ± 3.27 | 61.01 ± 2.56 | 65.95 ± 2.85 | 0.5459 ± 0.0379 |
| Mixup | 68.18 ± 2.32 | 68.82 ± 1.05 | 62.93 ± 1.63 | 69.18 ± 1.05 | 0.5886 ± 0.0153 |
| CORAL | 69.66 ± 1.35 | 71.67 ± 2.02 | 65.70 ± 1.91 | 71.67 ± 1.84 | 0.6207 ± 0.0243 |
| MMD | 70.67 ± 0.87 | 72.04 ± 1.63 | 65.93 ± 1.01 | 71.94 ± 1.32 | 0.6278 ± 0.0200 |
| DANN | 67.37 ± 2.70 | 68.87 ± 3.43 | 62.44 ± 2.71 | 68.71 ± 2.93 | 0.5849 ± 0.0441 |
| MLDG | 31.34 ± 6.55 | 30.32 ± 3.76 | 16.78 ± 5.34 | 20.31 ± 6.53 | 0.1119 ± 0.0638 |
| GroupDRO | 64.45 ± 4.27 | 65.70 ± 4.55 | 59.45 ± 5.23 | 65.47 ± 5.16 | 0.5416 ± 0.0613 |
| RSC | 66.83 ± 1.81 | 69.05 ± 2.51 | 62.00 ± 1.73 | 68.39 ± 1.96 | 0.5845 ± 0.0316 |
| ANDMask | 68.38 ± 1.20 | 69.82 ± 2.05 | 63.76 ± 0.98 | 69.71 ± 1.38 | 0.5980 ± 0.0239 |
| VREx | 67.14 ± 2.01 | 68.02 ± 2.17 | 62.68 ± 2.14 | 68.06 ± 2.14 | 0.5722 ± 0.0288 |
| URM | 65.63 ± 3.70 | 67.41 ± 3.33 | 60.66 ± 4.29 | 67.11 ± 3.66 | 0.5641 ± 0.0456 |
| Fishr | 66.37 ± 3.25 | 67.56 ± 4.49 | 61.88 ± 3.64 | 67.76 ± 3.83 | 0.5681 ± 0.0562 |
| Fish | 62.68 ± 2.73 | 64.80 ± 4.15 | 57.18 ± 3.43 | 64.35 ± 3.52 | 0.5305 ± 0.0496 |
| LAG | 69.43 ± 3.31 | 70.08 ± 4.23 | 64.20 ± 4.16 | 70.23 ± 4.16 | 0.6050 ± 0.0534 |
| DDLearn | 20.21 ± 0.34 | 37.94 ± 0.49 | 11.37 ± 0.69 | 21.22 ± 1.03 | 0.0053 ± 0.0085 |
| SupCon | 69.90 ± 0.37 | 72.08 ± 1.34 | 65.29 ± 1.48 | 71.51 ± 1.31 | 0.6266 ± 0.0146 |
| PCL | 69.13 ± 0.62 | 70.19 ± 1.05 | 64.19 ± 0.43 | 70.25 ± 0.71 | 0.6043 ± 0.0124 |
| SleepDG | <u>71.31 ± 0.44</u> | <u>72.73 ± 0.64</u> | <u>66.69 ± 0.57</u> | <u>72.62 ± 0.62</u> | <u>0.6371 ± 0.0084</u> |
| Ours | **72.16 ± 0.55** | **74.40 ± 0.81** | **68.15 ± 0.83** | **74.08 ± 0.73** | **0.6571 ± 0.0099** |

**Table 11:** Sleep leave-one-dataset-out selected-checkpoint performance with **SHHS** as the held-out target dataset. Values are mean ± sample standard deviation across seeds $\{0, 42, 96, 123, 3407\}$. Bal Acc / Acc / F1 metrics are percentages; $\kappa$ is reported on its raw scale. Best is bold and second best is underlined for each metric column.

| Method | Bal. Acc. | Acc. | F1-Macro | F1-Weighted | $\kappa$ |
|---|---|---|---|---|---|
| ERM | 65.77 ± 3.02 | 72.66 ± 3.42 | 63.23 ± 3.73 | 73.49 ± 3.18 | 0.6206 ± 0.0434 |
| ERM++ | 66.06 ± 1.42 | 71.03 ± 1.04 | 62.13 ± 1.43 | 72.11 ± 1.05 | 0.6029 ± 0.0144 |
| Mixup | 66.75 ± 2.96 | 72.82 ± 2.89 | 63.83 ± 3.11 | 74.14 ± 2.57 | 0.6242 ± 0.0370 |
| CORAL | 68.38 ± 1.64 | **76.84 ± 1.34** | **67.14 ± 1.08** | **77.20 ± 1.21** | **0.6753 ± 0.0164** |
| MMD | 68.07 ± 2.20 | 75.38 ± 1.08 | 65.89 ± 1.85 | 76.16 ± 1.27 | 0.6574 ± 0.0153 |
| DANN | 65.13 ± 1.54 | 74.12 ± 1.32 | 63.46 ± 1.57 | 74.56 ± 1.09 | 0.6381 ± 0.0154 |
| MLDG | 30.22 ± 9.46 | 44.18 ± 15.24 | 21.32 ± 11.62 | 32.07 ± 19.35 | 0.2062 ± 0.1948 |
| GroupDRO | 66.77 ± 0.37 | 74.16 ± 1.21 | 64.73 ± 0.79 | 75.28 ± 0.98 | 0.6411 ± 0.0137 |
| RSC | 65.33 ± 1.72 | 74.01 ± 2.86 | 63.77 ± 2.65 | 74.42 ± 2.57 | 0.6349 ± 0.0360 |
| ANDMask | 65.18 ± 2.21 | 74.04 ± 1.32 | 63.54 ± 1.97 | 74.38 ± 1.49 | 0.6346 ± 0.0200 |
| VREx | 65.14 ± 1.68 | 73.50 ± 1.98 | 63.25 ± 1.70 | 73.99 ± 1.68 | 0.6301 ± 0.0243 |
| URM | 63.10 ± 2.58 | 72.10 ± 1.75 | 61.64 ± 1.68 | 72.84 ± 1.59 | 0.6109 ± 0.0218 |
| Fishr | 66.15 ± 0.96 | 75.84 ± 0.38 | 65.86 ± 0.62 | 76.29 ± 0.35 | 0.6568 ± 0.0051 |
| Fish | 65.36 ± 0.75 | 74.46 ± 1.19 | 63.99 ± 0.79 | 74.95 ± 0.84 | 0.6432 ± 0.0149 |
| LAG | 67.19 ± 2.34 | 76.18 ± 2.67 | 66.12 ± 2.75 | 76.58 ± 2.29 | 0.6663 ± 0.0351 |
| DDLearn | 20.19 ± 0.34 | 41.08 ± 0.52 | 11.98 ± 0.70 | 24.24 ± 1.08 | 0.0054 ± 0.0096 |
| SupCon | 67.75 ± 1.50 | 73.96 ± 1.15 | 65.01 ± 1.26 | 75.13 ± 1.06 | 0.6385 ± 0.0156 |
| PCL | 66.70 ± 1.67 | 73.74 ± 0.99 | 64.35 ± 1.23 | 74.62 ± 1.13 | 0.6331 ± 0.0148 |
| SleepDG | 66.93 ± 1.21 | 73.04 ± 2.41 | 63.81 ± 2.54 | 73.91 ± 2.25 | 0.6247 ± 0.0284 |
| Ours | **68.63 ± 0.96** | 75.15 ± 1.46 | 66.18 ± 0.72 | 76.10 ± 1.11 | 0.6545 ± 0.0161 |

**Table 12:** Sleep leave-one-dataset-out selected-checkpoint performance with **HMC** as the held-out target dataset. Values are mean ± sample standard deviation across seeds $\{0, 42, 96, 123, 3407\}$. Bal Acc / Acc / F1 metrics are percentages; $\kappa$ is reported on its raw scale. Best is bold and second best is underlined for each metric column.

| Method | Bal. Acc. | Acc. | F1-Macro | F1-Weighted | $\kappa$ |
|---|---|---|---|---|---|
| ERM | 69.87 ± 1.82 | 74.55 ± 1.47 | 70.48 ± 2.16 | 73.85 ± 1.72 | 0.6605 ± 0.0198 |
| ERM++ | 72.64 ± 0.59 | **75.77 ± 0.52** | 72.75 ± 0.74 | 75.61 ± 0.65 | **0.6804 ± 0.0065** |
| Mixup | 68.90 ± 1.26 | 73.49 ± 0.83 | 69.47 ± 1.78 | 72.59 ± 1.18 | 0.6455 ± 0.0107 |
| CORAL | 68.71 ± 2.38 | 74.49 ± 1.60 | 69.50 ± 2.43 | 73.17 ± 2.14 | 0.6564 ± 0.0244 |
| MMD | 68.04 ± 1.26 | 73.79 ± 1.16 | 68.70 ± 0.88 | 72.32 ± 1.18 | 0.6469 ± 0.0173 |
| DANN | 69.33 ± 0.76 | 74.06 ± 0.57 | 70.10 ± 0.94 | 73.19 ± 0.73 | 0.6526 ± 0.0074 |
| MLDG | 23.55 ± 6.81 | 22.59 ± 7.82 | 10.73 ± 7.12 | 11.58 ± 9.05 | 0.0435 ± 0.0785 |
| GroupDRO | 67.25 ± 1.44 | 72.55 ± 1.56 | 67.79 ± 1.50 | 71.33 ± 1.62 | 0.6316 ± 0.0214 |
| RSC | 69.93 ± 1.89 | 74.49 ± 1.60 | 70.42 ± 2.06 | 73.65 ± 1.81 | 0.6596 ± 0.0214 |
| ANDMask | 69.13 ± 1.64 | 72.76 ± 0.83 | 69.70 ± 1.40 | 72.64 ± 1.26 | 0.6389 ± 0.0131 |
| VREx | 68.82 ± 1.23 | 73.89 ± 0.80 | 69.61 ± 1.19 | 72.98 ± 0.86 | 0.6502 ± 0.0100 |
| URM | 65.43 ± 2.03 | 70.94 ± 1.75 | 66.05 ± 1.98 | 69.62 ± 1.89 | 0.6090 ± 0.0242 |
| Fishr | 69.85 ± 1.76 | 73.68 ± 1.03 | 70.27 ± 1.74 | 73.22 ± 1.43 | 0.6504 ± 0.0150 |
| Fish | 62.95 ± 2.27 | 69.62 ± 1.95 | 63.52 ± 2.82 | 67.66 ± 2.43 | 0.5872 ± 0.0279 |
| LAG | 68.12 ± 2.38 | 72.96 ± 1.53 | 69.31 ± 1.91 | 72.40 ± 1.77 | 0.6374 ± 0.0235 |
| DDLearn | 21.52 ± 0.95 | 37.72 ± 0.63 | 13.64 ± 1.68 | 22.22 ± 1.49 | 0.0253 ± 0.0159 |
| SupCon | 71.09 ± 0.83 | 75.18 ± 0.94 | 71.91 ± 1.00 | 74.75 ± 0.99 | 0.6693 ± 0.0123 |
| PCL | 70.82 ± 0.56 | 74.83 ± 0.52 | 71.73 ± 0.50 | 74.48 ± 0.53 | 0.6644 ± 0.0068 |
| SleepDG | 70.96 ± 0.87 | 74.58 ± 0.91 | 71.64 ± 0.93 | 74.34 ± 1.00 | 0.6623 ± 0.0125 |
| Ours | **72.77 ± 0.87** | 75.59 ± 0.77 | **73.03 ± 0.75** | **75.64 ± 0.74** | 0.6782 ± 0.0103 |

# K Detailed HAROOD Per-Task Results

This section reports detailed per-task results for the HAROOD cross-person scenarios under the five-seed repetition protocol. In addition to balanced accuracy, the tables include accuracy and Cohen's $\kappa$ where available. Each table reports mean $\pm$ sample standard deviation over seeds $\{0, 42, 96, 123, 3407\}$.

## K.1 Significance-aware ranks

**Table 13:** Detailed significance-aware ranks by dataset-specific task. Rank is computed from paired seed-level tests across five seeds; lower is better.

| Method | T0 | T1 | T2 | T3 | T4 | Avg. |
|---|---|---|---|---|---|---|
| **DSADS** | | | | | | |
| ERM | 3 | 4 | 8 | 5 | N/A | 5.00 |
| ERM++ | 3 | 11 | 7 | 5 | N/A | 6.50 |
| Mixup | **1** | 3 | 2 | **1** | N/A | 1.75 |
| CORAL | 8 | 2 | 4 | 3 | N/A | 4.25 |
| MMD | 6 | 9 | 2 | 4 | N/A | 5.25 |
| DANN | 7 | 4 | 3 | 2 | N/A | 4.00 |
| MLDG | 18 | 18 | 18 | 18 | N/A | 18.00 |
| GroupDRO | 11 | 2 | 4 | 4 | N/A | 5.25 |
| RSC | 3 | 2 | 2 | 3 | N/A | 2.50 |
| ANDMask | 2 | 4 | 2 | 4 | N/A | 3.00 |
| VREx | 4 | 3 | 8 | 6 | N/A | 5.25 |
| URM | **1** | **1** | **1** | 9 | N/A | 3.00 |
| Fishr | 3 | 4 | 8 | 4 | N/A | 4.75 |
| Fish | 7 | 3 | 5 | **1** | N/A | 4.00 |
| LAG | 5 | **1** | 3 | 4 | N/A | 3.25 |
| DDLearn | 18 | 18 | 18 | 18 | N/A | 18.00 |
| SupCon | **1** | **1** | **1** | **1** | N/A | **1.00** |
| PCL | 4 | 8 | 4 | 4 | N/A | 5.00 |
| Ours | **1** | **1** | **1** | **1** | N/A | **1.00** |
| **USC** | | | | | | |
| ERM | 5 | 3 | 5 | 4 | N/A | 4.25 |
| ERM++ | 11 | 3 | 14 | 4 | N/A | 8.00 |
| Mixup | 3 | 3 | 3 | 2 | N/A | 2.75 |
| CORAL | 2 | 2 | 3 | **1** | N/A | 2.00 |
| MMD | 10 | 5 | 3 | 4 | N/A | 5.50 |
| DANN | 6 | 3 | 7 | 3 | N/A | 4.75 |
| MLDG | 19 | 19 | 19 | 19 | N/A | 19.00 |
| GroupDRO | 6 | 3 | 4 | 3 | N/A | 4.00 |
| RSC | **1** | **1** | **1** | 2 | N/A | 1.25 |
| ANDMask | 3 | 3 | 5 | 12 | N/A | 5.75 |
| VREx | 5 | 3 | 4 | 5 | N/A | 4.25 |
| URM | 2 | 3 | 2 | **1** | N/A | 2.00 |
| Fishr | 5 | 3 | 3 | 3 | N/A | 3.50 |
| Fish | 2 | 3 | 3 | **1** | N/A | 2.25 |
| LAG | 11 | 4 | 5 | 10 | N/A | 7.50 |
| DDLearn | 18 | 18 | 18 | 18 | N/A | 18.00 |
| SupCon | 4 | 3 | 2 | 3 | N/A | 3.00 |
| PCL | 5 | 5 | 5 | 5 | N/A | 5.00 |
| Ours | **1** | **1** | **1** | **1** | N/A | **1.00** |

*Continued on next page*

| Method | T0 | T1 | T2 | T3 | T4 | Avg. |
|---|---|---|---|---|---|---|
| **HAR** | | | | | | |
| ERM | 3 | 4 | 6 | 5 | 3 | 4.20 |
| ERM++ | 2 | 2 | 2 | 2 | 2 | 2.00 |
| Mixup | 2 | **1** | 2 | **1** | 2 | 1.60 |
| CORAL | 4 | 10 | 5 | 2 | 2 | 4.60 |
| MMD | 4 | 5 | 2 | 3 | **1** | 3.00 |
| DANN | 9 | **1** | 3 | 3 | 3 | 3.80 |
| MLDG | 19 | 19 | 19 | 19 | 19 | 19.00 |
| GroupDRO | 4 | 8 | 3 | 4 | 2 | 4.20 |
| RSC | 2 | 4 | 2 | 2 | 2 | 2.40 |
| ANDMask | 6 | 9 | 5 | 2 | 3 | 5.00 |
| VREx | 4 | 2 | 6 | 2 | 7 | 4.20 |
| URM | 11 | **1** | 5 | 3 | 4 | 4.80 |
| Fishr | 3 | 4 | 6 | 3 | 2 | 3.60 |
| Fish | 2 | 4 | 6 | 4 | 2 | 3.60 |
| LAG | 3 | 5 | 2 | 3 | **1** | 2.80 |
| DDLearn | 18 | 18 | 18 | 18 | 18 | 18.00 |
| SupCon | **1** | 2 | 2 | **1** | 2 | 1.60 |
| PCL | 2 | **1** | 3 | 3 | 3 | 2.40 |
| Ours | **1** | **1** | **1** | **1** | **1** | **1.00** |
| **EMG** | | | | | | |
| ERM | **1** | 2 | 7 | 4 | N/A | 3.50 |
| ERM++ | 2 | 2 | 2 | 2 | N/A | 2.00 |
| Mixup | 2 | 3 | 6 | 7 | N/A | 4.50 |
| CORAL | 2 | 2 | 2 | **1** | N/A | 1.75 |
| MMD | 2 | 2 | 2 | 2 | N/A | 2.00 |
| DANN | 6 | 6 | 11 | 6 | N/A | 7.25 |
| MLDG | 19 | 19 | 19 | 19 | N/A | 19.00 |
| GroupDRO | 2 | 2 | 2 | 2 | N/A | 2.00 |
| RSC | **1** | 2 | **1** | **1** | N/A | 1.25 |
| ANDMask | 2 | 2 | 2 | 6 | N/A | 3.00 |
| VREx | 2 | 2 | 3 | 4 | N/A | 2.75 |
| URM | 2 | **1** | 2 | 2 | N/A | 1.75 |
| Fishr | **1** | 5 | 7 | 4 | N/A | 4.25 |
| Fish | **1** | 2 | **1** | **1** | N/A | 1.25 |
| LAG | 2 | 10 | 7 | 11 | N/A | 7.50 |
| DDLearn | 17 | 16 | 18 | 17 | N/A | 17.00 |
| SupCon | 2 | 2 | 2 | 4 | N/A | 2.50 |
| PCL | 6 | 4 | 3 | 6 | N/A | 4.75 |
| Ours | **1** | **1** | **1** | **1** | N/A | **1.00** |
| **WESAD** | | | | | | |
| ERM | 3 | **1** | 2 | **1** | N/A | 1.75 |
| ERM++ | 3 | 3 | 3 | 4 | N/A | 3.25 |
| Mixup | 3 | 2 | 2 | 2 | N/A | 2.25 |
| CORAL | 2 | 4 | 5 | 8 | N/A | 4.75 |
| MMD | 2 | 2 | 4 | **1** | N/A | 2.25 |
| DANN | **1** | 4 | 9 | 7 | N/A | 5.25 |
| MLDG | 13 | 19 | 18 | 13 | N/A | 15.75 |
| GroupDRO | 4 | 4 | 3 | 13 | N/A | 6.00 |
| RSC | **1** | 3 | 4 | 4 | N/A | 3.00 |
| ANDMask | **1** | 4 | 3 | **1** | N/A | 2.25 |
| VREx | **1** | **1** | 3 | 2 | N/A | 1.75 |
| URM | 2 | **1** | 3 | 6 | N/A | 3.00 |
| Fishr | 3 | **1** | **1** | **1** | N/A | 1.50 |
| Fish | 5 | 3 | 15 | 12 | N/A | 8.75 |
| LAG | 3 | 3 | **1** | 3 | N/A | 2.50 |
| DDLearn | 2 | 12 | 8 | 6 | N/A | 7.00 |
| SupCon | **1** | **1** | 2 | **1** | N/A | 1.25 |
| PCL | 3 | 3 | 2 | **1** | N/A | 2.25 |
| Ours | **1** | **1** | **1** | **1** | N/A | **1.00** |

*Continued on next page*

| Method | T0 | T1 | T2 | T3 | T4 | Avg. |
|---|---|---|---|---|---|---|
| **PAMAP** | | | | | | |
| ERM | 3 | **1** | 3 | 6 | N/A | 3.25 |
| ERM++ | 4 | 5 | 3 | 10 | N/A | 5.50 |
| Mixup | **1** | **1** | **1** | 2 | N/A | 1.25 |
| CORAL | 4 | 3 | 4 | 3 | N/A | 3.50 |
| MMD | 11 | 3 | 4 | 3 | N/A | 5.25 |
| DANN | 3 | 6 | 3 | 2 | N/A | 3.50 |
| MLDG | 19 | 19 | 19 | 19 | N/A | 19.00 |
| GroupDRO | 4 | 5 | 2 | 3 | N/A | 3.50 |
| RSC | **1** | **1** | **1** | 2 | N/A | 1.25 |
| ANDMask | 8 | 4 | 2 | 5 | N/A | 4.75 |
| VREx | 3 | 3 | 4 | 3 | N/A | 3.25 |
| URM | 3 | **1** | 2 | **1** | N/A | 1.75 |
| Fishr | 3 | **1** | 3 | 5 | N/A | 3.00 |
| Fish | 3 | 2 | 2 | 2 | N/A | 2.25 |
| LAG | 3 | 12 | 5 | 14 | N/A | 8.50 |
| DDLearn | 18 | 18 | 18 | 18 | N/A | 18.00 |
| SupCon | 3 | **1** | 2 | 2 | N/A | 2.00 |
| PCL | 3 | 5 | 3 | 4 | N/A | 3.75 |
| Ours | **1** | **1** | **1** | **1** | N/A | **1.00** |

### K.2 VARIANCE DECOMPOSITION

**Table 14:** Overall variance decomposition on HAROOD. Mean is the mean of seed-averaged target-task means over all dataset-specific cross-person tasks. Std(task) measures variability across target tasks. Var(seed) and Std(seed) are the pooled within-task seed variance and its standard-deviation scale. Mean, Std(task), and Std(seed) are in percentage points; Var(seed) is in squared percentage points.

| **Method** | **Mean ↑** | **Std(task) ↓** | **Var(seed) ↓** | **Std(seed) ↓** |
|---|---|---|---|---|
| ERM | 71.35 | 18.36 | 12.17 | 3.49 |
| ERM++ | 71.03 | 18.67 | 4.20 | 2.05 |
| Mixup | 72.54 | 18.65 | 7.90 | 2.81 |
| CORAL | 71.15 | 20.14 | 8.42 | 2.90 |
| MMD | 70.87 | 19.61 | 14.24 | 3.77 |
| DANN | 68.47 | 20.22 | 24.57 | 4.96 |
| MLDG | 27.35 | 8.65 | 324.50 | 18.01 |
| GroupDRO | 70.65 | 20.56 | 5.56 | 2.36 |
| RSC | 72.85 | 19.63 | 8.42 | 2.90 |
| ANDMask | 71.09 | 18.98 | 16.87 | 4.11 |
| VREx | 71.71 | 18.24 | 8.78 | 2.96 |
| URM | 72.09 | 19.04 | 13.90 | 3.73 |
| Fishr | 71.61 | 17.94 | 13.28 | 3.64 |
| Fish | 69.37 | 21.34 | 28.09 | 5.30 |
| LAG | 70.84 | 18.47 | 9.60 | 3.10 |
| DDLearn | 54.56 | 16.18 | 11.29 | 3.36 |
| SupCon | <u>74.42</u> | 17.06 | 4.66 | 2.16 |
| PCL | 71.61 | 18.36 | 6.00 | 2.45 |
| Ours | **77.43** | 15.81 | 4.48 | 2.12 |

**Table 15:** Dataset-level variance decomposition on HAROOD. For each method and dataset, Mean is the mean of seed-averaged target-task means; Std(task) is the standard deviation across target tasks; Var(seed) and Std(seed) are the pooled within-task seed variance and its standard-deviation scale. Mean, Std(task), and Std(seed) are in percentage points; Var(seed) is in squared percentage points.

| Method | Dataset | Mean ↑ | Std(task) ↓ | Var(seed) ↓ | Std(seed) ↓ |
|---|---|---|---|---|---|
| ERM | DSADS | 81.86 | 3.26 | 5.01 | 2.24 |
| | USC | 70.01 | 7.75 | 1.45 | 1.20 |
| | HAR | 92.12 | 3.96 | 0.58 | 0.76 |
| | EMG | 74.19 | 4.68 | 8.98 | 3.00 |
| | WESAD | 38.72 | 6.43 | 52.14 | 7.22 |
| | PAMAP | 65.97 | 14.24 | 7.76 | 2.79 |
| ERM++ | DSADS | 79.56 | 5.11 | 11.44 | 3.38 |
| | USC | 68.95 | 8.90 | 2.42 | 1.56 |
| | HAR | 92.48 | 3.96 | 0.85 | 0.92 |
| | EMG | 76.19 | 5.39 | 0.94 | 0.97 |
| | WESAD | 38.36 | 6.43 | 7.31 | 2.70 |
| | PAMAP | 65.30 | 14.88 | 3.10 | 1.76 |
| Mixup | DSADS | 85.36 | 2.38 | 3.40 | 1.85 |
| | USC | 70.67 | 7.83 | 1.64 | 1.28 |
| | HAR | 93.21 | 3.67 | 0.99 | 1.00 |
| | EMG | 72.73 | 4.63 | 13.20 | 3.63 |
| | WESAD | 39.20 | 7.06 | 26.08 | 5.11 |
| | PAMAP | 68.90 | 13.78 | 3.82 | 1.96 |
| CORAL | DSADS | 82.32 | 3.33 | 5.24 | 2.29 |
| | USC | 70.87 | 7.92 | 4.94 | 2.22 |
| | HAR | 92.14 | 4.32 | 0.54 | 0.73 |
| | EMG | 76.90 | 5.38 | 3.23 | 1.80 |
| | WESAD | 33.43 | 4.35 | 33.42 | 5.78 |
| | PAMAP | 65.97 | 14.89 | 5.11 | 2.26 |
| MMD | DSADS | 81.12 | 3.94 | 10.64 | 3.26 |
| | USC | 69.54 | 7.00 | 2.57 | 1.60 |
| | HAR | 92.21 | 3.91 | 0.99 | 0.99 |
| | EMG | 76.70 | 5.31 | 2.56 | 1.60 |
| | WESAD | 34.70 | 3.66 | 67.51 | 8.22 |
| | PAMAP | 65.62 | 14.94 | 4.46 | 2.11 |
| DANN | DSADS | 81.87 | 2.06 | 8.30 | 2.88 |
| | USC | 69.11 | 8.89 | 5.20 | 2.28 |
| | HAR | 91.85 | 3.69 | 1.47 | 1.21 |
| | EMG | 63.79 | 4.05 | 111.11 | 10.54 |
| | WESAD | 32.56 | 1.36 | 21.78 | 4.67 |
| | PAMAP | 65.80 | 15.59 | 5.35 | 2.31 |
| MLDG | DSADS | 30.00 | 2.85 | 901.45 | 30.02 |
| | USC | 23.05 | 2.73 | 32.62 | 5.71 |
| | HAR | 41.86 | 3.20 | 867.51 | 29.45 |
| | EMG | 17.45 | 0.47 | 3.73 | 1.93 |
| | WESAD | 25.04 | 0.08 | 0.03 | 0.18 |
| | PAMAP | 23.05 | 4.68 | 5.90 | 2.43 |
| GroupDRO | DSADS | 81.73 | 3.46 | 7.82 | 2.80 |
| | USC | 70.15 | 8.36 | 1.87 | 1.37 |
| | HAR | 92.10 | 4.10 | 0.55 | 0.74 |
| | EMG | 77.03 | 5.71 | 1.70 | 1.30 |
| | WESAD | 31.67 | 5.31 | 17.66 | 4.20 |
| | PAMAP | 65.89 | 13.28 | 4.98 | 2.23 |
| RSC | DSADS | 83.83 | 4.46 | 3.31 | 1.82 |
| | USC | 73.65 | 6.83 | 2.37 | 1.54 |
| | HAR | 92.79 | 3.83 | 0.47 | 0.68 |
| | EMG | 77.20 | 5.29 | 8.18 | 2.86 |
| | WESAD | 35.37 | 3.08 | 34.93 | 5.91 |
| | PAMAP | 69.28 | 14.81 | 3.23 | 1.80 |

*Continued on next page*

| Method | Dataset | Mean ↑ | Std(task) ↓ | Var(seed) ↓ | Std(seed) ↓ |
|---|---|---|---|---|---|
| ANDMask | DSADS | 82.74 | 4.36 | 7.56 | 2.75 |
| | USC | 69.81 | 8.43 | 0.93 | 0.96 |
| | HAR | 91.81 | 4.51 | 0.62 | 0.79 |
| | EMG | 75.00 | 4.74 | 6.27 | 2.50 |
| | WESAD | 37.03 | 4.26 | 79.59 | 8.92 |
| | PAMAP | 64.98 | 14.42 | 10.31 | 3.21 |
| VREx | DSADS | 81.86 | 3.84 | 4.83 | 2.20 |
| | USC | 69.94 | 7.72 | 2.17 | 1.47 |
| | HAR | 92.07 | 3.87 | 0.90 | 0.95 |
| | EMG | 75.48 | 5.29 | 3.44 | 1.85 |
| | WESAD | 39.47 | 6.94 | 39.15 | 6.26 |
| | PAMAP | 66.34 | 14.65 | 4.16 | 2.04 |
| URM | DSADS | 83.27 | 5.01 | 7.07 | 2.66 |
| | USC | 71.49 | 7.53 | 3.95 | 1.99 |
| | HAR | 91.16 | 3.67 | 3.17 | 1.78 |
| | EMG | 76.59 | 5.60 | 9.41 | 3.07 |
| | WESAD | 36.50 | 6.02 | 56.78 | 7.54 |
| | PAMAP | 68.77 | 15.16 | 5.72 | 2.39 |
| Fishr | DSADS | 81.97 | 3.10 | 5.18 | 2.28 |
| | USC | 70.20 | 7.63 | 1.57 | 1.25 |
| | HAR | 92.12 | 4.09 | 0.70 | 0.84 |
| | EMG | 74.09 | 4.44 | 9.06 | 3.01 |
| | WESAD | 40.14 | 6.50 | 57.98 | 7.61 |
| | PAMAP | 66.02 | 14.30 | 8.33 | 2.89 |
| Fish | DSADS | 83.03 | 3.54 | 4.11 | 2.03 |
| | USC | 71.79 | 7.79 | 3.54 | 1.88 |
| | HAR | 92.20 | 3.87 | 0.54 | 0.73 |
| | EMG | 67.50 | 2.15 | 132.64 | 11.52 |
| | WESAD | 28.79 | 4.51 | 23.59 | 4.86 |
| | PAMAP | 67.22 | 13.94 | 10.98 | 3.31 |
| LAG | DSADS | 83.51 | 4.64 | 3.27 | 1.81 |
| | USC | 68.62 | 7.80 | 3.07 | 1.75 |
| | HAR | 92.35 | 3.98 | 0.86 | 0.93 |
| | EMG | 71.49 | 4.37 | 9.39 | 3.06 |
| | WESAD | 39.96 | 8.73 | 35.19 | 5.93 |
| | PAMAP | 63.71 | 14.10 | 7.98 | 2.82 |
| DDLearn | DSADS | 61.35 | 5.39 | 12.30 | 3.51 |
| | USC | 46.98 | 3.99 | 8.39 | 2.90 |
| | HAR | 80.69 | 2.91 | 9.12 | 3.02 |
| | EMG | 50.52 | 2.15 | 15.69 | 3.96 |
| | WESAD | 34.24 | 2.49 | 3.29 | 1.81 |
| | PAMAP | 47.06 | 10.25 | 19.51 | 4.42 |
| SupCon | DSADS | <u>86.61</u> | 3.04 | 3.19 | 1.79 |
| | USC | 70.89 | 7.66 | 1.26 | 1.12 |
| | HAR | <u>93.42</u> | 3.51 | 0.44 | 0.67 |
| | EMG | 76.88 | 5.92 | 2.41 | 1.55 |
| | WESAD | <u>46.10</u> | 6.98 | 18.20 | 4.27 |
| | PAMAP | 67.87 | 14.47 | 3.52 | 1.88 |
| PCL | DSADS | 81.37 | 2.94 | 3.13 | 1.77 |
| | USC | 69.80 | 7.35 | 1.24 | 1.12 |
| | HAR | 92.83 | 3.60 | 0.58 | 0.76 |
| | EMG | 74.47 | 5.42 | 3.91 | 1.98 |
| | WESAD | 39.72 | 9.66 | 22.58 | 4.75 |
| | PAMAP | 66.19 | 14.30 | 5.89 | 2.43 |
| Ours | DSADS | **88.35** | 3.77 | 2.83 | 1.68 |
| | USC | **74.62** | 6.96 | 0.85 | 0.92 |
| | HAR | **94.48** | 3.09 | 0.37 | 0.61 |
| | EMG | **80.46** | 5.51 | 0.41 | 0.64 |
| | WESAD | **50.82** | 7.59 | 20.57 | 4.54 |
| | PAMAP | **71.56** | 13.47 | 2.91 | 1.71 |

### K.3 PER-TASK METRICS

**Table 16:** Per-task results on DSADS using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean $\pm$ sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

| **Balanced accuracy (%)** Method | T0 | T1 | T2 | T3 |
|---|---|---|---|---|
| ERM | 82.40 ± 2.29 | 85.15 ± 1.99 | 82.54 ± 2.96 | 77.35 ± 1.44 |
| ERM++ | 83.38 ± 1.07 | 80.83 ± 2.23 | 81.98 ± 3.21 | 72.06 ± 5.42 |
| Mixup | 85.97 ± 1.57 | 87.26 ± 1.08 | 86.32 ± 1.70 | 81.89 ± 2.66 |
| CORAL | 79.60 ± 2.20 | 85.28 ± 2.36 | 85.12 ± 2.83 | 79.30 ± 1.58 |
| MMD | 79.11 ± 3.58 | 83.18 ± 3.39 | 85.47 ± 3.89 | 76.72 ± 1.78 |
| DANN | 80.20 ± 2.93 | 83.89 ± 2.77 | 83.39 ± 2.94 | 80.00 ± 2.88 |
| MLDG | 30.28 ± 30.49 | 31.84 ± 32.71 | 32.01 ± 29.66 | 25.89 ± 26.95 |
| GroupDRO | 79.40 ± 0.77 | 85.39 ± 3.88 | 83.92 ± 3.40 | 78.22 ± 2.01 |
| RSC | 82.18 ± 2.54 | 86.64 ± 2.06 | 88.18 ± 1.12 | 78.34 ± 1.12 |
| ANDMask | 83.16 ± 1.68 | 83.63 ± 3.10 | 87.34 ± 2.73 | 76.83 ± 3.22 |
| VREx | 81.94 ± 2.12 | 85.54 ± 2.35 | 83.41 ± 0.84 | 76.54 ± 2.94 |
| URM | 84.02 ± 3.26 | 87.00 ± 2.68 | 86.06 ± 2.99 | 75.99 ± 1.25 |
| Fishr | 82.40 ± 2.29 | 85.13 ± 1.99 | 82.62 ± 3.05 | 77.71 ± 1.49 |
| Fish | 80.89 ± 1.93 | 87.00 ± 1.57 | 84.89 ± 0.64 | 79.32 ± 3.14 |
| LAG | 81.43 ± 1.38 | 88.27 ± 3.00 | 86.31 ± 1.30 | 78.04 ± 0.72 |
| DDLearn | 56.61 ± 3.48 | 63.52 ± 2.60 | 67.96 ± 3.85 | 57.32 ± 3.94 |
| SupCon | 85.39 ± 0.40 | 89.03 ± 1.56 | 89.16 ± 2.78 | 82.88 ± 1.56 |
| PCL | 81.87 ± 1.64 | 83.16 ± 0.98 | 83.39 ± 2.44 | 77.08 ± 1.69 |
| Ours | **86.25 ± 0.77** | **90.66 ± 1.11** | **92.30 ± 2.42** | **84.18 ± 1.90** |
| **Accuracy (%)** Method | T0 | T1 | T2 | T3 |
| ERM | 82.40 ± 2.29 | 85.15 ± 1.99 | 82.54 ± 2.96 | 77.35 ± 1.44 |
| ERM++ | 83.38 ± 1.07 | 80.83 ± 2.23 | 81.98 ± 3.21 | 72.06 ± 5.42 |
| Mixup | 85.97 ± 1.57 | 87.26 ± 1.08 | 86.32 ± 1.70 | 81.89 ± 2.66 |
| CORAL | 79.60 ± 2.20 | 85.28 ± 2.36 | 85.12 ± 2.83 | 79.30 ± 1.58 |
| MMD | 79.11 ± 3.58 | 83.18 ± 3.39 | 85.47 ± 3.89 | 76.72 ± 1.78 |
| DANN | 80.20 ± 2.93 | 83.89 ± 2.77 | 83.39 ± 2.94 | 80.00 ± 2.88 |
| MLDG | 30.28 ± 30.49 | 31.84 ± 32.71 | 32.01 ± 29.66 | 25.89 ± 26.95 |
| GroupDRO | 79.40 ± 0.77 | 85.39 ± 3.88 | 83.92 ± 3.40 | 78.22 ± 2.01 |
| RSC | 82.18 ± 2.54 | 86.64 ± 2.06 | 88.18 ± 1.12 | 78.34 ± 1.12 |
| ANDMask | 83.16 ± 1.68 | 83.63 ± 3.10 | 87.34 ± 2.73 | 76.83 ± 3.22 |
| VREx | 81.94 ± 2.12 | 85.54 ± 2.35 | 83.41 ± 0.84 | 76.54 ± 2.94 |
| URM | 84.02 ± 3.26 | 87.00 ± 2.68 | 86.06 ± 2.99 | 75.99 ± 1.25 |
| Fishr | 82.40 ± 2.29 | 85.13 ± 1.99 | 82.62 ± 3.05 | 77.71 ± 1.49 |
| Fish | 80.89 ± 1.93 | 87.00 ± 1.57 | 84.89 ± 0.64 | 79.32 ± 3.14 |
| LAG | 81.43 ± 1.38 | 88.27 ± 3.00 | 86.31 ± 1.30 | 78.04 ± 0.72 |
| DDLearn | 56.61 ± 3.48 | 63.52 ± 2.60 | 67.96 ± 3.85 | 57.32 ± 3.94 |
| SupCon | 85.39 ± 0.40 | 89.03 ± 1.56 | 89.16 ± 2.78 | 82.88 ± 1.56 |
| PCL | 81.87 ± 1.64 | 83.16 ± 0.98 | 83.39 ± 2.44 | 77.08 ± 1.69 |
| Ours | **86.25 ± 0.77** | **90.66 ± 1.11** | **92.30 ± 2.42** | **84.18 ± 1.90** |
| **Cohen's $\kappa$** Method | T0 | T1 | T2 | T3 |
| ERM | 0.81 ± 0.02 | 0.84 ± 0.02 | 0.82 ± 0.03 | 0.76 ± 0.02 |
| ERM++ | 0.82 ± 0.01 | 0.80 ± 0.02 | 0.81 ± 0.03 | 0.71 ± 0.06 |
| Mixup | 0.85 ± 0.02 | 0.87 ± 0.01 | 0.86 ± 0.02 | 0.81 ± 0.03 |
| CORAL | 0.78 ± 0.02 | 0.84 ± 0.02 | 0.84 ± 0.03 | 0.78 ± 0.02 |
| MMD | 0.78 ± 0.04 | 0.82 ± 0.04 | 0.85 ± 0.04 | 0.75 ± 0.02 |
| DANN | 0.79 ± 0.03 | 0.83 ± 0.03 | 0.82 ± 0.03 | 0.79 ± 0.03 |
| MLDG | 0.26 ± 0.32 | 0.28 ± 0.35 | 0.28 ± 0.31 | 0.22 ± 0.28 |
| GroupDRO | 0.78 ± 0.01 | 0.85 ± 0.04 | 0.83 ± 0.04 | 0.77 ± 0.02 |
| RSC | 0.81 ± 0.03 | 0.86 ± 0.02 | 0.88 ± 0.01 | 0.77 ± 0.01 |
| ANDMask | 0.82 ± 0.02 | 0.83 ± 0.03 | 0.87 ± 0.03 | 0.76 ± 0.03 |
| VREx | 0.81 ± 0.02 | 0.85 ± 0.02 | 0.82 ± 0.01 | 0.75 ± 0.03 |
| URM | 0.83 ± 0.03 | 0.86 ± 0.03 | 0.85 ± 0.03 | 0.75 ± 0.01 |
| Fishr | 0.81 ± 0.02 | 0.84 ± 0.02 | 0.82 ± 0.03 | 0.76 ± 0.02 |
| Fish | 0.80 ± 0.02 | 0.86 ± 0.02 | 0.84 ± 0.01 | 0.78 ± 0.03 |
| LAG | 0.80 ± 0.01 | 0.88 ± 0.03 | 0.86 ± 0.01 | 0.77 ± 0.01 |
| DDLearn | 0.54 ± 0.04 | 0.61 ± 0.03 | 0.66 ± 0.04 | 0.55 ± 0.04 |
| SupCon | 0.85 ± 0.00 | 0.88 ± 0.02 | 0.89 ± 0.03 | 0.82 ± 0.02 |
| PCL | 0.81 ± 0.02 | 0.82 ± 0.01 | 0.82 ± 0.03 | 0.76 ± 0.02 |
| Ours | **0.85 ± 0.01** | **0.90 ± 0.01** | **0.92 ± 0.03** | **0.83 ± 0.02** |

**Table 17:** Per-task results on USC using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean ± sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

| **Balanced accuracy (%)** Method | T0 | T1 | T2 | T3 |
|---|---|---|---|---|
| ERM | 67.95 ± 1.38 | 77.10 ± 1.56 | 59.98 ± 0.55 | 75.01 ± 1.06 |
| ERM++ | 65.61 ± 1.93 | 77.19 ± 1.97 | 57.95 ± 1.11 | 75.05 ± 0.92 |
| Mixup | 67.26 ± 1.79 | 78.06 ± 0.76 | 61.26 ± 0.36 | 76.08 ± 1.62 |
| CORAL | 69.20 ± 3.20 | 77.74 ± 1.73 | 60.38 ± 1.00 | 76.17 ± 2.35 |
| MMD | 66.53 ± 1.61 | 75.95 ± 1.72 | 61.09 ± 1.34 | 74.59 ± 1.72 |
| DANN | 64.92 ± 2.16 | 78.58 ± 1.65 | 58.84 ± 1.88 | 74.08 ± 3.15 |
| MLDG | 19.44 ± 5.91 | 25.13 ± 5.92 | 22.42 ± 7.49 | 25.20 ± 2.10 |
| GroupDRO | 66.92 ± 1.79 | 77.89 ± 1.21 | 59.86 ± 1.33 | 75.93 ± 1.01 |
| RSC | **72.53 ± 1.30** | 80.91 ± 0.92 | 64.79 ± 2.20 | 76.38 ± 1.45 |
| ANDMask | 67.71 ± 0.70 | 78.74 ± 0.89 | 59.13 ± 1.41 | 73.65 ± 0.67 |
| VREx | 67.69 ± 1.50 | 77.31 ± 1.57 | 60.09 ± 1.41 | 74.67 ± 1.41 |
| URM | 70.12 ± 1.82 | 76.81 ± 2.67 | 61.40 ± 1.92 | 77.64 ± 1.29 |
| Fishr | 67.88 ± 1.37 | 77.26 ± 1.26 | 60.48 ± 0.76 | 75.19 ± 1.49 |
| Fish | 69.42 ± 2.11 | 77.46 ± 1.82 | 61.80 ± 2.05 | 78.46 ± 1.47 |
| LAG | 65.49 ± 1.73 | 76.69 ± 1.56 | 59.24 ± 1.23 | 73.07 ± 2.31 |
| DDLearn | 46.37 ± 2.19 | 51.63 ± 3.57 | 42.00 ± 1.65 | 47.93 ± 3.65 |
| SupCon | 67.53 ± 1.20 | 78.17 ± 1.16 | 61.73 ± 0.75 | 76.15 ± 1.29 |
| PCL | 67.71 ± 1.24 | 76.76 ± 0.87 | 60.39 ± 0.36 | 74.34 ± 1.60 |
| Ours | 71.45 ± 0.84 | **81.65 ± 0.60** | **66.44 ± 1.21** | **78.95 ± 0.94** |
| **Accuracy (%)** Method | T0 | T1 | T2 | T3 |
| ERM | 69.79 ± 1.94 | 78.85 ± 1.38 | 62.86 ± 0.64 | 78.97 ± 0.96 |
| ERM++ | 68.37 ± 1.99 | 79.33 ± 1.66 | 61.95 ± 1.12 | 79.88 ± 0.88 |
| Mixup | 69.23 ± 1.66 | 79.67 ± 0.44 | 66.20 ± 1.07 | 80.14 ± 1.29 |
| CORAL | 71.25 ± 3.66 | 79.44 ± 1.72 | 63.58 ± 1.41 | 78.73 ± 4.02 |
| MMD | 68.32 ± 2.77 | 77.91 ± 1.49 | 63.60 ± 0.92 | 80.42 ± 0.64 |
| DANN | 67.83 ± 2.25 | 80.05 ± 1.32 | 62.95 ± 2.52 | 79.29 ± 2.71 |
| MLDG | 21.42 ± 8.70 | 27.32 ± 6.30 | 24.81 ± 10.23 | 25.07 ± 5.47 |
| GroupDRO | 68.70 ± 1.67 | 79.72 ± 1.18 | 63.26 ± 1.98 | 80.87 ± 0.72 |
| RSC | **74.02 ± 1.36** | 81.91 ± 1.12 | 67.41 ± 1.87 | 79.83 ± 1.79 |
| ANDMask | 69.71 ± 1.01 | 80.13 ± 0.68 | 62.15 ± 1.30 | 78.03 ± 1.53 |
| VREx | 69.55 ± 2.54 | 79.30 ± 1.30 | 63.16 ± 1.36 | 78.45 ± 0.86 |
| URM | 71.62 ± 2.28 | 78.49 ± 2.26 | 65.15 ± 2.60 | 81.73 ± 1.07 |
| Fishr | 69.42 ± 1.98 | 79.03 ± 1.16 | 63.73 ± 1.39 | 79.05 ± 0.81 |
| Fish | 70.68 ± 2.43 | 79.09 ± 1.71 | 64.52 ± 2.50 | 81.91 ± 1.17 |
| LAG | 68.37 ± 1.78 | 78.92 ± 1.12 | 62.52 ± 0.92 | 77.98 ± 2.26 |
| DDLearn | 51.08 ± 2.64 | 55.92 ± 4.61 | 47.71 ± 2.08 | 55.23 ± 3.01 |
| SupCon | 69.44 ± 2.38 | 79.67 ± 0.66 | 64.34 ± 1.03 | 80.21 ± 1.49 |
| PCL | 70.10 ± 2.62 | 78.64 ± 0.69 | 63.13 ± 0.53 | 79.02 ± 2.13 |
| Ours | 73.55 ± 0.80 | **82.81 ± 0.89** | **69.35 ± 1.41** | **82.33 ± 1.23** |
| **Cohen's $\kappa$** Method | T0 | T1 | T2 | T3 |
| ERM | 0.67 ± 0.02 | 0.77 ± 0.02 | 0.59 ± 0.01 | 0.76 ± 0.01 |
| ERM++ | 0.65 ± 0.02 | 0.77 ± 0.02 | 0.58 ± 0.01 | 0.77 ± 0.01 |
| Mixup | 0.66 ± 0.02 | 0.78 ± 0.00 | 0.63 ± 0.01 | 0.78 ± 0.01 |
| CORAL | 0.68 ± 0.04 | 0.77 ± 0.02 | 0.60 ± 0.01 | 0.76 ± 0.04 |
| MMD | 0.65 ± 0.03 | 0.76 ± 0.02 | 0.60 ± 0.01 | 0.78 ± 0.01 |
| DANN | 0.65 ± 0.02 | 0.78 ± 0.01 | 0.59 ± 0.03 | 0.77 ± 0.03 |
| MLDG | 0.14 ± 0.08 | 0.20 ± 0.07 | 0.18 ± 0.10 | 0.18 ± 0.05 |
| GroupDRO | 0.66 ± 0.02 | 0.78 ± 0.01 | 0.60 ± 0.02 | 0.79 ± 0.01 |
| RSC | **0.71 ± 0.01** | 0.80 ± 0.01 | 0.64 ± 0.02 | 0.77 ± 0.02 |
| ANDMask | 0.67 ± 0.01 | 0.78 ± 0.01 | 0.58 ± 0.01 | 0.75 ± 0.02 |
| VREx | 0.66 ± 0.03 | 0.77 ± 0.01 | 0.59 ± 0.01 | 0.76 ± 0.01 |
| URM | 0.69 ± 0.02 | 0.76 ± 0.02 | 0.62 ± 0.03 | 0.80 ± 0.01 |
| Fishr | 0.66 ± 0.02 | 0.77 ± 0.01 | 0.60 ± 0.01 | 0.77 ± 0.01 |
| Fish | 0.68 ± 0.03 | 0.77 ± 0.02 | 0.61 ± 0.03 | 0.80 ± 0.01 |
| LAG | 0.65 ± 0.02 | 0.77 ± 0.01 | 0.59 ± 0.01 | 0.75 ± 0.02 |
| DDLearn | 0.45 ± 0.03 | 0.50 ± 0.05 | 0.42 ± 0.02 | 0.48 ± 0.04 |
| SupCon | 0.65 ± 0.02 | 0.78 ± 0.01 | 0.61 ± 0.01 | 0.77 ± 0.02 |
| PCL | 0.67 ± 0.03 | 0.76 ± 0.01 | 0.60 ± 0.01 | 0.77 ± 0.02 |
| Ours | 0.71 ± 0.01 | **0.81 ± 0.01** | **0.66 ± 0.02** | **0.80 ± 0.01** |

**Table 18:** Per-task results on HAR using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean ± sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

**Balanced accuracy (%)**

| Method | T0 | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|
| ERM | 92.82 ± 1.13 | 88.44 ± 0.86 | 88.05 ± 0.37 | 97.57 ± 0.29 | 93.72 ± 0.82 |
| ERM++ | 93.13 ± 0.91 | 88.54 ± 1.09 | 88.63 ± 1.16 | 97.88 ± 0.39 | 94.19 ± 0.86 |
| Mixup | 93.56 ± 0.59 | 89.16 ± 1.27 | 90.18 ± 1.55 | 98.26 ± 0.73 | 94.87 ± 0.22 |
| CORAL | 92.73 ± 0.78 | 87.49 ± 1.06 | 88.22 ± 0.88 | 97.81 ± 0.28 | 94.45 ± 0.33 |
| MMD | 92.41 ± 1.11 | 88.19 ± 1.18 | 88.80 ± 1.00 | 97.71 ± 0.44 | 93.95 ± 1.06 |
| DANN | 91.55 ± 1.16 | 89.68 ± 0.75 | 87.27 ± 1.96 | 96.75 ± 1.16 | 94.02 ± 0.48 |
| MLDG | 42.41 ± 29.90 | 40.41 ± 27.92 | 37.49 ± 29.21 | 46.17 ± 29.48 | 42.83 ± 30.68 |
| GroupDRO | 92.52 ± 0.97 | 87.42 ± 0.76 | 88.81 ± 0.64 | 97.61 ± 0.37 | 94.13 ± 0.83 |
| RSC | 93.56 ± 0.78 | 88.45 ± 0.88 | 89.63 ± 0.68 | 97.96 ± 0.55 | 94.36 ± 0.43 |
| ANDMask | 92.51 ± 0.47 | 86.95 ± 1.13 | 87.91 ± 1.07 | 98.04 ± 0.22 | 93.64 ± 0.66 |
| VREx | 92.17 ± 1.47 | 89.00 ± 1.26 | 87.97 ± 0.44 | 97.68 ± 0.35 | 93.54 ± 0.68 |
| URM | 90.44 ± 2.04 | 88.44 ± 2.50 | 87.26 ± 1.36 | 96.18 ± 1.66 | 93.51 ± 0.90 |
| Fishr | 92.82 ± 1.13 | 88.07 ± 1.01 | 88.21 ± 0.23 | 97.79 ± 0.33 | 93.70 ± 1.03 |
| Fish | 92.58 ± 0.99 | 88.33 ± 0.58 | 88.36 ± 0.80 | 97.14 ± 0.80 | 94.57 ± 0.33 |
| LAG | 92.86 ± 0.75 | 87.98 ± 0.93 | 88.94 ± 1.21 | 97.74 ± 0.21 | 94.22 ± 1.17 |
| DDLearn | 79.83 ± 2.01 | 78.66 ± 2.39 | 77.61 ± 2.58 | 84.56 ± 4.46 | 82.80 ± 3.05 |
| SupCon | 94.22 ± 1.22 | 89.91 ± 0.44 | 90.13 ± 0.55 | 98.36 ± 0.46 | 94.47 ± 0.15 |
| PCL | 93.40 ± 0.93 | 89.25 ± 1.22 | 89.38 ± 0.40 | 97.80 ± 0.34 | 94.32 ± 0.51 |
| Ours | **95.33 ± 0.25** | **91.05 ± 0.78** | **91.98 ± 1.05** | **98.83 ± 0.11** | **95.20 ± 0.23** |

**Accuracy (%)**

| Method | T0 | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|
| ERM | 92.83 ± 1.21 | 88.29 ± 0.91 | 88.22 ± 0.29 | 97.23 ± 0.34 | 93.26 ± 0.91 |
| ERM++ | 93.15 ± 0.99 | 88.43 ± 1.13 | 88.83 ± 1.14 | 97.56 ± 0.45 | 93.75 ± 0.89 |
| Mixup | 93.62 ± 0.64 | 89.10 ± 1.31 | 90.01 ± 1.51 | 98.00 ± 0.84 | 94.35 ± 0.24 |
| CORAL | 92.77 ± 0.80 | 87.36 ± 1.11 | 88.38 ± 0.84 | 97.48 ± 0.33 | 94.00 ± 0.32 |
| MMD | 92.48 ± 1.11 | 88.05 ± 1.23 | 88.89 ± 0.90 | 97.35 ± 0.49 | 93.61 ± 1.01 |
| DANN | 91.63 ± 1.10 | 89.55 ± 0.76 | 87.42 ± 1.95 | 96.37 ± 1.27 | 93.76 ± 0.39 |
| MLDG | 41.86 ± 30.62 | 40.69 ± 28.14 | 36.24 ± 29.99 | 45.20 ± 29.89 | 41.51 ± 31.42 |
| GroupDRO | 92.56 ± 1.01 | 87.27 ± 0.75 | 88.84 ± 0.61 | 97.30 ± 0.41 | 93.68 ± 0.78 |
| RSC | 93.64 ± 0.74 | 88.36 ± 0.85 | 89.54 ± 0.69 | 97.64 ± 0.64 | 93.95 ± 0.43 |
| ANDMask | 92.48 ± 0.49 | 86.84 ± 1.19 | 88.16 ± 0.98 | 97.81 ± 0.22 | 93.25 ± 0.66 |
| VREx | 92.21 ± 1.52 | 88.87 ± 1.30 | 88.19 ± 0.49 | 97.34 ± 0.38 | 93.11 ± 0.76 |
| URM | 90.35 ± 2.21 | 88.32 ± 2.48 | 87.24 ± 1.36 | 95.93 ± 1.67 | 93.16 ± 0.96 |
| Fishr | 92.83 ± 1.21 | 87.90 ± 1.03 | 88.38 ± 0.20 | 97.47 ± 0.39 | 93.26 ± 1.09 |
| Fish | 92.60 ± 1.13 | 88.19 ± 0.61 | 88.45 ± 0.74 | 96.73 ± 0.98 | 94.11 ± 0.29 |
| LAG | 92.94 ± 0.77 | 87.87 ± 0.95 | 89.00 ± 1.04 | 97.40 ± 0.22 | 93.80 ± 1.20 |
| DDLearn | 79.70 ± 2.04 | 79.09 ± 2.44 | 78.90 ± 2.50 | 86.29 ± 4.00 | 84.09 ± 2.46 |
| SupCon | 94.29 ± 1.24 | 89.84 ± 0.47 | 90.07 ± 0.59 | 98.12 ± 0.49 | 94.11 ± 0.12 |
| PCL | 93.41 ± 0.95 | 89.19 ± 1.21 | 89.43 ± 0.27 | 97.49 ± 0.39 | 93.90 ± 0.57 |
| Ours | **95.39 ± 0.26** | **90.98 ± 0.75** | **91.81 ± 1.08** | **98.66 ± 0.12** | **94.75 ± 0.27** |

**Cohen's $\kappa$**

| Method | T0 | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|
| ERM | 0.91 ± 0.01 | 0.86 ± 0.01 | 0.86 ± 0.00 | 0.97 ± 0.00 | 0.92 ± 0.01 |
| ERM++ | 0.92 ± 0.01 | 0.86 ± 0.01 | 0.87 ± 0.01 | 0.97 ± 0.01 | 0.92 ± 0.01 |
| Mixup | 0.92 ± 0.01 | 0.87 ± 0.02 | 0.88 ± 0.02 | 0.98 ± 0.01 | 0.93 ± 0.00 |
| CORAL | 0.91 ± 0.01 | 0.85 ± 0.01 | 0.86 ± 0.01 | 0.97 ± 0.00 | 0.93 ± 0.00 |
| MMD | 0.91 ± 0.01 | 0.86 ± 0.01 | 0.87 ± 0.01 | 0.97 ± 0.01 | 0.92 ± 0.01 |
| DANN | 0.90 ± 0.01 | 0.87 ± 0.01 | 0.85 ± 0.02 | 0.96 ± 0.02 | 0.92 ± 0.00 |
| MLDG | 0.31 ± 0.36 | 0.29 ± 0.33 | 0.25 ± 0.35 | 0.36 ± 0.35 | 0.32 ± 0.36 |
| GroupDRO | 0.91 ± 0.01 | 0.85 ± 0.01 | 0.87 ± 0.01 | 0.97 ± 0.00 | 0.92 ± 0.01 |
| RSC | 0.92 ± 0.01 | 0.86 ± 0.01 | 0.87 ± 0.01 | 0.97 ± 0.01 | 0.93 ± 0.01 |
| ANDMask | 0.91 ± 0.01 | 0.84 ± 0.01 | 0.86 ± 0.01 | 0.97 ± 0.00 | 0.92 ± 0.01 |
| VREx | 0.91 ± 0.02 | 0.87 ± 0.02 | 0.86 ± 0.01 | 0.97 ± 0.00 | 0.92 ± 0.01 |
| URM | 0.88 ± 0.03 | 0.86 ± 0.03 | 0.85 ± 0.02 | 0.95 ± 0.02 | 0.92 ± 0.01 |
| Fishr | 0.91 ± 0.01 | 0.85 ± 0.01 | 0.86 ± 0.00 | 0.97 ± 0.00 | 0.92 ± 0.01 |
| Fish | 0.91 ± 0.01 | 0.86 ± 0.01 | 0.86 ± 0.01 | 0.96 ± 0.01 | 0.93 ± 0.00 |
| LAG | 0.92 ± 0.01 | 0.85 ± 0.01 | 0.87 ± 0.01 | 0.97 ± 0.00 | 0.93 ± 0.01 |
| DDLearn | 0.76 ± 0.02 | 0.75 ± 0.03 | 0.75 ± 0.03 | 0.83 ± 0.05 | 0.81 ± 0.03 |
| SupCon | 0.93 ± 0.02 | 0.88 ± 0.00 | 0.88 ± 0.01 | 0.97 ± 0.01 | 0.93 ± 0.00 |
| PCL | 0.92 ± 0.01 | 0.87 ± 0.01 | 0.87 ± 0.00 | 0.97 ± 0.00 | 0.93 ± 0.01 |
| Ours | **0.94 ± 0.00** | **0.89 ± 0.01** | **0.90 ± 0.01** | **0.98 ± 0.00** | **0.94 ± 0.00** |

**Table 19:** Per-task results on EMG using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean $\pm$ sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

| **Balanced accuracy (%)** Method | T0 | T1 | T2 | T3 |
|---|---|---|---|---|
| ERM | 70.73 ± 2.18 | 81.09 ± 2.60 | 72.75 ± 2.66 | 72.19 ± 4.16 |
| ERM++ | 69.36 ± 0.87 | 82.55 ± 1.08 | 76.43 ± 0.98 | 76.41 ± 0.92 |
| Mixup | 67.99 ± 3.90 | 78.44 ± 4.46 | 74.36 ± 1.31 | 70.15 ± 3.99 |
| CORAL | 70.17 ± 2.00 | 83.30 ± 1.82 | 76.62 ± 1.82 | 77.49 ± 1.51 |
| MMD | 70.39 ± 2.06 | 83.38 ± 1.32 | 76.20 ± 1.81 | 76.84 ± 1.00 |
| DANN | 58.71 ± 11.10 | 68.52 ± 12.18 | 64.65 ± 8.72 | 63.26 ± 9.84 |
| MLDG | 18.15 ± 3.31 | 17.29 ± 1.39 | 17.13 ± 0.68 | 17.22 ± 1.27 |
| GroupDRO | 69.83 ± 1.81 | 83.80 ± 1.04 | 77.50 ± 0.79 | 76.99 ± 1.35 |
| RSC | 70.78 ± 2.69 | 83.64 ± 1.58 | 78.04 ± 3.82 | 76.35 ± 2.90 |
| ANDMask | 70.63 ± 1.15 | 81.56 ± 2.97 | 75.12 ± 3.69 | 72.70 ± 1.16 |
| VREx | 69.53 ± 0.55 | 82.14 ± 0.95 | 76.52 ± 1.21 | 73.71 ± 3.33 |
| URM | 69.75 ± 1.60 | 83.04 ± 4.98 | 78.47 ± 1.79 | 75.08 ± 2.65 |
| Fishr | 70.73 ± 2.18 | 80.63 ± 2.71 | 72.75 ± 2.66 | 72.26 ± 4.13 |
| Fish | 65.66 ± 7.17 | 67.11 ± 15.36 | 70.60 ± 9.89 | 66.64 ± 12.06 |
| LAG | 66.83 ± 3.47 | 76.21 ± 4.07 | 74.04 ± 2.58 | 68.90 ± 1.52 |
| DDLearn | 49.12 ± 1.10 | 53.64 ± 4.38 | 49.03 ± 5.09 | 50.31 ± 4.06 |
| SupCon | 70.46 ± 1.24 | 84.56 ± 1.93 | 77.64 ± 0.99 | 74.85 ± 1.85 |
| PCL | 68.23 ± 1.62 | 81.20 ± 1.65 | 75.62 ± 1.51 | 72.84 ± 2.83 |
| Ours | **74.14 ± 0.90** | **87.52 ± 0.65** | **80.85 ± 0.32** | **79.35 ± 0.54** |
| **Accuracy (%)** Method | T0 | T1 | T2 | T3 |
| ERM | 70.80 ± 2.14 | 81.20 ± 2.63 | 73.09 ± 2.85 | 72.26 ± 4.04 |
| ERM++ | 69.41 ± 0.87 | 82.63 ± 1.08 | 76.88 ± 0.98 | 76.48 ± 0.93 |
| Mixup | 68.07 ± 3.96 | 78.55 ± 4.46 | 74.80 ± 1.42 | 70.26 ± 3.99 |
| CORAL | 70.23 ± 2.04 | 83.35 ± 1.81 | 77.05 ± 1.78 | 77.57 ± 1.49 |
| MMD | 70.44 ± 2.08 | 83.44 ± 1.33 | 76.62 ± 1.74 | 76.85 ± 1.01 |
| DANN | 58.74 ± 11.23 | 68.54 ± 12.46 | 64.85 ± 8.94 | 63.19 ± 10.03 |
| MLDG | 17.98 ± 3.28 | 17.19 ± 1.50 | 17.05 ± 0.88 | 17.00 ± 1.20 |
| GroupDRO | 69.89 ± 1.83 | 83.85 ± 1.06 | 77.90 ± 0.75 | 77.05 ± 1.34 |
| RSC | 70.84 ± 2.70 | 83.74 ± 1.60 | 78.40 ± 3.86 | 76.34 ± 2.92 |
| ANDMask | 70.70 ± 1.14 | 81.57 ± 2.98 | 75.41 ± 3.85 | 72.68 ± 1.20 |
| VREx | 69.62 ± 0.52 | 82.19 ± 0.96 | 76.93 ± 1.18 | 73.77 ± 3.23 |
| URM | 69.83 ± 1.62 | 83.12 ± 4.99 | 78.88 ± 1.77 | 75.13 ± 2.63 |
| Fishr | 70.80 ± 2.14 | 80.72 ± 2.76 | 73.09 ± 2.85 | 72.34 ± 4.01 |
| Fish | 65.75 ± 7.08 | 67.12 ± 15.42 | 70.86 ± 10.33 | 66.72 ± 12.02 |
| LAG | 66.91 ± 3.51 | 76.33 ± 4.11 | 74.48 ± 2.40 | 69.02 ± 1.45 |
| DDLearn | 49.28 ± 1.09 | 53.98 ± 4.42 | 50.02 ± 4.95 | 50.52 ± 4.14 |
| SupCon | 70.57 ± 1.29 | 84.66 ± 1.90 | 78.04 ± 0.98 | 74.89 ± 1.84 |
| PCL | 68.27 ± 1.68 | 81.32 ± 1.61 | 76.03 ± 1.39 | 72.90 ± 2.77 |
| Ours | **74.23 ± 0.94** | **87.58 ± 0.61** | **81.19 ± 0.29** | **79.33 ± 0.55** |
| **Cohen's $\kappa$** Method | T0 | T1 | T2 | T3 |
| ERM | 0.65 ± 0.03 | 0.77 ± 0.03 | 0.68 ± 0.03 | 0.67 ± 0.05 |
| ERM++ | 0.63 ± 0.01 | 0.79 ± 0.01 | 0.72 ± 0.01 | 0.72 ± 0.01 |
| Mixup | 0.62 ± 0.05 | 0.74 ± 0.05 | 0.70 ± 0.02 | 0.64 ± 0.05 |
| CORAL | 0.64 ± 0.02 | 0.80 ± 0.02 | 0.72 ± 0.02 | 0.73 ± 0.02 |
| MMD | 0.65 ± 0.02 | 0.80 ± 0.02 | 0.72 ± 0.02 | 0.72 ± 0.01 |
| DANN | 0.51 ± 0.13 | 0.62 ± 0.15 | 0.58 ± 0.11 | 0.56 ± 0.12 |
| MLDG | 0.02 ± 0.04 | 0.01 ± 0.02 | 0.01 ± 0.01 | 0.01 ± 0.02 |
| GroupDRO | 0.64 ± 0.02 | 0.81 ± 0.01 | 0.73 ± 0.01 | 0.72 ± 0.02 |
| RSC | 0.65 ± 0.03 | 0.80 ± 0.02 | 0.74 ± 0.05 | 0.72 ± 0.04 |
| ANDMask | 0.65 ± 0.01 | 0.78 ± 0.04 | 0.70 ± 0.05 | 0.67 ± 0.01 |
| VREx | 0.64 ± 0.01 | 0.79 ± 0.01 | 0.72 ± 0.01 | 0.69 ± 0.04 |
| URM | 0.64 ± 0.02 | 0.80 ± 0.06 | 0.75 ± 0.02 | 0.70 ± 0.03 |
| Fishr | 0.65 ± 0.03 | 0.77 ± 0.03 | 0.68 ± 0.03 | 0.67 ± 0.05 |
| Fish | 0.59 ± 0.09 | 0.61 ± 0.18 | 0.65 ± 0.12 | 0.60 ± 0.14 |
| LAG | 0.60 ± 0.04 | 0.72 ± 0.05 | 0.69 ± 0.03 | 0.63 ± 0.02 |
| DDLearn | 0.39 ± 0.01 | 0.45 ± 0.05 | 0.40 ± 0.06 | 0.41 ± 0.05 |
| SupCon | 0.63 ± 0.01 | 0.82 ± 0.02 | 0.74 ± 0.01 | 0.70 ± 0.02 |
| PCL | 0.62 ± 0.02 | 0.78 ± 0.02 | 0.71 ± 0.02 | 0.67 ± 0.03 |
| Ours | **0.69 ± 0.01** | **0.85 ± 0.01** | **0.77 ± 0.00** | **0.75 ± 0.01** |

**Table 20:** Per-task results on WESAD using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean ± sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

| **Balanced accuracy (%)** | | | | |
|---|---|---|---|---|
| Method | T0 | T1 | T2 | T3 |
| ERM | 29.61 ± 2.46 | 43.16 ± 6.10 | 38.78 ± 6.71 | 43.34 ± 10.97 |
| ERM++ | 29.10 ± 2.98 | 42.77 ± 1.07 | 38.93 ± 2.45 | 42.67 ± 3.63 |
| Mixup | 29.85 ± 3.21 | 39.81 ± 4.49 | 40.15 ± 5.76 | 46.99 ± 6.38 |
| CORAL | 30.29 ± 2.78 | 39.08 ± 1.40 | 34.61 ± 3.08 | 29.75 ± 10.70 |
| MMD | 29.37 ± 3.04 | 37.70 ± 5.77 | 35.60 ± 2.01 | 36.13 ± 14.95 |
| DANN | 32.23 ± 2.89 | 34.56 ± 5.38 | 31.96 ± 3.68 | 31.51 ± 6.03 |
| MLDG | 25.16 ± 0.36 | 25.00 ± 0.00 | 25.00 ± 0.00 | 25.00 ± 0.00 |
| GroupDRO | 27.47 ± 4.14 | 36.23 ± 4.78 | 36.28 ± 3.41 | 26.69 ± 4.37 |
| RSC | 33.69 ± 6.31 | 39.46 ± 4.04 | 32.45 ± 5.96 | 35.87 ± 6.93 |
| ANDMask | 34.25 ± 6.30 | 36.69 ± 7.92 | 34.00 ± 6.90 | 43.16 ± 12.98 |
| VREx | 29.76 ± 6.01 | 44.54 ± 6.73 | 39.16 ± 3.78 | 44.43 ± 7.80 |
| URM | 31.43 ± 4.86 | 45.22 ± 6.31 | 34.90 ± 7.70 | 34.46 ± 10.21 |
| Fishr | 30.44 ± 2.71 | 44.19 ± 5.28 | 42.54 ± 8.81 | 43.38 ± 10.91 |
| Fish | 27.70 ± 2.58 | 35.31 ± 7.14 | 27.17 ± 2.47 | 24.97 ± 5.53 |
| LAG | 27.49 ± 3.71 | 41.02 ± 4.23 | 43.82 ± 7.64 | 47.50 ± 7.12 |
| DDLearn | 32.32 ± 1.97 | 33.01 ± 1.57 | 33.78 ± 2.10 | 37.87 ± 1.55 |
| SupCon | 36.29 ± 4.81 | 47.37 ± 4.60 | 47.91 ± 3.54 | 52.82 ± 4.00 |
| PCL | 28.68 ± 2.83 | 38.76 ± 4.65 | 39.18 ± 5.58 | 52.25 ± 5.44 |
| Ours | **40.77 ± 6.52** | **51.00 ± 2.17** | **52.37 ± 2.26** | **59.14 ± 5.47** |
| **Accuracy (%)** | | | | |
| Method | T0 | T1 | T2 | T3 |
| ERM | 37.12 ± 2.36 | 45.13 ± 5.48 | 43.53 ± 3.62 | 48.20 ± 13.95 |
| ERM++ | 35.18 ± 4.08 | 47.91 ± 2.50 | 44.34 ± 1.94 | 53.09 ± 3.57 |
| Mixup | 40.57 ± 2.31 | 42.65 ± 7.39 | 39.43 ± 6.96 | 54.33 ± 8.99 |
| CORAL | 37.94 ± 4.52 | 38.93 ± 2.61 | 37.31 ± 3.85 | 36.77 ± 14.05 |
| MMD | 38.78 ± 2.56 | 35.54 ± 4.21 | 38.05 ± 3.23 | 47.07 ± 14.76 |
| DANN | 40.32 ± 5.70 | 37.69 ± 5.21 | 37.26 ± 8.12 | 38.83 ± 10.71 |
| MLDG | 23.97 ± 3.50 | 24.35 ± 9.87 | 22.63 ± 0.00 | 23.53 ± 1.10 |
| GroupDRO | 35.00 ± 4.57 | 37.06 ± 2.64 | 32.63 ± 5.93 | 30.76 ± 5.85 |
| RSC | 42.01 ± 5.62 | 41.38 ± 5.85 | 38.72 ± 8.19 | 42.93 ± 11.83 |
| ANDMask | 43.23 ± 2.61 | 42.15 ± 11.27 | 38.77 ± 9.69 | 50.49 ± 15.04 |
| VREx | 38.29 ± 2.48 | 49.47 ± 8.23 | 45.02 ± 2.55 | 51.41 ± 11.65 |
| URM | 39.73 ± 5.45 | 51.44 ± 5.03 | 41.47 ± 7.05 | 33.10 ± 9.58 |
| Fishr | 37.95 ± 3.36 | 45.28 ± 4.86 | 43.92 ± 6.91 | 48.32 ± 13.77 |
| Fish | 30.67 ± 7.25 | 34.37 ± 8.36 | 24.99 ± 7.74 | 22.32 ± 11.27 |
| LAG | 36.38 ± 4.58 | 41.84 ± 5.50 | 48.02 ± 4.32 | 52.66 ± 8.39 |
| DDLearn | 41.51 ± 3.21 | 38.47 ± 4.51 | 44.19 ± 2.38 | 46.18 ± 2.72 |
| SupCon | 41.18 ± 4.91 | 49.79 ± 6.45 | 49.85 ± 5.33 | 54.66 ± 9.11 |
| PCL | 36.48 ± 2.04 | 39.50 ± 6.35 | 43.51 ± 6.25 | 57.84 ± 7.13 |
| Ours | **46.78 ± 7.93** | **54.12 ± 3.60** | **54.23 ± 4.96** | **64.48 ± 6.16** |
| **Cohen's $\kappa$** | | | | |
| Method | T0 | T1 | T2 | T3 |
| ERM | 0.06 ± 0.03 | 0.26 ± 0.07 | 0.20 ± 0.09 | 0.27 ± 0.18 |
| ERM++ | 0.05 ± 0.05 | 0.27 ± 0.02 | 0.21 ± 0.04 | 0.30 ± 0.06 |
| Mixup | 0.07 ± 0.04 | 0.22 ± 0.08 | 0.21 ± 0.07 | 0.34 ± 0.11 |
| CORAL | 0.08 ± 0.04 | 0.19 ± 0.03 | 0.15 ± 0.04 | 0.09 ± 0.18 |
| MMD | 0.07 ± 0.04 | 0.15 ± 0.07 | 0.16 ± 0.01 | 0.18 ± 0.25 |
| DANN | 0.13 ± 0.06 | 0.14 ± 0.08 | 0.11 ± 0.07 | 0.11 ± 0.09 |
| MLDG | 0.00 ± 0.01 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| GroupDRO | 0.04 ± 0.05 | 0.14 ± 0.06 | 0.14 ± 0.06 | 0.02 ± 0.06 |
| RSC | 0.13 ± 0.09 | 0.21 ± 0.07 | 0.13 ± 0.10 | 0.18 ± 0.13 |
| ANDMask | 0.15 ± 0.07 | 0.18 ± 0.13 | 0.15 ± 0.12 | 0.29 ± 0.19 |
| VREx | 0.07 ± 0.08 | 0.30 ± 0.11 | 0.22 ± 0.05 | 0.31 ± 0.15 |
| URM | 0.12 ± 0.08 | 0.31 ± 0.08 | 0.16 ± 0.11 | 0.10 ± 0.12 |
| Fishr | 0.07 ± 0.04 | 0.27 ± 0.06 | 0.23 ± 0.11 | 0.27 ± 0.18 |
| Fish | 0.05 ± 0.05 | 0.14 ± 0.09 | 0.03 ± 0.03 | -0.00 ± 0.09 |
| LAG | 0.05 ± 0.06 | 0.22 ± 0.07 | 0.28 ± 0.08 | 0.33 ± 0.11 |
| DDLearn | 0.11 ± 0.03 | 0.12 ± 0.03 | 0.14 ± 0.03 | 0.20 ± 0.03 |
| SupCon | 0.17 ± 0.07 | 0.32 ± 0.07 | 0.32 ± 0.05 | 0.38 ± 0.09 |
| PCL | 0.06 ± 0.04 | 0.20 ± 0.08 | 0.23 ± 0.06 | 0.40 ± 0.09 |
| Ours | **0.25 ± 0.10** | **0.37 ± 0.04** | **0.38 ± 0.04** | **0.49 ± 0.08** |

**Table 21:** Per-task results on PAMAP using five seeds $\{0, 42, 96, 123, 3407\}$ , each task is a leave-one-out subject groups for this dataset, more details in Appendix G. Each entry is mean ± sample standard deviation across seeds. Balanced accuracy and accuracy are reported in percent; $\kappa$ is reported on its raw scale. Best is bold and second-best is underlined within each metric column.

| **Balanced accuracy (%)** Method | T0 | T1 | T2 | T3 |
|---|---|---|---|---|
| ERM | 76.88 ± 1.46 | 68.42 ± 3.45 | 45.26 ± 3.39 | 73.34 ± 2.35 |
| ERM++ | 76.64 ± 1.28 | 67.62 ± 1.93 | 43.68 ± 2.16 | 73.25 ± 1.54 |
| Mixup | 78.15 ± 2.19 | **72.12 ± 1.23** | 48.60 ± 2.08 | 76.75 ± 2.16 |
| CORAL | 75.89 ± 1.11 | 68.19 ± 2.57 | 44.27 ± 1.10 | 75.52 ± 3.38 |
| MMD | 74.02 ± 1.35 | 68.06 ± 2.08 | 43.85 ± 1.26 | 76.55 ± 3.18 |
| DANN | 76.98 ± 1.42 | 65.98 ± 1.62 | 43.69 ± 3.31 | 76.53 ± 2.41 |
| MLDG | 27.69 ± 3.22 | 23.15 ± 2.60 | 16.62 ± 2.02 | 24.75 ± 1.55 |
| GroupDRO | 73.84 ± 2.93 | 67.60 ± 1.66 | 46.63 ± 2.12 | 75.50 ± 2.03 |
| RSC | **81.04 ± 1.19** | 69.62 ± 1.87 | 48.29 ± 2.33 | 78.18 ± 1.61 |
| ANDMask | 75.44 ± 1.53 | 66.05 ± 2.03 | 44.27 ± 5.21 | 74.15 ± 2.76 |
| VREx | 76.91 ± 0.65 | 68.25 ± 2.48 | 45.09 ± 1.79 | 75.10 ± 2.62 |
| URM | 77.58 ± 1.22 | 72.03 ± 2.21 | 46.48 ± 3.00 | 78.99 ± 2.75 |
| Fishr | 76.79 ± 1.51 | 68.35 ± 3.48 | 45.23 ± 3.36 | 73.73 ± 2.77 |
| Fish | 76.35 ± 1.93 | 69.08 ± 1.49 | 46.97 ± 5.31 | 76.50 ± 3.13 |
| LAG | 74.56 ± 3.71 | 65.57 ± 1.12 | 43.31 ± 2.45 | 71.40 ± 3.30 |
| DDLearn | 53.20 ± 4.07 | 46.62 ± 4.50 | 32.79 ± 3.46 | 55.64 ± 5.41 |
| SupCon | 77.87 ± 1.00 | 69.42 ± 2.40 | 46.94 ± 2.51 | 77.23 ± 1.02 |
| PCL | 77.53 ± 1.34 | 67.89 ± 1.83 | 45.58 ± 2.40 | 73.74 ± 3.55 |
| Ours | 80.68 ± 1.31 | 71.11 ± 1.69 | **52.67 ± 2.34** | **81.77 ± 1.27** |
| **Accuracy (%)** Method | T0 | T1 | T2 | T3 |
| ERM | 85.64 ± 1.19 | 76.64 ± 2.71 | 53.01 ± 5.90 | 82.47 ± 1.84 |
| ERM++ | 86.22 ± 0.98 | 75.62 ± 2.89 | 47.81 ± 5.11 | 83.50 ± 2.74 |
| Mixup | 86.51 ± 2.77 | 79.02 ± 0.83 | 55.70 ± 3.43 | 85.40 ± 2.65 |
| CORAL | 85.29 ± 0.82 | 76.00 ± 1.82 | 48.86 ± 3.15 | 85.43 ± 1.49 |
| MMD | 84.29 ± 0.94 | 76.62 ± 1.62 | 48.15 ± 3.87 | 85.76 ± 1.80 |
| DANN | 86.42 ± 0.70 | 74.21 ± 2.40 | 51.17 ± 4.64 | 85.10 ± 1.68 |
| MLDG | 35.95 ± 6.28 | 35.04 ± 6.79 | 25.66 ± 2.39 | 36.74 ± 3.72 |
| GroupDRO | 83.81 ± 2.07 | 75.91 ± 1.21 | 53.43 ± 3.74 | 85.23 ± 0.98 |
| RSC | **88.37 ± 1.23** | 77.81 ± 1.50 | 52.97 ± 4.83 | 85.07 ± 2.57 |
| ANDMask | 85.03 ± 2.66 | 74.80 ± 1.68 | 50.35 ± 7.53 | 81.62 ± 2.78 |
| VREx | 86.55 ± 0.75 | 76.36 ± 1.95 | 51.52 ± 2.46 | 83.61 ± 1.46 |
| URM | 86.42 ± 1.93 | **79.70 ± 2.11** | 52.70 ± 5.33 | 84.77 ± 1.41 |
| Fishr | 85.52 ± 1.14 | 76.72 ± 2.67 | 52.74 ± 5.61 | 82.93 ± 1.81 |
| Fish | 83.53 ± 3.60 | 77.49 ± 1.90 | 51.65 ± 7.38 | 85.49 ± 2.40 |
| LAG | 84.77 ± 3.68 | 74.10 ± 1.11 | 49.19 ± 3.63 | 82.07 ± 2.86 |
| DDLearn | 69.99 ± 4.65 | 62.49 ± 4.16 | 38.76 ± 6.62 | 71.23 ± 4.06 |
| SupCon | 87.17 ± 0.89 | 78.01 ± 2.19 | 52.38 ± 2.66 | **86.95 ± 0.28** |
| PCL | 86.51 ± 1.63 | 76.59 ± 1.40 | 51.43 ± 3.96 | 83.26 ± 2.61 |
| Ours | 87.90 ± 1.01 | 79.34 ± 1.90 | **60.49 ± 3.43** | 86.03 ± 0.80 |
| **Cohen's $\kappa$** Method | T0 | T1 | T2 | T3 |
| ERM | 0.84 ± 0.01 | 0.74 ± 0.03 | 0.47 ± 0.07 | 0.80 ± 0.02 |
| ERM++ | 0.84 ± 0.01 | 0.72 ± 0.03 | 0.42 ± 0.06 | 0.81 ± 0.03 |
| Mixup | 0.85 ± 0.03 | 0.76 ± 0.01 | 0.50 ± 0.04 | 0.84 ± 0.03 |
| CORAL | 0.83 ± 0.01 | 0.73 ± 0.02 | 0.43 ± 0.03 | 0.84 ± 0.02 |
| MMD | 0.82 ± 0.01 | 0.74 ± 0.02 | 0.42 ± 0.04 | 0.84 ± 0.02 |
| DANN | 0.85 ± 0.01 | 0.71 ± 0.03 | 0.45 ± 0.05 | 0.83 ± 0.02 |
| MLDG | 0.27 ± 0.07 | 0.25 ± 0.07 | 0.15 ± 0.03 | 0.28 ± 0.03 |
| GroupDRO | 0.82 ± 0.02 | 0.73 ± 0.01 | 0.48 ± 0.04 | 0.83 ± 0.01 |
| RSC | **0.87 ± 0.01** | 0.75 ± 0.02 | 0.47 ± 0.05 | 0.83 ± 0.03 |
| ANDMask | 0.83 ± 0.03 | 0.71 ± 0.02 | 0.44 ± 0.08 | 0.79 ± 0.03 |
| VREx | 0.85 ± 0.01 | 0.73 ± 0.02 | 0.46 ± 0.03 | 0.82 ± 0.02 |
| URM | 0.85 ± 0.02 | **0.77 ± 0.02** | 0.47 ± 0.06 | 0.83 ± 0.02 |
| Fishr | 0.83 ± 0.01 | 0.74 ± 0.03 | 0.47 ± 0.06 | 0.81 ± 0.02 |
| Fish | 0.81 ± 0.04 | 0.75 ± 0.02 | 0.46 ± 0.08 | 0.84 ± 0.03 |
| LAG | 0.83 ± 0.04 | 0.71 ± 0.01 | 0.43 ± 0.04 | 0.80 ± 0.03 |
| DDLearn | 0.66 ± 0.05 | 0.57 ± 0.05 | 0.31 ± 0.07 | 0.67 ± 0.05 |
| SupCon | 0.82 ± 0.04 | 0.74 ± 0.03 | 0.42 ± 0.08 | **0.84 ± 0.01** |
| PCL | 0.85 ± 0.02 | 0.73 ± 0.02 | 0.46 ± 0.04 | 0.81 ± 0.03 |
| Ours | 0.86 ± 0.01 | 0.77 ± 0.02 | **0.56 ± 0.04** | 0.84 ± 0.01 |

# L Expanded Ablation Results

## L.1 Per-task component ablations

**Table 22:** Detailed EMG ablation table. Columns correspond to target tasks. Each cell reports the best balanced accuracy (%) achieved by that ablation family, averaged over seeds. The Avg column first averages over tasks within each seed and then reports mean ± std across seeds.

| Ablation | Task 0 | Task 1 | Task 2 | Task 3 | Avg |
|---|---|---|---|---|---|
| Base | **74.40 ± 1.01** | 86.74 ± 0.57 | 80.49 ± 0.70 | **78.96 ± 0.35** | **80.15 ± 0.27** |
| InfoNCE | 70.13 ± 1.28 | 83.46 ± 2.07 | 77.55 ± 1.37 | 73.53 ± 2.60 | 76.17 ± 0.21 |
| RewInfoNCE | 71.74 ± 1.20 | 83.24 ± 2.28 | 78.66 ± 1.85 | 75.62 ± 0.72 | 77.32 ± 0.88 |
| NoHard | 72.56 ± 1.93 | 84.61 ± 1.59 | 80.47 ± 0.60 | 77.96 ± 0.74 | 78.90 ± 0.48 |
| Top0.60 | 72.49 ± 1.31 | 86.17 ± 0.57 | 80.28 ± 1.13 | 78.65 ± 0.45 | 79.40 ± 0.15 |
| Top0.90 | 73.41 ± 1.55 | 85.96 ± 0.74 | 80.03 ± 0.87 | 77.97 ± 1.15 | 79.34 ± 0.58 |
| HighAlign | 73.22 ± 1.64 | **87.14 ± 1.59** | **80.50 ± 0.19** | 76.95 ± 1.80 | 79.45 ± 0.80 |
| NoAlign | 69.26 ± 1.94 | 80.25 ± 1.38 | 76.22 ± 2.69 | 71.14 ± 1.70 | 74.22 ± 0.29 |

**Table 23:** Detailed PAMAP ablation table. Columns correspond to target tasks. Each cell reports the best balanced accuracy (%) achieved by that ablation family, averaged over seeds. The Avg column first averages over tasks within each seed and then reports mean ± std across seeds.

| Ablation | Task 0 | Task 1 | Task 2 | Task 3 | Avg |
|---|---|---|---|---|---|
| Base | **79.42 ± 0.79** | **71.37 ± 1.58** | **52.35 ± 0.97** | **81.70 ± 1.77** | **71.21 ± 0.50** |
| InfoNCE | 76.52 ± 1.21 | 67.41 ± 0.54 | 45.91 ± 3.95 | 76.69 ± 1.09 | 66.64 ± 1.21 |
| RewInfoNCE | 77.04 ± 2.51 | 69.50 ± 2.79 | 50.36 ± 2.81 | 81.27 ± 0.48 | 69.55 ± 1.21 |
| NoHard | 79.35 ± 1.68 | 68.51 ± 1.53 | 50.86 ± 1.80 | 79.16 ± 1.99 | 69.47 ± 1.36 |
| Top0.90 | 78.75 ± 2.09 | 70.04 ± 2.35 | 52.01 ± 1.46 | 80.22 ± 1.70 | 70.26 ± 0.95 |
| HighAlign | 78.73 ± 1.53 | 68.46 ± 0.86 | 50.62 ± 0.64 | 78.56 ± 1.18 | 69.09 ± 0.68 |
| NoAlign | 78.10 ± 2.30 | 66.23 ± 2.24 | 43.44 ± 2.31 | 73.53 ± 0.77 | 65.32 ± 0.96 |

## L.2 Target-plan component ablation

We isolate the contributions of identity (ID), same-class cross-domain (CD), and same-class same-domain (SD) relations in the target plan. All four variants retain identity support, and different-class pairs receive zero target mass. We evaluate on EMG and PAMAP2 with seeds $\{0, 42, 96\}$. For each seed, we first average balanced accuracy over the four held-out-participant tasks, then report the mean and sample standard deviation across seeds.

**Table 24:** Target-plan component ablation. Balanced accuracy (%) is reported as mean ± sample standard deviation over three seeds. The full-method row is the reference run for this controlled comparison.

| Target construction | EMG | PAMAP2 |
|---|---|---|
| ID only | $72.63 \pm 0.24$ | $64.72 \pm 1.47$ |
| ID + SD | $73.12 \pm 2.72$ | $64.70 \pm 1.54$ |
| ID + CD | $79.76 \pm 0.50$ | $70.36 \pm 3.03$ |
| Full Multa | $\mathbf{80.15 \pm 0.27}$ | $\mathbf{71.21 \pm 0.50}$ |

Adding CD relations to identity matching improves mean balanced accuracy by $7.13$ points on EMG and $5.64$ points on PAMAP2. SD relations alone provide little improvement, whereas adding them to ID + CD gives a further $0.38$ and $0.85$ points, respectively. These results identify cross-domain same-class matching as the main contributor, with within-domain same-class relations providing a complementary benefit.

# M ADDITIONAL SENSITIVITY AND ROBUSTNESS RESULTS

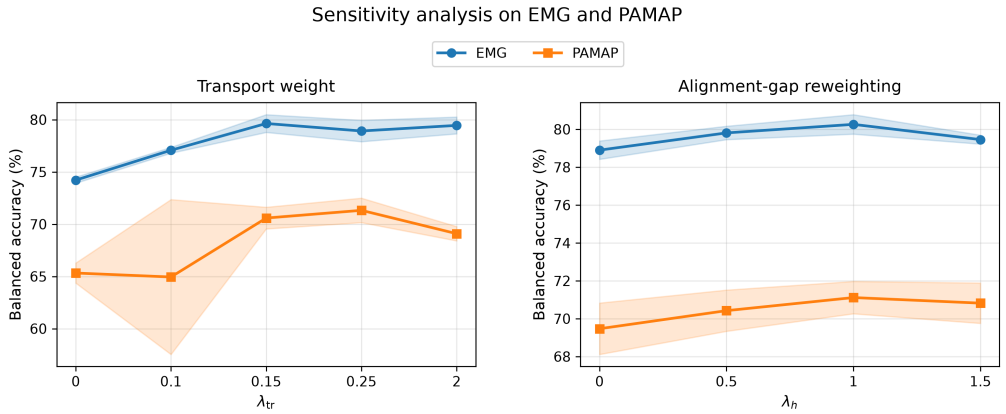


**Figure 2:** Dataset-level sensitivity analysis of MULTA on EMG and PAMAP for the two primary controls: transport weight $\lambda_{\text{tr}}$ and alignment-gap reweighting strength $\lambda_h$. We report mean balanced accuracy over three seeds; the shaded band denotes one standard deviation.

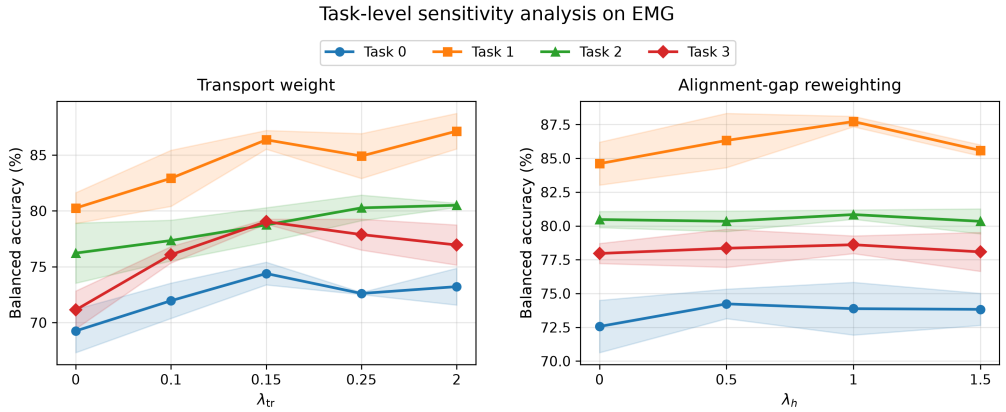


**Figure 3:** Task-level sensitivity analysis on EMG for $\lambda_{\text{tr}}$ and $\lambda_h$. EMG shows a relatively broad high-performance region, although Tasks 1–2 are more responsive to stronger transport alignment than Tasks 0 and 3.

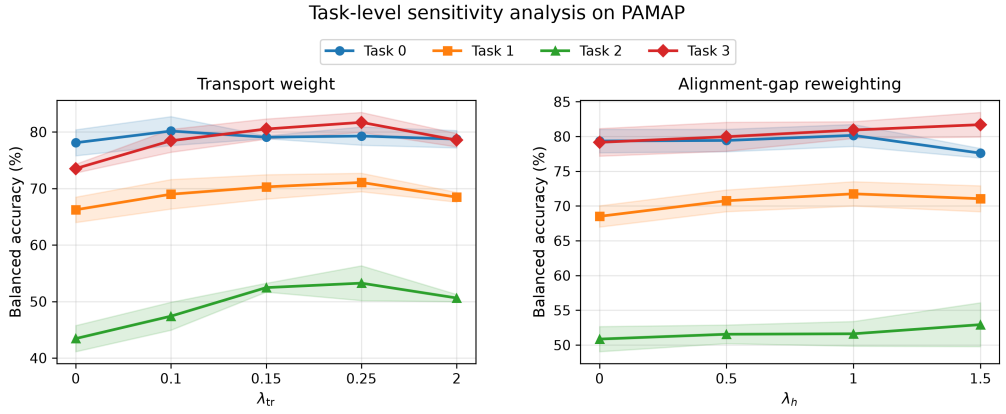


**Figure 4:** Task-level sensitivity analysis on PAMAP for $\lambda_{\text{tr}}$ and $\lambda_h$. Compared with EMG, PAMAP shows stronger task-to-task variation, especially for the transport weight.

| EMG target | $\lambda_{\text{tr}}$ | $\lambda_h$ |
|---|---|---|
| Task 0 | 0.15 / 74.40 ± 1.01 | 0.5 / 74.23 ± 1.09 |
| Task 1 | 2.0 / 87.14 ± 1.59 | 1.0 / 87.72 ± 0.39 |
| Task 2 | 2.0 / 80.50 ± 0.19 | 1.0 / 80.84 ± 0.34 |
| Task 3 | 0.15 / 79.04 ± 0.22 | 1.0 / 78.61 ± 0.65 |

**Table 25:** Task-level preferred values on EMG. Each cell reports *best value / mean±std* in balanced accuracy across seeds. EMG Tasks 1–2 are the clearest cases in which stronger transport alignment becomes beneficial.

| PAMAP target | $\lambda_{\text{tr}}$ | $\lambda_h$ |
|---|---|---|
| Task 0 | 0.1 / 80.16 ± 2.56 † | 1.0 / 80.16 ± 1.57 |
| Task 1 | 0.25 / 71.07 ± 1.63 | 1.0 / 71.75 ± 1.75 |
| Task 2 | 0.25 / 53.25 ± 3.07 | 1.5 / 52.93 ± 3.15 |
| Task 3 | 0.25 / 81.70 ± 1.77 | 1.5 / 81.70 ± 1.77 |

**Table 26:** Task-level preferred values on PAMAP. Each cell reports *best value / mean±std* in balanced accuracy across seeds. † indicates that the selected sensitivity point is based on two seeds rather than three.

| Variant | EMG mean | EMG Δ vs. base | PAMAP mean | PAMAP Δ vs. base |
|---|---|---|---|---|
| Base | 78.62 ± 0.19 | 0.00 ± 0.00 | 69.92 ± 0.88 | 0.00 ± 0.00 |
| Strong transport alignment ($\lambda_{\text{tr}} = 2$) | 76.74 ± 0.12 | -1.89 ± 0.30 | 66.38 ± 0.34 | -3.54 ± 0.68 |
| No gap reweighting ($\lambda_h = 0$) | 76.86 ± 0.37 | -1.76 ± 0.39 | 67.76 ± 1.71 | -2.16 ± 0.84 |
| No transport alignment ($\lambda_{\text{tr}} = 0$) | 74.22 ± 0.29 | -4.41 ± 0.44 | 65.32 ± 0.96 | -4.60 ± 0.10 |

**Table 27:** Dataset-level robustness variants. "Mean" denotes the task-averaged balanced accuracy after averaging within seed; "Δ vs. base" is measured relative to the matched base top-10 pool. Removing transport alignment causes the largest drop among the variants shown.

| Dataset | Target | Mean Δ vs. base | Mean best Δ | Mean worst Δ | Best observed Δ | Worst observed Δ |
|---|---|---|---|---|---|---|
| EMG | Task 0 | -2.54 ± 1.29 | -1.07 ± 1.09 | -4.38 ± 1.67 | 1.50 ± 1.65 | -8.95 ± 3.09 |
| EMG | Task 1 | -1.49 ± 0.80 | 0.29 ± 0.76 | -2.95 ± 0.96 | 3.03 ± 0.53 | -6.56 ± 2.17 |
| EMG | Task 2 | -0.90 ± 0.57 | 0.63 ± 0.56 | -2.36 ± 1.16 | 2.38 ± 0.49 | -5.03 ± 0.81 |
| EMG | Task 3 | -2.07 ± 1.04 | -0.53 ± 1.01 | -3.84 ± 1.28 | 1.89 ± 0.49 | -7.27 ± 2.64 |
| PAMAP | Task 0 | -1.37 ± 0.26 | 0.45 ± 0.55 | -3.65 ± 0.87 | 2.27 ± 0.64 | -6.91 ± 1.32 |
| PAMAP | Task 1 | -2.13 ± 0.68 | -0.11 ± 0.73 | -4.65 ± 1.51 | 1.21 ± 0.21 | -6.34 ± 1.81 |
| PAMAP | Task 2 | -2.27 ± 0.61 | -0.61 ± 1.09 | -3.83 ± 0.19 | 2.72 ± 2.16 | -5.37 ± 0.21 |
| PAMAP | Task 3 | -2.47 ± 1.07 | -0.93 ± 0.69 | -5.04 ± 2.20 | 0.90 ± 0.97 | -7.83 ± 2.01 |

**Table 28:** Target-level robustness summary over matched base/variant pairs. Mean Δ summarizes the average effect of the evaluated variants relative to the matched base pool, while the best- and worst-observed columns show the strongest improvement and strongest degradation found within each target family.

## N Qualitative Latent-Space Visualization

Figure 5 shows the evolution of a representative hard anchor-support case on SleepEDFx. Across epochs 1, 10, 20, 30, and 70, the anchor moves toward valid cross-domain supports, and the local neighborhood becomes more compact. This visualization is qualitative and is intended to illustrate the type of local cross-domain structure encouraged by MULTA; the quantitative claims in the main text are supported by the benchmark and ablation results.

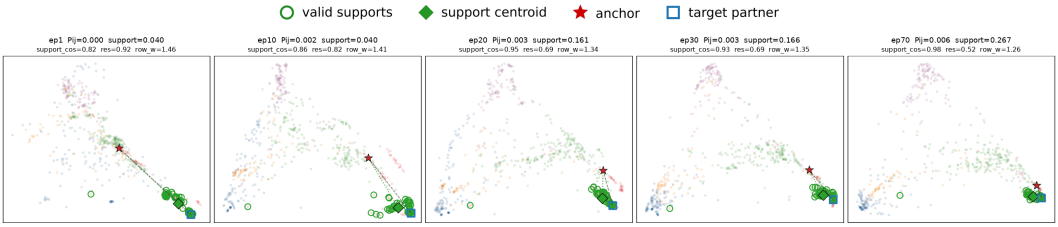


**Figure 5: Latent-space evolution of a representative hard pair on SleepEDFx.** The red star is the anchor, the blue square is the target partner, the green circles are valid cross-domain supports, and the green diamond is the support centroid. During training, the anchor moves toward the valid-support neighborhood, making the local cross-domain structure more compact and class-consistent.

## O More Learned Transport Visualization

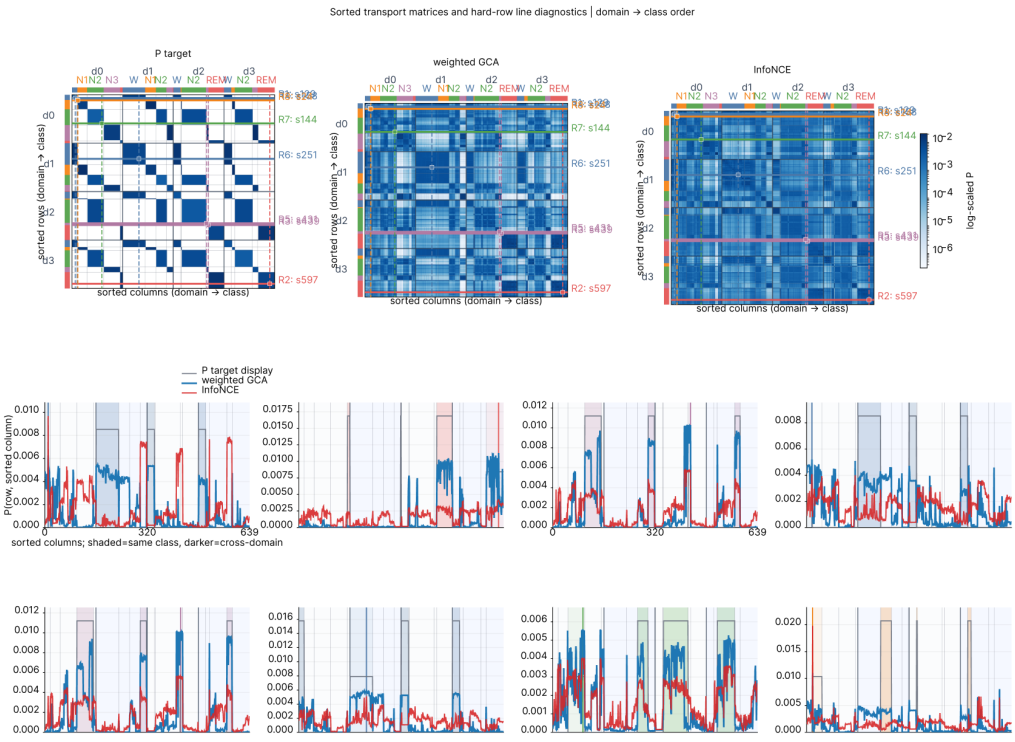


**Figure 6:** Domain-first view of the same transport matrices used in Figure 1. Rows and columns are sorted first by source domain and then by class, making cross-domain same-class relations appear across repeated class blocks in different domains. The top row compares the target plan, the learned plan of MULTA, and the InfoNCE affinity. The bottom panels show row-wise coupling profiles for representative hard anchors, where shaded regions indicate same-class columns and darker shading marks cross-domain same-class entries. Compared with InfoNCE, MULTA assigns more mass to same-class pairs across domains and reduces mass on different-class regions.

# P ALGORITHMIC DETAILS

Algorithm 1 summarizes the transport-matching objective for one source batch, using the notation defined in Sec. 3. The target plan and anchor weights are held fixed during differentiation, while gradients through the learned plan update the encoder. Finite-iteration effects are discussed in App. F.

**Algorithm 1** MULTA objective for one labeled source batch

**Require:** Encoder features $\{\mu_i\}_{i=1}^{B}$, labels $y$, source domains $d$, and method hyperparameters
**Ensure:** Weighted transport-matching loss $L_w$

Set uniform marginals $u \leftarrow B^{-1}\mathbf{1}$
Normalize features: $z_i \leftarrow \mu_i / \|\mu_i\|_2$
Construct affinities: $K_{ij} \leftarrow \exp(z_i^\top z_j / \tau)$
Compute the learned plan: $P_\theta \leftarrow \mathrm{Sinkhorn}(K; u, u)$
Construct the class–domain target $T$ from $(y, d)$ following App. A.1
Obtain pairing distributions: $t \leftarrow BT$, $q \leftarrow BP_\theta$
Compute missing same-class cross-domain probability: $\Delta_i^+ \leftarrow [\sum_j M_{ij}^+ (t_{ij} - q_{ij})]_+$
Compute different-class leakage: $\Delta_i^- \leftarrow [\sum_j M_{ij}^- (q_{ij} - t_{ij})]_+$
Combine gaps: $\zeta_i \leftarrow (\beta_+ \Delta_i^+ + \beta_- \Delta_i^-)/(\beta_+ + \beta_-)$
Normalize gaps within the batch and obtain $w_i = 1 + \lambda_h \hat{h}_i$ using Eq. (2)
Compute per-anchor matching loss: $\ell_i \leftarrow -\sum_j t_{ij} \log q_{ij}$
**return** $L_w \leftarrow \sum_i w_i \ell_i / \sum_i w_i$