STaRC: Spectral–Temporal Cross-Prediction and Bayesian Relational Calibration for Cross-Dataset Few-Shot Activity Recognition

Anonymous authors
Paper under double-blind review

Abstract

Cross-dataset few-shot human activity recognition (HAR) requires learning transferable motion representations across heterogeneous sensing conditions while adapting to target activities from few labeled examples. Existing approaches often treat cross-dataset representation learning and few-shot adaptation separately, leaving uncertainty in transferring heterogeneous base knowledge under scarce target supervision insufficiently modeled. We propose **STaRC**, a two-stage framework that integrates **Spectral–Temporal Cross-Prediction (STCP)** for pretraining with **Bayesian Relational Calibration (BRC)** for few-shot adaptation. We harmonize ten public IMU datasets into a common input format for self-supervised pretraining without requiring aligned activity labels. The resulting harmonized multi-dataset collection will be publicly released. STCP learns transferable representations through bidirectional cross-prediction between temporal and spectral views, capturing complementary temporal dynamics and spectral regularities beyond direct cross-view alignment. During few-shot adaptation, BRC probabilistically models relations between target features and base-class prototypes, samples multiple relational graphs to calibrate target representations, and integrates predictions from multiple base-knowledge experts through task-conditioned routing. This component transfers heterogeneous base knowledge while accounting for relational uncertainty induced by cross-dataset shift and limited supervision. Experiments across target datasets and shot settings demonstrate the effectiveness of both STCP and BRC. Under the same downstream calibration protocol, comparisons with representative self-supervised methods show the benefits of STCP pretraining. Comparisons with competitive few-shot baselines further support the effectiveness of BRC in transferring base knowledge to target episodes, particularly in low-shot settings. Ablation studies examine the effects of pretraining data scale, temporal and spectral learning objectives, and Bayesian calibration compared with point-estimate or no calibration. These results suggest that predictive representation learning and uncertainty-aware relational calibration provide a complementary approach to data-efficient cross-dataset HAR, with potential for personalized wearable sensing.

1 Introduction

Human Activity Recognition (HAR) supports a broad range of applications, including health monitoring, rehabilitation assessment, elder care, and human-computer interaction (Wang et al., 2019; Vrigkas et al., 2015). Compared with recognition systems that depend on cameras or microphones, inertial measurement unit (IMU) sensors continuously capture human motion through wearable devices while preserving visual privacy and remaining applicable across diverse environments (Antar et al., 2019; Li et al., 2021). In practical deployments, however, target activities, user populations, and sensing conditions often differ from the data available for training. Collecting sufficient annotations for every new environment also incurs substantial cost. These constraints motivate cross-dataset few-shot IMU HAR, which transfers knowledge from multiple existing datasets and recognizes activities in a target environment from only a few labeled examples.

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.

Cross-dataset few-shot IMU HAR involves two related challenges. First, IMU datasets differ in sampling frequency, sensor type and placement, channel configuration, activity execution, and label taxonomy (Yu et al., 2025b). These differences alter both temporal evolution and spectral structure, which limits the transferability of representations learned from heterogeneous datasets. Second, a small support set cannot reliably estimate target-class distributions, which motivates transferring statistics from data-rich base classes (Yang et al., 2021a). However, target-base similarities can exhibit different levels of uncertainty, and treating them as equally reliable may misguide the transfer of base knowledge (Zhang et al., 2021). Cross-dataset few-shot HAR therefore requires transferable motion features and uncertainty-aware calibration of base knowledge.

Recent advances in self-supervised time-series learning, masked modeling, and wearable pretraining have improved the transferability of IMU representations (Huang et al., 2025; Cheng et al., 2023; Hao et al., 2023; Ek et al., 2025). Existing methods commonly learn general features through signal reconstruction, instance discrimination, or cross-view consistency, but do not fully capture the predictive structure connecting temporal dynamics with spectral regularities. Few-shot HAR and cross-domain adaptation methods have explored metric learning, prototype refinement, attention mechanisms, and transductive inference (Deng et al., 2023; Liu et al., 2025a; Yang et al., 2021b). Most treat base-class prototypes and their relations to target samples as deterministic information, making it difficult to represent uncertain target-base alignment or select appropriate sources of base knowledge from an extremely small support set. Consequently, cross-dataset predictive representation learning and uncertainty-aware episodic calibration remain insufficiently integrated.

To address these challenges, we propose **STaRC**, a two-stage framework for cross-dataset few-shot activity recognition. Spectral-Temporal Cross-Prediction (**STCP**) learns transferable IMU representations through bidirectional prediction between temporal and spectral views during self-supervised pretraining on harmonized data from multiple datasets, without requiring aligned activity labels. Bayesian Relational Calibration (**BRC**) adapts these representations within each target episode by modeling uncertainty in target-base relations and adaptively weighting base-knowledge experts according to the target task. Together, the two stages support representation transfer across heterogeneous sensing conditions while reducing unreliable knowledge transfer under limited target supervision.

We evaluate STaRC across multiple target datasets and shot settings under a unified cross-dataset protocol, with comparisons against representative few-shot classification, HAR-specific transfer learning, and self-supervised pretraining methods. The results demonstrate the effectiveness and complementarity of STCP and BRC, while ablation studies assess the effects of pretraining data scale, STCP objectives, and Bayesian relational calibration.

Our main contributions are summarized as follows:

- We introduce **STaRC**, a two-stage framework for cross-dataset few-shot IMU activity recognition that couples spectral–temporal predictive pretraining with Bayesian relational calibration. STaRC addresses complementary challenges in transferring representations across datasets and adapting them to sparsely labeled target tasks.

- We develop **STCP**, a self-supervised pretraining approach based on bidirectional cross-prediction between temporal and spectral views. By learning to predict across these complementary views, STCP captures temporal dynamics and spectral structure without requiring aligned activity labels across datasets.

- We propose **BRC**, which performs Bayesian inference over target-base relations and represents their uncertainty through sampled relational graphs. These graphs guide episode-level feature calibration, with base-knowledge experts and task-conditioned routing supporting the adaptive aggregation of relevant base knowledge.

- We harmonize ten public IMU datasets into a unified pretraining resource and will release the resulting dataset to support reproducible research and further advances in cross-dataset IMU representation learning.

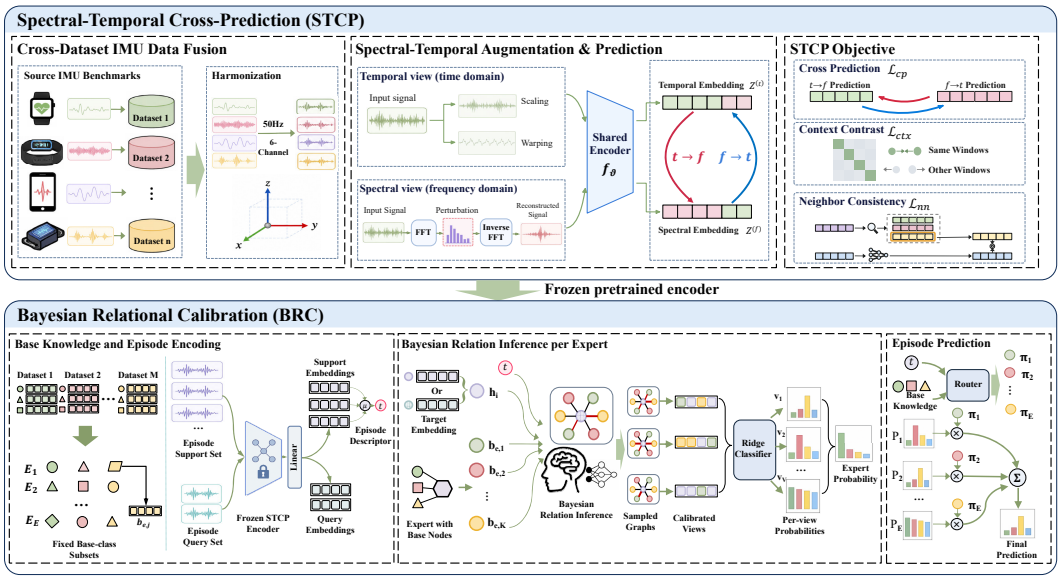


Figure 1: Overview of the proposed STaRC framework.

2 METHOD

We study cross-dataset few-shot activity recognition from IMU signals. Let $\mathcal{D}_{\text{pre}} = \{\mathbf{x}_i\}_{i=1}^{N_{\text{pre}}}$ denote an unlabeled pretraining collection assembled from heterogeneous IMU datasets, where $\mathbf{x}_i \in \mathbb{R}^{T \times C}$ is a sensor window of length T with C aligned channels. We set $C = 6$ for tri-axial accelerometer and gyroscope signals. The constituent datasets need not share an activity-label space. STCP receives only sensor windows during pretraining, whereas BRC subsequently accesses labeled base-class data to construct its base-knowledge bank.

The target task comprises a sequence of N-way K-shot episodes. Each episode is denoted by $\mathcal{E} = (\mathcal{S}, \mathcal{Q})$, where

$$\mathcal{S} = \left\{(\mathbf{x}^s_{c,k}, c)\right\}_{c=1,k=1}^{N,K}, \qquad \mathcal{Q} = \left\{(\mathbf{x}^q_{c,r}, c)\right\}_{c=1,r=1}^{N,Q_c}. \tag{1}$$

The support set $\mathcal{S}$ provides K labeled examples from each target class, and the query set $\mathcal{Q}$ contains Q_c examples per class. Support labels are available when predicting query classes. Query labels contribute only to episodic training or evaluation and remain unavailable during prediction.

As illustrated in Figure 1, STaRC comprises two stages. Spectral–Temporal Cross-Prediction (STCP) learns transferable motion representations from harmonized multi-dataset IMU signals through bidirectional prediction between temporal and spectral views. Bayesian Relational Calibration (BRC) subsequently adapts these representations by inferring uncertain target-base relations and sampling multiple relation graphs for episode-level feature calibration. An episode-conditioned router aggregates predictions from experts associated with different subsets of base knowledge.

2.1 SPECTRAL–TEMPORAL CROSS-PREDICTION

STCP learns cross-dataset motion representations through bidirectional prediction between temporal and spectral embeddings derived from harmonized IMU windows. This objective captures the predictive relationship between temporal evolution and spectral regularities while requiring no activity-label alignment across datasets.

Cross-dataset input harmonization. We harmonize ten public IMU datasets into a common input format. Signals are resampled to 50 Hz and segmented into fixed-length windows. For datasets containing multiple sensor locations, each IMU stream forms an individual pretraining sample. We retain tri-axial accelerometer and gyroscope measurements in a consistent six-channel order, discard other sensing modalities, and fill unavailable gyroscope channels with zeros.

Temporal and spectral view construction. For each harmonized window $\mathbf{x}_i$, STCP constructs temporal and spectral views:

$$\mathbf{x}_i^{(t)} = \mathcal{A}_t(\mathbf{x}_i), \qquad \mathbf{x}_i^{(f)} = \mathrm{Re}\left[\mathcal{F}^{-1}\left(\mathcal{A}_f(\mathcal{F}(\mathbf{x}_i))\right)\right]. \tag{2}$$

The temporal operator $\mathcal{A}_t$ applies transformations such as noise injection, amplitude scaling, permutation, temporal warping, resampling, and three-dimensional rotation. The spectral operator $\mathcal{A}_f$ applies frequency filtering and amplitude or phase perturbations. Here, $\mathcal{F}$ denotes a two-dimensional Fourier transform over the temporal and channel axes. Its inverse reconstructs a real-valued spectral view with the same $T \times C$ dimensions as the temporal view. Appendix B.3 specifies the complete transformations and their parameters.

Shared encoding and bidirectional cross-prediction. The shared encoder produces $\mathbf{Z}_i^{(r)} = f_\vartheta(\mathbf{x}_i^{(r)}) = [\mathbf{z}_{i,1}^{(r)}, \ldots, \mathbf{z}_{i,L}^{(r)}] \in \mathbb{R}^{L \times D}$ for $r \in \{t, f\}$, where L and D denote the sequence length and embedding dimension. For a direction $a \to b$, with $(a, b) \in \{(t, f), (f, t)\}$, the context module g summarizes $\mathbf{Z}_{i,1:\tau_a}^{(a)}$ at a sampled position τ_a. The predictor P estimates K_p embeddings of the other view, denoted by $\widehat{\mathbf{Z}}_i^{(b \leftarrow a)} = P(g(\mathbf{Z}_{i,1:\tau_a}^{(a)}))$. The two directions sample their context positions independently.

At offset $k \in \{1, \ldots, K_p\}$, the corresponding embedding $\mathbf{z}_{i,\tau_a+k}^{(b)}$ provides the prediction target. The prediction from the same sample forms the positive pair, while predictions from other minibatch samples provide contrasting candidates. Defining the dot-product score as $s(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$, the bidirectional cross-prediction loss is

$$\mathcal{L}_{\mathrm{cp}} = -\frac{1}{2BK_p} \sum_{(a,b)} \sum_{i=1}^{B} \sum_{k=1}^{K_p} \log \frac{\exp\left(s\left(\mathbf{z}_{i,\tau_a+k}^{(b)}, \widehat{\mathbf{z}}_{i,k}^{(b \leftarrow a)}\right)\right)}{\sum_{j=1}^{B} \exp\left(s\left(\mathbf{z}_{i,\tau_a+k}^{(b)}, \widehat{\mathbf{z}}_{j,k}^{(b \leftarrow a)}\right)\right)}, \tag{3}$$

where B is the minibatch size. This bidirectional objective requires temporal and spectral contexts to recover complementary structures from each other.

STCP further introduces two auxiliary objectives to complement cross-prediction. The context contrast loss $\mathcal{L}_{\mathrm{ctx}}$ operates on the projected context embeddings of the temporal and spectral views. Embeddings derived from the same IMU window form positive pairs, while embeddings from other windows in the minibatch provide negative pairs. These correspond to the diagonal and off-diagonal entries of the cross-view similarity matrix, respectively. By increasing the relative similarity of same-window pairs, this objective aligns contextual information across the two views while preserving distinctions between different windows.

The neighbor consistency loss $\mathcal{L}_{\mathrm{nn}}$ extends this supervision to neighboring representations retrieved from a memory bank. One view queries the bank through its projected context embedding and retrieves the nearest neighbor according to cosine similarity. A prediction head transforms the context embedding of the other view, and a contrastive objective encourages the resulting prediction to agree with the retrieved neighbor. The same procedure is applied with the two views exchanged. This objective encourages cross-view consistency with neighboring representations beyond the two augmentations of an individual window, without requiring activity labels.

The complete pretraining objective combines cross-prediction, context contrast, and neighbor consistency:

$$\mathcal{L}_{\mathrm{STCP}} = \lambda_1 \mathcal{L}_{\mathrm{cp}} + \lambda_2 \mathcal{L}_{\mathrm{ctx}} + \lambda_3 \mathcal{L}_{\mathrm{nn}}, \tag{4}$$

where λ_1, λ_2, and λ_3 balance the three objectives. Appendix B.4 provides the detailed loss definitions and weighting coefficients. After pretraining, the shared encoder serves as the frozen representation backbone for BRC.

2.2 Bayesian Relational Calibration

Although STCP produces transferable motion representations, the limited support set in a target episode provides insufficient evidence to reliably identify which base knowledge should be transferred. BRC addresses this challenge by modeling posterior distributions over target-base relations.

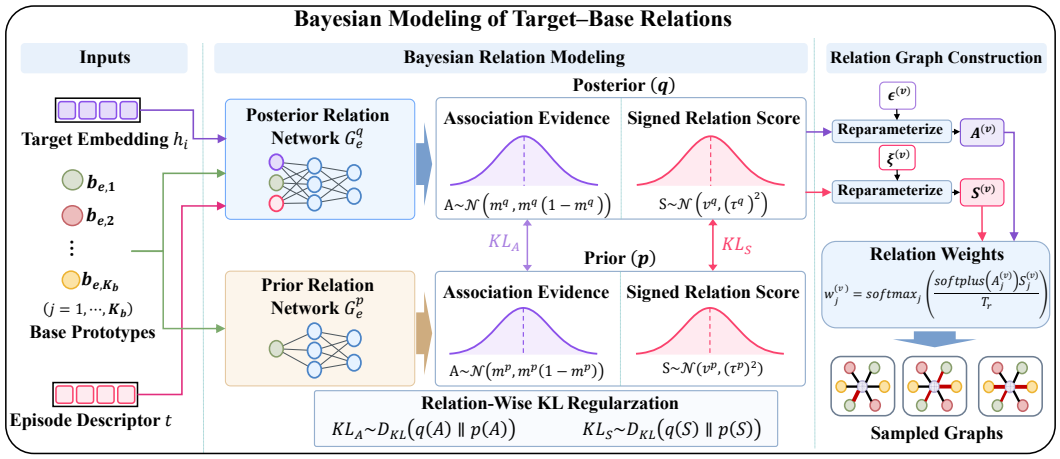


Figure 2: Bayesian relation inference within each expert. An episode-conditioned posterior models association evidence A and signed relation scores S, regularized against a base-conditioned prior through separate KL terms. Reparameterized posterior samples produce weighted target-base relation graphs.

Samples from these posteriors induce multiple plausible relation graphs for calibrating target representations with expert-specific base knowledge, while base-conditioned priors regularize relational inference during episodic training. Different experts operate on predefined, potentially overlapping subsets of base knowledge, and an episode-conditioned router combines their predictions according to the current support set.

Base knowledge and episode conditioning. BRC freezes the pretrained encoder f_ϑ and maps its representations into a relation space through a trainable linear layer ϕ_θ. For each support or query sample $\mathbf{x}_i$, the resulting target embedding is $\mathbf{h}_i = \phi_\theta(f_\vartheta(\mathbf{x}_i))$. To represent the available base knowledge, BRC averages the pretrained representations within each labeled base class r and applies the same linear layer to obtain $\mathbf{b}_r = \phi_\theta\big(|\mathcal{D}_r^b|^{-1} \sum_{\mathbf{x} \in \mathcal{D}_r^b} f_\vartheta(\mathbf{x})\big)$, where $\mathcal{D}_r^b$ denotes the samples of class r in the pretraining datasets. These prototypes summarize class-level base knowledge and serve as the base nodes for relational inference. Expert e receives a predefined subset $\mathcal{B}_e = \{\mathbf{b}_{e,j}\}_{j=1}^{K_b}$, which specifies the base knowledge available to that expert when calibrating target embeddings. The current episode is summarized by the episode descriptor $\mathbf{t} = |\mathcal{S}|^{-1} \sum_{\mathbf{x}_i \in \mathcal{S}} \mathbf{h}_i$. This descriptor aggregates information across the support set and provides a shared task context for inferring relations between individual target embeddings and base prototypes.

Binomial-inspired relational prior and posterior. Inspired by the view that human relational reasoning can emerge from the unconscious accumulation of weak binary association judgments (Huang et al., 2020; Liu & Jia, 2023), BRC describes each target-base relation through a binomial-inspired latent process. Each judgment corresponds to a Bernoulli event, and the accumulated evidence from n independent judgments follows $C_{e,i,j} \sim \text{Binomial}(n, \lambda_{e,i,j})$, where $\lambda_{e,i,j}$ denotes the activation probability. This formulation considers numerous judgments with vanishing activation probabilities, allowing weak events to accumulate into relational evidence. Direct inference, however, requires handling both the large number of judgments and their near-zero activation probabilities. BRC therefore approximates relation sampling with a continuous Gaussian proxy $A \sim \mathcal{N}(m, m(1-m))$, where $m = n\lambda \in (0, 1/2)$. This parameterization uses the aggregate quantity m rather than estimating n and λ separately. allowing relational evidence to be sampled through reparameterization. The additional discrepancy between this proxy and the Gaussian matching the binomial mean and variance admits a small finite bound. Specifically, the KL divergence from the proxy to the moment-matched Gaussian is bounded above by $(\log 2 - 1/2)/2$, as proved in Appendix B.1.

BRC further introduces a Gaussian variable $S \sim \mathcal{N}(\nu, \tau^2)$ to represent the signed relation score, where ν specifies the expected score and τ captures variation across sampled relation graphs. The

two variables jointly determine how each target-base relation contributes to the subsequent computation of graph weights. As illustrated in Figure 2, the posterior network conditions on the target embedding, base-class prototype, and episode descriptor, allowing relational inference to account for both the individual target-base pair and the current support task. The prior network depends only on the base-class prototype and provides a reference for the inferred relation:

$$\begin{aligned} \left(m^q_{e,i,j}, \nu^q_{e,i,j}, \tau^q_{e,i,j}\right) &= \mathcal{G}^q_e\left([\mathbf{h}_i \| \mathbf{b}_{e,j} \| \mathbf{t}]\right), \\ \left(m^p_{e,j}, \nu^p_{e,j}, \tau^p_{e,j}\right) &= \mathcal{G}^p_e(\mathbf{b}_{e,j}), \end{aligned} \tag{5}$$

where $\mathcal{G}^q_e$ and $\mathcal{G}^p_e$ denote the posterior and prior relation networks of expert e, respectively. BRC draws posterior samples to construct relation graphs for feature calibration. Suppressing the expert and target-base indices, the samples for view v are

$$\begin{aligned} A^{(v)} &= m^q + \sqrt{m^q(1-m^q)}\,\epsilon^{(v)}, \\ S^{(v)} &= \nu^q + \tau^q \xi^{(v)}, \qquad \epsilon^{(v)}, \xi^{(v)} \sim \mathcal{N}(0,1), \end{aligned} \tag{6}$$

where $\epsilon^{(v)}$ and $\xi^{(v)}$ are independent standard Gaussian variables. Each sampled pair supplies the association evidence and signed score for one relation view. During episodic training, separate KL terms regularize the posterior distributions of A and S against their corresponding base-conditioned priors. These terms constrain deviations in association evidence and signed relation scores, respectively, while allowing the posterior to adapt to the target embedding and support-task context.

Uncertainty-aware relational calibration. BRC samples multiple relation graphs for each target-expert pair, with each graph providing one view of the uncertain target-base relations. In the v-th graph, the sampled association evidence and signed relation weights are normalized over the base prototypes to calibrate the target representation:

$$\begin{aligned} w^{(v)}_{e,i,j} &= \mathrm{softmax}_j\left(\frac{\mathrm{softplus}(A^{(v)}_{e,i,j}) S^{(v)}_{e,i,j}}{T_r}\right), \\ \mathbf{f}^{(v)}_{e,i} &= \omega \mathbf{h}_i + (1-\omega)\sum_{j=1}^{K_b} w^{(v)}_{e,i,j} \mathbf{b}_{e,j}. \end{aligned} \tag{7}$$

Here, T_r is the temperature parameter for relation-weight normalization. The normalized weights specify the relative contributions of the base prototypes, and ω balances their weighted aggregation with the original target embedding. Each sampled graph therefore produces a calibrated representation $\mathbf{f}^{(v)}_{e,i}$ that combines target information with a particular mixture of base knowledge. Variation across the sampled relation weights propagates target-base uncertainty into multiple calibrated views for subsequent episodic prediction.

Episodic prediction and learning. Each expert fits an episode-specific ridge classifier to all calibrated support views and their labels. The classifier predicts probabilities for the calibrated query views, which BRC averages across relation samples to obtain $P_{e,i}(c)$. An episode-conditioned router assigns each expert a weight π_e according to the episode descriptor and the expert's base knowledge. A target-prototype branch computes class probabilities from distances to class prototypes formed from uncalibrated support embeddings. The final prediction combines the two branches as $P_i(c) = (1-\eta)\sum_e \pi_e P_{e,i}(c) + \eta P^{\mathrm{proto}}_i(c)$, where η controls the contribution of the target-prototype branch.

BRC regularizes relational inference through two posterior-to-prior Gaussian KL divergences. For association evidence, $\mathrm{KL}(q_A \| p_A)$ constrains the episode-conditioned evidence distribution relative to its base-conditioned reference. For signed relation scores, $\mathrm{KL}(q_S \| p_S)$ penalizes deviations in both the expected score and its uncertainty. Together, these terms constrain changes in the inferred relation distributions under limited episodic supervision. Let $\overline{\mathcal{K}}_A$ and $\overline{\mathcal{K}}_S$ denote their respective averages over support and query relations across all experts and base prototypes. BRC minimizes

$$\mathcal{L}_{\mathrm{BRC}} = -\frac{1}{|\mathcal{Q}|}\sum_{i \in \mathcal{Q}} \log P_i(y_i) + \lambda_A \overline{\mathcal{K}}_A + \lambda_S \overline{\mathcal{K}}_S. \tag{8}$$

The first term supervises the final episode predictions, while λ_A and λ_S control the two regularization terms. Query labels supervise episodic training and are not required at inference.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We construct the harmonized pretraining collection from UCI HAR (Reyes-Ortiz & Parra, 2013), WISDM (Weiss, 2019), PAMAP2 (Reiss, 2012), REALDISP (Banos & Amft, 2012), DaLiAc (Leutheuser et al., 2013), HAPT (Reyes-Ortiz et al., 2016), MHEALTH (Banos & Saez, 2014), MotionSense (Malekzadeh et al., 2018), RealWorld2016 (Sztyler & Stuckenschmidt, 2016), and USC-HAD (Zhang & Sawchuk, 2012). We evaluate few-shot recognition on PAMAP2 and UCI HAR. PAMAP2 is partitioned by activity class, with target classes separated from those used for training and meta-learning. UCI HAR contains six classes shared across all splits, which comprise different samples from the same classes. We evaluate 8-way PAMAP2 and 6-way UCI HAR episodes with $K \in \{1, 2, 4, 8, 16\}$ support examples and ten query examples per class over 1,000 episodes. The single-data setting draws pretraining data from one dataset, whereas the full-data setting adopts the complete harmonized collection. Further implementation details are provided in Appendix C.1.

3.2 COMPARISON EXPERIMENTS

Baselines and comparison protocol. We compare STaRC with two groups of methods corresponding to its two stages. The pretraining baselines include BYOL (Grill et al., 2020), SimCLR (Chen et al., 2020), SimSiam (Chen & He, 2020), TS-TCC (Eldele et al., 2021), TFC (Huang et al., 2025), TimeDART (Wang et al., 2025), and CrossHAR (Hong et al., 2024). The few-shot baselines include Matching Networks (MN) (Vinyals et al., 2016), Prototypical Networks (Snell et al., 2017), FRN (Wertheimer et al., 2021), CDN4 (Li et al., 2025), MFSIN (Yu et al., 2025a), SWAT (Liu et al., 2025b), and UNEM (Zhou et al., 2025). When their architectures permit component replacement, pretraining methods share the BRC adaptation stage and few-shot methods receive representations from the STCP encoder. Methods with an inseparable native prediction pipeline retain their original prediction head under the same episodic evaluation protocol. We report results under single-data and full-data settings.

In the single-data setting, both encoder pretraining and episodic training use PAMAP2 data for the PAMAP2 and UCI HAR evaluations. PAMAP2 reserves distinct activity classes for testing, allowing evaluation on unseen classes within the same dataset. For UCI HAR, its six activity classes offer limited scope for a class-disjoint split. We therefore train on PAMAP2 and evaluate on UCI HAR, which also measures cross-dataset transfer. In the full-data setting, STaRC and the component-matched baselines pretrain on the harmonized ten-dataset collection, conduct episodic training on the target benchmark's validation split, and evaluate on its test split.

Results. Table 1 shows that STaRC achieves the highest accuracy among the compared methods on both datasets across all shot settings. Moving from the single-data to the full-data setting improves its PAMAP2 accuracy by approximately 10% on average. The broader pretraining collection exposes the encoder to more varied motion and sensing conditions.

Within the full-data setting, time-series pretraining methods such as TS-TCC and TFC are competitive, indicating the value of objectives tailored to sensor signals. With the downstream stage held consistent in the component-matched comparisons, STaRC's stronger results suggest that STCP's bidirectional spectral-temporal prediction learns more transferable representations. Few-shot methods including FRN and UNEM also perform well in several settings. STaRC's advantage over these methods supports BRC's probabilistic treatment of target-base relations when transferring base-class knowledge to a target episode.

Full-data training does not benefit every baseline. For example, TimeDART (Wang et al., 2025) performs worse on PAMAP2 than in the single-data setting. This may reflect changes in the available base classes or difficulty separating activity representations learned from heterogeneous datasets.

Table 1: Few-shot classification accuracy (%) on PAMAP2 and UCI HAR. The best and second-best results within each data setting are shown in bold and underlined, respectively.

Setting	Group	Method	PAMAP2 (8-way)					UCI HAR (6-way)				
			1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
Single	Pretraining	BYOL 2020	49.30	55.37	58.95	62.52	60.53	42.56	43.19	44.23	45.17	44.94
		TS-TCC 2021	46.16	53.39	55.88	56.81	58.76	41.10	41.80	43.75	45.14	45.23
		CrossHAR 2024	42.04	44.15	46.05	47.75	47.91	47.32	52.70	57.23	60.08	61.66
		TimeDART 2025	50.46	56.18	61.47	64.67	66.39	36.77	39.82	43.44	45.97	48.53
	Few-shot	MN 2016	43.22	48.90	51.10	55.68	65.61	42.09	41.32	39.38	42.50	41.73
		FRN 2021	45.25	52.18	58.76	65.72	71.71	39.99	44.55	49.53	53.02	55.03
		MFSIN 2025a	40.94	43.67	47.19	44.98	47.19	36.49	40.55	32.63	23.21	29.39
		CDN4 2025	50.90	56.44	60.98	66.44	69.30	46.31	47.86	50.06	50.67	50.20
		UNEM 2025	48.08	53.83	57.05	60.12	62.60	42.13	45.47	47.33	51.31	51.05
	Ours	STaRC	**53.72**	**58.74**	**66.26**	**71.48**	**76.65**	**48.24**	**57.07**	**63.32**	**66.18**	**67.83**
Full	Pretraining	SimCLR 2020	54.67	59.58	65.01	66.52	69.35	48.81	55.41	58.75	62.14	65.66
		BYOL 2020	47.67	55.39	61.08	65.34	68.05	57.04	62.00	65.72	67.46	68.01
		SimSiam 2020	48.64	54.10	61.01	64.29	67.50	56.87	61.87	65.62	66.98	67.53
		TS-TCC 2021	58.42	66.93	70.64	70.91	73.01	69.71	75.58	79.53	81.13	81.84
		CrossHAR 2024	57.06	63.82	68.42	70.56	71.94	54.78	60.08	64.24	66.72	67.96
		TFC 2025	56.43	65.67	67.18	72.55	72.94	66.11	72.07	76.35	77.58	78.64
		TimeDART 2025	43.43	48.75	53.79	57.33	59.50	63.27	71.22	75.50	77.26	78.51
	Few-shot	MN 2016	41.36	50.21	54.01	55.38	60.83	56.57	70.23	77.23	82.66	83.87
		ProtoNet 2017	57.05	63.31	66.05	69.09	68.74	56.66	62.12	66.49	68.86	70.61
		FRN 2021	52.77	64.17	73.31	75.96	76.13	52.69	59.58	64.50	67.28	69.00
		MFSIN 2025a	39.03	45.63	41.61	46.92	49.82	61.20	63.80	64.28	66.39	65.64
		CDN4 2025	55.52	63.66	68.52	69.99	74.98	70.69	76.95	80.72	82.22	82.91
		SWAT 2025b	53.34	55.44	57.37	58.15	59.45	66.43	70.87	74.52	76.14	78.45
		UNEM 2025	57.01	66.92	72.50	76.29	79.93	70.96	78.08	79.76	82.27	79.30
	Ours	STaRC	**62.29**	**71.08**	**77.48**	**81.44**	**82.86**	**73.27**	**79.65**	**83.25**	**84.40**	**84.83**

Table 2: Ablation results on PAMAP2 (8-way). Panel (a) lists the number of pretraining samples and datasets. All results are classification accuracy (%).

(a) Pretraining data scale

Samples (datasets)	PAMAP2, K-shot 1	2	4	8	16
14,790 (1)	53.72	58.74	66.26	71.48	76.65
53,837 (3)	59.99	65.20	71.95	74.54	79.18
210,790 (6)	61.30	71.02	74.73	77.95	80.92
336,743 (9)	**63.24**	67.50	73.46	78.19	80.75
561,119 (10)	62.29	**71.08**	**77.48**	**81.44**	**82.86**

(b) STCP objective

Objective	PAMAP2, K-shot 1	2	4	8	16
Spectral only	53.67	60.54	72.04	71.66	76.30
No cross-prediction	56.37	66.83	75.45	76.95	79.80
Temporal only	60.21	**71.45**	77.33	79.77	81.52
Full STCP (ours)	**62.29**	71.08	**77.48**	**81.44**	**82.86**

(c) Relational calibration

Calibration	PAMAP2, K-shot 1	2	4	8	16
None	54.11	62.02	66.21	69.98	73.44
Point estimate	61.73	68.92	76.38	80.51	82.20
Bayesian (ours)	**62.29**	**71.08**	**77.48**	**81.44**	**82.86**

3.3 Ablation Studies

We examine pretraining data scale, the STCP prediction objective, and BRC relation modeling on PAMAP2 under the 8-way few-shot setting. Table 2 reports classification accuracy across five shot levels.

Pretraining data scale. Table 2(a) examines how the size of the harmonized pretraining collection affects recognition of PAMAP2 test activities. Accuracy improves overall as more datasets are included. The nine-dataset setting leads at 1-shot, and the full collection leads at the other four shot levels. Some results decline when the collection grows from six to nine datasets, suggesting that new sources can introduce different signal distributions and sensing conditions. Aligning sampling frequencies and channels reduces input-format differences, but effective transfer still requires a pretraining model that generalizes across the remaining distribution shifts.

Spectral-temporal cross-prediction. Table 2(b) compares full STCP with prediction between two temporal views, prediction between two spectral views, and a two-view objective without cross-prediction. Temporal-only prediction consistently outperforms spectral-only prediction, indicating

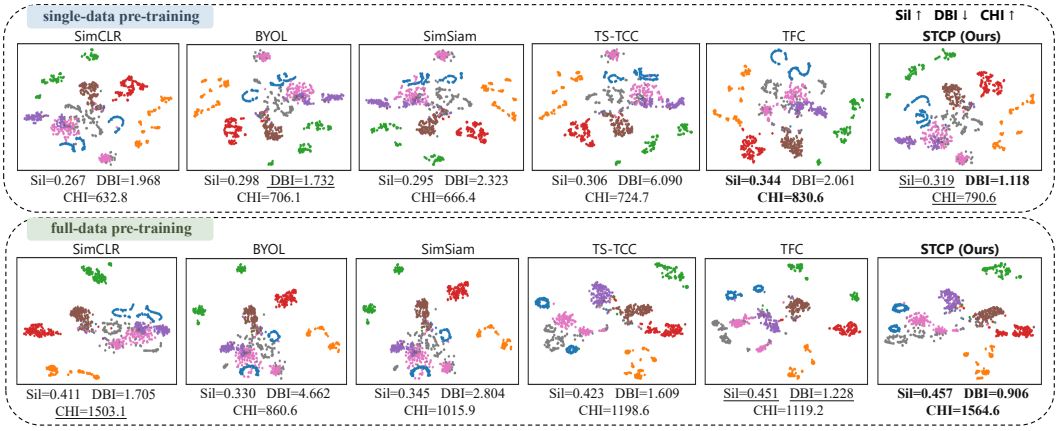


Figure 3: t-SNE visualizations of pretrained features from the held-out PAMAP2 test classes. Higher Silhouette and CHI values, together with lower DBI values, indicate greater class separation relative to within-class dispersion.

that temporal dynamics play a stronger role in this IMU recognition setting. Full STCP achieves the highest accuracy in four of the five shot settings and remains close to the best result at 2-shot; it also outperforms the variant without cross-prediction at every shot level. These results indicate that spectral regularities provide information beyond the temporal view and improve recognition when combined with temporal dynamics through cross-prediction.

Bayesian relational calibration. Table 2(c) compares three levels of target-base relational calibration. With no relational calibration, Meta-ProtoNet classifies queries from support-derived prototypes. Point-estimate calibration adjusts target features using a single estimate of each target-base relation. Bayesian calibration instead infers relation distributions and samples multiple graphs to account for uncertainty during feature calibration. The Bayesian variant achieves the highest accuracy at every shot level, supporting the benefits of both relational calibration and uncertainty modeling.

3.3.1 Visualization of Pretrained Feature Spaces

We visualize pretrained features from the held-out activity classes in the PAMAP2 test set under the single-data and full-data settings. In the single-data setting, each encoder is pretrained on PAMAP2; in the full-data setting, pretraining uses the harmonized ten-dataset collection. All methods use the same test samples, feature normalization, PCA reduction, and t-SNE settings. Figure 3 reports the Silhouette coefficient (Rousseeuw, 1987), Davies–Bouldin index (DBI) (Davies & Bouldin, 1979), and Calinski–Harabasz index (CHI) (Caliński & Harabasz, 1974) alongside each visualization.

In the single-data setting, STCP achieves the lowest DBI and ranks second on Silhouette and CHI. With full multi-dataset pretraining, it ranks first on all three metrics. The improved separation of held-out activities suggests that STCP learns more discriminative motion features from heterogeneous IMU datasets, consistent with its objective of predicting between temporal and spectral views.

4 Conclusion

We introduced STaRC for cross-dataset few-shot IMU activity recognition. By coupling spectral–temporal predictive pretraining with uncertainty-aware relational calibration, STaRC addresses both representation transfer across heterogeneous sensing conditions and the reliable use of base knowledge under scarce target supervision. Results on PAMAP2 and UCI HAR show that the two components provide complementary benefits across few-shot settings. These findings suggest that predictive representation learning and probabilistic relational calibration offer a useful direction for data-efficient wearable activity recognition in new sensing environments. We will release the harmonized ten-dataset IMU pretraining resource to support reproducible research on cross-dataset HAR.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to search for recent related work, and to provide ideas for the design of ablation experiments. We did not use generative AI tools to prove mathematical claims. We also did not use generative AI tools to directly generate full sections of the paper. And the use of AI tools for generating synthetic datasets is not applicable to this work.

We have reviewed all AI-assisted work. Specifically, all AI-polished text was read and checked by the authors to ensure correctness. The authors verified that the AI-retrieved related work is real and relevant, and all citations to related work were manually imported and formatted by the authors using official bibliographic information where available. The authors also reviewed the experimental ideas and any related code suggested with LLM assistance and verified their correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available benchmark datasets. Datasets used in this study do not contain patient-identifiable information, such as names, addresses, or other metadata that could directly identify individuals. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

We provide the main STaRC implementation and preprocessing code for a subset of the datasets in the supplementary materials. The appendix reports the hyperparameters used in our experiments. Together with the method description in the main text, these materials support reproduction of our experimental results.

REFERENCES

Anindya Das Antar, Masud Ahmed, and Md Atiqur Rahman Ahad. Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 134–139. IEEE, 2019.

Garcia Rafael Banos, Oresti and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TW22.

Toth Mate Banos, Oresti and Oliver Amft. REALDISP Activity Recognition Dataset. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5GP6D.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Dongzhou Cheng, Lei Zhang, Can Bu, Hao Wu, and Aiguo Song. Learning hierarchical time series data augmentation invariances via contrastive supervision for human activity recognition. *Knowledge-Based Systems*, 276:110789, 2023. doi: 10.1016/j.knosys.2023.110789.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

Shizhuo Deng, Zhubao Guo, Da Teng, Boqian Lin, Dongyue Chen, Tong Jia, and Hao Wang. Self-relation attention networks for weakly supervised few-shot activity recognition. *Knowledge-Based Systems*, 276:110720, 2023. doi: 10.1016/j.knosys.2023.110720.

Yadolah Dodge. De Moivre–Laplace theorem. In *The Concise Encyclopedia of Statistics*, pp. 356–358. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_268. URL https://doi.org/10.1007/978-0-387-32833-1_268.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Sannara Ek, Riccardo Presotto, Gabriele Civitarese, François Portet, Philippe Lalanda, and Claudio Bettini. Comparing self-supervised learning techniques for wearable human activity recognition. *CCF Transactions on Pervasive Computing and Interaction*, 2025. doi: 10.1007/s42486-024-00182-9. URL https://doi.org/10.1007/s42486-024-00182-9.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Dandan Guo, Long Tian, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6996–7010. Curran Associates, Inc., 2022. doi: 10.52202/068431-0507. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2e19dab94882bc95ed094c4399cfda02-Paper-Conference.pdf.

Shilei Hao, Zhihai Wang, Afanasiev D. Alexander, Jidong Yuan, and Wei Zhang. Micos: Mixed supervised contrastive learning for multivariate time series classification. *Knowledge-Based Systems*, 260:110158, 2023. doi: 10.1016/j.knosys.2022.110158.

Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–26, 2024. doi: 10.1145/3659597.

Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pp. 4531–4541. PMLR, 2020.

Zhenzhen Huang, Jiukai Deng, Shengzhi Wang, Chaogang Tang, and Shuo Xiao. Tfc: Time–frequency contrasting network for wearable-based human activity recognition. *Knowledge-Based Systems*, 319:113373, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113373. URL https://www.sciencedirect.com/science/article/pii/S0950705125004204.

Heike Leutheuser, Dominik Schuldhaus, and Bjoern M. Eskofier. Hierarchical, multi-sensor based classification of daily life activities: Comparison with state-of-the-art algorithms using a benchmark dataset. *PLOS ONE*, 8(10):e75196, 2013. doi: 10.1371/journal.pone.0075196. URL https://doi.org/10.1371/journal.pone.0075196.

Jie Li, Xiaofeng Liu, Zhelong Wang, Hongyu Zhao, Tingting Zhang, Sen Qiu, Xu Zhou, Huili Cai, Rongrong Ni, and Angelo Cangelosi. Real-time human motion capture based on wearable inertial sensor networks. *IEEE Internet of Things Journal*, 9(11):8953–8966, 2021.

Xiaoxu Li, Shuo Ding, Jiyang Xie, Xiaochen Yang, Zhanyu Ma, and Jing-Hao Xue. Cdn4: A cross-view deep nearest neighbor neural network for fine-grained few-shot classification. *Pattern Recognition*, 163:111466, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111466. URL https://www.sciencedirect.com/science/article/pii/S0031320325001268.

Jie Liu, Chongben Tao, Zhongwei Shen, Cong Wu, Tianyang Xu, Xizhao Luo, Feng Cao, Zhen Gao, Zufeng Zhang, and Sai Xu. Dual attention focus network for few-shot skeleton-based action recognition. *Knowledge-Based Systems*, 330:114549, 2025a. doi: 10.1016/j.knosys.2025.114549.

Tian Liu, Huixin Zhang, Shubham Parashar, and Shu Kong. Few-shot recognition via stage-wise retrieval-augmented finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2023.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 2nd International Workshop on Privacy by Design in Distributed Systems*, pp. 2:1–2:6, 2018. doi: 10.1145/3195258.3195260.

Shenghuan Miao and Ling Chen. Goat: A generalized cross-dataset activity recognition framework with natural language supervision. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–28, 2024. doi: 10.1145/3699736.

Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

Anguita Davide Ghio Alessandro Oneto Luca Reyes-Ortiz, Jorge and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C54S4K.

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. doi: 10.1016/j.neucom.2015.07.085.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–9, 2016. doi: 10.1109/PERCOM.2016.7456521.

Megha Thukral, Harish Haresamudram, and Thomas Plötz. Cross-domain har: Few-shot transfer learning for human activity recognition. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–35, 2025. doi: 10.1145/3704921.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

Daoyu Wang, Mingyue Cheng, Zhiding Liu, and Qi Liu. TimeDART: A diffusion autoregressive transformer for self-supervised time series representation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62627–62651. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25r.html.

Yan Wang, Shuang Cang, and Hongnian Yu. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, 137:167–190, 2019.

Qingxin Wei, Jiaming Huang, Yi Gao, and Wei Dong. One model to fit them all: Universal imu-based human activity recognition with llm-assisted cross-dataset representation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–22, 2025. doi: 10.1145/3749509.

Gary Weiss. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset . UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5HK59.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8012–8021, 2021.

Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13390–13399, 2020.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021a.

Shuo Yang, Songhua Wu, Tongliang Liu, and Min Xu. Bridging the gap between few-shot and many-shot learning via distribution calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9830–9843, 2021b.

Liyun Yu, Ziyu Guan, Wei Zhao, Yaming Yang, and Jiale Tan. Few-shot fine-grained image classification based on multi-scale selecting and feature interacting. *Knowledge-Based Systems*, 330:114743, 2025a. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.114743. URL https://www.sciencedirect.com/science/article/pii/S0950705125017812.

Xinhui Yu, Arvin Tashakori, Martin J. McKeown, and Z. Jane Wang. A heterogeneous federated learning framework for human activity recognition. *Knowledge-Based Systems*, 326:114008, 2025b. doi: 10.1016/j.knosys.2025.114008.

Mi Zhang and Alexander A. Sawchuk. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the ACM Conference on Ubiquitous Computing Workshop on Situation, Activity and Goal Awareness*, pp. 1036–1043, 2012.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Shih-Fu Chang. Uncertainty-aware few-shot image classification. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3420–3426. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/471. URL https://doi.org/10.24963/ijcai.2021/471. Main Track.

Long Zhou, Fereshteh Shakeri, Aymen Sadraoui, Mounir Kaaniche, Jean-Christophe Pesquet, and Ismail Ben Ayed. UNEM: UNrolled Generalized EM for Transductive Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. CVPR 2025.

A Related Work

A.1 Self-Supervised and Cross-Dataset Representation Learning for IMU Signals

Self-supervised learning is particularly relevant to IMU-based HAR because wearable devices generate abundant sensor streams, whereas activity annotation remains costly. General-purpose methods such as SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), SimSiam (Chen & He, 2020), and NNCLR (Dwibedi et al., 2021) established contrastive, non-contrastive, and nearest-neighbor objectives, while related self-supervised paradigms have also been evaluated for time-series and IMU representation learning (Ek et al., 2025). For time-series and wearable signals, TS-TCC combines temporal and contextual contrasting (Eldele et al., 2021), whereas TFC aligns representations across the temporal and frequency domains (Huang et al., 2025). Masked autoencoding further improves self-supervised representation learning for wearable HAR by reconstructing masked sensor segments (Ek et al., 2025). Recent research has also examined generative time-series pretraining (Wang et al., 2025). These methods primarily learn representations through instance discrimination, cross-view alignment, or signal reconstruction, while the bidirectional predictive relationship between temporal dynamics and spectral regularities remains insufficiently modeled.

Differences in sampling frequency, sensor configuration, placement, and label taxonomy motivate representation learning across heterogeneous IMU datasets. CrossHAR combines masked reconstruction and contrastive learning within a hierarchical self-supervised framework to improve cross-dataset generalization (Hong et al., 2024). GOAT incorporates natural-language supervision to establish semantic correspondences across datasets (Miao & Chen, 2024), while OneHAR introduces LLM-assisted information for universal IMU representation learning (Wei et al., 2025). These studies advance cross-dataset IMU learning through reconstruction, contrastive learning, or cross-modal semantic alignment. STCP focuses on the predictive structure shared by temporal and spectral views and learns cross-dataset motion representations through bidirectional cross-prediction.

A.2 Few-Shot Recognition and Uncertainty-Aware Calibration

Existing few-shot methods primarily follow metric learning, feature reconstruction, and transductive inference. Matching Networks learn support-conditioned matching functions (Vinyals et al., 2016), while Prototypical Networks classify queries according to their distances from class prototypes (Snell et al., 2017). FRN reconstructs query feature maps from support features (Wertheimer et al., 2021), CDN4 performs cross-view nearest-neighbor matching (Li et al., 2025), and MFSIN selects discriminative features at multiple scales and models support–query feature interactions for fine-grained few-shot classification (Yu et al., 2025a). SWAT adopts stage-wise retrieval-augmented adaptation (Liu et al., 2025b), and UNEM unfolds generalized expectation maximization into a learnable transductive inference procedure (Zhou et al., 2025). Within activity recognition, prior studies have explored self-relation attention (Deng et al., 2023), dual-attention adaptation (Liu et al., 2025a), and cross-domain few-shot transfer (Thukral et al., 2025). Although these methods improve support-query alignment, most infer class relations from deterministic distances, prototypes, or attention weights, which cannot adequately represent relational uncertainty when the support set contains very few samples.

Distribution calibration improves target-class estimation by transferring statistics from data-rich base classes. Distribution Calibration selects similar base classes according to feature distance and transfers their means and variances (Yang et al., 2021a). Adaptive Distribution Calibration subsequently learns transfer weights between target samples and base classes through hierarchical optimal transport (Guo et al., 2022). DPGN models instance-level and distribution-level relations through graph propagation (Yang et al., 2020), while Uncertainty-Aware Few-Shot Classification estimates the reliability of query-prototype similarities (Zhang et al., 2021). Despite these advances, calibration methods commonly derive knowledge-transfer decisions from a single relation estimate based on limited support observations, such as neighbor selection, transfer weights, or graph edges. Under substantial cross-dataset shifts, estimation errors can introduce mismatched base knowledge and distort target representations during calibration. BRC treats target-base relations within each expert's predefined base-class subset as latent random variables, infers their posterior distributions, and samples multiple plausible relation graphs to propagate relational uncertainty into target feature

calibration. An episode-conditioned router then assigns weights to expert predictions according to the current support episode.

B Appendix for Method

B.1 Gaussian Relational Proxy and Its Bound

B.1.1 Binomial Motivation and Gaussian Proxy

Inspired by the view that relational reasoning can emerge from the unconscious accumulation of weak binary judgments (Huang et al., 2020), we model each target-base relation through a sequence of independent Bernoulli events. For a fixed relation, we suppress the expert, target, and base-class indices and write

$$B_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda), \qquad C_n = \sum_{k=1}^{n} B_k \sim \text{Binomial}(n, \lambda), \tag{9}$$

where n denotes the number of judgments and λ their activation probability. Following (Liu & Jia, 2023), this construction considers an increasing number of judgments, each with an activation probability approaching zero, allowing individually weak events to accumulate into relational evidence.

The classical Gaussian approximation provides a continuous representation of a binomial variable by matching its mean and variance. For fixed $\lambda \in (0, 1)$, the De Moivre–Laplace theorem establishes convergence of the standardized binomial variable to a standard Gaussian as $n \to \infty$ (Dodge, 2008). When λ varies with n, a sufficient condition for the corresponding normal limit is $n\lambda(1-\lambda) \to \infty$. Writing the aggregate mean as $m = n\lambda$, the moment-matched reference Gaussian is

$$H_{m,n} = \mathcal{N}(n\lambda,\, n\lambda(1-\lambda)) = \mathcal{N}\Big(m,\, m\left(1 - \frac{m}{n}\right)\Big). \tag{10}$$

Moment matching defines this reference distribution for every admissible parameter pair, whereas the accuracy of the normal approximation depends on the parameter regime.

Motivated by this approximation and the relational modeling approach of prior work, BRC adopts a continuous Gaussian proxy parameterized by the aggregate quantity m, rather than estimating n and λ separately:

$$A \sim Q_m, \qquad Q_m = \mathcal{N}(m, m(1-m)), \qquad 0 < m < \frac{1}{2}. \tag{11}$$

The proxy retains the aggregate mean and expresses its variance entirely through m. This parameterization supports differentiable relation sampling through reparameterization and closed-form Gaussian KL regularization.

The adopted proxy should be distinguished from the classical normal limit. In particular, taking $n \to \infty$ while keeping $m = n\lambda$ fixed in $(0, 1/2)$ does not satisfy the variance-growth condition above. We therefore use Q_m as a continuous Gaussian approximation for relational modeling, without asserting asymptotically exact recovery of the binomial distribution in this regime. The analysis below establishes the parameterization of this proxy and bounds its additional discrepancy from $H_{m,n}$. This bound concerns the two Gaussian distributions, rather than the complete approximation error between C_n and A; moreover, $H_{m,n}$ is a mathematical reference distribution, not the learned relational prior.

B.1.2 KL-Based Parameterization of Association Evidence

We justify the parameterization of the association evidence through a KL projection onto the Gaussian proxy family $Q_x = \mathcal{N}(x, x(1-x))$, where $0 < x < 1/2$. Given an auxiliary Gaussian distribution $R = \mathcal{N}(\mu, \sigma^2)$ with $\mu < 1/2$ and $\sigma^2 > 0$, we consider

$$m_* = \arg\min_{0<x<1/2} F(x), \qquad F(x) = D_{\text{KL}}(Q_x \| R). \tag{12}$$

Here, R provides the reference for the parameterizationr.

Proposition. The objective in Eq. equation 12 has a unique minimizer

$$m_* = \frac{1+\ell-\sqrt{1+\ell^2}}{2} = \frac{\ell}{1+\ell+\sqrt{1+\ell^2}}, \qquad \ell = \frac{2\sigma^2}{1-2\mu} > 0, \tag{13}$$

and this minimizer satisfies $0 < m_* < 1/2$.

Proof. Using the KL divergence between two univariate Gaussian distributions, we obtain

$$F(x) = \frac{1}{2}\left[\log\frac{\sigma^2}{x(1-x)} + \frac{x(1-x)+(x-\mu)^2}{\sigma^2} - 1\right]. \tag{14}$$

Since $x(1-x)+(x-\mu)^2 = (1-2\mu)x+\mu^2$, differentiation gives

$$\begin{aligned} F'(x) &= \frac{1}{2}\left[-\frac{1}{x} + \frac{1}{1-x} + \frac{1-2\mu}{\sigma^2}\right], \\ F''(x) &= \frac{1}{2}\left[\frac{1}{x^2} + \frac{1}{(1-x)^2}\right] > 0. \end{aligned} \tag{15}$$

Thus, F is strictly convex on $(0, 1/2)$. Setting $F'(x) = 0$ and clearing the positive denominators yields

$$x^2 - (1+\ell)x + \frac{\ell}{2} = 0, \qquad \ell = \frac{2\sigma^2}{1-2\mu}. \tag{16}$$

For $\ell > 0$, we have $\ell < \sqrt{1+\ell^2} < 1+\ell$. Consequently, the smaller root lies in $(0, 1/2)$, while the larger root exceeds 1. The smaller root is therefore the unique stationary point in the admissible interval and, by strict convexity, its unique global minimizer. Rationalizing its numerator gives the second expression in Eq. equation 13.

B.1.3 A Bound on the Additional Gaussian Discrepancy

The discrepancy between the proxy $Q_m = \mathcal{N}(m, m(1-m))$ and the moment-matched reference $H_{m,n} = \mathcal{N}(m, m(1-m/n))$ can be quantified. These distributions share the same mean and differ only in their variances. The following result holds for every admissible m, independently of how it is parameterized.

Proposition. For every integer $n \geq 1$ and $m \in (0, 1/2)$,

$$0 \leq D_{\mathrm{KL}}(Q_m \| H_{m,n}) \leq \frac{-\log(1-m)-m}{2} < \frac{\log 2 - 0.5}{2}. \tag{17}$$

Proof. Since the two Gaussian distributions have identical means, their KL divergence reduces to

$$D_{\mathrm{KL}}(Q_m \| H_{m,n}) = \frac{1}{2}\left[\log\frac{1-m/n}{1-m} + \frac{1-m}{1-m/n} - 1\right]. \tag{18}$$

Define the variance ratio

$$r = \frac{1-m}{1-m/n}. \tag{19}$$

For $n \geq 1$ and $0 < m < 1/2$, we have $1-m \leq r \leq 1$. Writing $h(r) = (r - 1 - \log r)/2$, Eq. equation 18 becomes $D_{\mathrm{KL}}(Q_m \| H_{m,n}) = h(r)$. Moreover,

$$h'(r) = \frac{1}{2}\left(1 - \frac{1}{r}\right) \leq 0, \qquad 0 < r \leq 1. \tag{20}$$

Therefore, h is nonincreasing on this interval, and

$$0 = h(1) \leq h(r) \leq h(1-m) = \frac{-\log(1-m)-m}{2}. \tag{21}$$

To obtain a uniform bound, let $g(m) = [-\log(1-m) - m]/2$. Its derivative satisfies

$$g'(m) = \frac{m}{2(1-m)} > 0, \qquad 0 < m < \frac{1}{2}. \tag{22}$$

Table 3: Cross-dataset IMU corpus and harmonization procedure. "IMUs" denotes the number of sensor locations available in the original dataset. Each retained IMU stream is represented by three accelerometer and three gyroscope channels. The processed-window counts describe the corpus before the training split is selected.

Dataset	Rate	IMUs	Classes	Processed Windows	Harmonization
UCI HAR	50 Hz	1	6	10,299	Pad the 128-step signals to 256 steps
WISDM	20 Hz	2	18	78,264	Temporal resampling to the common resolution; remove Folding Clothes
PAMAP2	100 Hz	3	18	20,503	Downsample to 50 Hz and retain accelerometer–gyroscope streams
REALDISP	50 Hz	9	33	95,183	Retain six-channel IMU streams; remove Rope Jumping
DaLiAc	200 Hz	4	12	86,376	Downsample to 50 Hz; remove Vacuuming and Rope Jumping
HAPT	50 Hz	1	12	13,291	Segment into fixed-length IMU windows
MHEALTH	50 Hz	3	12	40,893	Select accelerometer and gyroscope measurements
MotionSense	50 Hz	1	6	14,130	Align accelerometer and gyroscope streams
RealWorld2016	50 Hz	6	8	224,376	Organize recordings by sensor location and IMU stream
USC-HAD	50 Hz	1	12	28,089	Segment the synchronized accelerometer–gyroscope signals

Thus, g is strictly increasing, and

$$g(m) < \lim_{u \uparrow 1/2} g(u) = \frac{\log 2 - 0.5}{2}. \tag{23}$$

Combining these inequalities proves Eq. equation 17. Thus, we obtain an approximate Gaussian proxy $\mathcal{N}(m, m(1-m))$ for the binomial distribution without taking $n \to \infty$ and $\lambda \to 0$.

B.2 Cross-Dataset IMU Harmonization

We construct the cross-dataset pretraining corpus from ten public human activity recognition datasets: UCI HAR Reyes-Ortiz & Parra (2013), WISDM Weiss (2019), PAMAP2 Reiss (2012), REALDISP Banos & Amft (2012), DaLiAc Leutheuser et al. (2013), HAPT Reyes-Ortiz et al. (2016), MHEALTH Banos & Saez (2014), MotionSense Malekzadeh et al. (2018), RealWorld2016 Sztyler & Stuckenschmidt (2016), and USC-HAD Zhang & Sawchuk (2012). These datasets cover diverse body locations, sensing devices, activity taxonomies, and collection protocols. As summarized in Table 3, the source datasets range from one to nine IMU locations, contain between 6 and 33 activity classes, and are collected at sampling rates from 20 Hz to 200 Hz.

We first map recordings to a shared temporal scale of 50 Hz. Signals collected above 50 Hz are downsampled, whereas lower-rate recordings are temporally resampled. Each recording is then divided into fixed-length windows of 256 time steps, corresponding to approximately 5.12 seconds at the common sampling rate. For datasets with shorter predefined windows, zero padding is used to preserve the common temporal dimension.

For every sensor location, we retain the tri-axial accelerometer and tri-axial gyroscope measurements and form a canonical six-channel stream,

$$\tilde{\mathbf{v}}_m \in \mathbb{R}^{6\times 256}, \tag{24}$$

where m indexes an IMU location. Magnetometer and dataset-specific auxiliary channels are excluded. Multi-IMU recordings are organized as separate six-channel streams during pretraining, allowing the encoder to learn transferable motion patterns independently of dataset-specific sensor layouts.

The resulting full dataset contains 561,119 harmonized six-channel IMU windows used for cross-dataset self-supervised pretraining. Activity labels are retained for dataset partitioning and downstream few-shot evaluation, while the pretraining objective operates on the harmonized motion streams without requiring a shared activity label space across datasets.

B.3 Temporal and Spectral Augmentations

For each input IMU window, we select one transformation from each family to construct the temporal and spectral views. Temporal transformations operate directly on the signals, whereas spectral transformations modify their two-dimensional Fourier coefficients along the temporal and channel axes. The spectral view is reconstructed by taking the real part of the inverse transform, preserving the original $T \times C$ dimensions. Table 4 summarizes the transformations and their parameters.

Table 4: Temporal and spectral augmentations used in STCP.

Transformation	Operation and parameters
Temporal augmentations	
Identity	Keep the input unchanged.
Gaussian noise	Add noise drawn from $\mathcal{N}(0, 0.8^2)$.
Amplitude scaling	Multiply signals by factors drawn from $\mathcal{N}(2, 1.1^2)$, sampled at each time step and shared across channels.
Sign inversion	Multiply all signal values by -1.
Segment permutation	Divide the sequence into 1–9 random segments and shuffle their order.
Channel permutation	Reorder channels using a fixed random seed of 21.
Temporal reversal	Reverse the sequence along the temporal axis.
Temporal warping	Distort the time coordinates using cubic splines with four interior knots and perturbation standard deviation 0.2.
Resampling	Interpolate at three times the original temporal resolution, then extract T points with stride 2 from a random starting position.
3D rotation	Rotate each three-axis group around a random axis, with the angle drawn from $\mathcal{U}(-\pi, \pi)$.
Permutation with noise	Apply segment permutation followed by Gaussian noise with standard deviation 0.8.
Strong amplitude scaling	Apply amplitude scaling with factors drawn from $\mathcal{N}(2, 2^2)$.
Spectral augmentations	
Identity	Keep the input unchanged.
High-pass filtering	Remove the central rectangular frequency region with half-widths $(32, 2)$ frequency bins.
Low-pass filtering	Retain only the same central rectangular frequency region.
Phase perturbation	Add phase offsets drawn from $\mathcal{U}(-\pi, \pi)$.
Local amplitude–phase perturbation	Add $\mathcal{N}(0, 0.8^2)$ noise to amplitudes and $\mathcal{U}(-\pi, \pi)$ offsets to phases within a contiguous half-band along the temporal-frequency axis.
Global amplitude–phase perturbation	Apply the same amplitude and phase perturbations over the entire spectrum.

B.4 STCP Pretraining Objectives

STCP optimizes pretraining through three complementary objectives: bidirectional cross-prediction, context contrast, and neighbor consistency. The cross-prediction loss follows Eq. 3, and the view transformations are described in Appendix B.3.

Context representations and sampling. Let $\widetilde{\mathbf{Z}}_i^{(r)}$ denote the encoder embeddings normalized along the feature dimension, where $r \in \{t, f\}$. The context module g, projection head ψ, and neighbor prediction head h_{nn} produce

$$\begin{aligned} \mathbf{c}_i^{(r)} &= g\Big(\widetilde{\mathbf{Z}}_{i,1:\tau_r}^{(r)}\Big), \\ \mathbf{u}_i^{(r)} &= \psi(\mathbf{c}_i^{(r)}), \qquad \mathbf{p}_i^{(r)} = h_{\mathrm{nn}}(\mathbf{u}_i^{(r)}). \end{aligned} \tag{25}$$

The temporal and spectral branches independently sample τ_r from $\{1, \ldots, L - K_p\}$, with each sampled position shared across the minibatch. The two branches share the context, projection, and prediction modules. The neighbor prediction head h_{nn} is distinct from the cross-prediction module P, which predicts embeddings at subsequent offsets in the other view.

Context contrast. We first define a symmetric contrastive loss that is shared by the two auxiliary objectives. Given paired sets $U = \{\mathbf{u}_i\}_{i=1}^B$ and $V = \{\mathbf{v}_i\}_{i=1}^B$, concatenate their representations as $(\mathbf{a}_1, \ldots, \mathbf{a}_{2B}) = (\mathbf{u}_1, \ldots, \mathbf{u}_B, \mathbf{v}_1, \ldots, \mathbf{v}_B)$. The positive index is $\ell^+ = \ell + B$ for $\ell \le B$ and

Table 5: STCP objective settings.

Parameter	Description	Value
K_p	Cross-prediction steps	6
T_c	Auxiliary contrastive temperature	0.1
M	Memory-bank capacity	256
$(\lambda_1, \lambda_2, \lambda_3)$	Loss weights	$(1.0, 1.0, 0.5)$

$\ell^+ = \ell - B$ otherwise. Using cosine similarity sim and temperature $T_c > 0$, define

$$\mathcal{C}(U, V) = -\frac{1}{2B} \sum_{\ell=1}^{2B} \log \frac{\exp(\mathrm{sim}(\mathbf{a}_\ell, \mathbf{a}_{\ell^+})/T_c)}{\sum_{\substack{k=1 \\ k \neq \ell}}^{2B} \exp(\mathrm{sim}(\mathbf{a}_\ell, \mathbf{a}_k)/T_c)}. \tag{26}$$

The denominator includes representations from both sets while excluding ℓ itself. For context contrast, the paired representations originate from the temporal and spectral views of the same window:

$$\mathcal{L}_{\mathrm{ctx}} = \mathcal{C}(U^{(t)}, U^{(f)}), \qquad U^{(r)} = \{\mathbf{u}_i^{(r)}\}_{i=1}^B. \tag{27}$$

Representations from other windows serve as negatives. The objective therefore aligns same-window contexts across views while distinguishing different windows.

Neighbor consistency. A shared memory bank $\mathcal{M} = \{\mathbf{m}_j\}_{j=1}^M$ stores detached projected representations. Each view retrieves its nearest neighbor using cosine similarity:

$$j_i^{(r)} = \arg \max_{1 \le j \le M} \mathrm{sim}(\mathbf{u}_i^{(r)}, \mathbf{m}_j), \qquad \mathbf{n}_i^{(r)} = \mathrm{sg}\left(\mathbf{m}_{j_i^{(r)}}\right), \tag{28}$$

where sg denotes stop-gradient. The prediction from one view is paired with the neighbor retrieved by the other view. Writing $P^{(r)} = \{\mathbf{p}_i^{(r)}\}_{i=1}^B$ and $N^{(r)} = \{\mathbf{n}_i^{(r)}\}_{i=1}^B$, we obtain

$$\mathcal{L}_{\mathrm{nn}} = \frac{1}{2}\left[\mathcal{C}(P^{(t)}, N^{(f)}) + \mathcal{C}(P^{(f)}, N^{(t)})\right]. \tag{29}$$

Each contrastive term uses the minibatch predictions and their retrieved neighbors; the entire memory bank is used for retrieval rather than as the loss denominator. The bank is initialized with normalized Gaussian random vectors. Both views query the pre-update bank, after which detached spectral projections overwrite older entries. Neither retrieval nor bank updates require activity labels.

Overall objective and settings. STCP minimizes $\mathcal{L}_{\mathrm{STCP}} = \lambda_1 \mathcal{L}_{\mathrm{cp}} + \lambda_2 \mathcal{L}_{\mathrm{ctx}} + \lambda_3 \mathcal{L}_{\mathrm{nn}}$, with nonnegative weighting coefficients. After pretraining, the auxiliary heads are discarded and the shared encoder is frozen for BRC.

B.5 Network Architecture and Implementation Details

We describe the network architecture and implementation configuration used for the PAMAP2 experiments. The principal settings are summarized in Table 6.

Encoder and prediction heads. The shared encoder processes six-channel IMU windows of length 256. A convolutional patch embedding layer uses a temporal patch size of 16 and stride 8, producing 31 patch tokens. A learned class token and positional embeddings are added before three Transformer blocks, each with embedding dimension 256, eight attention heads, and a feed-forward hidden dimension of 1024. During STCP pretraining, the context module summarizes a token prefix using a four-layer Transformer with dimension 100 and four attention heads. Six offset-specific linear heads map the context representation to 256-dimensional cross-view predictions. The auxiliary context projector has dimensions $100 \to 128 \to 64$, while the neighbor prediction head uses $64 \to 64 \to 64$. Both apply batch normalization and ReLU after the first linear layer.

Table 6: Principal network and inference settings for PAMAP2.

Configuration	Value
Encoder depth / attention heads	$3/8$
Encoder embedding dimension	256
Context depth / attention heads	$4/4$
Context embedding dimension	100
Cross-prediction steps, K_p	6
Relation feature dimension	64
Relation-network hidden dimension	512
Number of experts / prototypes per expert	$5/5$
Calibrated views per support/query sample	$100/1$
Feature-fusion coefficient, ω	0.5
Relation-weight temperature, T_r	1.0
Classifier temperature	0.15
Prototype mixture coefficient, η	0.5

Feature projection and base knowledge. After pretraining, the encoder is frozen, and the context module and pretraining heads are discarded. The 32 output tokens are flattened into an 8192-dimensional feature vector for each IMU. Features from the three PAMAP2 IMUs are concatenated and mapped to a 64-dimensional relation space through a trainable linear layer. The source bank contains 20 dataset-qualified base-class prototypes, each computed by averaging frozen-encoder features from the corresponding source training samples. Each single-IMU source prototype is repeated three times to match the target feature dimension before applying the same projection layer. Five experts use predefined, overlapping subsets of five source prototypes. The episode descriptor is the mean projected support representation.

Prior and posterior relation networks. Each expert has separate posterior and prior networks. The posterior concatenates the target embedding, base-class prototype, and episode descriptor, giving a 192-dimensional input. It uses an MLP with dimensions $192 \to 512 \to 512 \to 4$, whose raw outputs parameterize the association evidence and signed relation score. The prior receives only the 64-dimensional base-class prototype and uses an MLP with dimensions $64 \to 512 \to 3$. Both networks use SiLU activations between linear layers. Posterior samples generate calibrated feature views, while the prior provides the reference distributions for KL regularization.

Classification and expert aggregation. Each expert fits an episode-specific ridge classifier using all calibrated support views and predicts class probabilities for the calibrated query views. The classifier includes a bias term and uses ridge regularization of 0.01. The router compares linearly transformed episode and expert descriptors through cosine similarity and applies softmax over all five experts. A complementary prototype branch computes class means from the uncalibrated projected support features and predicts using negative Euclidean distances. Its probabilities are combined with the routed expert probabilities using an equal mixture. During episodic training, the feature projection, relation networks, and router are optimized while the encoder remains frozen.

C Appendix for Experiments

C.1 Experimental Protocol and Implementation Details

Dataset partitions. The preprocessing pipeline supports class-based partitions for PAMAP2, WISDM, REALDISP, MHEALTH, and USC-HAD. Selected activity classes are reserved for testing, while the remaining classes provide pretraining and meta-training data, with their samples further divided using an 80:20 ratio.

UCI HAR uses sample-based rather than class-based partitioning. We retain its official test split and divide the official training samples into pretraining and meta-training subsets using an 80:20 ratio with random seed 2024. All subsets contain the same six activity classes. DaLiAc, HAPT, MotionSense, and RealWorld2016 are used as additional pretraining sources without separate few-shot evaluation splits.

Table 7: Main hyperparameters for STCP and the P03 BRC configuration.

Parameter	Value
STCP pretraining	
Optimizer	Adam
Learning rate	10^{-4}
Batch size	512
Pretraining epochs	20
Prediction steps K_p	6
Contrastive temperature	0.1
Memory-bank capacity	256
BRC	
Optimizer	Adam
Learning rate	10^{-4}
Meta-training epochs	20
Training episodes per epoch	100
Projected feature dimension	64
Relation-network hidden dimension	512
Number of experts	5
Base prototypes per expert	5
Support / query calibration views	100 / 1
Ridge regularization coefficient	0.01
Residual fusion coefficient ω	0.5
Prototype mixing coefficient η	0.5
Relation / routing temperature	1.0 / 1.0
Classifier temperature	0.15
KL weights (λ_A, λ_S)	$(0.05, 0.06)$
Random seed	1

Table 8: Hardware and software environment.

Component	Configuration
GPU	NVIDIA GeForce RTX 4090
Python	3.10.19
PyTorch	2.1.2+cu121
CUDA (PyTorch build)	12.1
cuDNN	8.9.2
NumPy	1.26.3

Episode construction. Each episode selects N classes from the corresponding split without replacement. For each selected class, we randomly sample $K + 10$ distinct examples, assigning K to the support set and the remaining ten to the query set. Thus, an episode contains NK support examples and $10N$ query examples, with no shared samples between the two sets. Query examples are shuffled after sampling. We use $N = 8$ for PAMAP2 and $N = 6$ for UCI HAR, with $K \in \{1, 2, 4, 8, 16\}$. Each setting is evaluated over 1,000 episodes.

Hyperparameters and implementation environment. Table 7 summarizes the STCP and BRC settings. The main experiments were conducted primarily on NVIDIA GeForce RTX 4090 GPUs. The software environment is summarized in Table 8.

Pretraining data-scale ablation. The five data scales use nested dataset collections: (1) PAMAP2; (3) PAMAP2, UCI HAR, and DaLiAc; (6) the three-dataset collection plus WISDM, REALDISP, and HAPT; (9) the six-dataset collection plus MHEALTH, MotionSense, and USC-HAD; and (10) the nine-dataset collection plus RealWorld2016. These collections contain 14,790, 53,837, 210,790, 336,743, and 561,119 pretraining windows, respectively.

C.2 Baseline Methods

We compare STaRC with self-supervised representation learning methods and few-shot learning methods, which are summarized below:

Self-supervised representation learning methods.

- **SimCLR** Chen et al. (2020) uses contrastive learning between augmented views to learn visual representations from unlabeled images.
- **BYOL** Grill et al. (2020) learns visual representations through self-distillation between online and target networks, without requiring negative pairs.
- **SimSiam** Chen & He (2020) employs a Siamese architecture with stop-gradient learning to acquire visual representations without negative pairs or a momentum encoder.
- **TS-TCC** Eldele et al. (2021) combines temporal and contextual contrasting to learn discriminative time-series representations from differently augmented views.
- **CrossHAR** Hong et al. (2024) combines sensor-oriented data augmentation with hierarchical self-supervised pretraining to improve human activity recognition across datasets.
- **TFC** Huang et al. (2025) learns complementary temporal and frequency representations through contrastive objectives for wearable-sensor human activity recognition.
- **TimeDART** Wang et al. (2025) integrates autoregressive modeling and denoising diffusion to capture global dependencies and local patterns for time-series forecasting and classification.

Few-shot learning methods.

- **Matching Networks (MN)** Vinyals et al. (2016) use attention-based matching between query and support examples to recognize novel classes from limited labeled data.
- **Prototypical Networks (ProtoNet)** Snell et al. (2017) represent classes through support-feature means and classify queries by their distances to these prototypes for few-shot recognition.
- **FRN** Wertheimer et al. (2021) classifies queries through class-specific reconstruction of their feature maps from support features for few-shot image recognition.
- **MFSIN** (Yu et al., 2025a) extracts multi-scale features, uses attention-based tokenization to select informative regions,and models interactions between support and query representations for fine-grained few-shot image classification.
- **CDN4** Li et al. (2025) performs cross-view nearest-neighbor matching of local features to address intra-class variation in fine-grained few-shot image classification.
- **SWAT** Liu et al. (2025b) uses stage-wise retrieval-augmented fine-tuning of pretrained vision-language models to improve few-shot visual recognition.
- **UNEM** Zhou et al. (2025) uses an unrolled generalized expectation-maximization algorithm to jointly infer query labels in transductive few-shot classification.

These baselines cover contrastive and predictive representation learning, metric-based recognition, feature interaction and reconstruction, transductive inference, and retrieval-augmented adaptation.

C.3 Analysis of Expert Routing Behavior

To examine how the episode-conditioned router distributes weights across experts, we conduct an auxiliary routing analysis using a three-expert BRC. Each expert represents a predefined subset of base-class knowledge, and the router assigns weights according to the support-set episode descriptor and the experts' base knowledge. The same checkpoint is used across all five datasets, with 200 episodes sampled from each dataset. This analysis examines routing behavior only and does not report classification accuracy.

We report three complementary statistics in Table 9. Top-1 frequency records how often an expert receives the largest routing weight, whereas mean routing weight measures its average contribution

Table 9: Routing statistics for a three-expert BRC configuration over 200 episodes per dataset. Top-1 frequency measures how often an expert receives the largest routing weight. Mean routing weight averages each expert's weight across episodes, while mean maximum weight measures the average largest weight within an episode.

Dataset	Episode Protocol	Episodes	Top-1 Frequency			Mean Routing Weight			Mean Maximum Weight
			E_1	E_2	E_3	E_1	E_2	E_3	
PAMAP2	8-way 1-shot	200	0.455	0.260	0.285	0.388	0.287	0.325	0.762
UCI HAR	6-way 1-shot	200	0.480	0.325	0.195	0.341	0.426	0.233	0.617
USC-HAD	6-way 1-shot	200	0.220	0.195	0.585	0.180	0.387	0.433	0.612
HAPT	6-way 1-shot	200	0.185	0.565	0.250	0.263	0.432	0.305	0.592
MotionSense	6-way 1-shot	200	0.280	0.350	0.370	0.276	0.366	0.359	0.626

to the routed expert mixture. Mean maximum weight summarizes how strongly the router concentrates on the leading expert within individual episodes.

The routing patterns vary across datasets, and no expert receives all top-1 selections on any evaluated dataset. Expert E_1 is most frequently ranked first on PAMAP2 and UCI HAR, with frequencies of 0.455 and 0.480, respectively. Expert E_3 leads on USC-HAD with a frequency of 0.585, while E_2 leads on HAPT with a frequency of 0.565. MotionSense exhibits a more balanced distribution of top-1 selections across the three experts.

Top-1 frequency and mean routing weight capture different aspects of this behavior. On UCI HAR, E_1 is most frequently ranked first, but E_2 receives the largest average weight. Similarly, on USC-HAD, E_2 has the lowest top-1 frequency yet retains a substantial mean weight of 0.387. These results illustrate that an expert can contribute substantially to the soft mixture without being the most frequently selected leading expert.

The mean maximum weight ranges from 0.592 to 0.762, indicating varying degrees of within-episode routing concentration. PAMAP2 exhibits the strongest concentration, whereas HAPT shows a more distributed allocation. Overall, the results reveal non-constant, dataset-dependent routing patterns while retaining contributions from multiple experts on average.