DOMAIN GENERALIZATION UNDER SAMPLING PATTERN SHIFTS IN IRREGULAR TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Irregularly sampled multivariate time series (ISMTS) are prevalent in real-world applications, where both observation times and available measurements can vary substantially across domains. While recent models increasingly exploit such sampling information for prediction, its robustness under sampling pattern shifts remains underexplored. We introduce HAR-C, to the best of our knowledge the first controlled benchmark for sampling pattern shifts in ISMTS, and show that sampling shifts alone can substantially degrade performance, induce sampling-specific shortcuts, and remain challenging for existing domain generalization (DG) methods. Motivated by these findings, we propose PRISM, a DG framework that first learns complementary feature-centric and sampling-centric representations without task labels, and subsequently performs robust supervised training across diverse sampling variations to discourage brittle shortcut reliance. Extensive experiments on controlled and real-world ISMTS benchmarks demonstrate that PRISM consistently improves robustness to unseen sampling shifts over existing methods. Our code is available at `https://anonymous.4open.science/r/PRISM`.

1 INTRODUCTION

Irregularly sampled multivariate time series (ISMTS) are prevalent across various domains, such as electronic health records in hospitals and sensor observations on wearable devices (Azoff, 1994; Ismail Fawaz et al., 2019). Such irregularities stem from various factors, including naturally fluctuating data collection intervals, asynchronous sampling across different features, and uncertainties in observation timestamps (Eckner, 2012; Little & Rubin, 2019). Establishing predictive models in such domains is of significant practical importance yet particularly challenging, since modern deep learning architectures are built upon inductive biases that assume regularly structured inputs (*e.g.*, images, videos, and speech) (Bronstein et al., 2017; Rubanova et al., 2019; Shukla & Marlin, 2021), without explicitly taking into account irregular sampling patterns.

In response to these challenges, earlier approaches largely followed the convention of fitting ISMTS to existing models (*e.g.*, RNNs and Transformers) by forcing irregular observations onto a regular grid through temporal quantization and imputation (Lipton et al., 2015; Che et al., 2018). This incorporates aligning observations to regular intervals and imputing unobserved observations. However, such forced regularization distorts the original observation times, degrading temporal precision (Rubanova et al., 2019), while imputation introduces estimation errors at unobserved time points. More importantly, this process discards valuable information about *how observations were acquired*; for instance, deteriorating patients are often monitored more frequently and undergo more tests, making measurement frequency informative of patient condition (Che et al., 2018). These limitations have motivated a shift toward models that directly process irregular time series in their native form, including observation sets (values, feature types, and timestamps) (Horn et al., 2020; Shukla & Marlin, 2021; Tipirneni & Reddy, 2022), image-like representations (Li et al., 2023), and token sequences with temporal information (Tan et al., 2024; Tang et al., 2025). In essence, modern irregular time series models not only recover the underlying signal, but also preserve and exploit information contained in the observation process.

While this paradigm has substantially advanced ISMTS modeling, its robustness under real-world distribution shifts has yet to be fully scrutinized. Indeed, sampling patterns frequently vary across domains (*e.g.*, hospitals may differ in vital-sign monitoring frequency or laboratory-test availability),

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.

potentially turning useful sampling cues into brittle shortcuts. To systematically evaluate their impact, we introduce HAR-C, to the best of our knowledge the first controlled benchmark dedicated to sampling pattern shifts in ISMTS, built on human activity recognition (HAR) (Anguita et al., 2013) while preserving the underlying feature distribution and observation budget. Our analysis reveals that (i) sampling shifts alone substantially degrade performance, with larger shifts causing greater degradation; (ii) motivated by the predictive nature of sampling patterns in real-world ISMTS, we synthetically control their correlation with class labels and find that models increasingly follow sampling-associated labels even when they conflict with the underlying signal; (iii) real-world sampling processes vary substantially across domains while *carrying nontrivial predictive information*, creating conditions for such shortcuts to become unreliable; and (iv) existing domain generalization (DG) methods (Sun & Saenko, 2016; Sagawa et al., 2020; Lu et al., 2023; Yang et al., 2023) remain vulnerable when these shortcuts are stable across source domains but change only at test time.

These observations expose a fundamental dilemma: *sampling information is too useful to discard, yet too domain-dependent to trust unconditionally*; enforcing invariance may remove legitimate predictive cues, whereas unrestricted exploitation encourages brittle dependence on the acquisition process. To reconcile this tension, we propose PRISM, a model-agnostic DG framework that leverages sampling information while discouraging shortcut reliance. Since label supervision can directly encourage such shortcuts, PRISM first pre-trains two complementary representations without task labels. Specifically, two independently parameterized encoders learn a feature-centric representation through feature value reconstruction and a sampling-centric representation through sampling pattern reconstruction. During labeled training, PRISM constructs multiple stochastic sampling variations for each sample and minimizes the worst classification loss across them, discouraging brittle shortcut reliance while retaining sampling information that remains useful across variations. Extensive experiments on controlled and real-world ISMTS benchmarks demonstrate improved robustness to unseen sampling shifts over existing DG methods.

In summary, our contributions are threefold:

- We introduce HAR-C, a controlled benchmark dedicated to sampling pattern shifts in ISMTS, and provide an in-depth analysis showing that sampling shifts alone degrade performance, induce sampling-specific shortcuts, and remain challenging for existing DG methods.
- Motivated by these findings, we propose PRISM, a DG framework that learns complementary feature-centric and sampling-centric representations without task labels and robustly trains across stochastic sampling variations to prevent brittle shortcut reliance.
- Extensive experiments on controlled and real-world ISMTS benchmarks demonstrate that PRISM consistently improves robustness to unseen sampling shifts over existing DG methods, supported by comprehensive ablation and representation analyses.

2 In-depth Analysis of Sampling Pattern Shifts

In this section, we analyze how sampling pattern shifts affect prediction, induce shortcut reliance, arise in real-world data, and challenge existing domain generalization methods. Real-world time series often exhibit irregularities arising from opportunistic data collection, sensor failures, or event-driven observations, as illustrated in Figure 1, where feature observation rates in eICU (Pollard et al., 2018) vary substantially across hospitals. Importantly, the sampling process itself may vary across domains, changing *how* the same underlying signal is observed.

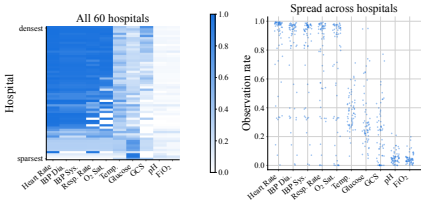

Figure 1: Hospital-level sampling pattern variations on eICU dataset.

2.1 HAR-C: A Controlled Benchmark for Sampling Pattern Shifts

Studying the isolated effect of sampling pattern shifts is difficult since changing the observation process often also changes the amount of available information or introduces broader distribution shifts. To disentangle these factors, we introduce *HAR-C*, a controlled benchmark built on the standard Human Activity Recognition (HAR) dataset (Anguita et al., 2013), where sampling patterns are manipulated while preserving the underlying signals and observation budget.

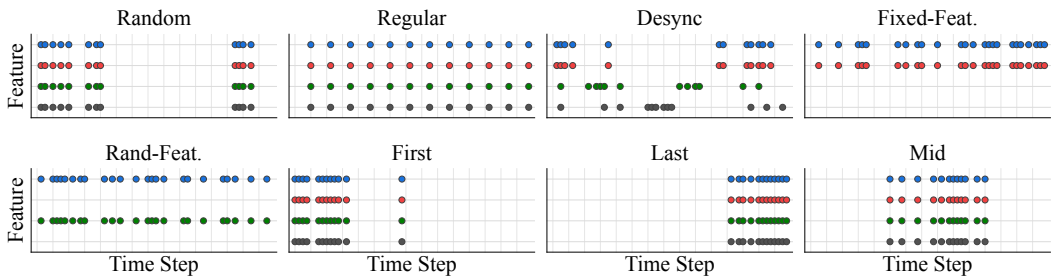

Figure 2: Overview of sampling pattern shifts in HAR-C. Starting from synchronized random sampling, HAR-C constructs diverse target sampling patterns while preserving the observation budget.

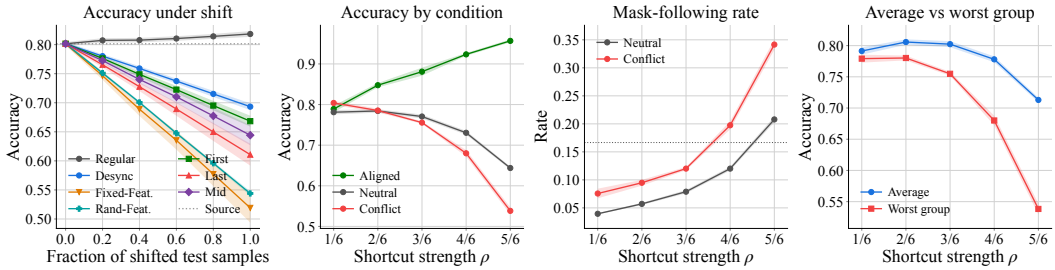

Figure 3: Analysis of sampling pattern shifts on HAR-C. The four panels show accuracy under increasing shift severity, accuracy under varying shortcut strength ρ, mask-following rate, and average versus worst-group accuracy.

Starting from regularly sampled HAR sequences, we retain 20% of the original observations to construct irregular inputs. Since the sensor channels in HAR are jointly recorded at each time point, the source domain, Random, uses synchronized random sampling across features. As illustrated in Figure 2, we construct diverse target shifts, including regular sampling (Regular, reference), desynchronization across features (Desync), heterogeneous feature-wise sampling (Fixed-Feat. and Rand-Feat.), and temporal concentration toward different parts of the sequence (First, Last, and Mid), while varying shift severity. Crucially, all sampling patterns preserve the same 20% observation budget, allowing us to separate the effect of sampling pattern shifts from that of a reduced observation budget.

2.2 Analysis of Sampling Pattern Shifts

Sampling pattern shifts alone substantially degrade prediction. We first examine how performance changes as an increasing fraction of test samples undergo each sampling shift. Throughout this analysis, we use a CNN with zero imputation as a simple and commonly used baseline for irregularly observed time series (Che et al., 2018; Kim et al., 2023; Lee et al., 2026). As shown in Figure 3, performance deteriorates under target sampling shifts despite preserving the same observation budget, with larger fractions of shifted samples generally causing greater degradation. The Regular process serves as a reference condition, whereas the remaining shifts reveal substantial sensitivity to *where* and *how* observations are collected. This behavior also persists across SeFT (Horn et al., 2020), mTAND (Shukla & Marlin, 2021), and GRU-D (Che et al., 2018) in Figure A8, indicating that the phenomenon is not architecture-specific.

Sampling patterns become shortcuts when correlated with labels. We next investigate whether models exploit sampling patterns as predictive shortcuts. We associate each of the six HAR classes with a distinct sampling pattern and vary the sampling–label association strength ρ from chance level $(1/6)$ to increasingly predictive settings. At test time, the sampling cue is aligned with the true label, neutral, or conflicting. As shown in Figure 3, increasing ρ improves aligned accuracy but sharply degrades conflicting accuracy, while the rising mask-following rate shows that predictions increasingly track the sampling-associated label. The widening average–worst-group gap further reveals failures hidden by strong average performance. A likely reason is that sampling patterns provide simple, directly observable cues, whereas class-relevant feature information requires integrating multiple noisy observations and their temporal dynamics.

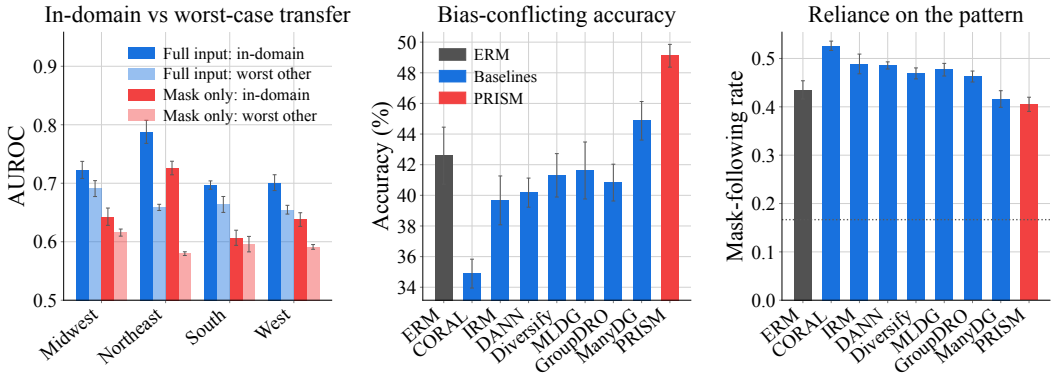

Figure 4: Analysis of sampling-specific shortcuts and domain generalization. The three panels show in-domain versus worst cross-region transfer on eICU, bias-conflicting accuracy on HAR-C, and mask-following rate across DG methods.

Real-world sampling processes are predictive, yet domain dependent. We next examine whether the same phenomenon arises in inherently irregular real-world electronic health records, using mortality prediction on eICU (Pollard et al., 2018). Since deteriorating patients are often monitored more frequently or receive additional measurements, the observation process can correlate with patient condition. Indeed, the leftmost panel of Figure 4 shows that observation masks alone achieve AUROC substantially above 0.5, indicating that *sampling patterns themselves contain predictive information* about mortality. However, this relationship is *unstable across environments*: observation frequencies vary substantially across hospitals (Figure 1), and mask-only prediction degrades markedly under cross-region transfer. Thus, sampling patterns can be both predictive and domain dependent—the two ingredients that make sampling-specific shortcuts brittle under domain shift.

Existing DG remains vulnerable to source-stable sampling shortcuts. Finally, we ask whether existing domain generalization methods can address this failure mode. We construct source environments in which the sampling–label association remains stable and predictive, but reverses in the unseen target domain. We evaluate bias-conflicting accuracy and the mask-following rate, measuring whether predictions follow the true class or the misleading sampling cue when the two conflict. As shown in Figure 4, existing DG methods provide only limited improvements over ERM while continuing to rely substantially on the sampling pattern. Under a strong sampling–label association, they perform well when the shortcut agrees with the label but degrade sharply when it becomes conflicting, suggesting that sampling patterns can remain a dominant shortcut under standard DG training.

3 Proposed Method: PRISM

This section elucidates PRISM, a model-agnostic DG framework for learning robust representations under sampling pattern shifts in ISMTS. Motivated by analysis in Section 2, PRISM first pretrains complementary feature-centric and sampling-centric representations through value reconstruction and sampling pattern reconstruction, respectively, without task labels (Section 3.2). Second, the two representations are jointly used for downstream prediction, where robust optimization across stochastically generated sampling variations discourages brittle reliance on sampling-specific shortcuts (Section 3.3). An overview of PRISM is illustrated in Figure 5.

3.1 Problem Setup

Before delving into the details of PRISM, we first formally define the problem setup of DG on ISMTS. Let $\mathcal{X}$ be the space of valid ISMTS inputs and $\mathcal{Y} = \{1, \ldots, C\}$ the label space for C-class classification. Each $\mathbf{x} \in \mathcal{X}$ is represented as $\mathbf{x} = (\boldsymbol{v}, \boldsymbol{m}, \boldsymbol{t})$, where, for a sequence with L time points and D features, $\boldsymbol{v} \in \mathbb{R}^{L \times D}$ denotes feature values, $\boldsymbol{m} \in \{0, 1\}^{L \times D}$ is the corresponding observation mask (0: missing, 1: observed), and $\boldsymbol{t} \in \mathbb{R}^{L \times 1}$ denotes observation times. Given labeled source data $\mathcal{D}_s \subseteq \mathcal{X} \times \mathcal{Y}$, our goal is to learn a predictor $F : \mathcal{X} \to \Delta^{C-1}$ that generalizes to an unseen target domain $\mathcal{D}_t \subseteq \mathcal{X} \times \mathcal{Y}$, whose sampling process may differ from those observed during training, where Δ^{C-1} denotes the probability simplex over C classes.

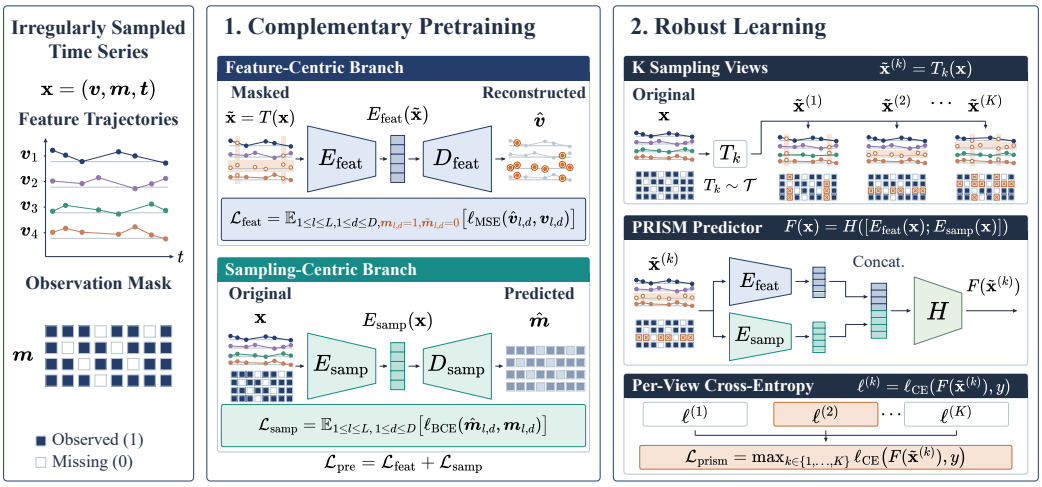

Figure 5: Overview of PRISM. Feature- and sampling-centric encoders are first pretrained through masked value and observation-mask reconstruction, then jointly trained using the worst classification loss across stochastic sampling variations.

3.2 Complementary Representation Pretraining

Our analysis in Section 2 reveals that sampling patterns are informative yet prone to shortcut reliance under task supervision. This creates a tension: learning feature representations invariant to sampling patterns may discard useful predictive cues, whereas learning a single representation from both feature values and sampling patterns may still rely on sampling-specific shortcuts. PRISM resolves this tension by pretraining two complementary representations without task labels, separately focusing on feature values and sampling patterns.

Given the backbone architecture of F, we instantiate two independently parameterized encoders: a feature-centric encoder E_{feat} and a sampling-centric encoder E_{samp}. This model-agnostic design applies across different ISMTS architectures. First, the feature-centric encoder is encouraged to capture rich information about the observed feature values. We sample a transformation $T \sim \mathcal{T}$ and construct a partially observed view $\tilde{\mathbf{x}} = T(\mathbf{x})$ by randomly hiding additional observations. From the remaining measurements, E_{feat} must retain sufficient information to recover the masked values. A lightweight decoder D_{feat} produces $\hat{\boldsymbol{v}} = D_{\text{feat}}(E_{\text{feat}}(\tilde{\mathbf{x}})) \in \mathbb{R}^{L\times D}$. The reconstruction loss is evaluated only on observations originally present but removed by T:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{1\le l\le L,\, 1\le d\le D,\, \boldsymbol{m}_{l,d}=1,\, \tilde{\boldsymbol{m}}_{l,d}=0} \left[\ell_{\text{MSE}}\left(\hat{\boldsymbol{v}}_{l,d}, \boldsymbol{v}_{l,d}\right)\right]. \tag{1}$$

This masked recovery task prevents simple input copying and encourages the representation to capture informative structure across the observed measurements.

The sampling-centric encoder instead preserves how the time series was observed. Since sampling patterns can themselves carry predictive information, a lightweight decoder D_{samp} reconstructs the original observation mask from $E_{\text{samp}}(\mathbf{x})$, yielding $\hat{\boldsymbol{m}} = D_{\text{samp}}(E_{\text{samp}}(\mathbf{x})) \in (0,1)^{L\times D}$. Although E_{samp} receives the full input, reconstructing m through a compact representation encourages it to summarize the observation structure rather than copy the input. Its reconstruction objective is

$$\mathcal{L}_{\text{samp}} = \mathbb{E}_{1\le l\le L,\, 1\le d\le D} \left[\ell_{\text{BCE}}\left(\hat{\boldsymbol{m}}_{l,d}, \boldsymbol{m}_{l,d}\right)\right]. \tag{2}$$

Unlike $\mathcal{L}_{\text{feat}}$, $\mathcal{L}_{\text{samp}}$ requires no additional masking, as its goal is to preserve rather than infer the sampling structure. Here, $\ell_{\text{MSE}}(a,b) = (a-b)^2$ and ℓ_{BCE} denote the mean-squared error and binary cross-entropy loss, respectively. Together, the two branches retain complementary views of the input—what is observed and how it is observed—without task labels shaping either representation. The two reconstruction objectives are equally weighted, $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{samp}}$, avoiding an additional balancing hyperparameter. After pretraining, the lightweight decoders D_{feat} and D_{samp} are discarded, leaving the two pretrained encoders for downstream prediction (He et al., 2022).

The augmentation distribution $\mathcal{T}$ is intentionally simple; it is not designed to mimic any particular target shift, and its parameters are selected using source-domain validation data only. A transformation

$T \sim \mathcal{T}$ independently samples timestep and feature drop ratios $p_t \sim \mathcal{U}(0, r_t)$ and $p_f \sim \mathcal{U}(0, r_f)$, and randomly masks the selected time points and features. The same augmentation distribution is reused to create diverse sampling variations for downstream robust learning.

3.3 Robust Learning Across Sampling Variations

The complementary representations learned during pretraining do not by themselves prevent shortcut reliance during downstream supervision. If simply concatenated, the classifier may still favor sampling cues that are highly predictive in the source data, while fixing their relative contributions is undesirable because their importance can vary across samples. We therefore combine both representations flexibly while discouraging dependencies that are fragile to sampling pattern changes.

For an input $\mathbf{x}$, the two representations are concatenated and passed through a shared classification head H: $F(\mathbf{x}) = H([E_{\text{feat}}(\mathbf{x}); E_{\text{samp}}(\mathbf{x})])$. Since the sampling process of the unseen target domain $\mathcal{D}_t$ is unknown, we expose each training sample to multiple plausible sampling variations. Specifically, we draw K transformations $T_k \sim \mathcal{T}$ and construct $\tilde{\mathbf{x}}^{(k)} = T_k(\mathbf{x})$ for $k \in \{1, \ldots, K\}$, each preserving the underlying instance and label while altering its sampling pattern.

For each sample, PRISM defines its robust classification loss as the worst loss among K sampling variations:

$$\mathcal{L}_{\text{prism}} = \max_{k \in \{1,\ldots,K\}} \ell_{\text{CE}} \left(F(\tilde{\mathbf{x}}^{(k)}), y \right), \tag{3}$$

where ℓ_{CE} denotes the cross-entropy loss. The objective is averaged over source samples, yielding sample-wise minimax optimization over sampling variations.

If the predictor relies on a sampling-specific shortcut, some augmented views can disrupt that cue and incur high loss. The worst-view objective thus prevents such failures from being averaged out, allowing the predictor to flexibly exploit feature and sampling information without an additional balancing hyperparameter or gating network.

4 Experiments

In this section, we thoroughly evaluate PRISM through the following research questions: (**RQ1**) Does PRISM robustly generalize to unseen sampling pattern shifts across controlled, real-world, and increasingly severe settings? (Section 4.2); (**RQ2**) Which design components are responsible for this robustness? (Section 4.3); (**RQ3**) Does the robustness generalize across backbone architectures while remaining practical? (Section 4.4)

4.1 Experimental setup

Datasets. We evaluate PRISM on three complementary benchmarks covering controlled, naturally occurring, and increasingly severe sampling shifts. HAR-C, introduced in Section 2.1, isolates sampling pattern shifts under a fixed observation budget, allowing us to evaluate their effect without confounding changes in the amount of observed information. eICU (Pollard et al., 2018) considers in-hospital mortality prediction under leave-one-region-out shifts across four regions, where sampling pattern shifts are intertwined with broader naturally occurring distribution shifts. PAM (Reiss & Stricker, 2012) contains multivariate sensor data from multiple individuals and follows (Zhang et al., 2022; Li et al., 2023) with leave-fixed-sensors-out shifts of increasing severity. Together, these benchmarks test whether robustness observed under controlled shifts transfers to more realistic and severe changes in the observation process. Detailed dataset statistics are provided in Table A4.

Domain generalization baselines. We compare PRISM with ERM and representative DG baselines: ARM (Zhang et al., 2021), CORAL (Sun & Saenko, 2016), DANN (Ganin et al., 2016), GroupDRO (Sagawa et al., 2020), IB-ERM (Ahuja et al., 2021), IRM (Arjovsky et al., 2019), Mixup (Zhang et al., 2018), MLDG (Li et al., 2018a), MMD (Li et al., 2018b), VREx (Krueger et al., 2021), Diversify (Lu et al., 2023), and ManyDG (Yang et al., 2023). These baselines span widely used general-purpose DG methods as well as approaches specifically designed for time series, covering diverse strategies including data augmentation, invariant representation learning, robust optimization, and meta-learning.

Table 1: Comparison of PRISM and existing DG methods on HAR-C. Models are evaluated on the source-matched Random condition and seven target sampling pattern shifts.

Algorithm	Random	Regular	Desync	Fixed-Feat.	Rand-Feat.	First	Last	Mid	Avg.
ERM	79.92±0.55	80.97±0.81	68.91±0.98	54.05±4.27	54.48±0.26	67.25±1.52	61.10±3.20	65.06±2.50	66.47±0.95
ARM	80.39±0.49	81.32±0.82	68.15±0.92	56.54±3.56	56.74±0.65	73.12±0.88	70.54±0.87	72.18±1.04	69.87±0.60
CORAL	80.05±0.54	81.51±0.67	70.26±1.22	53.97±3.49	54.91±0.48	67.12±1.90	60.34±3.60	67.10±2.86	66.91±0.88
DANN	81.94±0.24	82.72±0.73	70.30±0.90	51.82±4.97	54.41±0.53	69.21±1.25	60.59±2.66	68.12±1.20	67.39±0.98
GroupDRO	79.86±0.21	81.41±0.55	68.45±0.84	54.79±3.99	55.01±0.43	66.53±1.08	61.43±2.15	66.78±1.51	66.78±0.68
IB_ERM	80.51±0.45	81.45±0.84	68.94±0.97	53.87±4.17	56.10±0.53	64.53±1.34	61.17±3.46	66.10±2.81	66.58±1.26
IRM	80.53±0.41	81.55±0.47	68.27±0.64	53.52±4.62	53.74±0.65	67.35±1.66	61.16±2.31	66.02±1.81	66.52±1.11
Mixup	80.01±1.08	81.30±1.75	65.46±2.47	55.45±4.78	54.47±0.61	68.93±1.96	64.06±2.18	71.12±1.19	67.60±1.17
MLDG	84.01±0.29	86.08±0.18	75.38±0.68	58.85±3.98	56.54±0.44	74.66±0.87	71.22±0.72	74.84±0.74	72.70±0.54
MMD	80.96±0.24	83.02±0.68	70.34±0.94	54.36±4.51	54.32±0.63	69.23±1.15	67.69±1.80	71.10±1.59	68.88±0.85
VREx	70.41±1.08	71.05±1.66	58.42±1.55	50.30±4.53	52.27±0.55	63.50±1.55	64.68±1.10	65.20±1.41	61.98±1.05
Diversify	82.76±0.31	84.19±0.58	71.04±1.14	53.71±4.64	53.86±0.46	70.88±0.91	65.31±1.28	71.70±0.65	69.18±0.75
ManyDG	82.18±0.49	84.85±0.63	73.06±0.88	51.66±5.33	55.92±0.80	66.44±2.45	61.87±1.80	68.64±1.96	68.08±1.16
PRISM	**84.08**±0.16	**86.49**±0.21	**82.81**±0.27	**74.90**±1.29	**74.72**±0.21	**81.28**±0.20	**79.22**±0.35	**81.02**±0.23	**80.57**±0.19

Implementation details. Following prior time series DG work (Lu et al., 2023), we primarily use a CNN backbone, with alternative backbones evaluated in Section D. We use $K = 4$ throughout, with dataset-specific sampling augmentations detailed in Table A5. All experiments are repeated ten times, and we report the mean±standard error. Further implementation details and configurations are provided in the appendix and at `https://anonymous.4open.science/r/PRISM`.

4.2 Main Results

Controlled sampling pattern shifts. To answer RQ1, we first evaluate PRISM on HAR-C, where only the sampling process changes while the observation budget is controlled. PRISM achieves the best performance across all eight sampling conditions, improving average accuracy from 72.70% for the strongest baseline to 80.57% (Table 1). Excluding the Random and Regular conditions, the gap widens further (78.99% vs. 68.58%). Importantly, under the source-matched Random condition, PRISM and MLDG perform nearly identically (84.08% vs. 84.01%), whereas substantially larger gaps emerge once the target sampling process differs from that seen during training. Relative to the strongest baseline, this suggests that the gains stem mainly from maintaining performance as the observation process changes, rather than from better fitting the source distribution. Consistent with this interpretation, PRISM also achieves the strongest unbiased and bias-conflicting accuracy when sampling-specific shortcuts are introduced (Table A6), while its mask-following rate remains comparable to ERM (Figure 4), suggesting that it makes better use of feature information rather than discarding sampling cues.

Overall, the controlled setting directly attributes the gains of PRISM to robustness against sampling pattern shifts.

Natural and increasingly severe sampling shifts. We next test whether this robustness extends beyond the controlled HAR-C setting. On eICU, PRISM achieves the strongest AUROC and AUPRC across all four target regions despite heterogeneous real-world shifts, including changes in sampling patterns (Table 2). In contrast, existing DG baselines provide only limited gains over ERM, suggesting that standard DG objectives do not sufficiently address variation in the observation process when it co-occurs with broader distribution shifts. The consistent gains of PRISM across regions therefore indicate that explicitly accounting for sampling variability remains beneficial even when sampling shifts cannot be isolated from other sources of domain variation.

On PAM, the difference becomes increasingly pronounced as the target sampling process departs further from the source. While performance remains relatively stable under mild sensor removal, PRISM maintains 81.97% accuracy when 50% of sensors are entirely unavailable, compared with 58.99% for the strongest baseline (Table A7). The growing gap with increasing removal severity suggests that the benefit of PRISM becomes larger precisely when the available observation structure differs most strongly from training. Together, eICU and PAM show that the robustness observed on HAR-C transfers to naturally occurring shifts and persists under substantially more severe changes in sampling.

Table 2: Comparison of PRISM and existing DG methods on eICU under leave-one-region-out evaluation. AUROC and AUPRC are reported for Northeast, West, South, and Midwest target regions.

Algorithm	Northeast		West		South		Midwest		Avg.	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
ERM	85.30±0.20	57.91±0.35	76.89±0.20	43.77±0.19	78.59±0.20	40.96±0.28	78.09±0.34	37.27±0.98	79.72±0.15	44.98±0.25
ARM	74.08±1.25	45.82±0.99	68.76±1.18	37.86±1.05	69.96±1.30	33.90±0.76	72.64±0.48	31.48±0.47	71.36±0.60	37.26±0.43
CORAL	85.08±0.24	58.16±0.27	76.99±0.29	43.75±0.33	79.47±0.18	41.97±0.20	78.58±0.14	39.25±0.25	80.03±0.08	45.78±0.11
DANN	83.28±0.31	55.97±0.40	77.90±0.18	44.64±0.35	77.38±0.29	40.04±0.36	77.19±0.31	35.29±0.57	78.94±0.16	43.98±0.24
GroupDRO	84.36±0.30	57.40±0.27	76.99±0.27	43.85±0.33	77.33±0.33	39.32±0.36	77.39±0.39	37.76±0.45	79.02±0.13	44.58±0.16
IB_ERM	85.38±0.22	57.81±0.29	76.91±0.45	43.69±0.54	78.41±0.30	40.79±0.42	77.69±0.39	36.49±0.96	79.60±0.11	44.70±0.23
IRM	85.11±0.21	57.62±0.38	75.92±0.65	42.84±0.70	78.42±0.21	40.99±0.35	77.42±0.41	37.17±0.60	79.22±0.18	44.65±0.21
Mixup	85.01±0.26	57.80±0.19	78.32±0.26	45.31±0.34	79.06±0.25	41.65±0.33	78.61±0.19	38.24±0.53	80.25±0.11	45.75±0.18
MLDG	85.81±0.21	58.95±0.31	77.57±0.30	44.49±0.49	78.87±0.11	41.08±0.30	78.31±0.24	38.25±0.41	80.14±0.07	45.70±0.14
MMD	83.99±0.58	55.97±0.57	75.71±0.59	43.02±0.71	72.66±1.29	37.19±0.84	77.08±0.41	37.57±0.52	77.36±0.46	43.44±0.40
VREx	81.52±1.01	52.80±1.04	73.06±0.80	39.04±1.32	76.26±0.28	38.03±0.40	75.73±0.22	34.62±0.54	76.64±0.32	41.12±0.45
Diversify	84.11±0.28	56.09±0.49	76.24±0.43	42.93±0.56	77.27±0.28	40.23±0.34	77.25±0.33	36.76±0.81	78.72±0.24	44.01±0.32
ManyDG	85.22±0.22	57.94±0.26	77.24±0.24	44.59±0.36	78.71±0.12	41.35±0.27	78.40±0.23	38.27±0.56	79.89±0.10	45.54±0.17
PRISM	**87.15**±0.09	**60.02**±0.15	**79.88**±0.09	**47.42**±0.19	**81.35**±0.13	**43.92**±0.14	**81.50**±0.18	**41.64**±0.28	**82.47**±0.04	**48.25**±0.08

Table 3: Ablation study of PRISM. Top: incremental construction from ERM. Bottom: component removal. Worst: lowest target-condition accuracy on HAR-C. eICU results are averaged over regions.

Variant	HAR-C		eICU	
	Avg.	Worst	AUROC	AUPRC
Incremental construction				
ERM	66.47±0.95	54.05±4.27	79.72±0.15	44.98±0.25
+ Complementary pretraining	72.02±0.43	56.29±3.73	81.87±0.07	47.58±0.15
+ Sampling variations	79.58±0.17	73.85±0.18	82.17±0.07	47.91±0.15
+ Worst-view objective (PRISM)	**80.57**±0.19	**74.72**±0.21	**82.47**±0.04	**48.25**±0.08
Component removal from PRISM				
w/o $\mathcal{L}_{\text{feat}}$	80.02±0.14	74.36±0.25	80.74±0.07	46.38±0.06
w/o $\mathcal{L}_{\text{samp}}$	80.47±0.15	74.50±1.19	81.94±0.08	47.23±0.16
w/o E_{feat}	79.41±0.16	72.88±1.27	81.19±0.05	46.79±0.13
w/o E_{samp}	80.02±0.14	73.96±0.15	82.27±0.06	47.65±0.15

4.3 Ablation Study

To answer RQ2, Table 3 builds PRISM incrementally from ERM and removes each component from the full model. Every component improves average performance, but their relative importance depends on the type of distribution shift. On HAR-C, removing sampling variations ($K = 1$) causes the largest degradation, while training on the same variations with an average loss already recovers performance from 72.02% to 79.58%. Using the worst-view objective further improves accuracy to 80.57%, showing that exposure to diverse sampling processes provides the dominant source of robustness, with minimax optimization preventing difficult variations from being averaged away. Removing either pretraining objective or encoder branch produces additional degradation, indicating that the complementary feature-centric and sampling-centric representations provide information beyond augmentation alone.

The relative effects differ on eICU (Table A9). Here, complementary pretraining contributes more strongly under naturally occurring cross-region shifts, while sampling variations and worst-view optimization still provide consistent improvements. Thus, the ablations suggest that the components of PRISM play complementary roles: sampling variations are particularly important when the target observation process changes explicitly, whereas representation pretraining becomes more influential when sampling shifts are entangled with broader real-world variation. Full ablation results are provided in Tables A8 and A9.

4.4 Further Analysis

Generalization across backbone architectures. To answer RQ3, we first examine whether the gains of PRISM depend on the CNN architecture used in the main experiments. As shown in Figure 6, the improvement persists across Medium CNN, Large CNN, SeFT (Horn et al., 2020), mTAND (Shukla & Marlin, 2021), and GRU-D (Che et al., 2018).

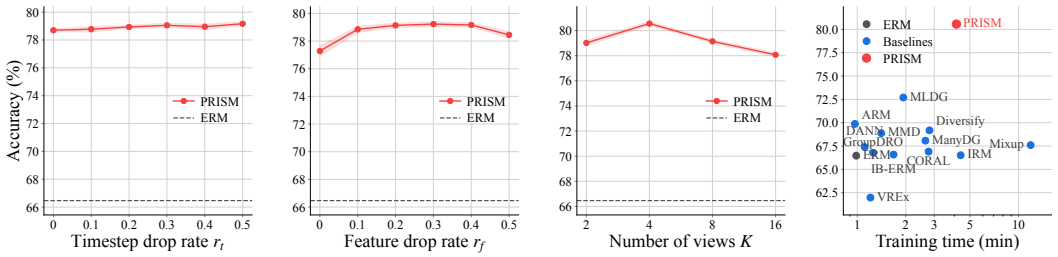

Figure 7: Sensitivity and efficiency analyses of PRISM. From left to right, we vary the timestep drop rate, feature drop rate, and number of sampling views K, then compare training time with HAR-C accuracy.

The difference is generally more pronounced under shifted sampling processes than under the source-matched condition, mirroring the pattern observed with the default backbone. This consistency across convolutional, recurrent, set-based, and attention-based architectures suggests that the robustness of PRISM arises from its training framework rather than a particular architectural inductive bias. Full results across all HAR-C sampling conditions are provided in Section D.

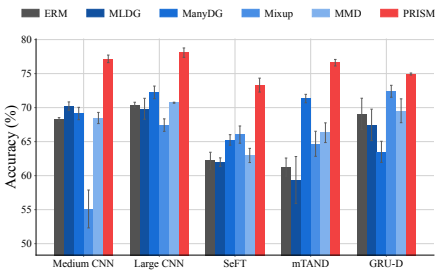

Figure 6: Backbone generalization on HAR-C. We compare PRISM with DG baselines under diverse backbones.

Representation analysis. We further examine whether the complementary pretraining objectives induce the intended information preferences. As shown in Figure A9, pretraining makes E_{feat} more informative about feature values (R^2 from 0.01 to 0.09), whereas E_{samp} becomes highly predictive of sampling patterns (AUROC from 0.73 to 0.97). Notably, E_{samp} retains almost no value information despite receiving the full input, indicating that it specializes in the observation structure rather than copying its input. After downstream supervision, these preferences partially blend, as expected when both representations are jointly optimized for prediction, but the two branches retain complementary tendencies. This supports the intended role of pretraining as encouraging complementary specialization rather than enforcing rigid disentanglement between feature and sampling information.

Sensitivity and efficiency. Finally, Figure 7 examines sensitivity to the augmentation strength and number of sampling variations. Moderate timestep and feature dropping already provide strong performance, while more aggressive perturbations yield diminishing returns. This indicates that PRISM does not require precise tuning of the augmentation distribution to obtain its robustness gains. Performance peaks at $K = 4$ and slightly decreases for larger K, possibly because the worst-view objective becomes overly pessimistic; we therefore use $K = 4$ by default. Although PRISM is more computationally expensive than standard ERM, its higher HAR-C accuracy yields a favorable accuracy–cost trade-off. Overall, the robustness of PRISM persists across architectures and broad hyperparameter choices without requiring excessive computation.

5 Conclusion

This paper studies domain generalization under sampling pattern shifts in irregularly sampled time series. We introduce a controlled benchmark showing that such shifts degrade performance, induce sampling-specific shortcuts, and challenge existing domain generalization methods. Motivated by these findings, we propose PRISM, which learns complementary feature-centric and sampling-centric representations and trains robustly over stochastic sampling variations. Experiments on controlled and real-world benchmarks show improved robustness to unseen sampling shifts, highlighting the need to account for sampling information that is predictive yet unstable across environments.

AI USE STATEMENT

Generative AI tools were used to assist with research ideation, code implementation, experimental design and analysis, and manuscript editing. All AI-assisted outputs were critically reviewed and verified by the authors, who take full responsibility for the methodology, experiments, results, and final content of this work.

ETHICS STATEMENT

Our experiments use publicly available or de-identified datasets, including the eICU Collaborative Research Database (Pollard et al., 2018). The eICU data are accessed and handled in accordance with the corresponding PhysioNet data use requirements, and no restricted patient-level data are shared with third-party services. Although our work focuses on robustness rather than clinical deployment, sampling patterns in electronic health records may reflect institution-specific patient and acquisition characteristics. Models exploiting such patterns may therefore inherit environment-specific biases and require careful validation before clinical use.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release the implementation of PRISM, including data preprocessing, HAR-C construction, sampling pattern augmentations, model configurations, training procedures, and evaluation code. The paper and appendix provide the experimental protocols, hyperparameters, and additional analyses required to reproduce the reported results. Our code is available at `https://anonymous.4open.science/r/PRISM`.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3438–3450, 2021.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 437–442, 2013. ISBN 978-2-87419-081-0.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL `https://arxiv.org/abs/1907.02893`.

E Michael Azoff. *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc., 1994. URL `https://dl.acm.org/doi/10.5555/561452`.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Proceedings of NeurIPS*, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html`.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018. URL `https://www.nature.com/articles/s41598-018-24271-9`.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Proceedings of NeurIPS*, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html`.

Andreas Eckner. A framework for the analysis of unevenly spaced time series data. *Preprint*, pp. 93, 2012. URL http://www.eckner.com/papers/unevenly_spaced_time_series_analysis.pdf.

Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad Javad Darvishi Bayazi, Pooneh Mousavi, Guillaume Dumas, and Irina Rish. WOODS: Benchmarks for out-of-distribution generalization in time series. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=mvftzofTYQ.

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. URL https://jmlr.org/papers/v17/15-239.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of ICML*, pp. 4353–4363. PMLR, 2020. URL https://proceedings.mlr.press/v119/horn20a.html.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019. URL https://link.springer.com/article/10.1007/s10618-019-00619-1.

SeungHyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*, pp. 16654–16667. PMLR, 2023.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5815–5826, 2021.

Kyung Hyun Lee, Donghwee Yoon, Hyunsun Lim, Ki-Byung Lee, and Yong Kyu Lee. Deep learning models for acute kidney injury prediction: multi-center external validation and evaluation under simulated continuous monitoring conditions. *npj Digital Medicine*, 2026.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of AAAI*, pp. 3490–3497, 2018a. URL https://aaai.org/papers/11596-learning-to-generalize-meta-learning-for-domain-generalization/.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of CVPR*, pp. 5400–5409, 2018b. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Li_Domain_Generalization_With_CVPR_2018_paper.html.

Zekun Li, Shiyang Li, and Xifeng Yan. Time series as images: Vision transformer for irregularly sampled time series. In *Proceedings of NeurIPS*, pp. 49187–49204, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a17c1eb808cf012065e9db47b7ca80d-Abstract-Conference.html.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260.

Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. In *Proceedings of ICLR*, 2023. URL https://openreview.net/forum?id=gUZWOE42l6Q.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018. URL https://www.nature.com/articles/sdata2018178.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of ISWC*, pp. 108–109. IEEE, 2012. URL https://ieeexplore.ieee.org/document/6246152.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Proceedings of NeurIPS*, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

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. In *Proceedings of ICLR*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. In *Proceedings of ICLR*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of ECCV (Workshops)*, pp. 443–450. Springer, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-49409-8_35.

Mingtian Tan, Mike A Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Proceedings of NeurIPS*, pp. 60162–60191, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ed5bf446f59e2c6646d23058c86424b-Abstract-Conference.html.

Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKIDD Explorations Newsletter*, 26(2):109–118, 2025. URL https://dl.acm.org/doi/abs/10.1145/3715073.3715083.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022. URL https://dl.acm.org/doi/10.1145/3516367.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. In *Proceedings of NeurIPS*, 1995. URL https://papers.nips.cc/paper_files/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Manydg: Many-domain generalization for healthcare applications. In *Proceedings of ICLR*, 2023. URL https://openreview.net/forum?id=lcSfirnflpW.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of ICLR*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Proceedings of NeurIPS*, pp. 23664–23678, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c705112d1ec18b97acac7e2d63973424-Abstract.html.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *Proceedings of ICLR*, 2022. URL https://openreview.net/forum?id=Kwm8I7dU-l5.

A Limitations and Future Work

Despite its effectiveness, PRISM has several limitations. Its sampling augmentations should ideally reflect the sampling pattern shifts encountered at test time, but such shifts are generally unknown in advance, making augmentation design challenging. Future work could learn generators of plausible sampling patterns or adapt the augmentation policy when prior knowledge about likely shifts is available. Moreover, in real-world settings, changes in the sampling process may be entangled with shifts in the underlying feature distribution, making it difficult to isolate and address sampling shifts alone. Understanding and handling such coupled shifts remains an important direction for future work.

B Related work

Predictive models for irregular time series. The machine learning community has dedicated significant efforts to advancing prediction models for irregular time series, evolving from data-centric to model-centric approaches. Classical imputation methods include simple statistical approaches like mean/zero imputation and forward-filling, and probabilistic methods like Gaussian Process regression (Williams & Rasmussen, 1995). These were followed by RNN variants like GRU-D (Che et al., 2018) and BRITS (Cao et al., 2018) that incorporated learnable interpolation mechanisms; yet, all these approaches often degraded temporal precision and overlooked valuable information in the sampling patterns. Modern approaches have shifted toward directly processing irregular time series in their native form, with Neural ODEs (Chen et al., 2018; Rubanova et al., 2019) treating them as continuous-time dynamic systems, and set-based architectures like SeFT (Horn et al., 2020), mTAND (Shukla & Marlin, 2021), and STraTS (Tipirneni & Reddy, 2022) handling them as observation sets with temporal features. Recent trends show diverse representation paradigms: vision-based approaches that reformulate time series as structured visual inputs (Li et al., 2023), methods leveraging language models for time series tasks (Tan et al., 2024; Tang et al., 2025), and graph-based frameworks that model sensor dependencies through neural message passing and temporal attention (Zhang et al., 2022). However, despite these advances, most existing approaches have not thoroughly examined their out-of-distribution generalization capabilities, particularly against sampling pattern shifts, critical as real-world applications often encounter significant variations in sampling patterns across different domains.

Modality-agnostic domain generalization. Domain generalization aims to learn robust models under distribution shifts by relaxing the *i.i.d. assumption* between training and test distributions, where data from the target domain is unavailable during training. Existing approaches broadly include data manipulation, representation learning, invariance-based learning, robust optimization, and meta-learning. Data manipulation methods increase source diversity through augmentations such as Mixup (Zhang et al., 2018), while representation learning aligns domains through adversarial training (DANN) (Ganin et al., 2016), moment matching (MMD) (Li et al., 2018b), or correlation alignment (CORAL) (Sun & Saenko, 2016). Methods targeting spurious correlations instead encourage predictors that remain stable across environments, including IRM (Arjovsky et al., 2019), VREx (Krueger et al., 2021), and IB_ERM (Ahuja et al., 2021). Robust optimization approaches such as GroupDRO (Sagawa et al., 2020) optimize worst-group performance, while MLDG (Li et al., 2018a) and ARM (Zhang et al., 2021) use meta-learning to improve generalization across domains. Despite their broad applicability, these methods are not tailored for sampling pattern shifts as a potentially predictive yet domain-dependent source of shortcut information in irregular time series.

Domain generalization on time series domain. Time series domain generalization requires sophisticated approaches to address both the unique characteristics of temporal data and their distribution shift patterns (Gagnon-Audet et al., 2023). Recent efforts have begun exploring this challenge across various contexts. WOODS (Gagnon-Audet et al., 2023) introduces comprehensive benchmarks for out-of-distribution generalization across diverse time series modalities, highlighting the unique challenges posed by temporal data. Diversify (Lu et al., 2023) approaches time series from a distribution perspective, learning generalized representations by characterizing latent distributions and their dynamic changes over time through adversarial training. ManyDG (Yang et al., 2023) addresses patient covariate shifts in healthcare applications by treating each patient as a distinct domain and explicitly removing patient-specific characteristics through orthogonal projection. However, these

methods have achieved only marginal performance improvements in practice (Yang et al., 2023), and most do not explicitly consider the sampling irregularity that is inherent in time series sampling patterns, leaving the challenge of domain generalization under varying sampling patterns largely unexplored.

C ALGORITHMS

This section provides the detailed training procedures of PRISM. Complementary representation pretraining is summarized in Algorithm 1, followed by sample-wise minimax learning across sampling variations in Algorithm 2.

Algorithm 1 Complementary Representation Pretraining in PRISM

Input: Source data $\mathcal{D}_s$, encoders $E_{\text{feat}}, E_{\text{samp}}$, decoders $D_{\text{feat}}, D_{\text{samp}}$, augmentation distribution $\mathcal{T}$

for each minibatch $\mathbf{x} \sim \mathcal{D}_s$ **do**
  Sample $T \sim \mathcal{T}$ and obtain $\tilde{\mathbf{x}} \leftarrow T(\mathbf{x})$
  $\hat{\boldsymbol{v}} \leftarrow D_{\text{feat}}(E_{\text{feat}}(\tilde{\mathbf{x}}))$
  $\mathcal{L}_{\text{feat}} \leftarrow \mathbb{E}_{\boldsymbol{m}_{l,d}=1,\, \tilde{\boldsymbol{m}}_{l,d}=0} \left[\ell_{\text{MSE}}(\hat{\boldsymbol{v}}_{l,d}, \boldsymbol{v}_{l,d})\right]$
  $\hat{\boldsymbol{m}} \leftarrow D_{\text{samp}}(E_{\text{samp}}(\mathbf{x}))$
  $\mathcal{L}_{\text{samp}} \leftarrow \mathbb{E}_{l,d} \left[\ell_{\text{BCE}}(\hat{\boldsymbol{m}}_{l,d}, \boldsymbol{m}_{l,d})\right]$
  Update $E_{\text{feat}}, E_{\text{samp}}, D_{\text{feat}}, D_{\text{samp}}$ by minimizing $\mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{samp}}$
end for
Discard D_{feat} and D_{samp}

Output: Pretrained E_{feat} and E_{samp}

Algorithm 2 Robust Learning Across Sampling Variations in PRISM

Input: Labeled source data $\mathcal{D}_s$, pretrained encoders $E_{\text{feat}}, E_{\text{samp}}$, classifier H, augmentation distribution $\mathcal{T}$, number of variations K

Define $F(\mathbf{x}) \leftarrow H([E_{\text{feat}}(\mathbf{x}); E_{\text{samp}}(\mathbf{x})])$
for each minibatch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{B} \sim \mathcal{D}_s$ **do**
  for $i = 1, \ldots, B$ **do**
    Sample $T_{i,1}, \ldots, T_{i,K} \sim \mathcal{T}$
    $\mathcal{L}_{\text{prism}}^{(i)} \leftarrow \max_{k \in \{1,\ldots,K\}} \ell_{\text{CE}}\left(F(T_{i,k}(\mathbf{x}_i)), y_i\right)$
  end for
  Update F by minimizing $\frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{\text{prism}}^{(i)}$
end for

Output: Trained predictor F

D ADDITIONAL EXPERIMENTS

This section provides additional dataset details, analyses of sampling pattern shifts, and extended evaluations complementing the main experiments.

D.1 DATASET DETAILS AND ADDITIONAL ANALYSIS

Table A4 summarizes the statistics and evaluation settings of HAR-C, eICU, and PAM. HAR-C is constructed by subsampling the original regularly observed HAR sequences at a fixed 20% observation rate, whereas eICU and PAM retain their naturally irregular observation patterns. The three benchmarks provide complementary settings: controlled sampling pattern shifts on HAR-C, naturally occurring cross-region shifts on eICU, and sensor-level sampling shifts on PAM.

Figure A10 visualizes representative HAR-C samples under the eight sampling conditions. Each condition retains the same number of observations while changing their temporal and feature-wise locations, illustrating how HAR-C isolates changes in the sampling process from changes in the observation budget.

Table A4: Dataset statistics for the three benchmarks. Observed fraction is the share of the $T \times F$ lattice that carries a measurement: on HAR-C it is injected by subsampling a fully observed signal at a 20% rate ($\lfloor 128 \times 0.2 \rfloor = 25$ of 128 steps, synchronised across channels), whereas on eICU and PAM it is the native charting or recording density. eICU is evaluated leave-one-region-out, so its split sizes depend on the held-out region; the Northeast fold is shown, with negative / positive mortality counts in parentheses. HAR-C holds out 20% of the training subjects for validation.

	HAR-C	**eICU**	**PAM**
Task	Activity recognition	In-hospital mortality	Activity recognition
# Classes	6	2	7
# Features F	6	11	12
Sequence length T	128	24	50
Time window	2.56 s	24 h	—
Observed fraction	20% (25/128 steps)	38.5% (native)	24.9% (native)
Distribution shift	8 sampling-pattern shifts	Leave-one-region-out	Leave-sensors-out
# Test conditions	8	4 regions	5 severities
Train	5,881	22,543 (19,992 / 2,551)	4,228
Val	1,471	5,636 (4,994 / 642)	1,056
Test	2,947	2,495 (2,166 / 329)	1,320
Total	10,299	30,674	6,604

Table A5: Sampling augmentation settings used for pretraining and downstream robust learning.

Dataset	Pretrain (r_t, r_f)	Downstream (r_t, r_f)	K
HAR-C	$(0.5, 0.4)$	$(0.5, 0.4)$	4
PAM	$(0.5, 0.4)$	$(0.5, 0.4)$	4
eICU	$(0.5, 0)$	$(0.2, 0)$	4

To verify that sensitivity to sampling pattern shifts is not specific to the CNN used in the main analysis, we repeat the shift-severity experiment across Medium CNN, Large CNN, SeFT (Horn et al., 2020), mTAND (Shukla & Marlin, 2021), and GRU-D (Che et al., 2018). As shown in Figure A8, performance generally deteriorates as the fraction of shifted test samples increases across architectures.

We also report the full sampling-specific shortcut results at $\rho = 0.9$ in Table A6. Across DG methods, accuracy is substantially lower when the sampling cue conflicts with the ground-truth label than when it is aligned, complementing the analysis in Section 2.2.

D.2 Additional Evaluation of PRISM

We further evaluate PRISM on PAM under leave-fixed-sensors-out and across diverse backbone architectures on HAR-C. On PAM, the fraction of entirely unobserved sensors is varied from 10% to 50%, and PRISM remains substantially more robust as the shift becomes more severe. On HAR-C, the gains of PRISM persist across Medium CNN, Large CNN, SeFT, mTAND, and GRU-D, indicating that its effectiveness is not tied to the default CNN backbone.

We provide the full per-condition ablation results on HAR-C in Table A8 and the per-region results on eICU under leave-one-region-out evaluation in Table A9, complementing the summary in Table 3. Both tables additionally include variants that remove a pretraining objective together with sampling variations. On eICU, removing any component degrades the average AUROC and AUPRC, with the feature-centric branch showing the largest effect, although the relative ranking of variants can vary across individual target regions.

Finally, we probe the information captured by E_{feat} and E_{samp} at random initialization, after pretraining, and after downstream training. Feature-value information is measured by R^2, while sampling-pattern information is measured by AUROC. The results show that complementary pretraining encourages the two encoders to emphasize different aspects of the input.

Table A6: Sampling-specific shortcut analysis on HAR-C at $\rho = 0.9$. At test time, the sampling cue is bias-aligned, unbiased, or bias-conflicting with the ground-truth label.

Algorithm	Bias-aligned	Unbiased	Bias-conflicting
ERM	93.99±2.44	58.00±0.69	42.59±1.86
ARM	77.51±1.30	57.61±1.35	44.36±1.56
CORAL	97.76±0.15	54.54±0.69	34.89±0.94
DANN	98.03±0.08	57.79±0.61	40.18±0.95
GroupDRO	97.48±0.18	57.67±0.76	40.83±1.20
IB_ERM	93.20±2.91	54.87±0.73	36.73±1.24
IRM	96.66±2.12	56.25±0.86	39.67±1.59
Mixup	91.23±3.14	52.17±1.09	30.29±1.72
MLDG	98.01±0.20	57.57±1.04	41.62±1.86
MMD	97.59±0.11	57.37±0.51	41.40±1.26
VREx	85.27±1.66	48.49±1.52	39.95±3.08
Diversify	**98.09±0.09**	56.83±0.86	41.31±1.42
ManyDG	93.55±0.79	59.17±0.81	44.87±1.26
PRISM	92.29±2.59	**64.98±0.83**	**49.11±0.74**

Table A7: Comparison of PRISM and existing DG methods on PAM under leave-fixed-sensors-out evaluation. A fixed fraction of sensors is entirely unobserved at test time, with the missing-sensor ratio varied from 10% to 50%.

Algorithm	10%	20%	30%	40%	50%	Avg.
ERM	85.46±0.47	79.67±1.54	67.82±3.54	57.24±4.85	50.33±4.66	68.11±2.39
ARM	85.63±0.43	80.10±1.75	69.86±3.96	59.39±4.45	53.35±5.04	69.66±2.55
CORAL	85.85±0.45	80.51±1.73	69.01±3.92	59.70±4.82	53.87±4.90	69.79±2.69
DANN	85.34±0.46	79.51±1.73	69.01±3.68	57.30±4.23	52.03±4.65	68.64±2.31
GroupDRO	85.65±0.32	79.47±1.59	67.52±3.23	57.31±3.93	51.26±4.15	68.24±2.16
IB_ERM	85.66±0.37	80.49±1.44	68.58±3.67	58.78±4.34	52.14±4.40	69.13±2.33
IRM	85.71±0.45	79.47±1.70	67.52±3.85	58.29±4.17	52.28±4.08	68.65±2.30
Mixup	**86.18±0.32**	81.61±1.42	72.55±3.20	65.19±3.70	58.99±4.54	72.90±2.24
MLDG	85.82±0.51	81.15±1.35	71.06±3.86	61.26±4.35	54.22±4.57	70.70±2.33
MMD	85.93±0.51	80.57±1.78	67.92±3.80	58.08±4.85	51.54±4.75	68.81±2.55
VREx	85.89±0.38	80.32±1.44	69.47±3.26	60.64±3.84	54.14±3.90	70.09±2.07
Diversify	85.47±0.36	80.20±1.76	67.63±3.05	58.01±4.36	52.33±4.61	68.73±2.18
ManyDG	85.08±0.57	80.04±1.29	68.82±3.52	59.52±4.44	52.27±4.85	69.15±2.47
PRISM	85.48±0.28	**85.36±0.31**	**84.42±0.49**	**83.25±1.05**	**81.97±1.01**	**84.10±0.48**

Table A8: Ablation study of PRISM on HAR-C. We ablate $\mathcal{L}_{\text{feat}}$, $\mathcal{L}_{\text{samp}}$, E_{feat}, E_{samp}, and the sampling-variation and worst-view components of downstream training.

Variant	Random	Regular	Desync	Fixed-Feat.	Rand-Feat.	First	Last	Mid	Avg.
ERM	79.92±0.55	80.97±0.81	68.91±0.98	54.05±4.27	54.48±0.26	67.25±1.52	61.10±3.20	65.06±2.50	66.47±0.95
w/o $\mathcal{L}_{\text{feat}}$ & w/o sampling variations	81.12±0.39	83.47±0.42	73.29±0.90	55.18±3.88	57.09±0.63	73.36±1.00	70.89±2.31	72.31±1.50	70.84±0.68
w/o $\mathcal{L}_{\text{samp}}$ & w/o sampling variations	82.86±0.25	85.61±0.33	78.93±0.55	58.59±3.49	58.32±0.57	75.47±0.62	66.17±1.44	70.67±0.91	72.08±0.70
Pretraining objective									
w/o $\mathcal{L}_{\text{feat}}$	83.47±0.20	85.73±0.15	82.50±0.28	74.60±1.15	74.36±0.25	80.66±0.20	78.87±0.23	79.98±0.29	80.02±0.14
w/o $\mathcal{L}_{\text{samp}}$	**84.08±0.13**	86.31±0.21	**83.07±0.16**	74.50±1.19	74.68±0.18	81.01±0.14	**79.27±0.24**	80.84±0.23	80.47±0.15
Encoder									
w/o E_{feat}	83.08±0.17	84.96±0.21	82.05±0.23	72.88±1.27	73.69±0.16	80.31±0.19	78.20±0.24	80.09±0.24	79.41±0.16
w/o E_{samp}	83.47±0.20	85.58±0.22	82.68±0.20	74.18±1.12	73.96±0.15	81.12±0.26	78.62±0.23	80.54±0.39	80.02±0.14
Downstream objective									
w/o sampling variations (K=1)	82.08±0.34	83.97±0.66	76.73±0.62	56.29±3.73	58.35±0.41	75.71±0.66	70.60±0.96	72.46±1.13	72.02±0.43
Average instead of worst-view loss	82.96±0.23	85.69±0.18	81.33±0.21	73.94±1.36	73.85±0.18	80.55±0.23	78.66±0.41	79.66±0.21	79.58±0.17
PRISM	**84.08±0.16**	**86.49±0.21**	82.81±0.27	**74.90±1.29**	**74.72±0.21**	**81.28±0.20**	79.22±0.35	**81.02±0.23**	**80.57±0.19**

Table A9: Ablation study of PRISM on eICU under leave-one-region-out evaluation. We evaluate the same pretraining, encoder, and downstream-objective ablations as on HAR-C across the four target regions.

Variant	Northeast		West		South		Midwest		Avg.	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
ERM (Vapnik, 1991)	85.30±0.20	57.91±0.35	76.89±0.20	43.77±0.19	78.59±0.20	40.96±0.28	78.09±0.34	37.27±0.98	79.72±0.15	44.98±0.25
w/o $\mathcal{L}_{\text{feat}}$ & w/o sampling variations	83.75±0.25	55.65±0.52	77.94±0.16	44.70±0.26	76.71±0.21	38.52±0.30	76.89±0.21	35.66±0.47	78.82±0.07	43.63±0.15
w/o $\mathcal{L}_{\text{samp}}$ & w/o sampling variations	86.04±0.10	58.48±0.24	78.12±0.11	44.86±0.25	79.41±0.18	41.41±0.32	79.23±0.15	37.85±0.34	80.70±0.06	45.65±0.15
Pretraining objective										
w/o $\mathcal{L}_{\text{feat}}$	85.08±0.09	57.62±0.27	78.69±0.19	45.61±0.25	79.42±0.08	42.05±0.12	79.76±0.16	40.25±0.20	80.74±0.07	46.38±0.06
w/o $\mathcal{L}_{\text{samp}}$	86.71±0.13	58.96±0.19	78.92±0.22	45.68±0.40	80.91±0.18	43.38±0.24	81.21±0.09	40.89±0.24	81.94±0.08	47.23±0.16
Encoder										
w/o E_{feat}	85.83±0.13	58.15±0.20	79.21±0.14	46.28±0.24	79.70±0.10	42.29±0.16	80.00±0.19	40.42±0.27	81.19±0.05	46.79±0.13
w/o E_{samp}	86.99±0.07	59.37±0.21	79.20±0.13	46.47±0.15	81.34±0.08	43.61±0.23	**81.56±0.19**	41.15±0.39	82.27±0.06	47.65±0.15
Downstream objective										
w/o sampling variations (K=1)	87.00±0.14	**60.48±0.33**	79.85±0.16	47.40±0.18	80.15±0.12	42.43±0.23	80.47±0.12	39.99±0.24	81.87±0.07	47.58±0.15
Average instead of worst-view loss	86.92±0.13	60.35±0.28	**80.21±0.22**	**47.77±0.29**	80.63±0.08	43.09±0.23	80.91±0.10	40.43±0.29	82.17±0.07	47.91±0.15
PRISM	**87.15±0.09**	60.02±0.15	79.88±0.09	47.42±0.19	**81.35±0.13**	**43.92±0.14**	81.50±0.18	**41.64±0.28**	**82.47±0.04**	**48.25±0.08**

Table A10: Architecture generalization on HAR-C. We compare PRISM and DG baselines across Medium CNN, Large CNN, SeFT, mTAND, and GRU-D under the eight sampling conditions.

Algorithm	Random	Regular	Desync	Fixed-Feat.	Rand-Feat.	First	Last	Mid	Avg.
Medium CNN									
ERM	77.78±0.17	81.25±0.57	66.81±0.95	54.98±3.12	55.85±0.48	69.82±0.81	65.24±1.24	69.09±0.91	67.60±0.38
MLDG	80.48±0.21	84.03±0.29	71.75±0.91	56.36±2.86	55.55±0.50	71.67±1.28	67.01±0.97	72.34±0.54	69.90±0.43
ManyDG	78.92±0.45	82.90±0.47	70.24±0.62	52.69±2.95	55.00±0.61	71.12±1.09	63.35±1.41	69.44±1.05	67.96±0.75
Mixup	79.61±0.32	59.91±4.40	45.72±2.87	45.48±3.00	42.73±2.06	60.02±1.67	40.90±2.61	53.13±2.42	53.44±1.87
MMD	77.94±0.33	80.97±0.57	67.86±0.63	56.98±2.78	55.80±0.44	69.27±0.83	63.39±1.08	68.21±0.89	67.55±0.33
PRISM	**82.79±0.17**	**84.26±0.17**	**79.48±0.27**	**74.97±0.82**	**74.55±0.45**	**80.45±0.33**	**77.36±0.36**	**79.50±0.35**	**79.17±0.18**
Large CNN									
ERM	80.71±0.80	83.86±0.83	75.17±0.77	57.76±3.41	57.42±0.46	60.46±1.22	58.64±0.75	68.42±0.92	67.80±0.60
MLDG	**83.43±0.51**	**85.44±0.37**	77.46±0.72	57.67±2.95	57.77±0.48	62.84±2.53	52.69±2.85	66.04±2.09	67.92±0.92
ManyDG	82.21±0.38	84.89±0.59	76.54±0.41	55.58±3.56	56.95±0.35	74.51±0.96	68.80±2.25	73.95±0.96	71.68±0.60
Mixup	83.04±0.36	74.84±2.73	66.86±2.04	58.42±3.28	55.16±0.93	60.84±1.46	46.39±1.16	64.22±1.39	63.72±0.95
MMD	81.70±0.19	84.52±0.44	75.53±0.65	58.85±2.45	57.55±0.48	60.75±1.70	60.54±2.18	69.04±1.93	68.56±0.70
PRISM	83.26±0.17	84.84±0.20	**81.60±0.37**	**76.15±0.78**	**75.16±0.32**	**80.50±0.34**	**77.56±0.40**	**80.24±0.35**	**79.92±0.18**
SeFT									
ERM	72.87±0.78	75.04±0.75	71.55±0.82	38.43±4.37	34.18±0.83	70.48±0.80	66.34±1.72	69.96±0.94	62.36±0.85
MLDG	71.27±1.17	73.26±1.05	70.65±1.20	39.80±2.88	39.60±1.09	69.53±0.98	67.79±1.36	69.04±1.15	62.62±0.90
ManyDG	72.63±0.57	74.67±0.63	72.01±0.61	52.90±2.48	44.60±0.84	69.51±0.55	68.05±0.82	69.81±0.63	65.52±0.43
Mixup	75.82±0.58	**77.93±0.43**	75.28±0.35	44.73±4.73	42.84±1.27	72.58±0.83	70.47±1.04	73.15±0.54	66.60±0.79
MMD	72.68±0.77	74.51±0.67	71.75±0.63	37.29±4.90	35.66±1.16	70.43±0.69	66.69±1.02	69.17±0.77	62.27±0.68
PRISM	**76.15±1.71**	76.98±1.71	**76.22±1.72**	**68.80±1.42**	**66.64±1.65**	**74.36±1.78**	**74.93±1.61**	**75.18±1.52**	**73.66±1.55**
mTAND									
ERM	74.28±0.35	75.68±0.48	42.07±1.17	37.76±3.16	37.40±0.88	72.63±0.47	71.04±0.39	72.29±0.32	60.39±0.64
MLDG	73.46±1.54	75.54±1.71	43.61±1.72	36.82±3.11	38.11±0.95	71.84±0.96	71.43±1.08	71.96±0.96	60.35±1.21
ManyDG	**86.94±0.16**	**89.52±0.22**	61.02±1.08	40.08±2.00	41.44±1.30	**83.43±0.23**	**81.84±0.17**	**82.52±0.35**	70.85±0.34
Mixup	78.00±0.96	80.39±0.91	47.07±1.05	36.68±3.16	38.77±0.61	75.71±0.69	73.62±0.68	75.33±0.59	63.20±0.77
MMD	78.09±1.21	80.40±1.10	46.18±1.18	43.23±1.62	38.95±0.80	74.83±0.83	73.78±0.85	74.76±0.84	63.78±0.75
PRISM	84.46±1.69	85.74±1.75	**67.81±1.67**	**76.38±1.71**	**76.45±1.39**	82.70±1.55	79.20±1.60	81.36±1.65	**79.26±1.59**
GRU-D									
ERM	81.30±0.43	83.95±0.51	77.63±0.57	48.65±3.22	46.92±1.22	75.20±0.73	71.98±1.25	76.25±0.91	70.23±0.77
MLDG	78.36±0.93	80.48±0.98	75.05±0.96	50.22±2.65	48.16±0.82	71.97±1.14	70.37±1.85	73.47±1.75	68.51±1.04
ManyDG	73.74±1.78	75.82±1.95	66.76±1.80	34.55±3.97	33.67±0.98	69.27±1.75	61.90±1.04	66.90±1.26	60.33±1.55
Mixup	83.35±0.38	85.59±0.40	78.01±0.63	50.54±2.56	50.30±0.64	79.32±0.41	76.78±0.51	79.95±0.32	72.98±0.40
MMD	81.09±0.60	84.33±0.50	77.35±0.33	48.13±3.10	47.85±1.06	75.67±0.47	71.74±1.37	75.67±0.88	70.23±0.73
PRISM	**86.84±1.57**	**87.58±1.63**	**81.21±1.38**	**77.34±1.92**	**77.15±1.36**	**84.65±1.64**	**79.85±1.64**	**82.80±1.38**	**82.18±1.51**

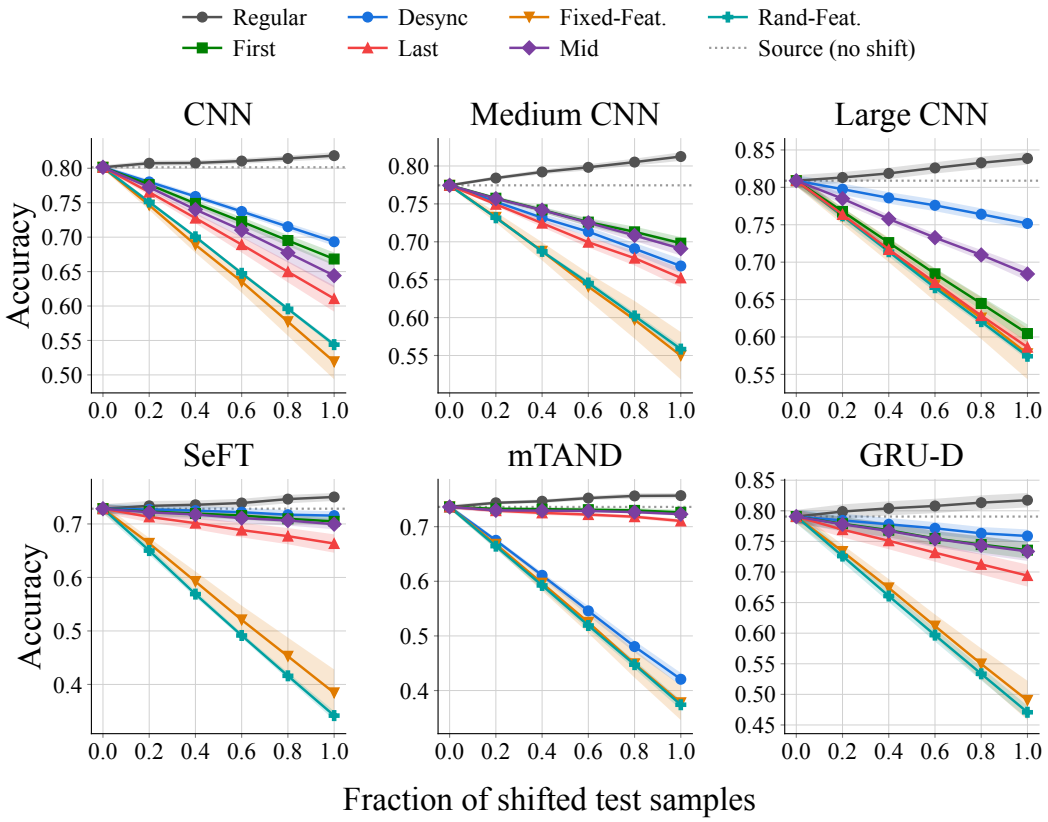

Figure A8: Further analysis of sampling pattern shifts across backbone architectures on HAR-C. We vary the fraction of shifted test samples for CNN, Medium CNN, Large CNN, SeFT, mTAND, and GRU-D.

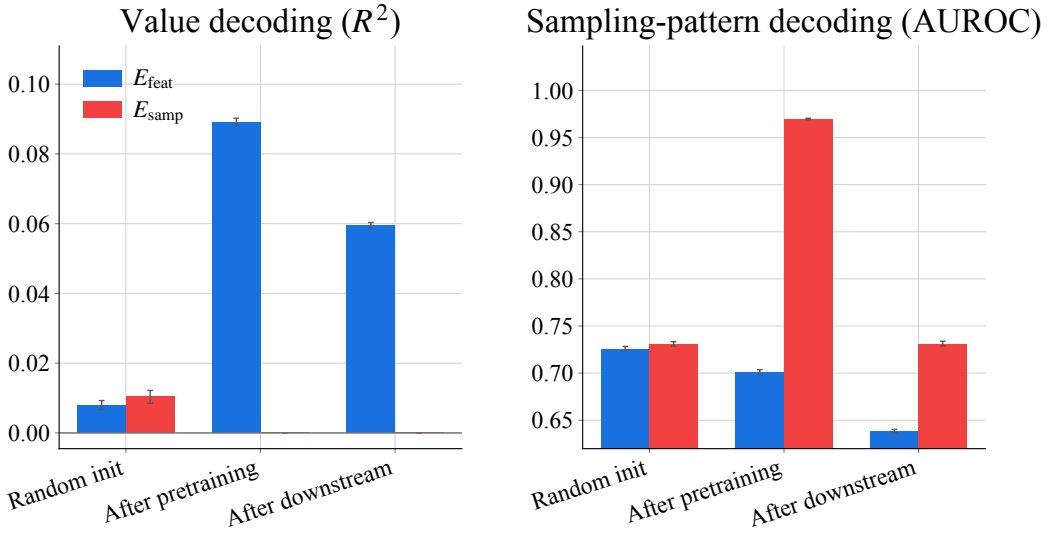

Figure A9: Representation analysis of the feature- and sampling-centric encoders. We measure feature-value information with R^2 and sampling-pattern information with AUROC at random initialization, after pretraining, and after downstream training.

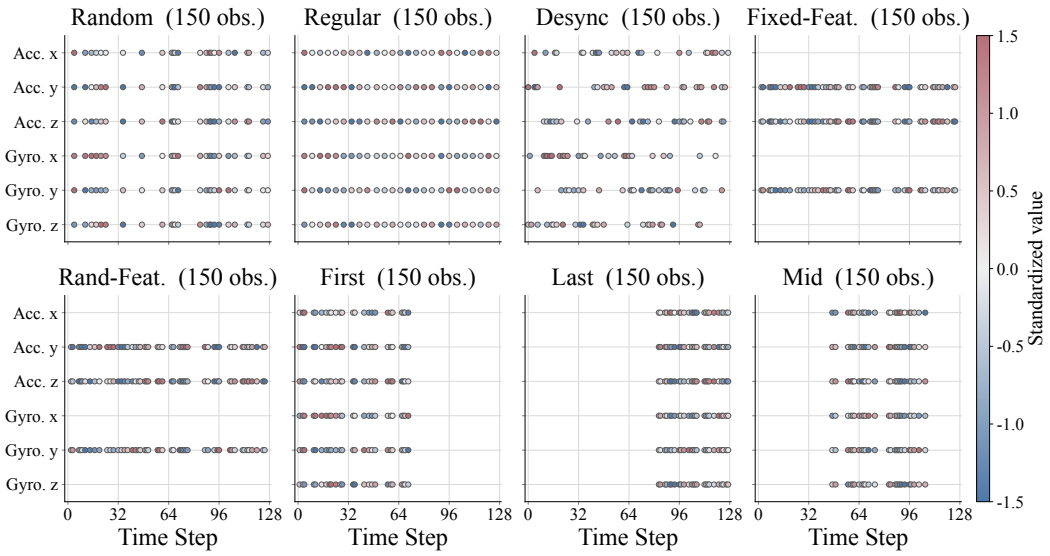

Figure A10: Examples of HAR-C observations under diverse sampling pattern shifts. Each condition retains 150 observations while varying their temporal and feature-wise locations; color indicates the standardized feature value.