HARP: LEARNING HETEROGENEITY-AWARE REPRESENTATIONS FOR PHYSIOLOGICAL HEALTH SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

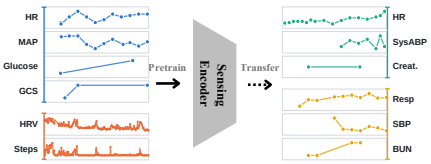
Health-sensing data supporting classification of clinical and everyday health states are naturally heterogeneous, where they differ in sensor sets, temporal resolutions, and missingness patterns across datasets. Despite advances in time-series pretraining, existing models mainly expect that target data is complete and regularly sampled. The transfer across real-world heterogeneous sensing configurations and the adaptation without target outcome labels remain challenging. We introduce **HARP**, a pretrained framework for health classification across heterogeneous sensing configurations. HARP uses masked-reconstruction pretraining over 50k participants with diverse sensor sets and observation length to enable learning from incomplete patterns. By combining physical duration patch embedding with canonical variable embedding, consistent temporal and variable meaning across datasets are preserved. For downstream classification, HARP supports both supervised adaptation and zero-shot transfer by a lightweight sensor-language bridge when target labels are unavailable. Across six clinical and wearable heterogeneous sensing datasets, HARP consistently outperforms state-of-the-art pretrained time-series models, improving average performance over the best pretrained model by 12% under linear probing and 10.3% under zero-shot transfer, while achieving competitive performance on three regularly sampled benchmarks. Beyond predictive performance, interpretability and scaling analyses show that HARP preserves temporal and variable meaning across various configurations and benefits from additional unlabeled pretraining data, supporting the learning of reusable heterogeneity-aware representations for real-world physiological sensing.

1 INTRODUCTION

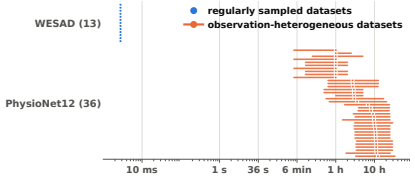
Clinical monitors and wearable devices produce multivariate physiological records, supporting a broad range of classification tasks across clinical care and everyday life, from identifying sepsis to recognizing well-being states (Harutyunyan et al., 2019; Abbaspourazad et al., 2024; Pillai et al., 2025). However, real-world health sensing datasets differ in how sensing signals are observed (Che et al., 2018; Horn et al., 2020; Zhang et al., 2024a; 2022). They record different variables, use different temporal resolutions, and cover different observation durations. For example, different studies use different sensor combinations, capturing distinct subsets of physiological variables. Even within a single dataset, measurements from different sensors may arrive asynchronously, and sensor streams may drop out for hours because of loose attachment. We refer to these differences as *heterogeneous sensing configurations* (Figure 1 (b)). This heterogeneity complicates modeling, as encoder designs suited to particular sensing variables and observation patterns may not transfer readily across datasets. This motivates a general-purpose sensing encoder that can learn from heterogeneous records and provide reusable representations across datasets (Figure 1 (a)).

Recent progress in general-purpose time-series pretraining offers a promising path toward reducing dataset-specific modeling (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). Time-series foundation models (TSFMs) such as MOMENT (Goswami et al., 2024) and Chronos-2 (Ansari et al., 2025) are pretrained across large time-series corpora, allowing learned temporal representations to be reused across domains. Time-series language models (TSLMs) further connect time-series representations with language to support downstream tasks that require cross-modal understanding, including natural-language reasoning, cross-modal retrieval, and zero-shot classification (Xie et al., 2025; Langer et al., 2026; Zhang et al., 2025; Li et al., 2025b).

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.



(a). Pretraining and transfer of a shared sensing encoder across datasets with different sensing configurations



(b). Variation in Observation Intervals

Method	Type	Cross-Dataset Pretraining	Heterogeneous Observations	Zero-Shot Classification
Chronos-2	TSFM	✓	✗	✗
MOMENT	TSFM	✓	✗	✗
NormWear	TSFM	✓	✗	✗
MIRA	TSFM	✓	✓	✗
ChatTS	TSLM	✓	✗	✓
OpenTSLM	TSLM	✓	✗	✓
SLIP	TSLM	✓	✗	✓
ATENet	Sup.	✗	✓	✗
PrimeNet	SSL	✗	✓	✗
LSM-2	SSL	✗	✓	✗
HARP (ours)	TSFM	✓	✓	✓

TSFM Time-series foundation models pretrained across datasets.
SSL Self-supervised pretraining on unlabeled records from each target dataset.
TSLM Time-series language models pretrained through sensor–text alignment.
HARP Cross-dataset encoder pretraining, followed by label-free bridge adaptation with frozen backbones.

(c). Comparison of Model Capabilities

Figure 1: **Motivation for heterogeneity-aware health sensing.** Heterogeneous sensing configurations differ in their variable sets (*cross-datasets*) and observation patterns, including sampling intervals, missingness, and observation durations (*heterogeneous observations*). **(a)**. A general-purpose sensing encoder. **(b)**. Intervals between consecutive measurements of a variable are fixed in WESAD but vary in P12. **(c)**. Representative methods compared on cross-dataset pretraining, heterogeneous observations, and zero-shot classification.

However, transferring existing models across heterogeneous sensing configurations remains challenging (Figure 1 (c)). Many are developed around regularly sampled inputs, using fixed-step representation learning (Nie et al., 2023) in which each step corresponds to a constant time interval, and thus may fail to capture physical measurement time under heterogeneous sampling resolutions with varying time intervals, a limitation especially consequential for health applications. For example, interpreting a change in intermittently measured glucose sensor for diabetes requires knowing when the readings were collected (Lutsker et al., 2026). Asynchronous measurements and observation gaps further complicate the interpretation of temporal patterns. Recent studies have started to address irregular and incomplete observations directly (Chowdhury et al., 2023; Xu et al., 2025). MIRA, for example, explicitly models irregular medical time series, but adopts a channel-independent representation learning that processes physiological variables separately for the forecasting task and also neglects the measurement time (Li et al., 2025a). Despite progress in general-purpose pretraining and irregular time-series modeling, existing pretrained encoders still show limited transfer performance on health classification tasks spanning across heterogeneous variable sets, temporal resolutions, and observation patterns as shown in Section 4.1.

Moreover, beyond the challenge of learning heterogeneity-aware representations, applying them to real-world health classification tasks requires adapting them to the target task. Common adaptation strategies, including linear probing and supervised fine-tuning, use labeled target examples to map sensing representations to outcome classes (Alain & Bengio, 2017; Kumar et al., 2022). In health applications, however, obtaining these labels may require clinical assessment, participant self-report, or dedicated study protocols (Shiffman et al., 2008; Ghassemi et al., 2020; Spathis et al., 2022). TSLMs offer a route to zero-shot classification via the power of the language backbone, but the development of this capability requires large-scale labeled time-text pairs and extra computation to align time-series encoder and language backbones during pretraining (Chen et al., 2026). Together, these challenges motivate our research question: *How can we learn transferable representations across heterogeneous sensing configurations and adapt them to health classification when target outcome labels are limited or unavailable?*

To address this gap, we introduce HARP, a pretrained framework for health classification across heterogeneous sensing configurations with a flexible adaptation mechanism. First, we propose a heterogeneity-aware sensing encoder that learns physical timing and variable identity embeddings, enabling consistent temporal and channel representations across diverse sensing configurations. The encoder is pretrained via masked reconstruction to yield domain-transferable representations. Second, we introduce lightweight adaptation strategies for both supervised and zero-shot transfer. For the zero-shot setting, we propose a sensor-language bridge that aligns sensing representations with automatically generated textual descriptions of unlabeled target records, enabling label-free classification without updating any pretrained weights.

We benchmark HARP against state-of-the-art (SOTA) pretrained models on six real-world health classification tasks spanning diverse heterogeneous sensing configurations. The results demonstrate

gains under supervised and zero-shot adaptation, while preserving competitive performance on three regularly sampled benchmarks. These findings highlight the transferability of our heterogeneity-aware representations and confirm that the strong performance of existing pretrained models on conventional benchmarks does not necessarily translate into effective transfer across heterogeneous sensing configurations. Our contributions are:

- We present HARP, a new pretrained encoder specially designed for heterogeneous health sensing data classification. The architecture explicitly embeds temporal and variable identities via reconstruction learning from incomplete observations.
- We explore and develop downstream adaptation strategies for different levels of target supervision, including a frozen-backbone sensor-language bridge to enable zero-shot transfer.
- Extensive results across real-world health tasks verify an average relative performance gain of 12% under linear probing and 10.3% under zero-shot transfer over the best SOTA pretrained models, while retaining competitive performance on regularly sampled benchmarks. We also analyze scalability, interpretability, and robustness to characterize HARP's transferability across heterogeneous sensing configurations.

2 Related Work

Time Series Foundation Models. Large-scale pretraining enables reusable time-series representations across datasets and classification tasks (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). MOMENT uses masked reconstruction to learn general-purpose temporal representations (Goswami et al., 2024), while Chronos-2 uses group attention to share information across related series and variables. For health sensing, modality-specific models learn transferable representations for EEG (Yoon et al., 2026) and ECG (Na et al., 2024), while NormWear (Luo et al., 2026) extends pretraining to multimodal physiological signals. However, these models generally assume data are complete and regularly sampled whereas real-world heterogeneous health data often are sampled irregularly, recorded at different rates, and contain substantial missingness. Consequently, existing representations may struggle to preserve temporal characteristics as shown in Figure 6. In contrast, our work preserves physical-time structure and temporal missingness when pretraining a shared encoder across incomplete records with different sensing configurations.

Learning from Irregular and Incomplete Time Series. Methods like ATENet, PrimeNet, and LSM-2 explicitly account for irregular observation times through supervised or self-supervised learning (SSL) (Shukla & Marlin, 2021; Lee et al., 2025). PrimeNet builds time-sensitive SSL pretraining prefix with unlabeled irregular series (Chowdhury et al., 2023). LSM-2 extends masked autoencoding to incomplete wearable recordings using an adaptive and inherited masking strategy (Xu et al., 2025). However, the encoders used by PrimeNet and LSM-2 are tied to predefined sensor sets, limiting their direct reuse across different variable sets. MIRA learns channel-independent representations of irregular medical time series, enabling reuse across different sensor sets. However, it does not explicitly model cross-variable dependencies that may benefit classification (Li et al., 2025a). Our shared encoder instead aligns variable identities across datasets and jointly models temporal and cross-variable dependencies, supporting transfer across different variable sets. Section 4.1 compares these methods through cross-dataset linear probing and in-domain SSL evaluation.

Sensor-Language Alignment for Zero-shot Transfer. Applying pretrained sensing encoders to downstream classification typically requires labeled target examples. Sensor-language alignment offers a route to zero-shot classification (Langer et al., 2026; Wang et al., 2025). For example, SensorLM and SLIP learn aligned sensing and language representations through paired sensor-text pretraining (Zhang et al., 2025; Chen et al., 2026). However, this alignment does not ensure effective transfer across heterogeneous sensing configurations as shown in Section 4.1. More importantly, existing approaches mainly develop such capability during pretraining requiring large-scale paired semantic-aware or labeled paired sensor-text corpora and substantial pretraining computation. HARP instead introduces a lightweight sensor-language bridge to align pretrained sensing representations with generated descriptions directly from unlabeled target records during the downstream adaptation phase, while keeping both the sensing and language backbones frozen. Compared with existing TSLMs, HARP achieves higher zero-shot AUROC on five of six clinical and wearable datasets (Table 2), demonstrating effective transfer capability across heterogeneous sensing configurations without target outcome labels.

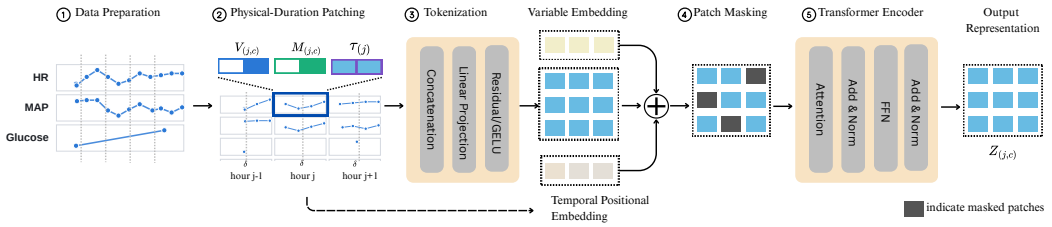


Figure 2: **Overview of the HARP sensing encoder.** Binned values, observation masks, and elapsed bin times are grouped into fixed-duration patches and embedded by a shared MLP. Temporal and variable identity embeddings are added before Transformer encoding.

3 METHOD

Overview. HARP is a two-stage framework for learning and adapting representations of heterogeneous health-sensing data. First, we pretrain a shared sensing encoder using masked reconstruction, explicitly encoding time and variable identity across different sensing configurations as shown in Figure 2. Second, before test-time inference, we adapt the frozen pretrained encoder to an unlabeled target dataset by contrastively training a lightweight sensor–language bridge on descriptions generated from its sensing records as shown in Figure 3.

3.1 ENCODER

The encoder maps each sensing record to a sequence of time-variable tokens by grouping asynchronous measurements into fixed-duration patches, each embedded with temporal and variable identity encodings. An observation mask is retained to preserve the original missingness structure. The encoder is pretrained via masked reconstruction to learn heterogeneity-aware representations that transfer across sensing configurations.

Data Preparation. Each sensing record contains measurements of C variables collected at potentially different times and sampling rates. For example, a record may contain heart-rate readings every minute and glucose measurements at only a few irregular time points. A general-purpose masked-reconstruction encoder often requires fixed-length inputs, so we preprocess asynchronous measurements into N uniform bins of width r by dividing a sensing record window of duration W (Harutyunyan et al., 2019; Che et al., 2018). Measurements of the same variable that fall in one bin are averaged, which yields a value matrix $\boldsymbol{V} \in \mathbb{R}^{N\times C}$. We also assign an observation mask $\boldsymbol{M} \in \{0,1\}^{N\times C}$ to each bin, where $M_{ic} = 1$ means at least one measurement in bin i is observed. Nothing is imputed, and the missing pattern and gaps of the original record remain. The choice of r for each dataset is given in Appendix A.5.

Physical-duration Patch Embedding. After binning asynchronous measurements, we group consecutive bins of each variable into patches, reducing the sequence length for more efficient temporal modeling (Nie et al., 2023). Existing TSFMs often group consecutive steps or bins into patches with a fixed number of steps. These patches would span different physical durations when sampling resolutions differ, so the same patch index can correspond to different measurement elapsed times across datasets. For example, 4 steps may cover 2.7 hours, while another 4 steps in a different dataset span 4.5 hours. We then define a patch as a fixed physical duration δ by grouping $p = \delta/r$ consecutive bins of each variable. (*e.g.*, a one-hour patch contains 6 ten-minute bins or 60 one-minute bins.)

For each patch j of variable c, we concatenate its bin values, observation masks, and bin start times into $\mathbf{x}_{j,c} = [V_{(j,c)} \,\|\, M_{(j,c)} \,\|\, \tau_{(j)}] \in \mathbb{R}^{3p}$. A shared MLP ϕ maps this input to a d-dimensional patch representation $\phi(\mathbf{x}_{j,c})$, thereby encoding its measurements, missingness, and timing. To accommodate different patch widths, we resize its input weights by interpolating separately the value, mask, and time blocks (Beyer et al., 2023; Chen et al., 2026). Together, this produces patch representations with a consistent physical duration and embedding dimension across temporal resolutions.

Temporal and Variable Identity Embedding. To preserve consistent meaning across sensing configurations including what sensor was measured, where observations are missing, and when the measurements are represented on the grid, each patch also receives a temporal and a variable identity embedding. First, we assign a learnable temporal positional embedding $\mathrm{pos}_t[j] \in \mathbb{R}^d$ by encoding

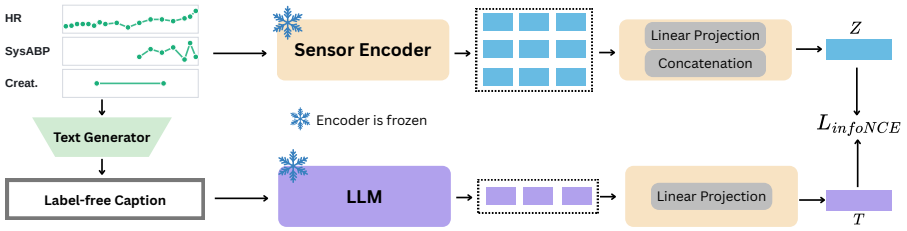


Figure 3: **Zero-shot transfer through the sensor–language bridge.** During the downstream classification stage, each unlabeled target training record is paired with a template-generated description; the frozen sensing encoder and frozen language model produce embeddings that are mapped by lightweight trained projections into a shared space and aligned with a symmetric contrastive loss.

each patch's exact time and add it to each patch representation. Because each patch spans the same physical duration, patch index j corresponds to elapsed start time $j\delta$. Therefore, we use this physical time as temporal position to learn the shared embedding across variables and datasets. These embeddings act like positional embeddings shared across heterogeneous sensing configurations, providing a consistent reference across records with different sampling resolutions and observation durations. For example, with one-hour patches, the patch starting at hour 30 receives the same positional embedding in both 48-hour and 60-hour records.

Second, datasets may record the same physiological variable under different names and channel orders, so channel positions alone cannot provide consistent variable identities (Ansari et al., 2025). We therefore propose a fixed mapping $a(c)$, manually defined from dataset variable definitions, to assign each variable an integer canonical ID. For example, SysABP in P12 and SBP in P19 share the same canonical ID for systolic blood pressure. Unseen variables will be identified as `unk`, details of vocabulary set handling are described in Appendix A.5.2. Therefore each identity ID receives a learned embedding $E_{\text{var}}[a(c)] \in \mathbb{R}^d$. Both temporal and variable embeddings are randomly initialized and learned during pretraining. We mask and flatten these tokens into a transformer encoder (Enc) across time and variable to keep dataset-specific temporal and variable dependency. The resulting token $Z_{j,c}$ carries its measurement content, temporal position, and variable identity.

$$Z_{j,c} = Enc(\phi(\mathbf{x}_{j,c}) + \text{pos}_t[j] + E_{\text{var}}[a(c)]) \in \mathbb{R}^d. \tag{1}$$

Reconstruction-based Pretraining Objectives. We pretrain the encoder via masked reconstruction. Sensing records are already naturally sparse, so we handle natural missingness and artificial masking separately to avoid forcing the encoder to reconstruct from empty patches. Concretely, a patch is treated as naturally missing ($m^{\text{nat}}_{(j,c)} = 1$) if all its bins are unobserved. Over the remaining observed patches, we randomly apply an artificial mask m^{art} across values, time, and variables. Masked embeddings are replaced with learned mask tokens, with 50% of tokens dropped in total. A lightweight decoder reconstructs the original patch values $\hat{V}_{(j,c)} \in \mathbb{R}^p$. The loss is computed only over artificially masked patches $\Omega = \{(j,c) : m^{\text{art}}_{j,c} = 1\}$, and within each patch only over observed bins, ensuring naturally missing values are never reconstruction targets:

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{(j,c)\in\Omega} \left\| \hat{V}_{(j,c)} - V_{(j,c)} \right\|^2_{M_{(j,c)}} \tag{2}$$

3.2 Zero-shot transfer through a sensor-language bridge

In addition to supervised adaptation using target labels, HARP supports zero-shot classification through a lightweight sensor–language bridge. After pretraining, we train this bridge on unlabeled target-training records paired with generated descriptions, while keeping both the sensing encoder and language backbone frozen. The trained bridge enables classification by comparing sensing representations with textual class descriptions, without requiring target outcome labels.

Label-free Caption construction. To provide alignment supervision and establish the sensor–language interface that enables zero-shot prediction, we pair each sensor record with its textual description. We adapt the statistical and structural captioning strategy of SensorLM (Zhang et al.,

2025) to irregular, incomplete records. Descriptions are computed from observed measurements before binning. Specifically, statistical descriptions summarize each observed variable's mean, range, and variability. Structural descriptions only report increasing or decreasing trends and extreme readings, together with their actual observation time. Unlike previous captioning methods, HARP's generator describes the measurements without using target labels and annotation-dependent semantic descriptions. Therefore, this yields sensor-language pairs without requiring regular sampling or labeled events. Rules for handling insufficient observations, and full caption templates and examples are provided in Appendix A.1.

Lightweight Frozen-backbone Alignment. Inspired by the contrastive alignment across different modalities (Radford et al., 2021; Elizalde et al., 2023; Zhai et al., 2022), we train lightweight projections on the outputs of the frozen sensing and language backbones to enable zero-shot transfer as shown in Figure 3. This confines target adaptation to the bridge, allowing the pretrained sensing and language representations to be reused. On the sensor side, we first obtain each variable representation by averaging over patches containing at least one observation, assigning a zero vector to a variable with no observations. A shared linear projection with GELU then transforms each variable representation. These representations are then concatenated in a fixed variable order and projected into a common embedding space. On the text side, a trainable linear head projects the frozen language model's last-token hidden state into the same space.

Contrastive Alignment and Prediction. We then train the bridge contrastively to assign higher similarity to matched record–caption pairs than to mismatched pairs within a batch. Let $\mathbf{z}_i$ and $\mathbf{t}_i$ denote the unit-normalized sensor and caption embeddings for the i-th pair in a batch of size B. We minimize the symmetric InfoNCE loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\alpha\,\mathbf{z}_i^\top\mathbf{t}_i)}{\sum_{j=1}^{B}\exp(\alpha\,\mathbf{z}_i^\top\mathbf{t}_j)} + \log\frac{\exp(\alpha\,\mathbf{t}_i^\top\mathbf{z}_i)}{\sum_{j=1}^{B}\exp(\alpha\,\mathbf{t}_i^\top\mathbf{z}_j)}\right], \tag{3}$$

After bridge training, we encode textual descriptions of the candidate classes (Appendix A.1) through the same text branch to define class prototypes $\{\mathbf{q}_k\}_{k=1}^{K}$. For a record x, prediction selects the prototype with the highest cosine similarity:

$$\widehat{y}(x) = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ \mathbf{z}(x)^\top\mathbf{q}_k. \tag{4}$$

For binary outcome evaluation, we use the cosine margin $s(x) = \mathbf{z}(x)^\top(\mathbf{q}_+ - \mathbf{q}_-)$ to rank records. All parameters remain fixed during prediction, and captions for test records are unnecessary.

4 Experiments and Results

Datasets. We evaluate on six clinical and wearable health-sensing datasets, spanning approximately 1.7k to 77.2k records and 4 to 36 variables per dataset. The tasks comprise in-hospital mortality prediction (MIMIC-III, MIMIC-IV, and P12), sepsis classification (P19), eight-class activity recognition (PAM), and five-class fatigue classification (PMData). We follow established configurations for the clinical datasets and PAM (Lee et al., 2025). HARP is pretrained without outcome labels on the training splits of MIMIC-IV and P12. Dataset statistics, preprocessing, and missingness characteristics are detailed in Appendix A.2.

Baselines. We compare with four groups of methods: time-series foundation models (Chronos-2, MOMENT, NormWear, and MIRA); time-series language models (OpenTSLM, SLIP, and ChatTS); open-weight language models (Gemma 3 and Llama 3.2); and self-supervised methods for irregular or incomplete sensing data (PrimeNet and LSM-2). We use ATENet as a strong supervised supervised reference. Model configurations and implementation details are provided in Appendix A.3.[1]

Metrics. We report AUROC on all six datasets, with AUPRC for binary tasks and macro-averaged F1 (MF1) for multiclass tasks where shown. Higher values indicate better performance. We evaluate over five folds and report the mean $\pm$ standard deviation across folds.

[1]Our code, pretrained checkpoints, and baseline implementations are available at `https://anonymous.4open.science/r/harp-1E55`.

Table 1: **Linear probing across heterogeneous health sensing benchmarks.** HARP achieves the highest mean AUROC among pretrained models across all six heterogeneous clinical and wearable benchmarks. **Bold** and underlined values indicate the best and second-best pretrained models. ATENet is a supervised reference excluded from the ranking. Gains is the improvement of HARP over the best baseline in each column.

Model	MIMIC-III		MIMIC-IV		P12		P19		PMData		PAM	
	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	MF1↑	AUROC↑	MF1↑
Supervised (reference):												
ATENet	85.14±0.99	39.37±3.59	87.18±0.52	49.40±1.28	85.92±0.68	53.83±1.85	84.02±1.38	41.16±3.02	57.31±3.21	17.54±3.62	99.07±0.48	94.80±1.60
Time-series foundation models (TSFMs):												
Chronos-2	78.27±0.80	26.91±1.15	80.49±0.36	35.46±1.75	72.83±1.83	37.33±1.59	81.31±2.96	40.67±4.07	56.87±3.58	19.61±0.71	90.67±0.64	89.79±0.52
MOMENT	73.84±1.02	19.54±1.32	75.54±0.56	26.24±0.65	67.35±0.82	28.92±1.92	81.78±1.62	42.12±3.83	54.19±3.45	17.15±1.90	86.95±2.14	87.87±1.44
NormWear	74.72±1.45	22.33±2.38	78.10±0.22	32.85±0.59	71.42±1.95	30.62±2.56	57.40±3.89	4.45±0.96	52.32±4.65	20.74±1.02	99.64±0.08	**96.37±0.86**
MIRA	80.84±0.97	28.78±1.85	85.18±0.55	43.63±1.70	78.67±0.97	41.54±2.52	80.81±2.20	27.87±1.68	55.37±5.23	19.03±1.98	96.23±0.97	92.41±1.09
Time-series language models (TSLMs):												
OpenTSLM	69.49±2.19	16.90±0.49	70.96±1.07	21.54±1.14	68.74±2.24	26.77±2.28	76.36±2.60	27.87±3.89	57.88±4.31	20.85±1.97	96.46±0.36	92.10±0.79
SLIP	75.76±0.62	22.33±0.93	78.38±0.32	32.16±1.11	69.63±2.00	28.18±1.50	81.31±2.49	39.07±2.74	54.11±3.26	19.99±1.09	96.86±0.69	93.33±0.51
HARP (Ours)	**87.27±0.66**	**42.97±1.87**	**88.26±0.24**	**51.94±0.90**	**85.23±1.81**	**51.64±2.96**	**87.80±2.60**	**50.55±3.53**	**60.04±4.38**	**21.06±1.75**	**99.66±0.09**	96.03±0.74
Gains	**+6.43**	**+14.19**	**+3.08**	**+8.31**	**+6.56**	**+10.10**	**+6.02**	**+8.43**	**+2.16**	**+0.21**	**+0.02**	**-0.34**

4.1 Effectiveness of Heterogeneity-Aware Representations

Cross-dataset Linear-probing. We first evaluate whether the pretrained encoder provides reusable representations for heterogeneous health-sensing classification. In this setting, HARP is pretrained without outcome labels on MIMIC-IV and P12, then evaluate on these two datasets and four additional datasets. For each target dataset, we freeze the encoder and train a linear classifier using target-training labels. As shown in Table 1, HARP achieves the highest mean AUROC among the pretrained models on all six datasets, its AUROC exceeds the best baseline by average absolute gain of 4.0. Averaged over all twelve metric columns, the relative gain is 12.0%. The gains over existing methods suggest that pretraining representations for heterogeneous health sensing benefits downstream transfer. The gains are largest on the clinical datasets and smaller on the two wearable datasets (Details in Appendix B.8). We additionally compare these pretrained baselines on three regularly sampled benchmarks under the same protocol (Appendix B.1). HARP retains competitive AUROC on these benchmarks while achieving larger gains overall under heterogeneous sensing. Together, these comparisons suggest that existing SOTA methods cannot deal with heterogeneous sensing, and HARP enables heterogeneity-aware representation learning transferable across real-world health sensing configurations.

In-domain SSL. To assess our encoder's effectiveness on heterogeneous health-sensing data without cross-dataset pretraining, we compare with PrimeNet and LSM-2, which are SSL methods designed for irregular or incomplete observations. Each method is pretrained without labels on the target-training split, and the frozen encoder is then evaluated with the same linear-probing protocol, training a linear classifier using target-training labels. As shown in Figure 4, HARP-SSL achieves the highest mean AUROC among SSL methods on all six datasets, exceeding LSM-2 by 3.5% points on average. It also surpasses the supervised ATENet reference on five datasets. These results show that the benefits of our encoder and pretraining objective also appear under in-domain pretraining, extending beyond the use of a shared pretraining corpus. Detailed results are provided in Appendix B.4.

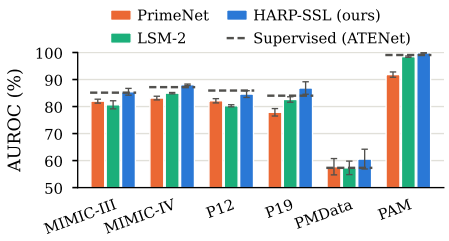


Figure 4: **In-domain linear-probing.** HARP achieves the highest mean classification AUROC among SSL methods on all six datasets.

4.2 Zero-shot Bridge

Zero-shot Classification. Here, zero-shot means *zero target labels* for training and adaptation. We fit the bridge on unlabeled target-training records paired with generated descriptions, keeping both pretrained backbones frozen. At test time, the trained bridge compares record embeddings with textual class descriptions. We compare with prompted LLMs and pretrained TSLMs (Implementation details are provided in Appendix A.4). Across all six datasets, HARP achieves a mean relative AUROC improvement of 10.3% over the strongest baseline for each dataset, ranking first on five

Table 2: **Zero-shot classification performance.** HARP achieves the highest mean zero-shot AUROC on five of six heterogeneous clinical and wearable benchmarks, demonstrating zero-shot transfer via a lightweight sensor–language alignment bridge.

Category	Model	MIMIC-III AUROC↑	MIMIC-IV AUROC↑	P12 AUROC↑	P19 AUROC↑	PMData AUROC↑	PAM AUROC↑
LLMs	Gemma3-270M	52.21±0.75	48.26±0.44	45.54±2.87	51.44±1.49	49.26±3.52	49.94±0.57
	Gemma3-1B	56.20±1.39	52.72±0.54	48.30±1.77	51.76±2.20	50.50±5.50	50.37±0.44
	Llama3.2-1B	55.92±1.15	52.22±0.67	47.18±1.48	55.16±2.72	49.25±1.48	49.61±0.37
	Llama3.2-3B	56.72±0.99	51.04±0.89	49.58±0.81	50.74±1.34	50.70±6.24	51.00±0.49
TSLMs	SLIP (Gemma3-270M)	55.59±0.49	53.52±0.66	55.34±1.52	53.20±2.85	49.14±3.45	**56.89±0.66**
	OpenTSLM (Llama-1B)	55.00±1.59	50.81±2.35	51.64±3.03	52.90±2.25	48.31±0.81	49.82±0.89
	ChatTS (Qwen-8B)	57.04±3.95	45.20±4.57	57.23±2.10	47.37±3.04	53.47±2.83	52.19±0.94
Zero-shot bridge	HARP (Gemma3-270M)	**67.88±1.64**	**64.03±2.39**	**62.09±3.73**	**63.11±1.65**	**54.58±4.94**	55.80±2.14
Gains		**+10.84**	**+10.51**	**+4.86**	**+7.95**	**+1.11**	**-1.09**

datasets (Table 2). Compared with existing LLMs and TSLMs with zero-shot capability, these results demonstrate that a lightweight bridge can also enable effective classification through textual class descriptions without target-label supervision. The slight drop in AUROC on PAM may be due to SLIP's pretraining on large-scale daily wearable and HAR datasets.

Bridge Ablations. We apply the same bridge to frozen TSFMs including MOMENT, Chronos-2, and MIRA encoders (Appendix B.2), demonstrating that the bridge can support label-free adaptation of other pretrained sensing representations. HARP achieves the highest mean AUROC on all six datasets under the same bridge, suggesting that the heterogeneity-aware representation remains important even when the alignment mechanism is shared. Besides, we compare contrastive alignment with the reconstruction-based alternative reported in Appendix B.3. Contrastive alignment achieves higher mean AUROC on all six datasets, with gains ranging from 3.50% to 16.11%. This suggests that learning to distinguish matched record-text pairs is better suited to the evaluated classification tasks than reconstructing descriptions, consistent with prediction by comparing a record against competing class descriptions (Zhang et al., 2024b).

4.3 Scalability

Having evaluated both shared-corpus and in-domain pretraining, we next examine how the composition and size of the pretraining corpus affect downstream performance.

Pretraining Dataset Scale. We pretrain on 25%, 50%, 75%, and 100% of the default MIMIC-IV + P12 corpus and evaluate the frozen encoder through linear probing. Figure 5 shows that performance generally increases with more data, demonstrating that additional unlabeled records improve the learned representations. Detailed results are in Appendix B.6.

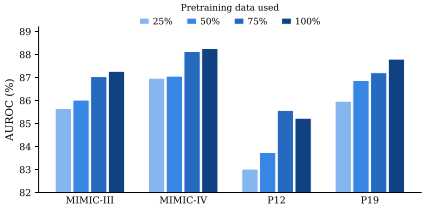


Figure 5: **Scalability performance.** More pretraining data generally improves downstream AUROC.

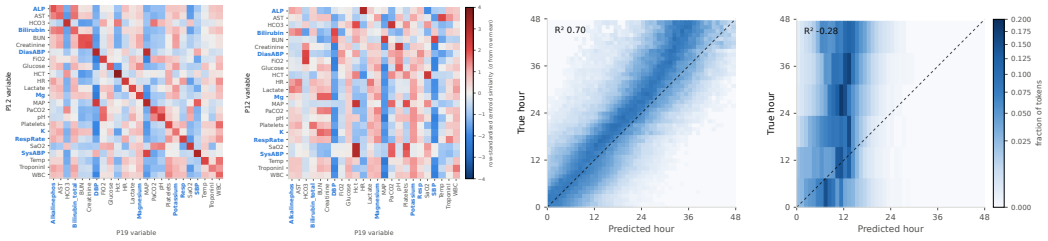
Pretraining Dataset Combinations. Varying the source datasets also changes downstream performance (Appendix B.5). A larger corpus is not uniformly better: MIMIC-IV + P12 uses 59.0k records yet achieves higher AUROC on P12 and PMData than MIMIC-III + MIMIC-IV with 83.4k records. One possible explanation is that MIMIC-IV + P12 exposes the encoder to more diverse variable sets and missingness patterns, supporting more transferable representations despite its smaller pretraining corpus. Together with the scaling results, this suggests that both the amount of pretraining data and the sensing configurations it covers matter for transferable representations.

4.4 Interpretability

We next examine whether the learned representations preserve variable correspondence and physical time across sensing configurations which may contribute to the performance improvements.

Variable Identity across Datasets. We compare HARP with Chronos-2 using channel-position embeddings. For each of the 24 variables shared by P12 and P19, we average the frozen encoder's observed-token outputs over test records within each dataset and compare these variable centroids by cosine similarity (Figure 6a,b). Retrieval measures the fraction of P12 variables whose nearest P19 centroid represents the same variable. HARP aligns learnt representations of the same physiological variable across datasets, achieving a mean retrieval rate of 0.82 across folds. In contrast, an encoder that only embeds channel position during pretraining retrieves 0.13, suggesting it cannot learn reusable representations of the same sensing variables. These results support the reuse of variable-specific knowledge across heterogeneous health sensing datasets.

Physical Time across Sampling Grids. We fit a ridge regressor to predict physical hour from frozen encoder outputs on MIMIC-III, evaluating hours 0-47. As shown in Figure 6c,d, HARP predictions concentrate within a narrow range of true physical hour. However, MOMENT tends to predict within a narrow range at first 14 hours. Together, these findings support HARP's ability to encode transferable physical-time information.



(a) HARP with E_{var} (b) Chronos2 without E_{var} (c) HARP with pos_t (d) MOMENT without pos_t

Figure 6: **HARP learns shared variables and time information. (a, b)**. Cosine similarity between average variable representations in P12 and P19. A stronger diagonal indicates better alignment of the same physiological variables across datasets. Blue labels mark variables with different names across datasets. **(c, d)**. Predicted versus true physical hour on MIMIC-III for HARP and MOMENT. Predictions closer to the dashed diagonal are more accurate.

4.5 Ablation Studies and Other Analysis

We ablate the key modules of HARP under the linear-probing setting, removing the explicit time encoding, or the channel (variable) identity embedding, and run the cross-domain linear probing again. Table 3 reports the results and average changes relative to the full model. We further examine sensitivity to bin width r and patch duration δ (Appendix A.5.1), finding that these choices should be considered jointly. To assess computational cost, we also compare parameter counts and pretraining time per epoch with baseline encoders (Appendix B.7).

Table 3: **Ablation performance.** Removing temporal encoding or variable identity reduces performance on all four clinical datasets. Avg Δ is the mean change in percentage points from the full model (shaded) across all eight metrics.

Model	Avg Δ	MIMIC-III		MIMIC-IV		P12		P19	
		AUROC$^\uparrow$	AUPRC$^\uparrow$	AUROC$^\uparrow$	AUPRC$^\uparrow$	AUROC$^\uparrow$	AUPRC$^\uparrow$	AUROC$^\uparrow$	AUPRC$^\uparrow$
HARP (full)	–	87.27±0.66	42.97±1.87	88.26±0.24	51.94±0.90	85.23±1.81	51.64±2.96	87.80±2.60	50.55±3.53
w/o explicit time	−3.46	85.68±0.36	38.72±1.42	86.33±0.72	45.84±2.77	80.88±1.96	44.08±3.57	86.32±2.41	50.13±4.24
w/o channel identity	−4.37	85.96±0.53	38.39±1.15	86.20±0.61	45.17±1.62	79.07±1.39	39.87±2.35	86.53±2.86	49.49±3.81

5 Discussions and Conclusions

Limitations. While promising, temporal binning limits fine-grained timing information, and variable alignment relies on a manually curated vocabulary mainly on clinical vocabulary and yields smaller gains on other domains. Our bridge currently supports zero-shot classification, future work could explore the language backbone's reasoning and generative capabilities during fine-tuning.

Conclusion. We present HARP, a framework for learning heterogeneity-aware representations of real-world health-sensing signals. By explicitly representing physical time, variable identity, and missingness, our shared encoder supports transfer across sensing configurations, complemented by a lightweight bridge for zero-shot adaptation. Experiments demonstrate effective supervised and zero-shot classification across clinical and wearable datasets.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, and to refine table layouts and formatting, and for code structure organizing. We have reviewed all AI-assisted work. Specifically, all AI-polished text was manually reviewed and revised by the authors to ensure accuracy and faithful representation of our contributions. We did not use generative AI tools for research ideation or execution, to draft sections of the paper, or to prove mathematical claims. The remaining required disclosure categories are not applicable to this work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide complete implementation details in the appendix, including model architecture specifications, hyperparameter configurations, training procedures, and dataset construction steps. All experimental settings required to reproduce the reported results are documented therein. Our code, pretrained checkpoints, and baseline implementations are available at `https://anonymous.4open.science/r/harp-1E55`.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. FlexiViT: One model for all patch sizes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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.

Yuliang Chen, Arvind Pillai, Yu Yvonne Wu, Tess Z Griffin, Lisa Marsch, Michael V Heinz, Nicholas C Jacobson, and Andrew Campbell. Learning transferable sensor models via language-informed pretraining. *arXiv preprint arXiv:2603.11950*, 2026.

Ranak Roy Chowdhury, Jiacheng Li, Xiyuan Zhang, Dezhi Hong, Rajesh K Gupta, and Jingbo Shang. PrimeNet: Pre-training for irregular multivariate time series. In *AAAI Conference on Artificial Intelligence*, 2023.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Joint Summits on Translational Science Proceedings*, 2020:191–200, 2020.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, 2020.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-LLM: Large language models for health prediction via wearable sensor data. In *Conference on Health, Inference, and Learning*, 2024.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, et al. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *International Conference on Machine Learning*, 2026.

Sangho Lee, Kyeongseo Min, Youngdoo Son, and Hyungrok Do. Adaptive time encoding for irregular multivariate time-series classification. In *Advances in Neural Information Processing Systems*, 2025.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical time series foundation model for real-world health data. In *Advances in Neural Information Processing Systems*, 2025a.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 354–379, 2025b.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 7(3):1–43, 2026.

Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, Eli Meirom, Eric P Xing, Gal Chechik, Hagai Rossman, and Eran Segal. A foundation model for continuous glucose monitoring data. *Nature*, 650:978–986, 2026.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Dimitris Spathis, Ignacio Perez-Pozuelo, Tomas I. Gonzales, Yu Yvonne Wu, Søren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments. *npj Digital Medicine*, 5, 2022. URL https://api.semanticscholar.org/CorpusID:253098961.

Yilin Wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, Lei Jia, Yuanxiang Li, and Zhongyu Wei. ITFormer: Bridging time series and natural language for multi-modal QA with large-scale multitask dataset. In *International Conference on Machine Learning*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18, 2025.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, et al. LSM-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Hyungjun Yoon, Seungjoo Lee, Yu Wu, XiaoMeng Chen, Taiting Lu, Freddy Yifei Liu, Taeckyung Lee, Hyeongheon Cha, Haochen Zhao, Gaoteng Zhao, Dongyao Chen, Cecilia Mascolo, Sung-Ju Lee, and Lili Qiu. Beyond hearing: Learning task-agnostic ExG representations from earphones via physiology-informed tokenization. In *International Conference on Learning Representations*, 2026.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60179–60196. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/zhang24bw.html.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024b.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, et al. SensorLM: Learning the language of wearable sensors. In *Advances in Neural Information Processing Systems*, 2025.

A Experimental Setup and Implementation Details

A.1 Caption Construction Details

Each training record is paired with a description made of two layers, following the statistical and structural captioning of SensorLM (Zhang et al., 2025). Both layers are computed from the observed sensing-record measurements at their original times, before binning.

Statistics report, for every observed variable, the mean, maximum, minimum, and standard deviation over its observed values. **Structure** reports local trends and extremes. Specifically, a least-squares line is fitted inside fixed 12-hour spans against the actual measurement times and kept as a trend clause when its rise over the span exceeds 15% of the variable's spread. An extreme is a reading whose robust z-score against the median (based on the median absolute deviation) exceeds 3.5, reported at the hour it was measured. SensorLM's semantic layer (activity, sleep, mood) is omitted by construction, since the only annotation on an ICU record is the outcome, and a description that carried it would break the zero-shot claim. **Insufficient observations.** To deal with high missing rate observations in heterogeneous sensing configuration, a variable with fewer than four observations receives no trend clause and only a statistics description, and variables that are never measured in a record are reported as having no observed values. These mechanisms are illustrated in Figure 7. We provide one example of a full caption below.

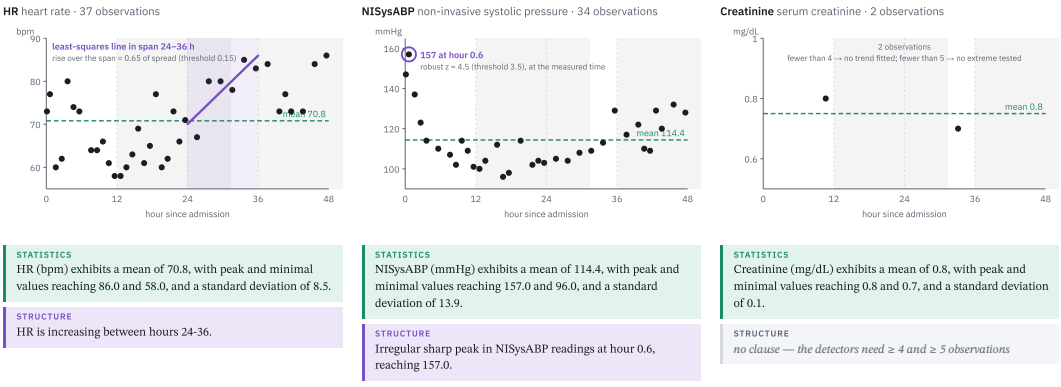


Figure 7: **Examples of generated captions of P12.**

Examples of full generated caption.

One full generated description, for P12 · subject 132539

STATISTICS — per organ system, over observed values only

For Cardiovascular, HR (bpm) exhibits a mean of 70.8, with peak and minimal values reaching 86.0 and 58.0, and a standard deviation of 8.5. NISysABP (mmHg) exhibits a mean of 114.4, with peak and minimal values reaching 157.0 and 96.0, and a standard deviation of 13.9. NIDiasABP (mmHg) mean, max, min, std are 50.1, 67.0, 39.0, 7.9. NIMAP (mmHg) exhibits a mean of 71.6, with peak and minimal values reaching 92.3 and 58.7, and a standard deviation of 9.1. For Respiratory, RespRate (breaths/min) shows a mean of 17.4, with range 24.0 to 12.0 and a standard deviation of 3.2. For Renal and metabolic, BUN (mg/dL) exhibits a mean of 10.5, with peak and minimal values reaching 13.0 and 8.0, and a standard deviation of 2.5. Creatinine (mg/dL) exhibits a mean of 0.8, with peak and minimal values reaching 0.8 and 0.7, and a standard deviation of 0.1. Urine (mL) exhibits a mean of 169.4, with peak and minimal values reaching 900.0 and 30.0, and a standard deviation of 167.6. Na (mEq/L) exhibits a mean of 136.5, with peak and minimal values reaching 137.0 and 136.0, and a standard deviation of 0.5. For Hematologic, WBC (cells/nL) exhibits a mean of 10.3, with peak and minimal values reaching 11.2 and 9.4, and a standard deviation of 0.9. Platelets (cells/nL) shows a mean of 203.0, with range 221.0 to 185.0 and a standard deviation of 18.0. HCT (%) mean, max, min, std are 32.5, 33.7, 30.3, 1.6. For Neurologic, GCS (3-15) shows a mean of 14.9, with range 15.0 to 14.0 and a standard deviation of 0.3. For Other, Temp (C) mean, max, min, std are 37.4, 38.2, 35.1, 0.9.

STRUCTURE — trends and extremes, at their measured hours

For Cardiovascular, HR is increasing between hours 24-36. NIMAP exhibits an increasing trend from hour 36 to 48. Irregular sharp peak in NISysABP readings at hour 0.6, reaching 157.0. For Respiratory, RespRate is increasing between hours 12-24. From hour 0 to 12, RespRate shows a decreasing trend. For Renal and metabolic, From hour 36 to 48, Urine shows a increasing trend. Urine exhibits an decreasing trend from hour 0 to 12. A sharp peak is detected for Urine at hour 0.1 (900.0). For Other, Irregular sharp drop in Temp readings at hour 0.6, reaching 35.6. Irregular sharp drop in Temp readings at hour 0.1, reaching 35.1. From hour 0 to 12, Temp shows a increasing trend.

393 words (cohort mean 431), ≈860 tokens; the 1536-token cap is not reached. Highlighted clauses are the ones drawn in Figure 7.

Class descriptions and test-time prediction. At test time, HARP takes a sensing record and the candidate class descriptions, with both backbones and the trained bridge frozen. The sensor branch encodes the record directly; the text branch encodes one sentence per class, using the same language backbone and projection as during bridge training. No description is generated for the test record. Prediction selects the nearest class embedding by cosine similarity (Eq. 4), as illustrated in Figure 8. For the binary clinical tasks, AUROC is computed from the margin between the two class embeddings; for the multiclass wearable tasks, the class embeddings are centered by their mean before the argmax. Class scores are calibrated by their rank among the training records' scores, without outcome labels. The class descriptions for the binary mortality task on P12 are shown below.

Class description — in-hospital mortality

DIED	The ICU monitoring summary indicates that the patient dies during the hospital stay.
SURVIVED	The ICU monitoring summary indicates that the patient survives to hospital discharge.

A.2 Dataset Details

Table 4 summarizes the six evaluation datasets, including the number of records, the number of recorded variables, the number of classes, the observation window, the bin width r, and the overall missing rate. The datasets differ substantially along every one of these axes, which is precisely the heterogeneity HARP is designed to absorb.

Preprocessing. MIMIC-III and MIMIC-IV[2] are processed following the mTAN pipeline (Shukla & Marlin, 2021), using the same 12 physiological variables. P12, P19, and PAM are used directly in

[2]MIMIC-III: `https://physionet.org/content/mimiciii/`;
MIMIC-IV: `https://physionet.org/content/mimiciv/`.

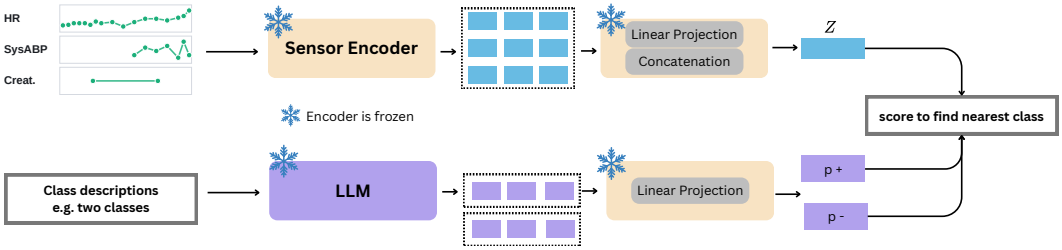


Figure 8: **Test-time zero-shot classification.** With both backbones and the trained bridge frozen, HARP matches a sensing record's embedding to the nearest textual class embedding. The text branch receives only the candidate class descriptions; no description of the test record is generated.

the processed form released with ATENet (Lee et al., 2025)[3], without further modification. For PMData[4], we follow the Health-LLM pipeline (Kim et al., 2024) to construct records from the wearable logs, but segment them into one-day windows rather than the one-week windows used there, so that each record covers a single day and supports the more fine-grained daily fatigue classification task.

Training splits. MIMIC-IV and P12 form the pretraining corpus and are therefore in-domain. The remaining four are fully held out. Train/val/test splits are 80/10/10 for P12, P19 and PAM, MIMIC-III and MIMIC-IV, and 60/20/20 for PMData.

Table 4: **Dataset statistics.** Summary of the six evaluation datasets. Missing rate is the fraction of empty bins after binning.

Dataset	Domain	#Records	#Train	#Variables	#Classes	Window	Bin r	Missing rate	Task
MIMIC-III	ICU	53,211	34.1k	12	2	48 h	10 min	94.5%	In-hospital mortality
MIMIC-IV	ICU	77,172	49.4k	12	2	48 h	10 min	92.2%	In-hospital mortality
P12	ICU	11,988	9.6k	36	2	48 h	10 min	96.2%	In-hospital mortality
P19	ICU	38,803	31.0k	34	2	60 h	60 min	79.4%	Sepsis
PAM	Wearable	5,333	4.3k	17	8	–	1 raw stpe	60.3%	Activity recognition
PMData	Wearable	1,722	1.0k	4	5	1 day	10 min	9.7%	Daily fatigue

A.3 Baseline Details

We group the baselines by the adaptation regime in which they are used. Unless stated otherwise, we use the publicly released checkpoint of each model, keep the encoder frozen, and train only the downstream classifier so that all methods are compared under an identical protocol.

ATENet (Lee et al., 2025) targets end-to-end classification of irregularly sampled multivariate time series. Its adaptive time encoding interpolates each variable onto a set of learned reference time points rather than a fixed grid, and temporal and inter-variable consistency regularizers encourage the representation to respect the dynamics within and across variables. It uses no pretraining and therefore serves as a fully supervised reference rather than a transferable baseline; we adopt the authors' architecture, hyperparameters, and dataset splits. Code: `https://github.com/shlee-labs/ATENet`.

Chronos-2 (Ansari et al., 2025) is a time-series foundation model for zero-shot forecasting. Extending the univariate Chronos family, it uses in-context learning with a group-attention mechanism that shares information across related series, so that univariate, multivariate, and covariate-informed forecasting are handled by one pretrained model without task-specific fine-tuning. We use the released 120M-parameter checkpoint and extract record-level representations from the frozen encoder. Code: `https://github.com/amazon-science/chronos-forecasting`.

MOMENT (Goswami et al., 2024) is a general-purpose time-series foundation model meant to serve forecasting, classification, anomaly detection, and imputation from one pretrained encoder. It is a

[3]P12, P19, PAM: `https://github.com/shlee-labs/ATENet`
[4]PMData: `https://datasets.simula.no/pmdata/`

T5-style encoder pretrained with masked patch reconstruction on the Time-series Pile, a large collection of public time series assembled by the authors, and it processes each channel of a multivariate input independently. We use the released 113M-parameter base checkpoint with the frozen encoder. Code: `https://github.com/moment-timeseries-foundation-model/moment`.

NormWear (Luo et al., 2026) is a foundation model for multivariate wearable physiological signals such as PPG, ECG, EEG, GSR, and IMU. It tokenizes each channel through a wavelet time–frequency representation, pretrains a masked autoencoder on 2.5M unlabeled segments from 18 public wearable datasets, and interleaves channel-aware fusion blocks with the encoder so that arbitrary channel combinations are encoded jointly. We use the released backbone checkpoint (about 136M parameters) with the frozen encoder; its separate text-fusion module for zero-shot inference is not used. Code: `https://github.com/Mobile-Sensing-and-UbiComp-Laboratory/NormWear`.

MIRA (Li et al., 2025a) is a medical time-series foundation model built for real-world clinical data with irregular sampling. It combines a continuous-time rotary positional encoding that embeds irregular timestamps, a frequency-specific mixture-of-experts layer that lets experts specialize across sampling regimes, and a neural-ODE block that extrapolates latent dynamics to arbitrary future times, and it is pretrained on a large clinical corpus for forecasting. Since no pretrained weights are publicly released, we follow the official implementation and pretrain it ourselves on the same corpus used for HARP, so that the comparison isolates the architecture rather than the pretraining data. Code: `https://github.com/microsoft/MIRA`.

OpenTSLM (Langer et al., 2026) is a family of time-series language models that add time series as a native modality of a pretrained LLM, so that several series of any length can be prompted and reasoned about in natural language. Its two variants inject the series either as soft-prompt tokens concatenated with the text or through Flamingo-style cross-attention, and both are trained with a curriculum that ends on chain-of-thought datasets for activity recognition, sleep staging, and ECG question answering. We use the released checkpoints trained through the ECG stage with a Llama-3.2-1B backbone: for linear probing we take the time-series encoder of the soft-prompt (SP) variant, and for zero-shot classification we use the full Flamingo variant with the same class descriptions given to HARP. Code: `https://github.com/StanfordBDHG/OpenTSLM`.

SLIP (Chen et al., 2026) learns transferable sensor representations through language-informed pretraining on more than 600K sensor–caption pairs. Its FlexMLP patch embedder shares weights across temporal resolutions and input lengths, which lets one encoder take heterogeneous sensor configurations, and the encoder is trained with contrastive alignment to captions together with sensor-conditioned captioning through a Gemma-3-270M language backbone. For linear probing we use its sensor encoder; for zero-shot classification we use the full model with the same class descriptions. Code: `https://github.com/yuc0805/SLIP`.

ChatTS (Xie et al., 2025) is a time-series multimodal LLM for understanding and reasoning over multivariate time series in natural language. It encodes each series into patches that are concatenated with the text tokens of the prompt, and it is trained on synthetic series generated from controllable attributes, which avoids the need for large human-annotated time-series–text corpora. We evaluate the released 8B model zero-shot with the same class descriptions (Appendix A.1). Code: `https://github.com/NetManAIOps/ChatTS`.

Gemma 3 (Gemma Team, 2025) and **Llama 3.2** (Grattafiori et al., 2024) are open-weight language models used as zero-shot references without any time-series-specific training: each test record is serialized as its generated textual description and included in the prompt together with the class descriptions, without target-label supervision. We evaluate several model sizes (Gemma 3 at 270M and 1B, Llama 3.2 at 1B and 3B) to separate backbone capacity from time-series understanding.

PrimeNet (Chowdhury et al., 2023) is a self-supervised pretraining method for irregular multivariate time series that needs no regular grid. It combines time-sensitive contrastive learning, whose augmentations respect the density of observations in time, with a reconstruction objective over masked observed values, so that the encoder learns from the irregular sampling pattern itself. Following its in-domain setting, we pretrain it separately on each target dataset and then apply the same linear-probing protocol. Code: `https://github.com/ranakroychowdhury/PrimeNet`.

Table 5: **Encoder configuration.** Architecture of the HARP encoder and decoder.

Component	Configuration
Bin width r / patch duration δ	10 min / 1 h ($p = 6$)
Patch embedder	MLP $3p \to 768 \to 192$ + skip
Temporal / variable tables	60×192 / 61×192
Encoder	4 blocks, $d = 192$, MLP 768
Decoder	2 blocks, $d = 128$, MLP 512
Parameters (encoder / total)	≈2.0M / 2.4M

LSM-2 (Xu et al., 2025) is a self-supervised model that learns directly from incomplete wearable sensor data without imputation. Its adaptive and inherited masking treats the naturally missing cells of a record as inherited masks alongside artificially sampled ones inside a masked-autoencoder framework, so that one model is pretrained and applied under real-world missingness; the original model was pretrained on 40M hours of day-long wearable data. As no official implementation is released, we follow the configuration described in the paper.

A.4 Implementation Details

Pretraining. Table 5 lists the encoder configuration. Records are binned at $r = 10$ min and grouped into one-hour patches ($p = 6$ bins). Each patch input $[V \,\|\, M \,\|\, \tau] \in \mathbb{R}^{3p}$ is embedded by the FlexMLP patch embedder, a two-layer MLP ($18 \to 768 \to 192$) with a linear skip, and receives temporal and variable identity embeddings from a 60×192 table (one row per elapsed hour, covering the longest 60-hour window) and a 61×192 table over the canonical vocabulary. Naturally missing and artificially masked patches are dropped together so that 50% of the tokens are hidden; the visible tokens pass through four pre-norm Transformer blocks of width 192 and a final LayerNorm, and the record representation is the mean over visible tokens. A lightweight decoder projects tokens to width 128, re-inserts a learned mask token at every hidden position together with decoder-side identity embeddings, applies two Transformer blocks of width 128, and reconstructs the p values of each patch. Every attention layer, in the encoder and the decoder, has 4 heads, and the MLP hidden width is four times the block width (768 and 512).

We optimize with AdamW (β_1=0.9, β_2=0.95, weight decay 10^{-4}), peak learning rate 1.2×10^{-3}, gradient-norm clipping at 1.0, and mixed precision. The learning rate warms up linearly over the first 5% of steps and then follows a cosine decay to zero. Batches hold 128 records from a single corpus. The loss is the mean squared error on observed cells inside artificially masked patches. The encoder has about 2M parameters and the decoder about 0.4M. All experiments, including pretraining, downstream adaptation, and the timings in Table 12, run on four NVIDIA RTX A6000 GPUs (48 GB each).

Linear-probing classification. During downstream evaluation, on top of the frozen record vector we fit the linear head of SLIP: a parameter-free BatchNorm (ϵ=10^{-6}, no affine), dropout, and one linear layer, trained with cross-entropy and AdamW for 50 epochs with gradient-norm clipping at 1.0. The learning rate warms up linearly over the first 5 epochs and then decays with a cosine schedule. For linear probing of the TSLMs, we use their time-series encoders as the pretrained encoders.

Zero-shot bridge. Here, zero-shot means classification without target-label supervision, while allowing bridge training on unlabeled target-training records. Both the sensing encoder and the language backbone remain frozen (Figure 3). The *sensing encoder* is read out per variable: for each of the C variables of the target dataset we average its observed tokens, giving a $C \times 192$ matrix per record, standardized per entry with statistics of the training records. The sensor side of the bridge is a linear map $192 \to 64$ shared across variables, a GELU, and one linear map from the concatenated $64C$ vector to the shared space of width 256. The *language model* is Gemma-3-270M, frozen; a description is encoded as the hidden state of its last token (640-d), standardized per dimension on the training descriptions, and mapped to the shared space by a trained linear layer. Both embeddings are unit normalized. The bridge is trained with the contrastive loss using AdamW (learning rate 10^{-3}, weight decay 10^{-2}), batch size 256, 40 epochs and a cosine learning-rate schedule.

Adaptation and inference. Bridge training uses only target-training records paired with descriptions generated from their measurements, without using outcome labels (Appendix A.1). The unlabeled training split also supplies the feature-standardization statistics and the reference scores for rank-based calibration. Test records are never used to fit the bridge, standardization, or calibration, and target outcome labels are used only for evaluation. At test time, all parameters, including the trained bridge, are frozen. As shown in Figure 8, HARP encodes the sensing record directly and compares it with the candidate class embeddings. Generated record descriptions are used for bridge training only; the text branch receives only class descriptions at inference.

Baseline inputs and target-data access. All methods in Table 2 share the same test-time task: predict a class from the test record and the candidate class descriptions, without parameter updates or access to the test label. Inputs follow each model's native modality. TSLMs receive the sensing record directly; SLIP and OpenTSLM use their released zero-shot evaluation procedures with class sentences or a task prompt specifying the candidate classes. Prompted LLMs receive the record's generated description together with the same class descriptions; this generated description is the textual serialization of the record. These baselines are not fitted on target records and do not access the target-training split. To compare encoders with the same access to unlabeled target-training data, Appendix B.2 additionally trains the same bridge on frozen MOMENT, Chronos-2, and MIRA representations under the same protocol as HARP.

A.5 Temporal Geometry and Variable Identity Mapping

A.5.1 Temporal Geometry: Cell Width and Token Duration

Two quantities define the temporal geometry of the encoder: the cell width r used to bin raw observations, and the physical duration δ of a token. This subsection reports the values we use and analyzes how much they matter.

Data preprocessing. The window W is 48 h for MIMIC-III, MIMIC-IV and P12, 60 h for P19 and 24 h for PMData; r is 10 min for MIMIC-III, MIMIC-IV, P12 and PMData and 60 min for P19, which is recorded hourly. PAM has no physical clock; its samples are placed on a pseudo-time axis with one sample per unit. Details are in Table 4.

Sensitivity to these choices. Figure 9 reports downstream AUROC on P12 with a frozen linear probe as δ varies, for three bin widths r. Two observations follow. First, every configuration falls within roughly two AUROC points of the others, so the encoder is not brittle to either choice. Second, shorter tokens are consistently better, and the penalty for long tokens grows as the bins become finer: at $\delta = 3$ h the 10-minute cell width loses about 1.6 points relative to the 60-minute one, since a token must then summarize 18 bins rather than 3. Short tokens are therefore our default.

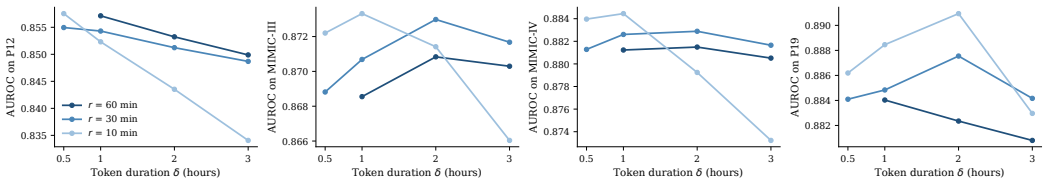


Figure 9: **Sensitivity to cell width and token duration.** Downstream AUROC on P12 (frozen linear probe) when the token duration δ and bin width r change during pretraining. Performance degrades gently with longer tokens for coarse bins and more steeply for fine ones, while all settings remain within about two AUROC points.

A.5.2 Curation of the Predefined Variable-ID Mapping

We construct a shared vocabulary from the variable definitions of the four clinical datasets: MIMIC-III, MIMIC-IV, P12, and P19. The vocabulary contains 52 canonical clinical variables and nine group-specific unknown entries, giving $E_{\text{var}} \in \mathbb{R}^{61\times d}$. Each entry has a fixed integer ID that identifies an embedding row; the IDs do not impose an ordering on clinical variables.

Alias curation. We collect the channel names, measurement definitions, and units from each dataset and manually associate equivalent measurements with a canonical entry. Dataset-specific names are

retained as explicit aliases. For example, SysABP in P12 and SBP in P19 both map to `sbp`. The resulting mapping is predefined and has no learned parameters; resolving a channel name requires an alias lookup, without text tokenization or a language encoder. Table 6 gives representative mappings.

Table 6: **Examples of representative mappings in the clinical vocabulary.** Aliases for the same measurement share a canonical entry; SpO2 and SaO2 remain separate.

Dataset-specific names	Canonical entry	Measurement
SysABP, SBP	`sbp`	Systolic blood pressure
RR, RespRate	`resp_rate`	Respiratory rate
ALP, Alkalinephos	`alkaline_phos`	Alkaline phosphatase
SpO2	`o2sat_pulse`	Pulse-oximetry oxygen saturation
SaO2	`sao2_abg`	Arterial blood-gas oxygen saturation

ID assignment and unseen variables. Each canonical variable receives a fixed ID, and each of the nine clinical groups has an additional unknown entry. For a dataset channel c, the alias mapping selects its canonical ID if that variable occurred in the pretraining corpus. A clinical variable outside the vocabulary or absent from the pretraining corpus is assigned its group's unknown ID. This also covers vocabulary entries whose individual embeddings received no pretraining updates. In Section 3.1, $a(c)$ denotes the resulting integer index, including this fallback.

Embedding learning. The selected vector $E_{\text{var}}[a(c)]$ is added to every patch of channel c. The embedding vectors are randomly initialized and optimized through reconstruction pretraining, while the alias mapping and ID assignments remain fixed. Consequently, records from different datasets update the same embedding whenever their channels share a canonical variable.

B Additional Analysis and Results

B.1 Linear Probing on Regularly Sampled Benchmarks

To verify that the gains under heterogeneous sensing do not come at the cost of performance on conventional data, we repeat the linear-probing protocol of Section 4.1 on three regularly sampled, complete benchmarks: HAR, Sleep-EDF, and PTB-XL (Table 7). HARP achieves the highest AUROC on Sleep-EDF and PTB-XL and is within one AUROC point of the best model on HAR. On macro-F1, NormWear leads on HAR and SLIP on Sleep-EDF and PTB-XL. The encoder therefore remains competitive on regularly sampled data, where its explicit handling of gaps and timing brings no additional benefit, while its advantage emerges under heterogeneous sensing configurations (Table 1). Read together, the two tables also show that strong performance on conventional benchmarks does not carry over: the same baselines that match or exceed HARP here fall well behind it once temporal resolution, observation patterns, and missingness vary across records and datasets. Their representations, learned from regularly sampled and complete inputs, are therefore not transferable to the heterogeneous setting, which is exactly the gap that HARP's heterogeneity-aware representation closes.

B.2 Applying the Sensor–Language Bridge to Other Pretrained Encoders

The sensor–language bridge is agnostic to the encoder it sits on: it only requires a frozen representation of a record. To verify this, we train the same bridge on top of pretrained time-series foundation models and evaluate zero-shot classification under the identical protocol, changing only the encoder. Table 8 reports the results. **Results.** Two conclusions follow. First, the bridge is a strong zero-shot mechanism in its own right: placed on generic TSFM encoders that were never aligned with text, it reaches 54 to 62 AUROC on the four clinical tasks, and bridged MOMENT exceeds the best prompted LLM or TSLM of Table 2 on all four of them without any time–text pretraining. Second, the encoder still matters: under the identical bridge, HARP is the best encoder on all six datasets, by 2.9 to 6.0 AUROC points over the strongest bridged baseline (4.3 on average). On the clinical tasks, the bridge on a generic encoder accounts for roughly four points of the zero-shot gain over prior TSLMs and the heterogeneity-aware encoder for roughly five more, so neither component

Table 7: **Linear probing on regularly sampled benchmarks.** All datasets are regularly sampled with complete observations. Following the protocol of Table 1, all pretrained encoders are frozen and only a linear classifier is trained on each target dataset.

Model	HAR		Sleep-EDF		PTB-XL	
	AUROC↑	MF1↑	AUROC↑	MF1↑	AUROC↑	MF1↑
Time-series foundation models (TSFMs):						
Chronos-2	98.47±0.14	88.29±0.72	88.22±1.18	61.36±2.08	78.73±0.85	48.40±0.83
MOMENT	98.08±0.14	87.80±0.65	87.23±1.34	58.98±2.25	81.69±0.73	51.34±1.10
NormWear	**99.58±0.12**	**95.27±0.99**	88.34±1.33	61.97±2.64	78.06±0.26	47.89±0.45
Time-series language models (TSLMs):						
OpenTSLM	97.97±0.13	83.39±0.52	85.39±1.26	54.87±2.24	74.63±0.24	41.62±1.07
SLIP	99.41±0.15	93.08±1.20	88.52±1.94	**62.40±3.74**	81.70±0.57	**52.01±0.52**
HARP (Ours)	98.77±0.18	88.41±0.68	**88.54±1.86**	61.72±3.56	**82.49±0.45**	48.22±1.23

Table 8: **Applying the sensor–language bridge to other pretrained encoders.** The same bridge is trained on top of each frozen pretrained encoder and evaluated zero-shot, with no target labels used at any stage. Rows differ only in the encoder that produces the sensing representation. Gains is the improvement of HARP over the best alternative encoder in each column.

Encoder (+ bridge)	MIMIC-III	MIMIC-IV	P12	P19	PMData	PAM
	AUROC↑	AUROC↑	AUROC↑	AUROC↑	AUROC↑	AUROC↑
MOMENT	61.37±1.36	55.06±1.48	59.15±3.16	58.71±2.01	51.35±2.18	52.48±1.05
Chronos-2	58.94±1.79	58.34±1.03	55.18±2.76	53.73±1.22	49.51±4.83	51.32±1.29
MIRA	61.91±2.72	56.98±1.33	57.21±2.30	54.24±1.16	50.24±5.79	51.89±1.31
HARP (Ours)	**67.88±1.64**	**64.03±2.39**	**62.09±3.73**	**63.11±1.65**	**54.58±4.94**	**55.80±2.14**
Gains	**+5.97**	**+5.69**	**+2.94**	**+4.40**	**+3.23**	**+3.32**

alone explains the result. All encoders remain close to chance on PMData and below the wearable-pretrained SLIP on PAM, indicating that zero-shot recognition of daily fatigue and activities from a few wearable channels is limited by the encoder's representation rather than by the bridge.

B.3 Bridge Ablations

HARP trains the sensor–language bridge with a contrastive (InfoNCE) objective that pulls each record embedding towards its own caption and pushes it away from the captions of other records in the batch. An alternative is to align the two modalities by reconstruction, as in MSiTF (Luo et al., 2026), where the sensing representation is trained to regenerate the textual description rather than to be discriminated among alternatives. Table 9 replaces our objective with this reconstruction-based alternative while keeping everything else fixed. The contrastive objective is substantially better on every dataset, which matches how the bridge is used at inference time: zero-shot prediction compares a record embedding against a set of class descriptions, so the bridge must learn to *discriminate* between candidate texts rather than to reproduce any one of them. A further difference concerns supervision. MSiTF as proposed in NormWear still depends on labeled records: its sensor–text alignment is trained by matching signal embeddings to template sentences built from the ground-truth labels of its training sets, so the label information enters the alignment before any "zero-shot" prediction is made. Our bridge never sees outcome labels; it aligns records to captions generated from the unlabeled target-training records themselves. In Table 9 both rows use these label-free captions, so the comparison isolates the objective, and the label-free setting is a strictly harder one than the original MSiTF protocol.

Table 9: **Bridge objective ablation.** Zero-shot classification when the contrastive objective used to train the sensor-language bridge is replaced by a reconstruction-based alternative.

Bridge objective	MIMIC-III	MIMIC-IV	P12	P19	PMData	PAM
	AUROC↑	AUROC↑	AUROC↑	AUROC↑	AUROC↑	AUROC↑
MSiTF (reconstruction)	51.77±5.47	52.59±10.68	52.97±5.23	56.57±2.19	51.08±1.40	48.48±2.38
InfoNCE (Ours)	**67.88±1.64**	**64.03±2.39**	**62.09±3.73**	**63.11±1.65**	**54.58±4.94**	**55.80±2.14**
Gains	**+16.11**	**+11.44**	**+9.12**	**+6.54**	**+3.50**	**+7.32**

Table 11: **Effect of pretraining dataset combinations on downstream performance.** Each row pretrains HARP on a different combination of source datasets under the same setting; downstream adaptation is identical across rows. The shaded row is the default configuration used in our main results.

Pretraining datasets	#Pretrain	MIMIC-III		MIMIC-IV		P12		P19		PMData		PAM	
		AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	MF1↑	AUROC↑	MF1↑
ATENet (reference)	–	85.14±0.99	39.37±3.59	87.18±0.52	49.40±1.28	85.92±0.68	53.83±1.85	84.02±1.38	41.16±3.02	57.31±3.21	17.54±3.62	99.07±0.48	94.80±1.60
P19	31.0k	77.70±1.23	24.28±2.04	78.46±1.25	29.36±2.38	70.74±2.68	26.89±1.35	87.09±2.56	49.07±4.61	54.04±5.46	22.18±1.76	99.58±0.11	95.18±1.04
MIMIC-III + P19	65.1k	86.30±0.30	39.91±1.61	87.55±0.55	48.71±1.43	84.35±0.33	48.81±2.97	88.68±2.29	52.01±3.71	55.60±4.68	22.03±3.08	98.40±0.39	88.73±1.91
MIMIC-III + P12	43.6k	86.40±1.22	40.28±3.25	86.85±0.94	47.90±2.35	83.76±0.88	46.54±3.68	87.24±1.82	48.56±1.70	53.27±3.70	20.25±1.45	99.70±0.08	96.13±0.48
MIMIC-IV + P19	80.4k	86.89±0.93	42.03±2.08	88.12±0.49	50.62±1.26	83.91±1.45	47.97±1.15	88.87±1.67	54.60±3.13	55.04±3.90	19.17±0.42	99.52±0.16	95.24±0.33
MIMIC-IV + P12 (default)	59.0k	87.27±0.66	42.97±1.87	88.26±0.24	51.94±0.90	85.23±1.81	51.64±2.96	87.80±2.60	50.55±3.53	60.04±4.38	21.06±1.75	99.66±0.09	96.03±0.74
MIMIC-III + MIMIC-IV	83.4k	87.41±0.88	43.29±2.23	88.31±0.39	52.00±1.80	84.32±0.53	50.33±1.47	86.99±2.12	48.13±6.31	53.09±4.68	20.76±2.37	99.67±0.07	95.44±0.43

B.4 In-Domain Self-Supervised Pretraining

Table 10 reports the full in-domain SSL comparison summarized in Figure 4, adding the secondary metric for each dataset (AUPRC for the binary clinical tasks and macro-F1 for the multiclass wearable tasks). HARP-SSL achieves the highest AUROC and the highest secondary metric among the self-supervised methods on all six datasets.

Table 10: **Comparison with in-domain SSL methods.** Self-supervised baselines designed for irregular sensing data are pretrained in-domain on each target dataset, and HARP-SSL applies our pretraining objective in the same in-domain setting. Figure 4 in the main text plots the AUROC columns. ATENet is a supervised reference.

Model	MIMIC-III		MIMIC-IV		P12		P19		PMData		PAM	
	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑	AUROC↑	MF1↑	AUROC↑	MF1↑
Supervised (reference):												
ATENet	85.14±0.99	39.37±3.59	87.18±0.52	49.40±1.28	85.92±0.68	53.83±1.85	84.02±1.38	41.16±3.02	57.31±3.21	17.54±3.62	99.07±0.48	94.80±1.60
In-domain SSL:												
PrimeNet	81.99±0.73	35.17±2.36	83.14±0.68	41.51±1.05	82.13±0.79	45.71±0.88	77.87±1.38	23.54±1.96	57.73±3.02	19.49±1.60	91.82±0.98	66.78±2.49
LSM-2	80.64±1.50	30.36±1.77	84.95±0.25	43.72±1.36	80.28±0.39	43.93±2.99	82.61±1.01	38.31±1.70	57.31±2.50	21.33±2.04	98.46±0.25	95.67±0.43
HARP-SSL (Ours)	**85.46±1.26**	**37.48±2.88**	**88.06±0.26**	**50.72±1.44**	**84.61±1.25**	**50.12±3.11**	**86.86±2.31**	**48.66±3.68**	**60.55±3.66**	**24.32±1.44**	**99.83±0.07**	**97.74±0.42**

B.5 Pretraining Dataset Combinations

We vary the source-dataset combination while keeping the downstream linear-probing protocol fixed as reported in Table 11 with AUPRC and macro-F1. The number of pretraining records is the sum of the training splits of the selected datasets. Adding MIMIC-III to P19 pretraining increases AUROC on five of the six datasets, but reduces PAM AUROC from 99.58 to 98.40. Thus, combining sources can improve transfer, but its benefit is task-dependent. No source combination achieves the highest AUROC on every benchmark: MIMIC-IV + P12 performs best on P12 and PMData, whereas MIMIC-IV + P19 performs best on P19. The larger MIMIC-III + MIMIC-IV corpus is strongest on the two MIMIC benchmarks but does not lead elsewhere. These results suggest that corpus composition should be considered alongside data scale. Because the combinations also differ in record count, this study does not isolate the effect of source diversity at a fixed data budget.

B.6 Per-Dataset Pretraining Data Scaling

Figure 10 reports the pretraining data scaling study of Section 4.3 for each benchmark separately, with an independent y-axis per panel so that per-dataset trends and their variance remain visible despite the wide differences in absolute performance across tasks.

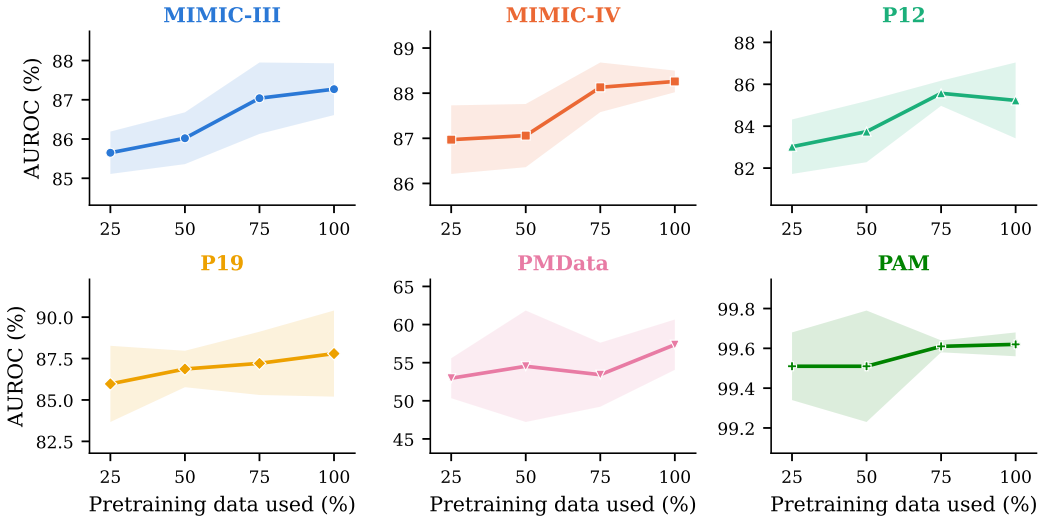


Figure 10: **Per-dataset effect of pretraining data scale.** Downstream AUROC on each of the six benchmarks when HARP is pretrained on randomly subsampled fractions of the default pretraining corpus (MIMIC-IV + P12) under the same setting. Shaded bands show ± 1 std. Note the independent y-axis ranges.

B.7 Model Size and Efficiency

Table 12 compares the number of parameters, inference time, and pretraining cost of HARP with the baseline pretrained encoders. HARP is among the smallest models, with about 2M parameters against more than 100M for MIRA, Chronos-2, and SLIP and about 340M for MOMENT; only the OpenTSLM time-series encoder is comparably small. It is also the fastest at inference, producing a record representation in 0.52 ms. Among the two models pretrained on our corpus, one pretraining epoch takes 65 s for HARP versus 15.2 min for MIRA, about 14 times longer, because MIRA's neural-ODE block requires repeated numerical integration during training.

Table 12: **Model size and efficiency.** Parameter counts of the model used to produce the record representation (the released model for the TSFMs, the time-series encoder alone for OpenTSLM and SLIP, and the encoder of HARP), inference time per record on the same hardware with each model's own batch size, and pretraining time per epoch for the two models we pretrain ourselves on the same corpus and hardware. The other baselines use released checkpoints, so no pretraining time is reported for them.

Model	#Params (M)	Inference (ms / record)	Pretrain time / epoch
MIRA	113.4	14.2	15.2 min
Chronos-2	115.8	25.4	–
MOMENT	341.2	104.7	–
OpenTSLM-enc	2.1	3.2	–
SLIP-enc	116.0	63.8	–
HARP (Ours)	2.0	0.52	65 s

B.8 Detailed Analysis of Cross-Dataset Linear Probing

Evaluation setting. Table 1 evaluates a shared encoder pretrained without outcome labels on the training splits of MIMIC-IV and P12. We distinguish three settings: evaluation within these pre-

training datasets, transfer to additional clinical datasets (MIMIC-III and P19), and transfer from clinical pretraining to wearable sensing (PAM and PMData). The latter four datasets are excluded from encoder pretraining. For every dataset, we freeze the encoder and train a linear classifier using target-training labels.

Under linear probing, HARP improves AUROC over the strongest pretrained baseline by 5.5 percentage points on average across the four clinical datasets, compared with 2.2 points on PMData and 0.02 points on PAM. On PAM, HARP and NormWear both achieve near-ceiling AUROC (99.66 and 99.64, respectively), leaving little room for further improvement on this metric. The wearable gains are also limited under zero-shot classification: HARP exceeds the strongest baseline on PMData by 1.1 points but trails SLIP on PAM by 1.1 points (Table 2).

These differences may partly reflect the pretraining domain and variable vocabulary. HARP learns exclusively from clinical records, and its manually curated vocabulary aligns physiological variables across clinical datasets (Appendix A.5.2). This vocabulary may provide less informative variable identities for wearable measurements outside its coverage. The baseline pretraining corpora also differ: NormWear is pretrained on wearable physiological signals, while SLIP uses a broader collection of sensor–text pairs spanning multiple domains (Luo et al., 2026; Chen et al., 2026). These differences may contribute to their competitive wearable performance. The results support the reuse of clinically pretrained representations for wearable classification, with smaller and task-dependent gains.