LEARNING TRANSFERABLE SENSOR MODELS VIA LANGUAGE-INFORMED PRETRAINING

Yuliang Chen [*], **Arvind Pillai, Yu Yvonne Wu, Tess Z. Griffin,**
Lisa Marsch, Michael V. Heinz, Nicholas C. Jacobson, Andrew Campbell
Dartmouth College

ABSTRACT

Sensing systems produce large scale unlabeled multivariate time series, therefore self supervised pretraining is a practical way to learn transferable representations. Yet many foundation models are trained for forecasting and can miss the semantic structure needed for classification and reasoning. Sensor language alignment improves semantic transfer, but existing methods often assume fixed sensor inputs, such as predefined channels, lengths, or temporal resolutions, which limits cross domain use. We introduce **SLIP** (**S**ensor **L**anguage **I**nformed **P**retraining), an open source framework that learns language aligned representations that generalize across diverse sensor configurations. SLIP combines contrastive alignment with sensor conditioned captioning, supporting both discriminative understanding and generative reasoning. By repurposing a pretrained decoder-only language model using cross attention and adding a flexible patch embedder, SLIP handles different temporal resolutions and variable length inference without additional retraining. Our experiments show that SLIP improves linear probing and zero-shot classification, as well as signal captioning and question answering.

Track: Research

1 INTRODUCTION

Ubiquitous sensors generate large multivariate time series, but labels are costly, making self supervised pretraining a practical way to learn transferable representations at scale. This has motivated many modality specific models across domains (Pillai et al., 2024; Saha et al., 2025; McKeen et al., 2025), yet these models often fail to transfer across sensor types, sampling rates, and task formats. In contrast, general purpose time series foundation models are typically trained for forecasting with reconstruction or regression objectives (Ansari et al., 2024; 2025; Liu et al., 2025; Luo et al., 2025), which can overlook the semantic structure needed for classification and reasoning (Figure 1). This gap has motivated recent work that brings language supervision into time series learning, including generalist approaches that map time series into pretrained language models (Kim et al., 2024; Jin et al., 2023; Xie et al., 2024). In parallel, sensor-text alignment and wearable focused studies (Ndir et al., 2025; Zhang et al., 2025; Luo et al., 2025) are less general across sensor types and temporal resolutions, and many do not directly target sensor based question answering.

We propose **SLIP**, a language-informed sensor encoder that aligns time series with text to support heterogeneous sensor configurations and temporal resolutions. SLIP repurposes a pretrained decoder only language model into a multimodal encoder-decoder by splitting it into a text encoder and a sensor conditioned decoder. SLIP introduces *FlexMLP* to adapt patch granularity across sequence lengths and sampling rates, enabling pretraining on 600K sensor caption pairs spanning over 1B time points, yielding $\mathbf{SLIP_{Base}}$. Across 11 datasets, $\text{SLIP}_{\text{Base}}$ achieves 76.57% average linear probing accuracy. We then apply supervised finetuning (SFT) to $\text{SLIP}_{\text{Base}}$ to equip it with sensor question answering (QA) capability, resulting in $\mathbf{SLIP_{SFT}}$. On four sensor QA benchmarks, SLIP_{SFT} achieves 64.83% average accuracy, and in captioning it attains a BERTScore of 0.887.

[*]Correspondence to: Yuliang Chen <yuliang.chen.gr@dartmouth.edu>.

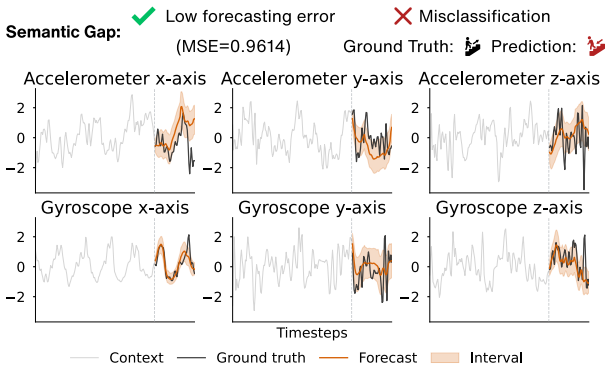


Figure 1: **Illustrated example of the forecasting–classification gap**. Chronos2 achieves accurate forecasting on UCI HAR with low error (MSE = 0.96), yet its learned representations lead to incorrect activity classification (walking downstairs vs. upstairs).

Study	Resolution Adaptive	Retrieval	QA	Open Source
Chronos (Ansari et al., 2024)	✗	✗	✗	✓
Chronos2 (Ansari et al., 2025)	✗	✗	✗	✓
Sundial (Liu et al., 2025)	✗	✗	✗	✓
SensorLM (Zhang et al., 2025)	✗	✓	✗	✗
Normwear (Luo et al., 2025)	✗	✓	✗	✓
ChatTS (Xie et al., 2024)	✗	✗	✓	✓
OpenTSLM (Langer et al., 2025)	✗	✗	✓	✓
SLIP (ours)	✓	✓	✓	✓

Table 1: **Capability comparison of sensor text modeling approaches.** We compare prior studies and SLIP across key capabilities, including temporal **resolution adaptive** sensing to handle different input sequence length and frequency, sensor text **retrieval** to capture semantic meaning, sensor **question answering (QA)**, and **open source** availability.

2 Related Work

Early sensor representation learning relied on large scale self supervised pretraining with masked reconstruction (Dong et al., 2023; Nie et al., 2023) or contrastive objectives (Zhang et al., 2022). General purpose time series foundation models are trained on diverse domains and perform well on forecasting, but their objectives do not always transfer to sensor classification (Ansari et al., 2024; 2025; Liu et al., 2025). Domain specific physiological foundation models improve performance within a modality, for example PPG focused work (Pillai et al., 2024; Saha et al., 2025), but remain less general across sensor types and tasks. Language based approaches either reprogram time series into pretrained language models (Jin et al., 2023; Hu et al., 2025) or learn sensor-specific text alignment with paired supervision (Zhang et al., 2025; Ndir et al., 2025). NormWear targets heterogeneous sensors but assumes fixed sampling rates and does not focus on question answering (Luo et al., 2025). Table 1 summarizes these gaps, and highlights that SLIP supports heterogeneous sensor domains and variable temporal resolutions while covering tasks from classification to question answering.

3 Methodology

SLIP is a conceptual extension of Contrastive-Captioner (CoCa) (Yu et al., 2022) for learning transferable language-aligned sensor representation for sensor-language applications that require both strong sensor understandings and contextual reasoning. SLIP is trained with paired $\langle X_s, X_t \rangle$ where X_s denotes the **sensor input** (a multivariate time series) and X_t is the **textual description** of X_s. As shown in Figure 2, SLIP comprises the following components:

Sensor Encoder $(X_s \mapsto Z_s)$ compresses high-volume sensor inputs to compact sensor embeddings Z_s – a sequence of continuous vectors analogous to tokens. Our sensor encoder is a 120M parameter Transformer (Vaswani et al., 2023; Nie et al., 2023). Following Sundial (Liu et al., 2025), we use Pre-LN and flash attention for stable and efficient training. Because temporal granularity (i.e. frequency, length) varies across sequences, we introduce a novel adaptive patch embedding called *FlexMLP* that supports variable patch sizes with no added parameters or compute. FlexMLP combines FlexiViT style weight resizing across patch sizes (Beyer et al., 2023) with the Chronos2 MLP patch-embedder that encodes time indices and an explicit missing value mask (Ansari et al., 2025). Given a target patch size, we split the raw signal into non overlapping patches and, for each patch, concatenate three aligned views: the signal values, a binary mask that marks missing samples, and the corresponding time indices. This concatenated patch vector is then mapped to a fixed dimensional token by an MLP whose input layer is shared across patch sizes by resizing its weight matrix to match the current patch dimensionality, while keeping the remaining projections unchanged. This weight resizing lets FlexMLP change granularity across sampling rates and window lengths without retraining (Algorithm 1). Larger patches reduce token count for long sequences, making full self attention over the entire multivariate series feasible. We concatenate patch tokens from all sensors into one 1D sequence and apply standard self attention, enabling global cross sensor and long range temporal interactions, while using 2D RoPE to preserve the underlying 2D structure (Su et al., 2023).

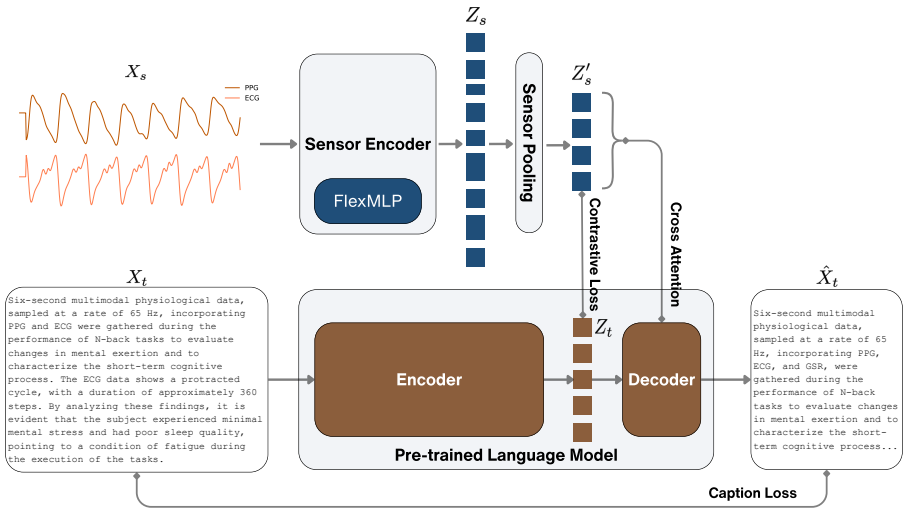


Figure 2: **Sensor-Language Informed Pretraining (SLIP) Architecture.**

Sensor Pooler. $(Z_s \mapsto Z'_s)$ is an attention pooling layer that compresses the variable length sensor token sequence into a fixed size representation Z'_s, filtering task irrelevant noise. Following CoCa, we use a single multi-head cross attention pooling layer with learnable query tokens: sensor tokens serve as keys and values, and the queries output a fixed number of pooled tokens set by the query count.

Text Encoder-Decoder. The text encoder $(X_t \mapsto Z_t)$ is a unimodal Transformer that produces text representations Z_t. The multimodal decoder $(\langle Z'_s, Z_t \rangle \mapsto \hat{X}_t)$ fuses Z_t with pooled sensor tokens Z'_s via cross attention to predict the target text $\hat{X}_t$, with both streams aligned before fusion. We initialize the text encoder from the first 12 layers of Gemma3-270M (Team et al., 2025) and the decoder from the final 6 layers, inserting a cross attention module into each of these six layers so text tokens can attend to sensor tokens during autoregressive decoding.

Training Objectives. We train SLIP by following the implementation of contrastive loss and caption loss in CoCa with equal weights.

Dataset. Pretraining SLIP requires large-scale paired time-series and text data, which is far less available than in vision–language settings. We start from community-released time series corpora (Liu et al., 2024; Luo et al., 2025) that provide diverse signals without aligned text, and we generate multi-level captions at statistical, structural, and semantic levels following the SensorLM recipe. To further increase pattern diversity, we augment this corpus with synthetic time series–text pairs from ChatTS (Xie et al., 2024). The resulting pretraining set contains over 600K samples and approximately one billion time points spanning energy, environment, health, IoT, nature, transportation, and web domains, with sampling rates ranging from seconds to months and varied sequence lengths. To reduce template repetition, we prompt Qwen2-7B-IT (Yang et al., 2024) to generate three paraphrases per caption and randomly sample one during training. As in ChatTS, we sample multivariate and univariate examples with a 2:1 ratio. Corpus statistics are summarized in Table 6.

4 Experiments

We evaluate three settings. For sensor-only classification, we use 11 datasets across four domains: activity recognition (WISDM, UCIHAR), clinical diagnosis (Stroke, Diabetes, Hypertension, Sleep Stage, and Heart Condition), stress prediction (WESAD, StudentLife), and urban sensing (Obstacles, BeijingAQI). Table 7 lists the sensor modalities and label taxonomies for all 11 datasets. On each dataset, we extract frozen representations from $\text{SLIP}_{\text{Base}}$ and train a linear classifier, which measures representation quality with minimal downstream capacity (Radford et al., 2021). We compare against self supervised baselines (SimMTM, TF-C) (Dong et al., 2023; Zhang et al., 2022), time series foundation models (NormWear, Sundial, Chronos, Chronos2), and sensor-language models (ChatTS).

For zero-shot sensor language understanding, we evaluate sensor text retrieval on the same 11 dataset suite using a CLIP style protocol (Radford et al., 2021) that tests whether sensor representations align with their textual descriptions without task specific supervision. Since most time series encoders are

Table 2: **Evaluation of the $\text{SLIP}_{\text{Base}}$ on 11 downstream sensor tasks across 4 domains compared against multiple baselines with linear probing (LP) and zero-shot (ZS).** We report mean and standard deviation of top-1 accuracy over 5-fold evaluation with different random seeds.

Eval	Model	WISDM	UCIHAR	Stroke	Diabetes	Hypertension	Sleep Stag.	Heart Cond.	WESAD	StudentLife	Obstacles	Beijing AQI
LP	Stat Feat	77.14 ± 0.25	85.62 ± 0.38	90.15 ± 0.00	82.22 ± 0.00	37.73 ± 1.30	77.33 ± 0.024	60.99 ± 0.33	68.52 ± 0.77	48.26 ± 1.37	81.69 ± 0.80	68.81 ± 0.51
	SimMTM	33.76 ± 0.08	70.52 ± 0.24	89.39 ± 0.00	82.22 ± 0.00	37.88 ± 1.07	41.76 ± 0.07	55.90 ± 0.08	62.51 ± 0.46	50.64 ± 0.69	36.37 ± 0.98	71.26 ± 0.26
	TFC	44.11 ± 0.11	59.95 ± 0.13	89.85 ± 0.61	80.74 ± 0.47	40.00 ± 0.57	44.88 ± 0.05	58.11 ± 0.25	55.43 ± 0.46	43.67 ± 1.70	33.66 ± 0.62	45.60 ± 0.46
	Normwear	70.82 ± 0.03	79.37 ± 0.32	90.45 ± 0.61	82.37 ± 0.30	40.45 ± 1.13	82.92 ± 0.18	69.53 ± 0.07	80.27 ± 0.49	49.17 ± 1.24	83.38 ± 0.32	74.40 ± 0.38
	Chronos2	75.18 ± 0.07	75.38 ± 0.51	87.12 ± 0.96	73.33 ± 3.18	38.79 ± 0.88	**83.74** ± 0.14	61.20 ± 0.19	67.26 ± 0.75	47.89 ± 0.69	84.14 ± 0.23	51.74 ± 0.63
	Chronos	81.19 ± 0.13	84.04 ± 0.30	88.48 ± 2.00	80.59 ± 1.36	45.30 ± 0.88	76.69 ± 0.14	**73.27** ± 0.18	66.46 ± 0.18	49.54 ± 1.00	75.45 ± 0.28	**80.48** ± 0.14
	Sundial Base	41.39 ± 0.12	55.84 ± 0.29	77.27 ± 0.83	72.00 ± 5.03	37.73 ± 3.16	70.19 ± 0.14	61.26 ± 0.12	59.64 ± 0.40	47.53 ± 2.20	68.04 ± 0.23	67.85 ± 0.26
	ChatTS	75.66 ± 0.00	56.24 ± 0.00	90.91 ± 0.00	82.22 ± 0.00	44.70 ± 0.00	80.65 ± 0.08	67.93 ± 0.21	77.04 ± 0.18	50.28 ± 0.90	77.75 ± 0.65	61.23 ± 1.84
	SLIP	**82.28** ± 0.10	**85.67** ± 0.26	**91.36** ± 0.67	**83.11** ± 0.55	**47.58** ± 0.57	82.62 ± 0.09	71.59 ± 0.19	**81.79** ± 0.61	**54.31** ± 1.58	**85.32** ± 0.35	76.66 ± 0.27
ZS	Normwear	3.91 ± 0.66	12.98 ± 2.15	89.40 ± 3.74	**82.22** ± 7.18	**36.38** ± 8.38	**39.09** ± 0.73	15.52 ± 1.26	16.58 ± 6.43	**43.12** ± 13.02	24.30 ± 2.03	2.40 ± 2.34
	SLIP	**7.45** ± 0.14	**27.26** ± 0.46	**90.91** ± 0.00	75.56 ± 2.65	35.76 ± 0.30	34.58 ± 0.17	**16.08** ± 0.19	**40.45** ± 0.87	42.02 ± 0.69	**30.08** ± 0.13	**22.25** ± 1.67

Table 3: **Evaluation of SLIP_{SFT} on Sensor QA benchmarks.** The TSQA dataset provides answers without reasoning traces, whereas the other datasets include answers with explicit reasoning. We report top-1 accuracy for all dataset.

Model	TSQA	HAR-CoT	Sleep-CoT	ECG-QA-CoT
OpenTSLM SP	11.96	0.55	5.91	1.11
OpenTSLM Flamingo	25.46	63.43	68.49	35.50
SLIP_{SFT}	**<u>83.60</u>**	**<u>64.35</u>**	**<u>74.19</u>**	**<u>37.18</u>**

Table 4: **Evaluation of SLIP_{SFT} against OpenTSLM using the same Gemma3-270M backbone on the M4 captioning dataset.** Higher is better for all metrics.

Model	BLEU4	METEOR	ROUGE-L	SBERTSim.	BERTScore
OpenTSLM SP	0.0026	0.0456	0.0255	0.1551	0.7250
OpenTSLM Flamingo	**<u>0.1141</u>**	0.3210	**<u>0.2894</u>**	0.7990	0.8858
$\text{SLIP}_{\text{Base}}$	0.0116	0.2440	0.1409	0.6276	0.8338
SLIP_{SFT}	0.1130	**<u>0.3814</u>**	0.2569	**<u>0.8691</u>**	**<u>0.8870</u>**

trained only on numerical signals and never see text, we focus the direct comparison on $\text{SLIP}_{\text{Base}}$ versus NormWear, which also uses language supervision.

For instruction following tasks, we evaluate question answering and captioning using established splits from OpenTSLM. For QA, we use HAR-CoT, Sleep-CoT, and ECG-QA-CoT with the original train validation test splits, and we also reformat TSQA into a multiple choice QA protocol (Zellers et al., 2018) to test basic sensor understanding. We benchmark SLIP_{SFT}, obtained by supervised finetuning $\text{SLIP}_{\text{Base}}$ with the captioning loss only, against OpenTSLM (soft prompting and Flamingo variants with a Gemma3 270M backbone) (Langer et al., 2025). To isolate the benefit of pretraining rather than extensive task specific training, we finetune each QA dataset independently for four epochs, using ten epochs for Sleep-CoT. For captioning, we use the M4 dataset from OpenTSLM, finetune on the M4 training split for four epochs, and evaluate on the M4 test split, reporting BLEU4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), SBERTSimilarity (Reimers & Gurevych, 2019), and BERTScore (Zhang et al., 2020).

5 RESULTS

$\text{SLIP}_{\text{Base}}$ achieves the best average linear probe and zero-shot accuracy across 11 classification datasets. Table 2 shows that $\text{SLIP}_{\text{Base}}$ beats the strongest baseline, NormWear (76.57 vs 73.01). $\text{SLIP}_{\text{Base}}$ also leads in zero-shot retrieval with accuracy of 38.40% vs 33.26% for NormWear. It is strongest on most of the tasks, but weaker than NormWear on clinical diagnosis (Stroke, Diabetes, Hypertension, and Sleep Stage). This is likely because NormWear is pretrained on matched high frequency clinical signals and aligned with ClinicalTinyLlama, which have broader clinical coverage and stronger domain specific priors for diagnosis oriented representations.

$\text{SLIP}_{\text{Base}}$ is a strong starting point for sensor QA with minimal finetuning. Table 3 shows SLIP_{SFT} outperforms OpenTSLM with soft prompting and Flamingo style training under the same language backbone, Gemma3-270M, despite OpenTSLM using curriculum learning for over 40 epochs per task while we finetune $\text{SLIP}_{\text{Base}}$ for fewer than 10 epochs.

$\text{SLIP}_{\text{Base}}$ generates semantically aligned captions without M4 training and improves with brief finetuning. In Table 4, $\text{SLIP}_{\text{Base}}$ already shows strong SBERTSimilarity and BERTScore. Finetuning on M4 yields SLIP_{SFT}, which better matches reference phrasing and improves semantic scores, while reaching n-gram scores close to OpenTSLM Flamingo, suggesting remaining gaps are often paraphrases rather than semantic errors.

Table 5: **Ablation studies results**. The default setting adopted by SLIP is marked in blue. We calculate ±delta within each group of ablations in comparison with the default setting.

	Sensor Perception (Accuracy)		Retrieval (Accuracy)		QA (Accuracy)	
(a) Training Objectives						
SLIP pretraining	77.14		39.36		64.83	
Caption-only	74.30	(-2.84)	25.12	(-14.24)	57.26	(-7.57)
Contrastive-only	74.77	(-2.37)	36.90	(-2.46)	48.03	(-9.98)
Random paired	62.15	(-14.99)	22.38	(-16.98)	35.08	(-29.75)
(b) Sensor Encoder Parameter Size						
$\text{SLIP}_{\text{Small}}$ (40M)	74.84	(-2.30)	26.71	(-12.65)	53.99	(-10.84)

	Sensor Perception (Accuracy)		Retrieval (Recall@1)		QA (Accuracy)	
(c) Cross-Sensor Learning						
Self-Attention w/ 2D RoPE	77.14		39.36		64.83	
Group Attention	77.04	(-0.10)	39.21	(-0.15)	58.01	(-6.82)
(d) FlexMLP						
w/o FlexMLP (*patch size* = 16)	74.16	(-2.98)	34.79	(-4.42)	61.38	(-3.45)
(e) Partial finetuning of text-encoder						
Fine-tune 4-layer	77.14		39.36		64.83	
Freeze	73.56	(-3.58)	35.68	(-3.68)	57.09	(-7.74)

5.1 Ablation Studies

We study key design choices in SLIP and report average top-1 classification accuracy, zero-shot classification accuracy (both over 11 datasets), and QA accuracy (over 4 datasets).

Training Objectives. We compare SLIP against single-objective variants in Table 5.1. Relative to a contrastive-only model, SLIP improves supervised classification (+2.37), zero-shot retrieval (+2.46), and QA (+9.98), indicating that the captioning objective provides complementary semantic supervision. SLIP also substantially outperforms a caption-only model in retrieval (+14.24) and QA (+7.57), suggesting that caption-only sensor embeddings remain weakly grounded in the input signals. We include SLIP trained with misaligned sensor-text pairs as a sanity check against the aligned setup.

Model Size. A smaller sensor encoder (40M) leads to a substantial drop in retrieval (–12.65) and QA (–10.84), while sensor perception is less affected (–2.3). This suggests that the smaller encoder preserves task-relevant features for linear separability but fails to learn an embedding geometry that supports cross-modal alignment. We attribute this to modality imbalance: when the sensor branch becomes the bottleneck, optimization primarily adjusts the text branch and cross-modal projections, providing weaker gradients to shape sensor embeddings into a discriminative, well-aligned space.

Cross-sensor Learning. Group attention shows negligible impact on sensor perception (–0.10) and zero-shot classification (–0.15), but causes a sharp drop on the univariate TSQA dataset (–21.94), reducing average SFT performance by 6.82. We attribute this to multivariate-dominated pretraining encouraging cross-channel shortcuts under group attention. When evaluated on univariate TSQA, this signal is absent, weakening long-range evidence aggregation within a single channel. Brief supervised finetuning does not correct this behavior. Given comparable classification performance, group attention remains a practical option when memory constraints make full attention infeasible.

FlexMLP. Using a fixed patch size (*e.g.*, 16) during pretraining and evaluation degrades performance, particularly for zero-shot classification (–4.42), where no task-specific head aggregates information across patches. Because patch size determines temporal resolution, datasets with different sampling frequencies (*e.g.*, hourly vs. second-level) favor different patch granularities.

Freezing Text Encoder. We evaluate a parameter-efficient variant of SLIP that freezes the text encoder and trains only the sensor encoder, projector, and multimodal decoder. Freezing the text encoder prevents mutual adaptation between modalities during contrastive learning, forcing the sensor encoder to match fixed text targets that may not reflect the underlying signals. As a result, sensor representations collapse toward a limited region of the feature space, leading to consistent degradation across downstream tasks.

6 Discussion

Overall, SLIP efficiently repurposes a decoder only language model into a multimodal encoder-decoder model. $\text{SLIP}_{\text{Base}}$ achieves state of the art performance from a single checkpoint across diverse sensor application domains. We also show that SLIP_{SFT} can adapt to complex sensor question answering tasks with little finetuning. Our work provides a unified sensor encoder pretrained with sensor text alignment, suggesting it can serve as a common starting point for sensor language models by adapting quickly to the language space with finetuning for downstream sensor language tasks.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

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, 2023. URL https://arxiv.org/abs/2212.08013.

Shing Chan, Hang Yuan, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition, 2024. URL https://arxiv.org/abs/2402.19229.

Song Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5RK5G.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7: 1–30, December 2006. ISSN 1532-4435.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling, 2023. URL https://arxiv.org/abs/2302.00861.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. Context-alignment: Activating and enhancing llm capabilities in time series. *arXiv preprint arXiv:2501.03747*, 2025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data, 2024. URL https://arxiv.org/abs/2401.06866.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. Opentslm: Time-series language models for reasoning over multivariate medical text- and time-series data, 2025. URL https://arxiv.org/abs/2510.02410.

Yu Liang, Zhi Chen, Guang Liu, Mohamed Elgendi, Zhihua Chen, Robin Ward, and Zhizheng Wang. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific Data*, 5:180020, 2018. doi: 10.1038/sdata.2018.20. URL https://doi.org/10.1038/sdata.2018.20.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models, 2024. URL https://arxiv.org/abs/2402.02368.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models, 2025. URL https://arxiv.org/abs/2502.00816.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals, 2025. URL https://arxiv.org/abs/2412.09758.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

Tidiane Camaret Ndir, Robin Tibor Schirrmeister, and Tonio Ball. Eeg-clip : Learning eeg representations from natural language descriptions, 2025. URL https://arxiv.org/abs/2503.16531.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

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

Mithun Saha, Maxwell A Xu, Wanting Mao, Sameer Neupane, James M Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025.

Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*, pp. 400–408, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3242969.3242985.

Vinicius M. A. Souza. Asphalt pavement classification using smartphone accelerometer and complexity invariant distance. *Engineering Applications of Artificial Intelligence*, 74:198–211, 2018. ISSN 0952-1976. doi: 10.1016/j.engappai.2018.06.003. URL https://www.sciencedirect.com/science/article/pii/S0952197618301349.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, November 2022. doi: 10.13026/kfzx-aw45. URL https://doi.org/10.13026/kfzx-aw45. Version 1.0.3.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben Zeev, and Andrew T. Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, 2014. Association for Computing Machinery. doi: 10.1145/2632048.2632054.

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

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

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. *arXiv preprint arXiv:2412.03104*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference, 2018. URL https://arxiv.org/abs/1808.05326.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency, 2022. URL https://arxiv.org/abs/2206.08496.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. Sensorlm: Learning the language of wearable sensors, 2025. URL https://arxiv.org/abs/2506.09108.

A DATASETS

We thank the prior work that collected and open-sourced dataset that used in this work, and we summarize it again below for the convenience of future researchers.

A.1 PRETRAINING DATASET

Inspired by SensorLM (Zhang et al., 2025), we automatically generate hierarchal caption (i.e. statistical, structural, semantic) of each multivariate sensor signal. The domain distribution of this sensor-paired dataset is shown in Table A.1. We open source this large dataset to support future research on sensor language models, and we also release the caption generation pipeline for creating large scale sensor language datasets.

Table 6: Category distribution.

	Health	Synthetic	Web	Nature	Energy	IoT	Environment	Transport
# of Samples	237050	105085	67865	32358	2743	2611	1082	28
Percent (%)	52.82	23.41	15.12	7.21	0.61	0.58	0.24	0.01
Sources	(Liu et al., 2024) (Luo et al., 2025) (Chan et al., 2024)	(Xie et al., 2024)	(Liu et al., 2024)	(Liu et al., 2024)	(Liu et al., 2024)	(Liu et al., 2024)	(Liu et al., 2024)	(Liu et al., 2024)

A.2 DOWNSTREAM DATASET

Table 7: Evaluation dataset details.

Dataset	Sensor	# Samples (Train \Test)	Freq.	# Cls	Label Names
WISDM (Weiss, 2019)	Accelerometer X, Y, Z	22396 / 5600	30Hz	18	Catch, Chips, Clap, Dribble, Drink, Fold, Jog, Kick, Pasta, Sandwich, Sit, Soup, Stair, Stand, Teeth, Type, Walk, Write
UCI-HAR (Reyes-Ortiz & Parra, 2013)	Accelerometer X, Y, Z Gyroscope X, Y, Z	1847 / 793	50Hz	5	Lay, Sit, Stand, Walk, Walking Upstairs, Walking Downstairs, Transition
PPG-CVA (Stroke) (Liang et al., 2018)	PPG	525 / 132	65Hz	2	Normal, Stroke
PPG-DM (Diabetes) (Liang et al., 2018)	PPG	522 / 135	65Hz	2	Normal, Diabetes
PPG-HTN (Hypertension) (Liang et al., 2018)	PPG	525 / 132	65Hz	4	Normal, Pre-hypertension, Stage-1, Stage-2
SleepEDF (Sleep stage) (Kemp et al., 2000)	EEG-Fpz-Cz, EEG-Pz-Oz	33599 / 8709	100Hz	5	Light spindle, Light, Deep, REM, Wake
PTB-XL (Heart cond.) (Wagner et al., 2022)	12-lead ECG	11320 / 1650	100Hz	5	Normal ECG, Myocardial Infarction, ST/T Change, Conduction Disturbance, Hypertrophy

Dataset	Sensor	# Samples (Train \Test)	Freq.	# Cls	Label Names
WESAD (Schmidt et al., 2018)	Chest Acc. X, Y, Z Chest ECG, EMG, EDA, Temp, Resp, Wrist Acc. X, Y, Z, Wrist BVP, EDA	882 / 223	700Hz	3	Neutral, Amusement, Stress
StudentLife (Wang et al., 2014)	Activity, Audio, Conversation, Phone Charge, Phone Lock, Time to deadline, Day of the week, Exam period, Sleep rating, Sleep duration	1074 / 109	Minute	3	Normal, Medium Stress, High Stress
AsphaltObstacles (Souza, 2018)	Acceleration magnitude	390 / 391	100Hz	4	Raised crosswalk, Raised markers, Speed bump, Vertical patch
Beijing AQI (Chen, 2017)	dew-point temperature, windspeed, PM25, PM10, NO2, SO2, CO	1168 / 293	Hour	4	Good, Moderate, Unhealthy for sensitive groups, Unhealthy, Hazardous

B Implementation Details

B.1 FlexMLP

We provide pseudo code for flexify a MLP patch-embedder in Algorithm 1

B.2 Patch Size Heuristic

Since we are performing joint training on datasets with varying sensor resolutions and sequence lengths, we define a frequency-based patch size heuristic following (Woo et al., 2024). We incorporate a broader range of window sizes to ensure that the number of tokens per sample remains roughly consistent across diverse datasets, thereby minimizing the computational overhead caused by excessive padding. The patch sizes are assigned based on data frequency as follows:

- **Daily:** 16, 32
- **Hourly:** 6, 8, 16, 24, 32, 64
- **Minute-level:** 16, 24, 128
- **Second-level:** 4, 6, 8, 12, 16, 20, 25, 32, 64, 128

B.3 Pretraining Implementation details

We pretrain for 40 epochs with batch size 72 on 4 NVIDIA H200 GPUs using AdamW (Loshchilov & Hutter, 2019) (weight decay = 0.05). The learning rate is warmed up to $2e^{-4}$ over 80k iterations and cosine-decayed to $1e^{-7}$. Sensor augmentations are disabled during pretraining to preserve sensor–text alignment, but standard augmentations (jittering, scaling, time flipping) are applied during supervised finetuning (Zhang et al., 2025). Additional implementation details are in the codebase.

Algorithm 1 Minimal FlexMLP pseudo-implementation.

```
class FlexMlp(nn.Module):
  def __call__(self, x, mask, time_index, base_patch=16):
    '''
    x, mask, time_index: (B, L)
    hidden_dim = 768
    mlp_dim = 3072
    '''

    x_p = patchify(x, patch_size)
    m_p = patchify(mask, patch_size)
    t_p = patchify(t, patch_size) # (B, num_patches, patch_size)
    z_p = concat([x_p, m_p, t_p], axis=-1) # (B, num_patches, patch_size*3)

    # Shared MLP weights trained at base_patch:
    w = self.param("w_mlp", (mlp_dim, base_patch*3))
    b = self.param("b_mlp", (mlp_dim,))
    w_res = self.param("w_res", (hidden_dim, base_patch*3))
    b_res = self.param("b_res", (hidden_dim,))

    # Flex trick: resize weights to match current patch size:
    w_p = resize(w, (mlp_dim, patch_size*3))
    w_res_p = resize(w_res, (hidden, patch_size*3))
    h = linear(z_p, w_p, b)
    r = linear(z_p, w_res_p, b_res)

    # Fixed output projection:
    w_out = self.param("w_out", (hidden_dim, mlp_dim))
    b_out = self.param("b_out", (hidden_dim,))
    h = linear(h, w_out, b_out)

    return MLP(h) + r
```

Notes: Changes to existing code highlighted via violet background.

B.4 Downstream Evaluation Details

Following the standard transfer learning evaluation protocol from prior work (He et al., 2021), we did not perform hyperparameter search (i.e. learning rate or weight decay). All models were frozen, and only a randomly initialized linear classifier was trained for 50 epochs with 5 warmup epochs using the AdamW optimizer, a base learning rate of 0.01, and a weight decay of 0.05. Supervised finetuning hyper-parameter details is shown in Table 8.

Table 8: **Hyper-parameters used in the supervised fine-tuning experiment.**

Hyper parameter	TSQA	HAR-CoT	Sleep-CoT	ECG-QA-CoT	M4 Caption
Optimizer	Adamw (Loshchilov & Hutter, 2019)				
Gradient clip	1.0				
LR decay schedule	Cosine Schedule Decaying to $1e^{-7}$				
Train Epoch	4	4	10	4	4
Train batch size	64	64	32	32	64
Warm up epochs	1	1	1	1	1
Weight decay rate	0.05	0.05	0.05	0.05	0.05

Table 9: **Pairwise Wilcoxon results.** SLIP compared with each baseline.

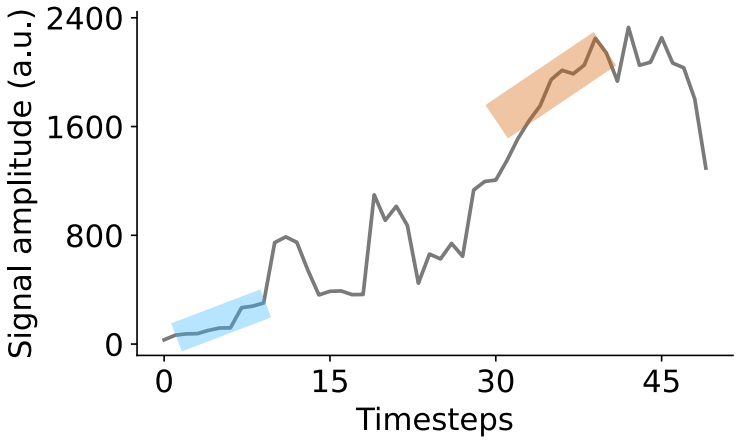
Eval	modelA	modelB	p_value	significant
LP	$SLIP_{Base}$	Stat Feat	0.000977	True
	$SLIP_{Base}$	SimMTM	0.000977	True
	$SLIP_{Base}$	TFC	0.000977	True
	$SLIP_{Base}$	Sundial Base	0.000977	True
	$SLIP_{Base}$	Normwear	0.001953	True
	$SLIP_{Base}$	Chronos2	0.001953	True
	$SLIP_{Base}$	Chronos	0.041992	True
	$SLIP_{Base}$	ChatTS	0.000977	True
ZS	$SLIP_{Base}$	Normwear	0.24	False

C Statistical Comparisons

Table 9 reports pairwise Wilcoxon signed rank tests (Demšar, 2006; Wilcoxon, 1945) comparing SLIP with each baseline across the 11 datasets, using paired per dataset results computed from Table 2. In the linear probing setting, SLIP differs from every baseline at the 0.05 level. In the zero-shot retrieval setting, SLIP versus Normwear is not significant, which is consistent with Normwear showing stronger alignment on the diagnosis tasks. We do not report statistical tests for the supervised fine tuning experiment because there are fewer than five datasets, following the guidance in (Demšar, 2006).

D ADDITION EXAMPLE OF QUESTION ANSWERING

Prompt:
You are an expert in time series analysis. Please generate a detailed caption for this time-series, describing it as accurately as possible



Generated Caption:
The time-series graph illustrates data points over a period of 50 units, showcasing a general upward trend with fluctuations. Initially, the series starts at a low point near 0 and gradually increases, reaching around 1000 by the 10th unit. A noticeable peak occurs around the 15th unit, followed by a decline, creating a trough near the 20th unit. From there, the series experiences a significant upward surge, peaking at approximately 2500 around the 35th unit. This is followed by minor fluctuations, maintaining a high level until the 45th unit...

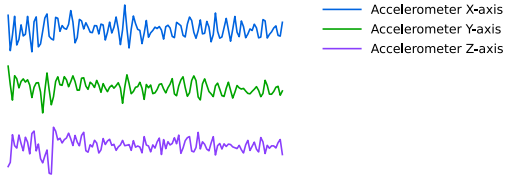
Figure 3: **A qualitative example of sensor captioning from the M4 dataset.** Highlighted segments denote salient local trends referenced in the generated caption.

Prompt:

```
Instructions:
 - Begin by analyzing the time series without assuming a
specific label.
 - Think step-by-step about what the observed patterns
suggest regarding movement intensity and behavior.
 - Write your rationale as a single, natural paragraph — do
not use bullet points, numbered steps, or section headings.
 - Do **not** mention any class label until the final
sentence.

 Possible activity labels are:
 biking, lying, running, sitting, standing, walking,
walking_down, walking_up.

 - Make sure that your last word is the answer. You MUST end
your response with "Answer: "
```



Prediction:

Rationale: The accelerometer data over the 2.56 second window shows significant variability across all three axes, indicating substantial movement. The **X-axis exhibits frequent peaks and troughs, suggesting lateral or forward-backward motion** that is consistent with dynamic activity. The **Y-axis also displays fluctuations**, though they are less pronounced, which might indicate less vertical movement or tilt. **The Z-axis shows moderate variability**, which could be associated with subtle changes in height or orientation. The overall pattern of the data **suggests a continuous and rhythmic motion, likely involving repetitive cycles** of movement. The presence of such variability and intensity across all axes is more characteristic of an activity involving consistent motion and balance adjustments rather than a static or minimal movement scenario. Therefore, the observed patterns are more aligned with the activity of biking. **Answer: biking.**

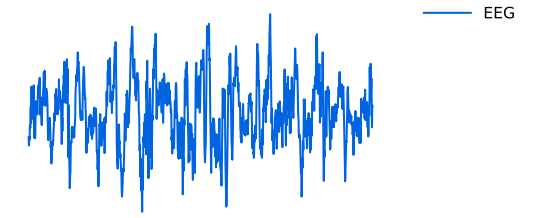
Ground Truth:

The accelerometer data over the 2.56 second window shows significant variability across all three axes, indicating substantial movement. The X-axis and Y-axis **exhibit frequent fluctuations with both positive and negative accelerations**, suggesting dynamic motion rather than a static position. The **Z-axis also shows variability, although with slightly less amplitude**, which could indicate vertical movement or changes in elevation. The overall pattern suggests a **repetitive, rhythmic motion consistent with a cycling activity,** where the body experiences continuous motion due to pedaling and shifts in balance. These characteristics are less likely to be observed in a stationary activity, which would typically show minimal changes in acceleration. Therefore, the observed patterns are more indicative of an active movement rather than stillness. **Answer: biking.**

Figure 4: A qualitative example of sensor question answering from the Har-CoT dataset.

Prompt:

Instructions:
- Analyze the data objectively without presuming a particular label.
- Reason carefully and methodically about what the signal patterns suggest regarding sleep stage.
- Write your reasoning as a single, coherent paragraph. Do not use bullet points, lists, or section headers.
- Only reveal the correct class at the very end.
- Never state that you are uncertain or unable to classify the data. You must always provide a rationale and a final answer.Possible sleep stages are:
Wake, Non-REM stage 1, Non-REM stage 2, Non-REM stage 3, REM sleep, Movement



Prediction:

Rationale: The time series data presented shows a pattern characteristic of certain sleep stages, distinguished by specific EEG features. The signal displays a relatively consistent amplitude and frequency, with **occasional bursts of higher amplitude activity** that can be observed throughout the interval. These bursts are indicative of the presence of **sleep spindles**, which are a hallmark of a particular sleep stage. The overall pattern lacks the high-frequency, low-amplitude activity typically associated with wakefulness, and instead demonstrates the more rhythmic and stable activity seen during sleep. The presence of these features strongly suggests that the EEG data corresponds to a sleep stage characterized by reduced consciousness and decreased sensory perception. **Answer: Non-REM stage 2**

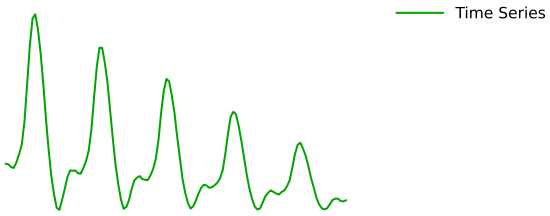
Ground Truth:

The time series data presented shows a pattern characteristic of certain sleep stages, distinguished by specific EEG features. The signal displays a rhythmic pattern with **periodic bursts of activity**, which may suggest the presence of **sleep spindles** typically associated with certain sleep stages. The amplitude and frequency of these bursts are consistent with known EEG patterns observed during sleep. Additionally, the overall variability and complexity of the signal, with its moderate amplitude fluctuations and rhythmic bursts, align with typical features observed during a particular stage of sleep. These characteristics help differentiate between the stages, and the presence of these specific patterns in the EEG data strongly indicates the sleep stage being observed. **Answer: Non-REM stage 2**

Figure 5: A qualitative example of sensor question answering from the Sleep-CoT dataset.

Prompt:

```
Instructions:
Given the following definitions: Constant volatility: The
time series shows relatively consistent fluctuation
magnitude throughout the period. Increased volatility: The
time series shows a rise in the magnitude of fluctuation
over time. Decreased volatility: The time series shows a
reduction in the magnitude of fluctuations over time. Select
one of the following answers that best describes the
provided time series:
(a) This time series has a constant volatility.
(b) This time series has an increased volatility.
(c) This time series has a decreased volatility.
Only answer (a), (b), or (c). Predict the Volatility Answer:
```



Prediction: C

Ground Truth: C

Figure 6: A qualitative example of Multiple Choice Question from the TSQA dataset.

Prompt:

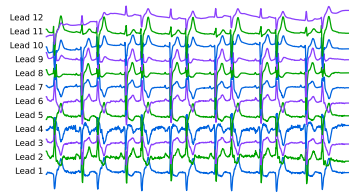
Clinical Context: 75-year-old male patient. 12-lead ECG. clinical recording. recorded with AT-6 C 5.5. Signal quality: baseline drift noted, static noise present, burst noise present, electrode artifacts present. extra beats detected during recording. pacemaker present.

Your task is to examine the ECG signal and answer the following medical question:

Question: Which diagnostic symptom does this ECG show, myocardial infarction in anterior leads or myocardial infarction in anteroseptal leads, excluding uncertain symptoms?

nstructions:
- Begin by analyzing the time series without assuming a specific answer.
- Think step-by-step about what the observed patterns suggest regarding the cardiac condition.
- Write your rationale as a single, natural paragraph — do not use bullet points, numbered steps, or section headings.
- Do **not** mention any final answer until the very end.
- Consider the ECG morphology, intervals, and any abnormalities that relate to the question.

Based on your analysis of the ECG data, select your answer from the following options:
complete left bundle branch block,
complete right bundle branch block,
digitalis effect, first degree av block,
...
...
...
- Make sure that your last word is the answer. You MUST end your response with "Answer: "



Prediction:

Side analysis for the 12-lead ECG of this 75-year-old male patient. The presence of a pacemaker and various artifacts such as baseline drift, static noise, burst noise, and electrode artifacts complicate the interpretation. However, focusing on the cardiac patterns, the leads V1 through V4, which represent the anteroseptal region, should be closely examined for signs of **myocardial infarction**. In these leads, significant ST-segment elevation or pathological Q waves would indicate myocardial infarction. The presence of extra beats may suggest arrhythmias, but the primary concern is the identification of ischemic changes. The absence of significant changes in the inferior leads (II, III, aVF) and **the lack of reciprocal changes in the anterior leads (V1-V3) further support the localization of the infarction to the anteroseptal region**. Given these observations, the ECG findings are consistent with myocardial infarction in the anteroseptal leads. Answer: myocardial infarction in anteroseptal leads. The ECG shows significant changes in the anteroseptal leads, specifically V1 to V4, which are indicative of myocardial infarction in this region. **Answer: myocardial infarction in anteroseptal leads**.

Ground Truth:

The ECG shows evidence of myocardial infarction, particularly focusing on the anterior leads. In this recording, the presence of ST-segment elevation in leads V1 through V4 is indicative of an anterior **myocardial infarction**. These leads correspond to the anterior wall of the heart, which is typically affected in such cases. **The baseline drift and noise present in the recording do not obscure the clear elevation of the ST segments in these leads**. The presence of a pacemaker and extra beats does not alter the interpretation of the ST-segment changes, which are crucial in diagnosing myocardial infarction. The absence of significant changes in the lateral leads (I, aVL, V5, V6) further supports the localization of the infarction to the anterior region rather than a lateral or subendocardial injury. Therefore, based on the observed patterns, the ECG is consistent with myocardial infarction in the anterior leads. **Answer: myocardial infarction in anterior leads**.

Figure 7: A qualitative failure example of sensor question answering from the ECG-QA-CoT dataset.