JETS: A Self-Supervised Joint Embedding Time Series Foundation Model for Behavioral Data in Healthcare

Erik Xie
Department of EECS
Massachusetts Institute of Technology
ejxie@mit.edu

Raquel Rodriguez Martinez, MD
Empirical Health
raquel.rodriguez@post.harvard.edu

Wyatt Chang
Empirical Health
wyatt.chang@empirical.health

Brandon Ballinger
Empirical Health
bmb@empirical.health

Abstract

Behavioral time series from wearable devices offer rich health insights but are often characterized by noise, missing values, and irregular sampling. While prior research on learning from physiological time series has focused on dense, regular, and low-level sensor data, self-supervised pre-training on high-level, behavioral data remains a key challenge. We propose JETS (Joint Embedding for Time Series), a masked pre-training framework designed to address these challenges in behavioral time series. JETS was pre-trained on a long-term dataset collected from real-world wearables, demonstrating robustness to noisy and incomplete measurements by distilling the data into a learned latent space. When fine-tuned and evaluated on downstream, individual-level diagnostic prediction tasks, JETS outperforms established baselines, validating the effectiveness of joint-embedding architectures for ubiquitous behavioral data and paving the way for new applications in personalized digital health.

1 Introduction

1.1 Background

The popularization of wearable devices has led to an abundance of long-term behavioral time series data (e.g., heart rate, sleep, activity), offering rich insights into individual health trajectories [12]. Modeling their temporal dynamics has crucial implications for applications such as early disease detection [16] and biological marker prediction [13], enabling health monitoring and possible prevention outside of traditional clinical settings. However, these data typically exist as Irregular Multivariate Time Series (IMTS) [11], characterized by high dimensionality, sparsity, and irregular sampling, due to real-world factors like intermittent device usage, sensor failures, variable recording frequencies, and different participation timelines [18, 4]. These properties challenge traditional time series models that often require dense, regularly sampled, fixed-length inputs. Furthermore, the scarcity and high cost of clinical labels [8] render fully supervised learning infeasible on a large scale, requiring semi- or self-supervised approaches to be adapted.

To address these challenges, we propose JETS (Joint Embedding for Time Series), a self-supervised learning framework to learn robust representations from physiological IMTS. Inspired by the Joint Embedding Predictive Architecture (JEPA) [2], JETS learns to predict the latent representations of time series segments from the visible context. We show that the resulting embeddings are versatile and can be effectively fine-tuned for downstream tasks such as disease prediction, highlighting the potential new applications of behavioral time series data in health.

1.2 Related Work

Self-Supervised Learning for Time Series: Self-supervised learning (SSL) is a dominant paradigm for learning from unlabeled time series [20]. Approaches include contrastive methods like TS2Vec [19], predictive methods that forecast future values [17, 22], and generative methods based on masked signal modeling [21]. While many generative models reconstruct raw signals, more recent joint-embedding predictive architectures (JEPAs) [2, 14], such as TS-JEPA [6], perform prediction in a learned representation space to capture more abstract, semantic features, but only for continuous, uni-variate data. JETS builds upon this JEPA framework, adapting it for the unique challenges of behavioral IMTS.

Wearable Foundation Models: Recent work has trained foundation models directly on short time-windows of low-level physiological signals from wearables. LSM-2 was trained on 40 M hours of multimodal wearable data (heart rate, accelerometry, electrodermal activity, temperature, altitude) with reconstructive and contrastive losses; [18]]; the Apple Heart and Movement Study trained a contrastive model on 20 million PPG and 3.75 million ECG recordings [[1]]; and DeepHeart used 57,675 person-weeks of heart-rate data with reconstructive losses [[3]].

Foundation Models for Behavioral Timeseries: Many clinically relevant physiological patterns (e.g., circadian rhythms) only emerge over longer timespans. The Wearable Behavior Model (WBM) [7] demonstrated that incorporating higher-level behavioral metrics, such as VO2Max and resting heart rate, improved accuracy on downstream diagnostic tasks. Like WBM, JETS builds on behavioral timeseries, but is designed for more constrained, lower-resolution, and mixed-source data (e.g., self-reports and sensor readings).

2 Joint Embedding for Behavioral Time Series

To our knowledge, our Joint Embedding for Time Series (JETS) framework represents the first application of a joint embedding architecture to long-horizon, irregularly-sampled multivariate time series (IMTS) from behavioral data.

2.1 Training Data

The study utilizes a longitudinal dataset comprising wearable device data collected from a cohort of 16,522 individuals, with a total of ~3 million person-days. For each individual, 63 distinct time series metrics were recorded at a daily or lower resolution. These metrics are categorized into five physiological and behavioral domains: cardiovascular health, respiratory health, sleep, physical activity, and general statistics. A more comprehensive statistical summary of the dataset is provided in the Appendix.

2.2 Model Architecture

The JETS framework consists of four primary components: a learnable tokenizer, a patch-based masking strategy, a dual-encoder system, and a predictor network. Each input instance is presented as a set of L observations $\{(t_i, v_i, m_i)\}$, corresponding to day, value, and metric type.

1. Tokenization: To handle irregular sampling, time difference is used instead of absolute time for the Mamba architecture (i.e. $\Delta t_i = t_i - t_{i-1}$) due to its state-space nature. Each of the three dimension of the triplets is passed through an embedding layer that maps it to the hidden dimension D. The resulting embeddings are combined to form a sequence of tokens $\mathbf{T} \in \mathbb{R}^{L,D}$.

2. Masking: Tokenized sequence $\mathbf{T}$ is divided into a fixed number of patches, and a high percent (i.e. 70%) of the patches are randomly removed from $\mathbf{T}$ to form $\mathbf{T}_{\text{ctx}}$. This approach has shown to be effective in the MAE framework [11].

3. Encoders: JETS uses a bidirectional Mamba context encoder (E_θ) that encodes unmasked tokens $\mathbf{T}_{\text{ctx}}$, and a target encoder (E_ϕ), with an identical structure, that encodes the full sequence $\mathbf{T}$. The target encoder's weights are trained as an exponential moving average of the context encoder's, as in the JEPA [2] and BYOL [9] frameworks, i.e. $\phi \leftarrow \tau\phi + (1-\tau)\theta$. The asymmetry has been shown to effectively prevent representation collapse.

4. Predictor: A small neural network P that takes the output from the context encoder $E_\theta(\mathbf{T}_{\text{ctx}})$, along with the positional embeddings (time and variable) of the masked patches, predicts the target representations for the masked positions. We use transformer decoder layers in JETS.

JETS is then trained with the MSE objective: $\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} ||P(E_\theta(\mathbf{T}_{\text{ctx}}), \text{pos}_j) - E_\phi(\mathbf{T})||_2^2$

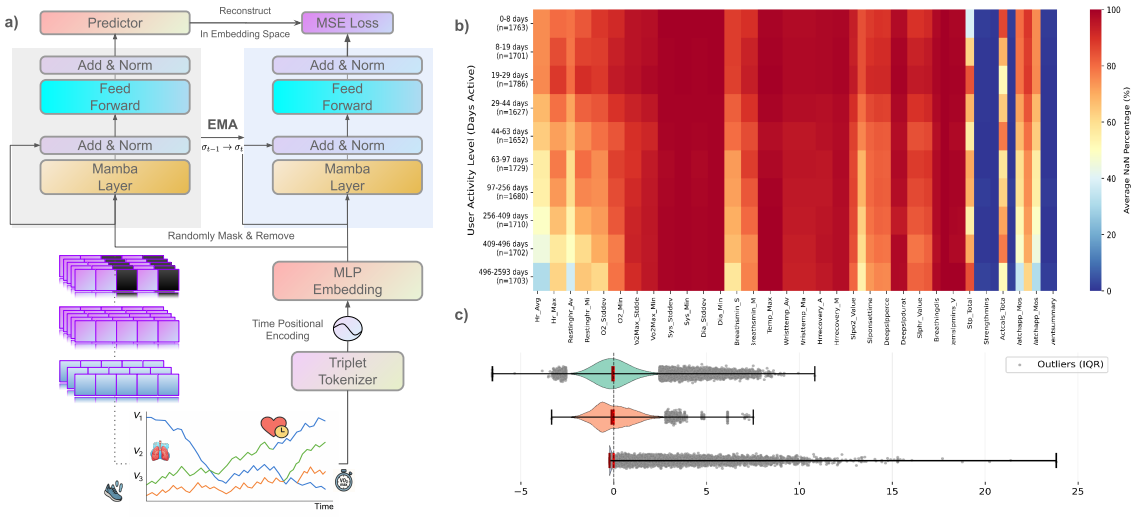

Figure 1: a) An overview of the model architecture using a joint embedding framework b) A summary of data missingness (NaN %), y: individuals binned by active duration, x: time series variable c) Distribution of example time series variables, y (top to bottom): average HR, VO_2max and wristTemp, x: z-score

3 Experiments

To evaluate the quality of the learned representations, we conduct linear probing experiments on two downstream tasks: disease diagnosis (self-reported) and biomarker prediction. For these tasks, the pretrained JETS encoder is frozen, and a single linear layer is trained on the available annotated labels. Implementation details are provided in the Appendix.

3.1 Baselines

Several baseline methods were selected to assess different aspects of our approach. To keep the comparison fair, an identical representation dimension of $D = 256$ is chosen for all models, and their parameter counts are kept as similar as possible.

1. Mean-pooling: As a simple sanity check, we employ a mean-pooling baseline. This method aggregates each of the 63 time series into a single 63-dimensional vector, providing a global summary of each user's data while discarding all temporal information.

2. Masked Autoencoder (MAE): To ablate the effect of our joint embedding pre-training, we implement an MAE baseline [11]. MAE is trained without the joint embedding objective using a transformer encoder and decoder architecture.

3. JETS-Former: A variant where the Mamba blocks in the encoder are replaced with bi-directional transformer blocks of a comparable size. This attention-based architecture is compared against our recurrent model, with other modules (besides the tokenizer, see Appendix) unchanged.

4. PrimeNet: A self-supervised algorithm that uses time-sensitive contrastive learning, specifically adapted to IMTS. We include PrimeNet to compare the effect of our reconstructive regime against its contrastive approach, as its been shown to out-perform a family of other contrastive algorithms (i.e. TS2Vec).

3.2 Results

15% of participants with self-reported medical history were chosen for evaluation.

Diagnosis Prediction: the task consists of predicting the presence or absence of specific medical conditions. Each condition constitutes a binary classification problem where the self-reported diagnosis serves as the ground truth label. We report the area under ROC (AUROC) and PRC (AUPRC) for the binary diagnosis classification task.

Biomarker Prediction: the model is trained to predict the value of continuous physiological markers (e.g., cholesterol), as a regression task. We report Mean Relative Error (MRE).

JETS, with a Mamba-based backbone, shows robust performance when compared to baseline models on the classification task (Table 1).

Table 1: Downstream Diagnosis Prediction. Left: AUROC (↑). Right: AUPRC (↑)

Target	**Mean-Pooling**		**JETS**		**MAE**		**JETS-Former**		**PrimeNet**	
ADHD or ADD	0.643	0.245	**0.668**	**0.260**	0.612	0.214	0.623	0.204	0.611	0.209
Asthma	0.673	**0.158**	**0.679**	0.149	0.598	0.105	0.616	0.120	0.619	0.149
Atrial flutter	0.495	0.003	**0.705**	**0.026**	0.428	0.004	0.576	0.006	0.604	0.006
Autism spectrum	0.658	0.099	0.650	0.080	0.610	0.072	0.588	0.058	**0.719**	**0.101**
Circadian rhythm	0.582	0.013	**0.654**	**0.019**	0.470	0.010	0.472	0.011	0.479	0.016
Depression	0.630	0.230	0.648	0.239	0.573	0.216	0.619	0.206	**0.656**	**0.272**
ME/CFS	0.607	0.012	**0.810**	**0.026**	0.385	0.004	0.458	0.004	0.580	0.005
Osteoporosis	0.749	**0.055**	0.758	0.050	0.648	0.028	0.585	0.038	**0.865**	0.042
POTS	0.678	0.233	0.731	0.307	0.630	0.028	0.680	0.276	**0.754**	**0.347**
Sick Sinus Syndrome	0.748	0.012	**0.868**	**0.125**	0.670	0.005	0.396	0.005	0.673	0.046
Substance abuse	0.589	**0.076**	**0.915**	0.047	0.613	0.064	0.700	0.026	0.757	0.053
Long Covid	0.631	0.047	**0.672**	**0.047**	0.521	0.022	0.512	0.022	0.587	0.005
Anxiety	0.643	0.301	0.675	**0.345**	0.592	0.260	0.641	0.271	**0.697**	0.345
Hypertension[1]	0.661	0.062	**0.868**	0.164	0.562	0.136	0.649	0.043	0.731	**0.272**

Within prediction of biomarkers, JETS had the highest performance across models, but overall accuracies were lower likely due to limitations in the number of sample used in training and evaluation (see Appendix).

Table 2: Biomarker prediction, MRE (↓)

Target	**Meaning-Pooling**	**JETS**	**MAE**	**JETS-Former**	**PrimeNet**
A1C	3.184	**3.167**	3.262	3.218	5.721
Glucose	0.083	**0.081**	0.082	**0.081**	0.335
HDL	1.576	**1.493**	1.568	1.618	1.645
LDL	3.561	3.363	3.692	3.554	**2.102**
hsCRP	2.809	2.353	1.959	**1.477**	1.585
Cholesterol	1.828	1.790	1.864	1.881	**0.619**

4 Discussion

In this study, we purposed an adaption of the Joint Embedding framework to pretraining on behavioral IMTS data. Through two benchmarks relevant to real-world applications, we showed that JETS was able to capture temporal relations robustly, highlighting its potential in facilitating personal health management outside of the traditional healthcare framework.

While JETS shows promise for behavioral time series, several limitations remain: we used only an MSE objective and one tokenization strategy, leaving contrastive losses (as in C-JEPA [15]) and alternative discretizations for future study; although results improved predictive metrics, fairness across subgroups and clinical utility should be assessed before real-world deployment.

[1] all blood pressure metrics were removed from inference data for this task to prevent data leakage

References

[1] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals, 2024.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

[3] Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H. Tison, Gregory M. Marcus, Jose M. Sanchez, Carol Maguire, Jeffrey E. Olgin, and Mark J. Pletcher. Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction, 2018.

[4] Ranak Roy Chowdhury, Jiacheng Li, Xiyuan Zhang, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. Primenet: Pre-training for Irregular Multivariate Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7184–7192, jun 26 2023.

[5] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024.

[6] Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embedding go temporal. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

[7] Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions, 2025.

[8] 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. AMIA Joint Summits on Translational Science*, 2020:191–200, may 30 2020. [Online; accessed 2025-07-30].

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[10] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 4 1982. [Online; accessed 2025-08-19].

[11] Zhangyi Hu, Jiemin Wu, Hua Xu, Mingqian Liao, Ninghui Feng, Bo Gao, Songning Lai, and Yutao Yue. Imts is worth time $\times$ channel patches: Visual masked autoencoders for irregular multivariate time series prediction, 2025.

[12] Andrew T. Jebb, Louis Tay, Wei Wang, and Qiming Huang. Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology*, 6, Jun 2015.

[13] Umapathi Krishnamoorthy, V Karthika, M K Mathumitha, Hitesh Panchal, Vijay Kumar S Jatti, and Abhinav Kumar. Learned prediction of cholesterol and glucose using ARIMA and LSTM models – A comparison. *Results in Control and Optimization*, 14:100362, 3 2024.

[14] Yann LeCun. A Path Towards Autonomous Machine Intelligence — openreview.net. `https://openreview.net/forum?id=BZ5a1r-kVsf`, 2022.

[15] Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning, 2024.

[16] Md Ifaj Hossan Omi, Md Atik Shams, Md Samiur Rahman, Ziaul Karim Asfi, Md Akhtaruzzaman Adnan, and Shahera Hossain. A personalized approach using lstm-based time-series analysis. *Activity, Behavior, and Healthcare Computing*, page 300, 2025.

[17] Zineb Senane, Lele Cao, Valentin Leonhard Buchner, Yusuke Tashiro, Lei You, Pawel Andrzej Herman, Mats Nordahl, Ruibo Tu, and Vilhelm von Ehrenheim. Self-Supervised Learning of Time Series Representation via Diffusion Process and Imputation-Interpolation-Forecasting Mask. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2560–2571. ACM, aug 24 2024.

[18] Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. Lsm-2: Learning from incomplete wearable sensor data, 2025.

[19] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series, 2022.

[20] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, and Shirui Pan. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects, 2024.

[21] Shubao Zhao, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, and Yi Wang. Himtm: Hierarchical multi-scale masked time series modeling with self-distillation for long-term forecasting, 2024.

[22] Shubao Zhao, Xinxing Zhou, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, Yi Wang, Yanlong Wen, and Xiaojie Yuan. Rethinking self-supervised learning for time series forecasting: A temporal perspective. *Knowledge-Based Systems*, 305:112652, 12 2024.

A Technical Appendices and Supplementary Material

A.1 Code and Data Availability

The dataset was originated from a mobile app, and de-identified before its use in this study. Individual participants agreed to a privacy policy, terms of service, and HIPAA notice (if receiving medical care). Due to HIPAA compliance, the authors are unable to release the training data. All code used in the study is available at: `https://anonymous.4open.science/r/JETS-1027`

A.2 Pre-training Details

JETS was pre-trained for 50 epochs using the following configuration. All experiments were implemented in PyTorch and ran on an Nvidia L4 GPU.

Implementation: Bi-directional Mamba2 layers were used in the encoder. For the encoder, we used 8 layers and 4 heads, with d_state = 64, d_conv = 4, expand = 2. For the predictor, we used 2 transformer decoder layers with the context embedding as context, and the learned positional embedding as queries. The predictor used n_head = 2, and a hidden_dim of $2 \cdot D$. Empirically, we found this predictor architecture led to more stable training than an MLP (with or without normalization).

Objective: We employed a masked modeling objective where mask ratio = 0.7, the masked tokens are subsequently removed from the input into the context encoder. The training loss was the Mean Squared Error (MSE) between the representations predicted by the predictor and the frozen representations (normalized) generated by the target encoder. Experiments replacing the MSE objective with cosine similarity showed no notable performance differences.

Optimizer: We used the AdamW optimizer with a learning rate of 1e-4 and a weight decay of 1e-5. All experiments were implemented in PyTorch and ran on an Nvidia L4 GPU.

Learning Rate Schedule: We applied linear warmup for 8 epochs. A Cosine Annealing scheduler was used to adjust the learning rate, started at 1e-5 and gradually decayed to a minimum of 1e-8 over the 50 epochs. In addition, we employed a linear schedule of the EMA momentum, from 0.996 to 1, as used in I-JEPA [2].

Batching and Regularization: The model was trained with a batch size of 16. To ensure training stability, we applied gradient clipping with a maximum L2-norm of 1.0.

Evaluation: At the end of each epoch, the model's performance was evaluated on the validation set. The model checkpoint with the lowest validation loss was saved for all downstream fine-tuning and evaluation tasks.

A.3 Linear Probing Details

All linear probes were trained and tested with a random 85-15 split, with 50 and 100 epochs of training, for diagnoses and biomarkers, respectively. The embeddings were mean-pooled along the time axis for all models.

Implementation: Each probe is a linear layer mapping from the embedding dimension to 1.

Objective: For diagnosis classification, we used binary cross-entropy with positive weighting computed by the ratio $N_{neg} : N_{pos}$. For biomarker prediction, we used mean absolute error to minimize the effect of potential self-reported outliers. All biomarker values were normalized to mean = 0 and std = 1 prior to training.

Optimizer: We used the Adam optimizer with a constant learning rate of 1e-4 and weight decay of 1e-2 for all biomarker and diagnosis probes.

A.4 Baseline Details

Masked Autoencoder: The core principle of Masked Autoencoders (MAE) involves a self-supervised reconstruction task where the model learns representations by directly predicting randomly masked portions of the input time series. For a fair comparison with our model, we aligned the MAE's architecture with JETS. Specifically, the MAE's encoder is identical to the JETS encoder, utilizing the same number of layers and attention heads. The decoder was intentionally designed to have a comparable complexity and parameter count to the predictor module in JETS, ensuring that the total number of trainable parameters across both models is nearly identical. This architectural parity allows us to isolate the performance differences attributable to the joint embedding objective. To maintain experimental consistency, we employed the same masking strategies, batch size (e.g., B=16), and evaluation metrics used for JETS.

JETS-Former: To investigate alternatives to state-space models like Mamba [5], we developed JETS-Former. This model explores whether the attention mechanism can provide a more effective representation for capturing long-range dependencies in behavioral time series. The core modification involves replacing the Mamba layers in both the encoder of the original JETS architecture with bi-directional transformer blocks. For each layer, we used n_head = 4, a feedforward dim of 4 $\times$ embedding_dim, GeLU activation, and 0.1 dropout. The number of layers was kept unchanged. In addition, transformers are permutation invariant and require positional encoding. For each triplet, we replaced Δt with t in the input and used a standard sinusoidal embedding to encode timestamps.

A.5 Data Details

Table 2 presents the details about our raw training data, and the portion of days each variable was available. Several pre-processing steps were taken prior to tokenization into triplets.

Filtering: We kept users with at least 300 total readings across all variables.

Normalization: The range of days was normalized to [0, 1], and each time series variable was normalized to a mean of 0 and a standard deviation of 1. Only training data was used to compute normalization statistics to prevent data leakage.

Chunking: A maximum observation length for each user was set to 5000, and for individuals with more than 5000 observations across all variable, they were chunked into new individuals.

Outliers: We removed outliers with z-scores greater than 8.0 for each individual variable. Several variables were heavily skewed by these outliers due to self-reported data or sensor errors, as shown in Table 2, which the removal step mitigated. Several baselines (MAE, TS2Vec) were unable to train without outlier removal. Specifically, we trained JETS without the outlier removal and observed comparable performance, which highlights the joint embedding strategies.

Table 3: Behavioral Time Series Pre-training Data

Variable (HR = heart rate)	Category	Sample rate	Collection Method	Avg. Avail. (%)	Mean
HR_avg	Cardiovascular	Continuous	Wearable	55.7	78.87 bpm
HR_stdDev	Cardiovascular	Continuous	Wearable	55.7	14.96 bpm
HR_max	Cardiovascular	Continuous	Wearable	35.2	127.57 bpm
HR_min	Cardiovascular	Continuous	Wearable	35.2	55.36 bpm
restingHR_avg	Cardiovascular	Daily	Wearable	50.4	65.18 bpm
restingHR_max	Cardiovascular	Daily	Wearable	32.2	65.29 bpm
restingHR_min	Cardiovascular	Daily	Wearable	32.3	64.57 bpm
oxygen_avg	Respiratory	Continuous	Wearable	30.5	96.0%
oxygen_stdDev	Respiratory	Continuous	Wearable	30.5	2.0%
oxygen_max	Respiratory	Continuous	Wearable	18.8	99.0%
oxygen_min	Respiratory	Continuous	Wearable	18.8	93.0%
vo2Max_avg	Cardiovascular	Sporadic	Wearable	8.2	33.94 mL/kg/min
vo2Max_stdDev	Cardiovascular	Sporadic	Wearable	8.2	0.10 mL/kg/min
vo2Max_max	Cardiovascular	Sporadic	Wearable	5.0	34.42 mL/kg/min
vo2Max_min	Cardiovascular	Sporadic	Wearable	5.0	34.28 mL
systolic_avg	Cardiovascular	Sporadic	Self-report	1.0	122.05 mmHg
systolic_stdDev	Cardiovascular	Sporadic	Self-report	1.0	3.79 mmHg
systolic_max	Cardiovascular	Sporadic	Self-report	0.5	128.68 mmHg
systolic_min	Cardiovascular	Sporadic	Self-report	0.5	117.09 mmHg
diastolic_avg	Cardiovascular	Sporadic	Self-report	1.0	76.85 mmHg
diastolic_stdDev	Cardiovascular	Sporadic	Self-report	1.0	2.54 mmHg
diastolic_max	Cardiovascular	Sporadic	Self-report	0.5	80.86 mmHg
diastolic_min	Cardiovascular	Sporadic	Self-report	0.5	73.08 mmHg
breathsMin_avg	Respiratory	During Sleep	Wearable	33.6	16.40 br/min
breathsMin_stdDev	Respiratory	During Sleep	Wearable	33.6	1.64 br/min
breathsMin_max	Respiratory	During Sleep	Wearable	20.8	20.70 br/min
breathsMin_min	Respiratory	During Sleep	Wearable	20.8	13.38 br/min
temp_avg	General	Sporadic	Self-report	0.4	40.19 °F
temp_max	General	Sporadic	Self-report	0.4	41.13 °F
temp_min	General	Sporadic	Self-report	35	38.71 °F
wristTemp_avg	General	During Sleep	Wearable	5.9	96.77 °F
wristTemp_stdDev	General	During Sleep	Wearable	5.9	0.03 °F
wristTemp_max	General	During Sleep	Wearable	5.9	96.80 °F
wristTemp_min	General	During Sleep	Wearable	5.9	96.73 °C
HRRecovery_avg	Cardiovascular	Per Workout	Wearable	4.6	29.15 bpm
HRRecovery_stdDev	Cardiovascular	Per Workout	Wearable	4.6	0.30 bpm
HRRecovery_max	Cardiovascular	Per Workout	Wearable	2.7	29.41 bpm
HRRecovery_min	Cardiovascular	Per Workout	Wearable	2.7	28.93 bpm
sleepOxygen	Respiratory	Daily	Wearable	15.7	95.0%
sleepDuration	Sleep	Daily	Wearable	38.4	397.28 min
sleepOnset	Sleep	Daily	Wearable	22.8	34.09 min
remSleepPercent	Sleep	Daily	Wearable	20.3	20.97%
deepSleepPercent	Sleep	Daily	Wearable	20.7	12.85%
remSleepDuration	Sleep	Daily	Wearable	3.0	83.77 min
deepSleepDuration	Sleep	Daily	Wearable	3.0	42.86 min
sleepQuality	Sleep	Daily	Wearable	17.6	0.80
sleepHR	Cardiovascular	Daily	Wearable	17.6	66.16 bpm
sleepHrv	Cardiovascular	Daily	Wearable	15.7	0.05 ms
breathingDisturb	Respiratory	Daily	Wearable	2.5	4.27 events/hr
steps	Activity	Daily	Wearable	28.6	120k
cardioMins	Activity	Daily	Wearable	98.8	30.14 min
strengthMins	Activity	Daily	Wearable	98.8	3.13 min
workoutTimeMins	Activity	Daily	Wearable	99.5	33.23 min
activityCals	Activity	Daily	Wearable	35.2	530.49 kcal
highIntensity	Activity	Daily	Wearable	98.8	0.32 min
mostRecentHRV	Cardiovascular	On-demand	Wearable	51.8	0.04 ms
mostRecentOxygen	Respiratory	On-demand	Wearable	30.3	96%
mostRecentHR	Cardiovascular	On-demand	Wearable	55.5	77.59 bpm
numHoursData	General	Daily	System Log	99.3	15.35 hours

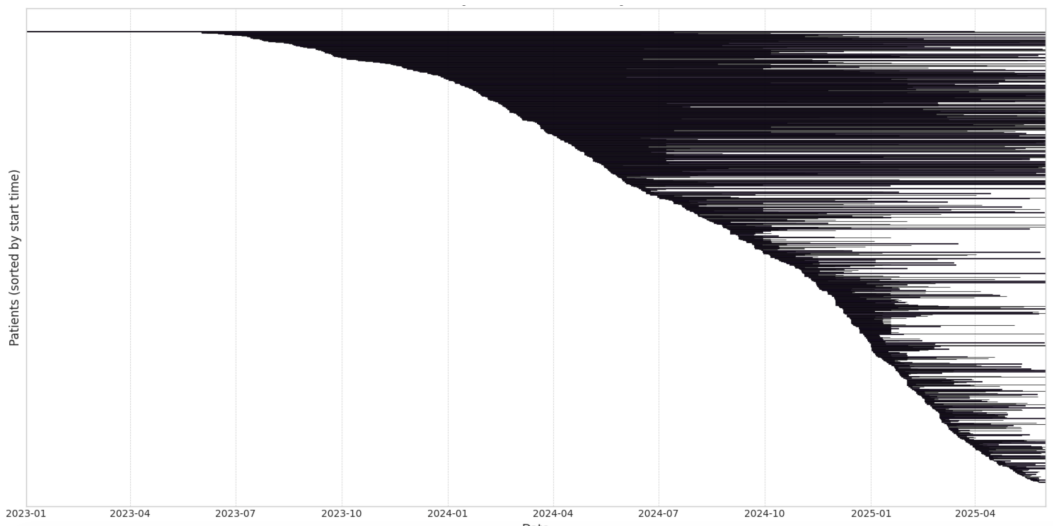

Figure 2: Participation timeline for a random subset of 1000 individuals in the JETS dataset, 2023-2025

Table 4: Finetuning Diagnosis Variables and Positive Rates

Target Variable	**Description**	**% Positive (All)**	**% Positive (Test)**
ADHD or ADD	Attention-Deficit/Hyperactivity Disorder	14.54%	14.61%
Asthma	Chronic Respiratory Disease	7.97%	7.57%
Atrial flutter	Rapid, Regular Heart Rhythm	0.27%	0.58%
Autism spectrum	Autism Spectrum Disorder (ASD)	4.52%	4.51%
Circadian rhythm	Circadian Rhythm Sleep-Wake Disorders	1.00%	0.93%
Depression	Major Depressive Disorder (MDD)	15.80%	17.46%
ME/CFS	Myalgic Encephalomyelitis/Chronic Fatigue Syndrome	0.46%	0.60%
Myocarditis	Inflammation of the Heart Muscle	0.27%	0.07%
Osteoporosis	Bone Density Loss Disease	1.33%	1.33%
POTS	Postural Orthostatic Tachycardia Syndrome	13.94%	16.33%
Sick Sinus Syndrome	Sinoatrial Node Dysfunction	0.27%	0.40%
Substance abuse	Substance Use Disorder (SUD)	0.73%	1.00%
Long Covid	Post-COVID conditions	2.12%	2.66%
Anxiety	Anxiety Disorders	20.25%	20.58%
Hypertension	High Blood Pressure	2.32%	3.12%

Table 5: Finetuning Biomarker Variable Distributions

Target Variable	**Description**	**Mean (Train)**	**Num. Avail. (Total)**
A1C	Glycated Hemoglobin Level	5.073	210
Glucose	Blood Glucose Level	92.594	301
HDL	High-Density Lipoprotein Level	54.105	272
LDL	Low-Density Lipoprotein level	104.675	238
hsCRP	High-sensitivity C-reactive Protein	3.947	168
Cholesterol	Total Cholesterol Level	189.251	278

A.6 Additional Metrics

Supplemental to the main section, we report 95% AUROC confidence intervals for JETS computed using the formula given by Hanley and McNeil [10]. We note that some of the targets were ultra-rare,

with very few users reported positive. This possibly contributed to the large variation in AUROC of the model.

Table 6: Diagnosis AUROC 95% Confidence Intervals. Left: Lower. Right: Upper

Target	**Lower**	**Upper**
ADHD or ADD	0.631	0.705
Asthma	0.629	0.729
Atrial flutter	0.531	0.879
Autism spectrum	0.585	0.715
Circadian rhythm	0.513	0.794
Depression	0.613	0.682
ME/CFS	0.657	0.963
Myocarditis	0.087	1.00
Osteoporosis	0.648	0.868
POTS	0.697	0.764
Sick Sinus Syndrome	0.703	1.00
Substance abuse	0.827	1.00
Long Covid	0.589	0.755
Anxiety	0.643	0.707
Hypertension	0.809	0.927

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the introduction are supported by experiments in section 3.2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the study are discussed in section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All aspects of model architecture and training setup are included in the Appendix, sections A.2, A.3, and A.4. Code is provided in section A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is available in Appendix section A.1 and includes both the model and the experiment setup. Data is unfortunately constrained by HIPAA policies and unable to be made public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details relevant to the results are included in the Appendix, sections A.2 through A.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The AUROC confidence intervals were included for JETS in the appendix under a normal distribution assumption.

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware specifications were included in the Appendix subsections A.2 and A.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and ensured the study conforms to the guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction and discussion sections the paper discuss the implication and impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The JETS pre-training framework is a method of extracting general representations from behavioral time series. The study doesn't pose any risk if the downstream applications are responsibly chosen, as in the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only existing code used in this study was from TS2Vec [19], which the original author open-sourced under the MIT license. Other models involved were implemented independently.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Anonymized code is released and properly documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM use did not contribute to any core components of this study.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.