MUST: MULTIMODAL UNIFIED SENSOR TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The development of general time-series foundation models is fundamentally constrained by the asynchronous, multi-rate nature of physiological data. Traditional early-fusion architectures attempt multimodal integration by forcing diverse streams onto rigid, synchronous interpolation grids. However, this artificial alignment incurs severe computational overhead, wastes parameter capacity on imputed noise, and causes cross-modal contamination where high-variance signals dominate. To overcome these limitations, we introduce Modality-Specific Early Tokenization coupled with semantic latent predictive alignment. Each sensor stream is processed by an independent patch encoder tailored to its native frequency, structurally isolating noise and bypassing artificial data imputation. By dynamically concatenating only valid active tokens, a shared transformer organically aligns multi-rate sequences with $\mathcal{O}(1)$ routing compute overhead. Evaluated on large-scale, real-world datasets comprising over 3.7 billion minutes of continuous sensor data from 50,000 users across nine modalities and 15 distinct activity classes, our framework consistently outperforms generic tokenization baselines across micro- and macro-level tasks. For fine-grained kinematics, it achieves a relative 18.2% Macro F1 score improvement on complex, long-tail activities. Furthermore, within the Digital Well Being (DWB) dataset, our architecture drives significant improvements in predicting socioeconomic factors related to health (69.7% vs. 65.7% average accuracy) while also achieving a higher average continuous tracking score across five physiological metrics (Pearson $r = 0.177$ vs. 0.154). Ultimately, we establish native modality tokenization as a highly resilient, architecture-agnostic primitive for multi-resolution time-series. To facilitate future research, we will open-source the DWB benchmark.

1 INTRODUCTION

The ubiquity of wearable sensors offers an unprecedented opportunity to capture continuous, 24/7 physiological and behavioral data. To leverage this wealth of information, the field has initiated a transformative shift from isolated heuristics toward general time-series foundation models. Recent breakthroughs, such as the Large Sensor Model (LSM), have proven that self-supervised representation learning can successfully scale to millions of hours of raw multimodal data. The ultimate vision of this domain is a single, holistic architecture capable of natively ingesting thousands of distinct sensor modalities, ranging from high-frequency heart rate signals to sporadic clinical lab results.

However, scaling these models via early-fusion and raw data imputation introduces fundamental engineering and representational bottlenecks. Unlike video or audio, sensor ecosystems are inherently asynchronous and multi-resolution. Modalities sample at vastly different timescales (e.g., 1 Hz continuous heart rate vs. 5-minute Heart Rate Variability), creating severe temporal alignment challenges. Furthermore, off-wrist periods and hardware packet drops create highly sparse missing blocks. While early resampling and zero-imputation onto a synchronous grid effectively solve generative reconstruction tasks, this approach suffers from quadratic scaling bottlenecks ($\mathcal{O}(N^2)$) when expanded to high-modality regimes. Crucially, forcing asynchronous data into a single unified frequency risks cross-modal interference; high-frequency biological variance (e.g., motion artifacts) can inadvertently pollute latent spaces during early fusion, leading to modality overshadowing and representation collapse.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

While modality-specific encoders have become standard practice in Computer Vision (e.g., parallel networks for RGB, depth, and semantic segmentation) and general multimodal frameworks (e.g., Audio/Video), these domains predominantly deal with synchronous, spatial, or cleanly aligned temporal data. Standard spatial encoders lack the mechanisms to handle the asynchronous, irregularly sampled rhythms inherent to biological sensors. Bringing the concept of modality isolation to multi-resolution time-series is not a trivial application; it requires a fundamental redesign of how we handle temporal boundaries, sequence concatenation, and masked attention.

To overcome the limitations of synchronous grids and bridge this gap in time-series representation learning, we propose a structural departure toward Modality-Specific Early Tokenization, coupled with semantic latent predictive alignment (JEPA). Instead of forcing rigid data alignment through offline resampling, each time-series modality is processed by its own independent patch encoder meticulously tailored to its native sampling frequency. This independent tokenization acts as a natural bandwidth bottleneck that structurally isolates high-frequency noise and seamlessly bypasses missing data blocks without artificial imputation. Ultimately, this allows a shared Vision Transformer (ViT) to organically align highly irregular, multi-rate temporal sequences in a unified semantic space.

Core Contributions

- **An Architecture-Agnostic Primitive for Native-Scale Preservation:** We introduce Modality-Specific Tokenization to completely bypass rigid synchronous interpolation grids. By preserving native sampling rates via independent patch encoders, this architecture-agnostic primitive improves downstream performance across different self-supervised backbones, yielding robust representations for both fine-grained kinematics (Activity Recognition) and macro-level physiological traits (e.g., Age, Depression).
- **Noise Resilience via Bottleneck Isolation:** We demonstrate that modality-specific bottlenecks isolate temporal artifacts and hardware dropouts. Preventing high-frequency noise from polluting the global token sequence prior to fusion ensures that joint representations are formed strictly from authentic physiological variance.
- **Dynamic, $\mathcal{O}(1)$ Asynchronous Alignment:** Instead of relying on computationally heavy placeholder imputation or combinatorial Mixture-of-Experts (MoE) routing, our architecture dynamically concatenates valid active tokens. This allows the shared ViT to organically align sparse, multi-rate streams, scaling compute natively with data availability.
- **Open-Source Benchmark for Diverse Clinical Tasks:** To democratize high-modality research, we publicly release the DWB dataset and our pre-training pipelines. This establishes the first open-source benchmark bridging continuous asynchronous streams with both high-frequency kinematic applications and low-resolution clinical endpoints under arbitrary real-world missingness.

2 Related Work

2.1 Sensor Multimodality for Self-supervised Learning

The closest related works are the Large Sensor Models (LSM) (Narayanswamy et al., 2026; Xu et al., 2025), which use a Masked Autoencoder (MAE) architecture (He et al., 2022) with early fusion of sensor modalities at the input level. However, their approach assumes that all modalities operate at the same regular sampling frequency and utilizes a modality-agnostic patching mechanism. This contrasts with our approach, which utilizes independent, modality-specific patch encoders to preserve native continuous frequencies without synchronous grids.

2.2 Sensor Multimodality for Supervised Learning

Recent architectures like FuseMoE (Han et al., 2024), Flex-MoE (Yun et al., 2024), and MAESTRO (Mohapatra et al., 2025) tackle arbitrary missing modalities for supervised learning via Sparse Mixture-of-Experts (MoE) routing. However, these introduce significant overhead: Flex-MoE scales combinatorially ($\mathcal{O}(2^{|M|}-1)$), FuseMoE relies on hidden-dimension-competing static placeholders, and MAESTRO uses supervised Symbolic Aggregate Approximation (SAX) (Lin et al., 2003), which discretizes and loses fine-grained continuous variance.

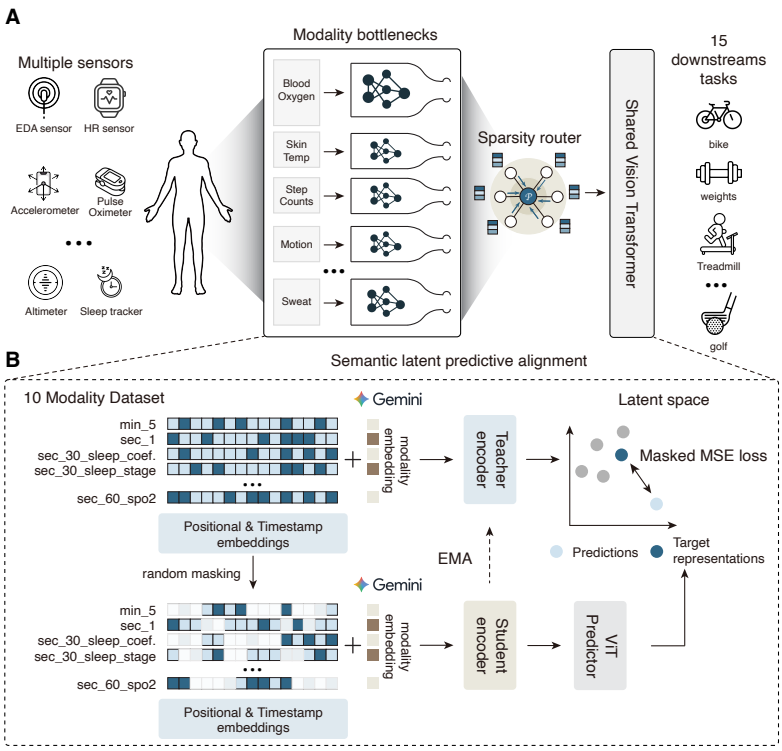

Figure 1: Modality-specific bottlenecks coupled with Multimodal Joint-Embedding Predictive Alignment (JEPA).

2.3 Multimodality More Broadly

In the multimodal domain more broadly, architectures like CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023) successfully align disparate modalities by projecting independent encodings into a shared semantic space. While highly effective for static or cleanly tokenized sequences, these frameworks often face combinatorial explosion when scaling beyond a few modalities. Similarly, advanced Computer Vision models employ parallel, modality-specific encoders to regularize learning (Girdhar et al., 2022), yet lack the native temporal mechanisms required for multi-rate, asynchronous data. Our work adapts the philosophy of modality isolation and bottleneck fusion (Nagrani et al., 2021), uniquely engineering it to solve the temporal irregularity and arbitrary missingness of physiological time-series.

3 Methodology

We formulate self-supervised learning on heterogeneous, multi-rate physiological sensors as a joint latent predictive alignment task. Let $\mathcal{S} = \{m_1, m_2, \ldots, m_M\}$ be a set of M distinct modalities, each sampled at its native frequency. To prevent the loss of granularity from blind resampling, we introduce Modality-Specific Bottlenecks coupled with Multimodal Joint-Embedding Predictive Alignment (JEPA) (Assran et al., 2023).

Rather than relying on mean-imputation or LOCF for missing data, we employ a native Score Penalty masking mechanism based on the valid fraction of each patch. Completely missing or structurally corrupted patches receive a heavy penalty, pushing them out of the unmasked context. For partially valid patches, invalid values are zero-imputed before encoding. Consequently, noisy patches are entirely excluded from the Student's sequence, ensuring the shared ViT routes strictly authentic physiological variance. Furthermore, corrupted target patches are excluded from the JEPA loss computation to prevent noisy gradient updates.

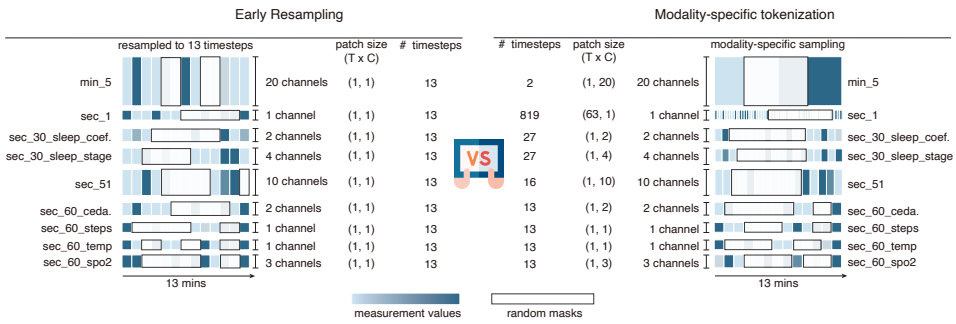

Figure 2: Processing of diverse multi-rate inputs (e.g., continuous EDA yielding 13 tokens vs. continuous PPG yielding 819 timesteps subsequently encoded into 13 tokens) using independent patch encoders tailored to intrinsic temporal shapes.

3.1 Modality-Specific Bottlenecks and Sparsity Routing

Each active modality m is processed by an independent mask-aware patch encoder tailored to its intrinsic temporal resolution. To establish a cohesive temporal scale before sequence fusion, most patch sizes are designed to encompass approximately one minute of physical time. Exceptions include sleep (30-second patches, to preserve its 27-step sequence indivisibility) and HRV (5-minute patches, to avoid upsampling).

Given input sequence $X_m \in \mathbb{R}^{T_m \times C_m}$, the encoders project these time-aligned patches into a shared embedding space D, producing sequence $E_m \in \mathbb{R}^{N_m \times D}$. Following validity-aware masking (Xu et al., 2025), only unmasked tokens from valid physiological blocks are retained.

As shown in Figure 1, each token is augmented with three embeddings: (1) a sinusoidal positional embedding, (2) a timestamp embedding (combining day of year, day of week, hour, and minute) to preserve absolute and periodic timing across missing intervals, and (3) a modality embedding. The modality embedding encodes semantic identity using a frozen Gemini text encoder on modality descriptions (Appendix Table 4). These allow the shared Transformer to retain chronological and modal identity after concatenation.

Dynamically concatenating only valid tokens along the global sequence dimension forms a unified, temporally-aligned sequence. This bypasses computationally heavy placeholder imputation and Mixture-of-Experts (MoE) routing, allowing the shared Student ViT to organically align multi-rate streams with $\mathcal{O}(1)$ routing compute.

3.2 Semantic Latent Alignment (JEPA)

Instead of reconstructing missing data in the input space, we enforce alignment directly in the high-dimensional latent space. A Joint Predictor receives the unmasked global context embeddings and predicts the latent target representations of each modality independently. For each masked target patch, a shared learnable target embedding—augmented with the corresponding positional embedding—specifies the prediction target location. The losses across individual modality streams are linearly summed and scaled by $1/d_{\text{proj}}$, where d_{proj} denotes the latent projection dimension.

3.3 Native Sparsity and Latent Alignment

Traditional pre-training relies on interpolation grids (e.g., forward-filling or raw-signal imputation) to align sparse, irregular asynchronous streams. We argue this introduces two critical failure modes:

1. **Latent space corruption (gradient poisoning):** Imputing raw gaps forces the model to predict synthetic patterns and noise, risking gradient shortcuts that corrupt the latent space rather than capturing true physiological variance.

2. **Compute waste:** Forcing all modalities onto a grid matching the highest-frequency sensor introduces massive computational overhead to process redundant imputed frames.

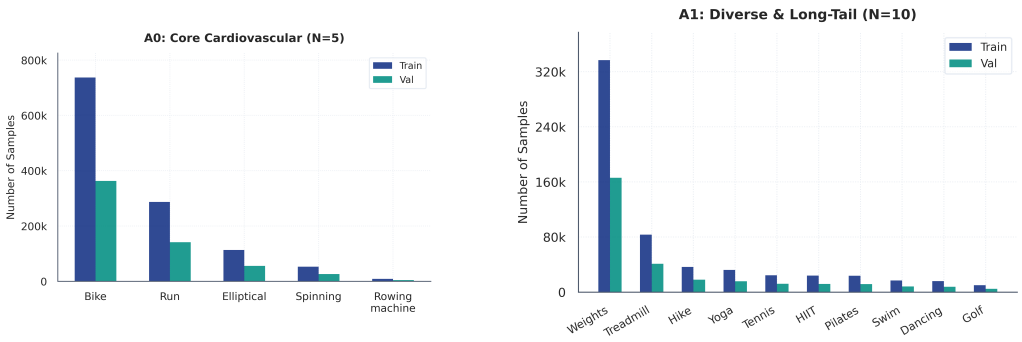

Figure 3: Distribution of A0 (Core) and A1 (Complex) activities across the validation set.

In contrast, our native sparsity approach operates strictly in event space. Tokens act as independent, variable-length temporal windows; during sensor dropouts, corrupted windows are routed out of the active context rather than imputed. The shared Transformer thus learns exclusively from valid physiological signals, maximizing both compute efficiency and gradient integrity.

4 Experimental Setup

4.1 Pre-training and Evaluation Datasets

We evaluate our architecture on large-scale, real-world physiological sensor datasets collected from wearable devices.

- **Pre-training Dataset:** The foundational models are pre-trained on a massive multimodal dataset comprising asynchronous streams from thousands of users, totaling millions of hours of sensor data. To evaluate our architecture's capacity for handling multi-rate temporal dynamics, the data encompasses 9 diverse modality streams: motion-derived kinematics, 1-minute step count, continuous EDA (cEDA), 1-second PPG-derived heart rate, 5-minute Heart Rate Variability (HRV), absolute SpO2, wrist skin temperature, sleep coefficient/wrist contact indicators, and sleep stage probabilities (detailed comprehensively in Appendix Table 3). All train, validation, and test splits are performed at the subject level, ensuring that no user's data appears in more than one split and preventing the leakage of user-specific physiological patterns.
- **Clinical and Demographic Tasks (DWB):** For macro-level clinical evaluations, we utilize the held-out test split (18,380 samples) of the Digital Well Being (DWB) dataset. The test cohort exhibits a diverse clinical profile (Mean age = 42.3, $\sigma = 12.0$; PHQ-8 mean = 4.46, $\sigma = 4.22$; GAD mean = 5.15, $\sigma = 4.60$). In accordance with study consent and privacy protocols, individual raw sensor traces are not explicitly depicted in the manuscript; the dataset is approved for public academic research under a CC-BY-NC 4.0 license (see Data Availability Statement).

4.2 Activity Class Categorization: Core (A0) vs. Complex (A1) Activities

To rigorously evaluate downstream activity recognition performance across different movement profiles and data availability, we group our 15 target activities into two evaluation sets:

- **A0 (Core Cardiovascular, N=5):** Continuous aerobic workouts (Running, Biking, Elliptical, Spinning, Rowing Machine). Contains 590,320 validation samples.
- **A1 (Diverse and Complex, N=10):** Specialized sports and multi-directional workouts (Treadmill, Weights, Hiking, Yoga, Tennis, Swimming, Pilates, Dancing, HIIT, Golf). Contains 297,673 validation samples.

4.3 Implementation Details

We implement our framework using the Modular Encoder Model Container in JAX/Flax. The Student and Teacher backbones are based on the standard Vision Transformer architecture (ViT-Base). Each of the active modality streams is assigned a specialized `MaskAwareTemporallyAware PatchEncoder` with native patch sizes matched to their underlying temporal sampling frequencies. To construct the global sequence, continuous raw sensor streams are chunked into non-overlapping **13-minute contextual windows**, striking a balance between localized activity tracking and computational efficiency. To provide the shared Vision Transformer with explicit cross-modal context, we encode the semantic descriptions of each sensor using the Gemini text encoder, which is kept frozen during training. These representations are projected into 768-dimensional semantic vectors and injected as Modality Embeddings alongside standard temporal positional encodings. The exact semantic prompts used for all 9 modalities are detailed in Appendix Table 4.

- **Masking Protocols:** We apply a 50% physical mask at the patch encoder level to simulate real-world data missingness. Additionally, for the JEPA objective, a target mask fraction of 50%–80% (50% for fine-grained kinematic pre-training and 80% for macro-level DWB evaluations) is used to define the semantic prediction tasks. Target patches can be selected either through uniform random masking across the sequence or through contiguous temporal block masking, allowing the model to learn from both dispersed and sustained missing regions.
- **Optimization:** The network is trained for $150,000$ steps using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.05. We employ a base learning rate of 1.0×10^{-4} (swept up to 2.5×10^{-4}) with a $7,500$-step (5%) linear warmup followed by a cosine decay schedule.
- **Momentum Teacher:** The Teacher network parameters are updated via an exponential moving average (momentum) starting at 0.9996 and decaying to 1.0.

5 Results & Analysis

We evaluate our proposed Modality-Specific Multimodal JEPA framework against both raw data baselines and standard early-resampling (synchronous grid) architectures. Evaluations are conducted via linear probing on frozen pre-trained representations.

5.1 Downstream Performance: Modality-Specific vs. Early Resampling

Person-Level Clinical, Demographic, and Lifestyle Tasks (DWB)
We evaluated the architecture on the open-source DWB dataset using a constrained 13-minute context window (Table 1). Under the JEPA backbone, the continuous tracking profile demonstrates functional specialization: Modality-Specific Tokenization drives gains on physiological and affective markers that rely on multi-rate signal integrity, including PSS Stress ($r = 0.1715$ vs. 0.1094, $+56.8\%$), Sleep Disturbance ($r = 0.1577$ vs. 0.1321, $+19.4\%$), and PHQ-8 Depression ($r = 0.1795$ vs. 0.1167, $+53.8\%$), while Early Resampling retains competitive performance on static demographic proxies such as Age ($r = 0.3245$ vs. 0.2985) and GAD Anxiety ($r = 0.0862$ vs. 0.0795). Crucially, on PSS Stress, Modality-Specific Tokenization recovers from the temporal blurring of Early Resampling ($r = 0.1715$ vs. 0.1094) to match the non-pretrained Raw Baseline ($r = 0.1710$), whereas the Raw Baseline achieves the highest correlation on PHQ-8 Depression ($r = 0.2421$)—reflecting both the timescale challenge of mapping 13-minute windows to retrospective 2-week depressive symptoms and the reliance of subjective affective states on unsmoothed raw signal variance. Across all five continuous endpoints, Modality-Specific Tokenization improves the overall average Pearson correlation under both JEPA ($r = 0.1773$ vs. 0.1538, $+15.3\%$) and MAE ($r = 0.2328$ vs. 0.1635, $+42.4\%$).

On structural categorical tasks, Modality-Specific Tokenization improves overall average accuracy under both JEPA (69.65% vs. 65.72%, $+6.0\%$) and MAE (62.93% vs. 61.07%, $+3.0\%$), though Early Resampling occasionally benefits specific aggregated targets (e.g., categorical Dep/Anx under JEPA or Medicaid under MAE). Notably, on Employment classification—a target heavily dependent on diurnal rhythms spanning 1-second activity bursts and 30-second sleep transitions—Modality-

Table 1: **Downstream performance on the DWB Test Cohort ($N = 18{,}380$).** Modality-Specific Tokenization (shaded) generally outperforms early resampling for both MAE and JEPA backbones. Brackets show the relative change over Early Resampling with the same backbone (gain / loss). **Bold**: best pre-trained result per backbone. †Baseline equals the majority-class proportion.

	Baseline	**MAE**		**JEPA**	
Task	(Raw)	Early Resamp.	**Modality-Spec. (Ours)**	Early Resamp.	**Modality-Spec. (Ours)**
Continuous Regression & Demographic Tracking (Pearson r ↑)					
Age	0.1629	0.2740	**0.4050** (+47.8%)	**0.3245**	0.2985 (−8.0%)
PHQ-8 (Depression)	0.2421	0.2100	**0.2317** (+10.3%)	0.1167	**0.1795** (+53.8%)
GAD (Anxiety)	0.0729	0.0930	**0.1604** (+72.5%)	**0.0862**	0.0795 (−7.8%)
PSS (Stress)	0.1710	0.1413	**0.2007** (+42.0%)	0.1094	**0.1715** (+56.8%)
Sleep Disturbance	0.0728	0.0993	**0.1664** (+67.6%)	0.1321	**0.1577** (+19.4%)
Average	0.1443	0.1635	**0.2328 (+42.4%)**	0.1538	**0.1773 (+15.3%)**
Categorical Person-Level Tasks (Accuracy % ↑)					
Employment†	66.28	66.28	**66.57** (+0.4%)	69.19	**72.38** (+4.6%)
Dep/Anx	53.62	**59.80**	58.29 (−2.5%)	**60.30**	57.79 (−4.2%)
Medicaid	67.15	**75.87**	69.19 (−8.8%)	73.26	**78.78** (+7.5%)
Race/Ethnicity	49.69	42.33	**57.67** (+36.2%)	60.12	**69.63** (+15.8%)
Average	59.19	61.07	**62.93 (+3.0%)**	65.72	**69.65 (+6.0%)**

†The 66.28% accuracy of the Raw Baseline and MAE Early Resampling equals the majority-class prior (66.28% employed in the DWB test split), so neither model beats the class prior. Modality-Specific JEPA reaches 72.38% because it learned useful diurnal features. | **Dep/Anx**: categorical threshold derived from PHQ-8/GAD severity scores. | Relative change = (Modality-Spec. − Early Resamp.) / Early Resamp.

Specific JEPA achieves 72.38% test accuracy, substantially outperforming both Early Resampling JEPA (69.19%) and the Raw Baseline (66.28% majority-class prior). Similar gains emerge in Race/Ethnicity (69.63% vs. 60.12%) and Medicaid (78.78% vs. 73.26%) classification (Table 1).

Multimodal Activity Recognition

Resolution vs. Modality Isolation: A natural question is whether the performance gains of Modality-Specific Tokenization stem from preserving native temporal resolution or from structural modality isolation. Constructing an early-fusion grid matched to the highest sensor frequency (1 Hz) would require upsampling all 44 channels, expanding the input to over 36,000 values per 13-minute window. This causes an intractable $\mathcal{O}(N^2)$ quadratic self-attention bottleneck while forcing sparse channels (e.g., 5-minute HRV or 30-second sleep) to be heavily over-interpolated. Modality-specific patch encoders resolve this trade-off by acting as specialized bandwidth bottlenecks: they preserve native intra-sensor dynamics within localized linear projections while projecting each stream into a compact, shared token sequence of uniform temporal duration.

As hypothesized, this architectural advantage is most pronounced in long-tail (A1) activities, where the Weighted F1 score under the JEPA backbone increased from 0.53 to 0.57 (+7.5% relative gain), and the unweighted Macro F1 increased from 0.33 to 0.39 (+18.2% relative gain, driving standout relative improvements on fine-grained classes such as Pilates (+366.7%) and Swim (+16.2%)). On common, data-rich A0 activities, both tokenization frameworks perform highly competitively (achieving 0.91 Weighted F1), indicating that while high-frequency data density can overcome early-fusion noise for simple movements, modality isolation is required to resolve complex long-tail kinematics.

Generalizability to Masked Autoencoders (MAE)
To prove that our tokenization bottleneck is an architecture-agnostic primitive, we applied it to a standard Masked Autoencoder (MAE) backbone. Using modality-specific encoders, the MAE achieves strong baseline performance, yielding a Weighted F1 of 0.86 on core (A0) activities and 0.51 on complex, long-tail (A1) activities. We then assessed the contribution of this native tokenization by replacing it with the generic tokenization early-resampling grid. While the early-resampling approach performs strongly on the simpler, data-rich A0 activities (0.90 Weighted F1 / 0.67 Macro F1), it suffers a sharp drop in prevalence-weighted fidelity on the more complex A1 tasks, where Weighted F1 drops from 0.51 to 0.45 (Table 2). Interestingly, under the MAE backbone, while Modality-

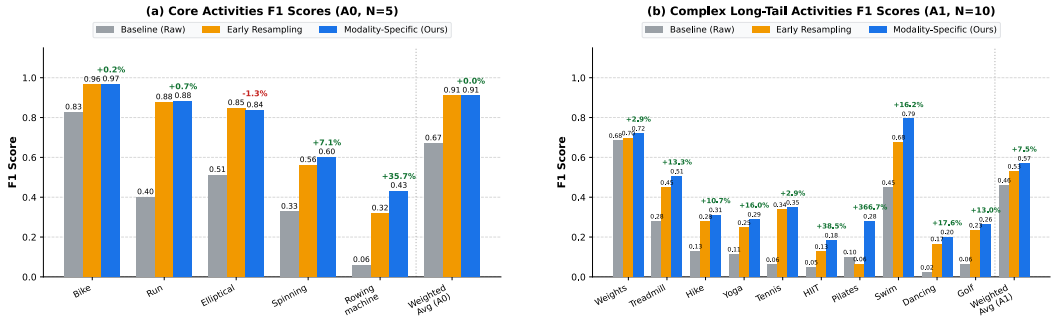

Figure 4: F1 scores comparing the no-pretraining baseline (logistic regression), early resampling, and our modality-specific tokenization for (a) core and (b) complex long-tail activities. Rightmost bars show sample-weighted average F1 scores. Annotations indicate relative percentage gains over early resampling, with bar values rounded to two decimal places for readability. Modality-specific tokenization improves weighted F1 by 7.5% (0.57 vs. 0.53) on long-tail activities, with notable gains for Pilates (+366.7%) and Swim (+16.2%), alongside steady gains on core cardiovascular activities such as Rowing machine (+35.7%).

Table 2: **Multimodal activity recognition on the Validation Cohort** ($N = 887{,}993$)**.** Generic early resampling performs adequately on data-rich Core (A0) activities. Modality-specific tokenization (shaded) prevents representation collapse on complex Long-Tail (A1) activities, most consistently with the JEPA backbone. Brackets show the relative change over Early Resampling with the same backbone (gain / loss). **Bold**: best per backbone.

	Core (A0)		Long-Tail (A1)	
Tokenization	W-F1 ↑	M-F1 ↑	W-F1 ↑	M-F1 ↑
MAE				
Early Resampling	**0.90**	**0.67**	0.45	**0.32**
Modality-Spec. (Ours)	0.86 (−4.4%)	0.64 (−4.5%)	**0.51 (+13.3%)**	0.29 (−9.4%)
JEPA				
Early Resampling	**0.91**	0.71	0.53	0.33
Modality-Spec. (Ours)	**0.91** (±0.0%)	**0.74** (+4.2%)	**0.57 (+7.5%)**	**0.39 (+18.2%)**

W-F1: weighted F1 | **M-F1**: macro F1 | **Core (A0)**: 5 cardiovascular activities ($N = 590{,}320$) | **Long-Tail (A1)**: 10 complex sport activities ($N = 297{,}673$) | Relative change = (Ours − Early Resamp.)/Early Resamp.

Specific Tokenization improves overall prevalence-weighted fidelity on A1 activities (Weighted F1 0.51 vs. 0.45, a +13.3% gain), unweighted Macro F1 declines from 0.32 to 0.29. This highlights an objective-dependent trade-off: pixel/patch-level generative reconstruction allocates modeling capacity proportionally to raw signal energy, favoring dominant motion patterns, whereas non-generative predictive alignment in latent space (JEPA) preserves balanced discriminative separation across long-tail classes (+18.2% Macro F1 gain). Similar improvements are observed on DWB continuous regression and categorical tasks (Table 1), again confirming the generalizability of our modality-specific encoders.

Comparing MAE and JEPA across Table 1 and Table 2 further illuminates the complementary strengths of generative reconstruction versus latent predictive alignment. Because MAE optimizes a token-level reconstruction loss directly in raw signal space, the encoder is forced to retain exact continuous amplitude, baseline magnitude, and morphological statistics—cues that strongly benefit continuous person-level regression ($r = 0.2328$ for Modality-Specific MAE vs. 0.1773 for JEPA). Conversely, because wearable streams contain substantial high-frequency motion and hardware artifacts, JEPA's prediction in abstract latent space discards low-level amplitude noise in favor of discrete semantic invariance. Consequently, Modality-Specific JEPA decisively outperforms Modality-Specific MAE across all categorical lifestyle tasks (69.65% vs. 62.93% average accuracy; e.g., 72.38% vs. 66.57% on Employment) as well as both Core (0.91 vs. 0.86 W-F1) and Long-Tail (0.57 vs. 0.51 W-F1; 0.39 vs. 0.29 M-F1) activity recognition.

5.2 Resilience to Temporal Sparsity and Missing Data

Real-world wearable data is characterized by severe temporal sparsity, hardware dropouts, and off-wrist periods. In contrast to traditional architectures that rely on synthetic zero- or mean-imputation to preserve rigid rectangular grids, our framework evaluates the viability of physical token excision. During pre-training, we apply a 50% physical validity-aware mask to simulate real-world sensor dropouts, dynamically excluding invalid patches from the token sequence. Under this 50% physical masking regime, the pre-trained representations achieve strong stability across diverse movement profiles: Modality-Specific JEPA maintains a 0.91 Weighted F1 on Core (A0) activities and a combined A0+A1 Weighted F1 of 0.73 (rising to 0.92 on A0 and 0.75 on combined A0+A1 under an 80% target mask), while Modality-Specific MAE maintains a 0.86 Weighted F1 on Core (A0) activities and 0.66 on combined A0+A1. This demonstrates that the shared Vision Transformer can learn coherent multimodal representations directly from authentic, partially observed physiological streams without requiring artificial imputation grids.

5.3 Cross-Modal Dynamics and Evaluation Settings

To verify that the joint embedding space successfully captures complementary variance across the 9 available sensor modalities, we compare our self-supervised representations against a non-pretrained raw baseline. Pre-training provides massive downstream headroom across both structural and demographic targets. Utilizing the raw baseline on the DWB dataset yielded substantially lower downstream performance on overall continuous ($r = 0.1443$ vs. 0.1773) and categorical (59.19% vs. 69.65%) benchmarks compared to our pre-trained Modality-Specific JEPA model (Table 1). For instance, Medicaid prediction dropped from an accuracy of 78.78% down to 67.15%, and Race/Ethnicity classification collapsed from 69.63% down to 49.69%. This confirms that the shared ViT successfully learns rich, cross-modal semantic alignment that linear probes on raw multi-rate signals cannot naturally resolve (see Appendix Figure 5 for qualitative t-SNE manifold visualizations).

6 Discussion & Conclusion

We present a scalable framework for time-series foundation models that resolves asynchronous timescales, temporal sparsity, and cross-modal contamination in physiological data. By coupling modality-specific tokenization with JEPA latent alignment, we demonstrate a core machine learning insight: independent patch encoders exhibit clear functional specialization over rigid synchronous grids, driving substantial gains on multi-rate dynamics and long-tail kinematics while synchronous grids remain competitive primarily on static, low-frequency endpoints.

An Architecture-Agnostic Primitive. Our ablations confirm that modality-specific bottlenecks serve as an architecture-agnostic primitive, yielding consistent downstream gains across both JEPA and MAE backbones. Dynamically concatenating only valid active tokens natively bypasses hardware missingness without complex MoE gating, achieving $\mathcal{O}(1)$ routing compute while minimizing the sequence length fed into the ViT's quadratic attention span.

The Hierarchy of Time: Micro-States vs. Macro-Phenotypes. Constrained 13-minute embeddings act as pristine extractors for physiological "micro-states," excelling on fine-grained tasks like Multimodal Activity Recognition. Conversely, longitudinal "macro-states" (e.g., Age, PHQ-8 Depression) depend on chronic behavioral rhythms; bridging this timescale gap will require hierarchical aggregation across multi-hour or multi-day horizons.

Limitations & Future Work. Pre-training across billions of minutes constrained exhaustive multi-seed variance profiling, and scaling concatenation to thousands of streams will eventually encounter quadratic attention limits, motivating linear state-space backbones. Future work will incorporate stochastic modality dropout (MoDrop), benchmark across public cohorts (WESAD, UK-Biobank, OpenMHC), and align these unified embeddings with Large Language Models (LLMs) for clinical diagnostics. Full DWB open-source release details and ethical safeguards against demographic profiling are provided in Appendix A.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Armand Joulin, Michael Rabbat, and Yann LeCun. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xinyu Han, Hoang Nguyen, Charles W Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Ricci, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pratyush Mohapatra, Y Sui, A Pandey, S Xia, and Q Zhu. Maestro: Adaptive sparse attention and robust learning for multimodal dynamic time series, 2025.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Girish Narayanswamy, Maxwell A. Xu, A. Ali Heydari, Samy Abdel-Ghaffar, Marius Guerard, Kara Vaillancourt, Zhihan Zhang, Jake Garrison, Levi Albuquerque, Dimitris Spathis, Hong Yu, Hamid Palangi, Xuhai "Orson" Xu, David G. T. Barrett, Joseph Breda, Jed McGiffin, Yubin Kim, Yuwei Zhang, Naghmeh Rezaei, Samuel Solomon, Karan Ahuja, Tim Althoff, Jake Sunshine, Ming-Zher Poh, Benjamin Yetton, Ari Winbush, Nicholas B. Allen, James M. Rehg, Isaac Galatzer-Levy, Yun Liu, John Hernandez, Anupam Pathak, Conor Heneghan, Yuzhe Yang, Ahmed A. Metwally, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Xin Liu, and Daniel McDuff. Towards a general intelligence and interface for wearable health data, 2026. URL https://arxiv.org/abs/2605.22759.

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

Max A Xu, Daniel McDuff, Xin Liu, et al. Lsm-2: Learning from incomplete wearable sensor data, 2025.

S Yun, I Choi, J Peng, Y Wu, J Bao, Q Zhang, J Xin, Q Long, and T Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

APPENDIX

A DATA AVAILABILITY & ETHICAL CONSIDERATIONS

Data Availability: The Digital Well Being (DWB) dataset is released under a CC-BY-NC 4.0 license. To strictly preserve privacy, distributed data consists of de-identified feature aggregations rather than raw trajectories. Access is restricted to non-commercial academic research via a formal request procedure requiring verified academic affiliation. Access forms and evaluation splits will be linked upon publication.

Ethical Considerations: Demographic and socioeconomic targets (e.g., Race/Ethnicity, Medicaid) from the DWB dataset were utilized strictly as evaluation proxies to assess the model's capacity to capture complex lifestyle determinants of health. This foundation model is intended solely for clinical and behavioral research. It is not designed for, and must not be deployed for, the physiological profiling, surveillance, or real-world inference of protected demographic classes or socioeconomic status.

Table 3: Temporal Scale, Data Sources, and Included Features (44 Total Channels across 9 Modalities)

Temporal Scale	**Source / Group Name**	**Features Included**
5-Minute	Heart Rate Variability (PPG)	Median RR, Mean RR, SDNN, RMSSD, pNN50, LF/HF Ratio, Spectral Entropy (20 channels)
1-Second	Continuous Heart Rate	Heart Rate in BPM (1 channel)
30-Second	Sleep & Wrist Contact	Sleep Coefficient, Is On Wrist (2 channels); Awake, Light Sleep, Deep Sleep, REM Sleep (4 channels)
1-Minute	Continuous EDA & Vitals	EDA Level, EDA Contact % (2 channels); Steps (1 channel); Wrist Temp (1 channel); Absolute SpO2, SpO2 Confidence, SpO2 Coverage (3 channels)
51-Sec (Approx)	Kinematic / Motion Features	Step Count, Jerk Auto, Log Energy, Axis Mean, Kurtosis, etc. (10 channels)

Table 4: Gemini-generated semantic descriptions utilized for Modality Embeddings. These descriptions are encoded via a frozen text encoder to provide the joint Vision Transformer with explicit semantic context for each multi-rate stream. The Modality Embeddings distinguish modalities after patch concatenation while preserving their semantic relationships, such that modalities with similar descriptions have more similar embeddings.

Modality Key	**Semantic Description**
`min_5`	PPG Heart Rate Variability (HRV) metrics over 5-minute windows
`sec_1`	1-second resolution Heart Rate (HR) series
`sec_30_sleep_coeff`	30-second sleep coefficient and wrist contact indicators
`sec_30_sleep_stage`	30-second categorical sleep stage classification probabilities
`sec_51`	51-second aggregated physical motion and kinematic features
`sec_60_ceda`	1-minute Continuous Electrodermal Activity (cEDA) level and contact %
`sec_60_steps`	1-minute aggregated step count
`sec_60_temp`	1-minute wrist skin temperature series
`sec_60_spo2`	1-minute pulse oximetry blood oxygen saturation

B QUALITATIVE ANALYSIS: LATENT SPACE CLUSTERING

To qualitatively assess the semantic quality of the learned representations, we visualized the global latent embeddings using t-SNE projections. As shown in Figure 5, the resulting plots reveal tight, well-separated clusters corresponding to distinct A0 activity classes. Compared to the early resampling and baseline, where overlapped regions among rowing machine, run and elliptical activities are noticeably large, our method yielded a clearer separation and granular representations. This indicates the multimodal integration facilitates the capture of more fine-grained phenotypes. While both tokenizations achieve an identical 0.91 Weighted F1 when linear probes are evaluated on Core (A0)

activities (Table 2), the geometric organization demonstrates that preserving native temporal rates maintains tighter, more coherent physiological manifolds. The primary quantitative discriminative advantage emerges on the long-tail (A1) action space, where early resampling induces representation collapse. The emergent sub-clusters may reveal personalized health insights uncovered through multiple sensors, and we leave their systematic interpretation and clinical validation for future work. This provides empirical evidence that modality-specific bottlenecks successfully align complementary features while avoiding cross-modal noise contamination.

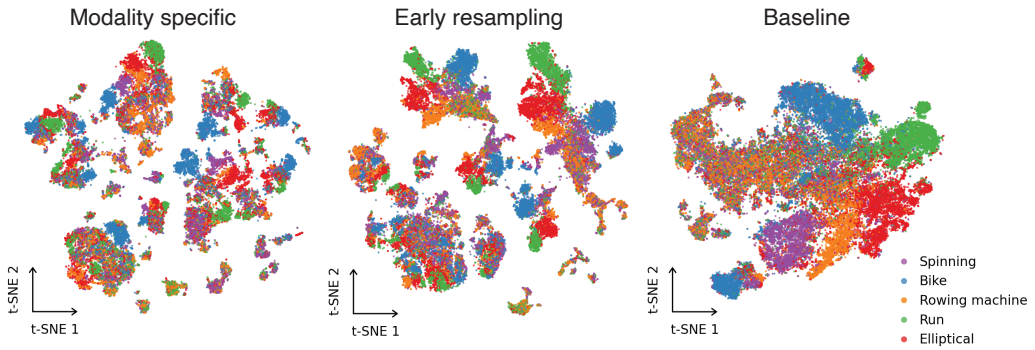

Figure 5: t-SNE projections of global latent embeddings across Core (A0) activities (Bike, Run, Elliptical, Spinning, Rowing Machine). Modality-Specific Tokenization yields tighter, well-separated cluster manifolds and reduces inter-class overlap compared to Early Resampling and non-pretrained baselines. Both Modality-Specific Tokenization and Early Resampling use the same JEPA backbone (ViT-Base model). Discrete color hues correspond to individual activity classes, highlighting distinct topological separation.

AI Assistance Statement

Following ICLR 2027 guidelines, we disclose the use of Large Language Models (LLMs) and AI assist tools during the development of this work. AI coding assistants were utilized during the research execution phase to aid in the software implementation of the JAX/Flax architecture. Additionally, a frozen LLM (Gemini) was utilized programmatically within the architecture to generate the semantic text descriptions of the sensor modalities, which were subsequently projected into the Modality Embeddings (detailed in Appendix Table 4). Finally, general-purpose LLMs were utilized during the drafting phase to aid in formatting, editing, and polishing the final manuscript text. The authors take full responsibility for the accuracy, originality, and integrity of all content within this paper.