GlucoFM: A Dual-Stream Foundation Model for Continuous Glucose Monitoring

Zechen Li[1,2] **Keerthana Natarajan**[1] **Weizhi Zhang**[1] **Menglian Zhou**[1]
Simon A. Lee[1] **Yuwei Zhang**[1] **Maxwell A. Xu**[1] **Zeinab Esmaeilpour**[1]
Flora D. Salim[2] **Mark Malhotra**[1] **Lindsey Sunden**[1] **Shwetak Patel**[1]
Yuzhe Yang[1,†] **Ahmed A. Metwally**[1,†]
[1]Google Research [2]University of New South Wales, Sydney
[†]Co-last authors
Correspondence to: {zechenl, yuzheyang, aametwally}@google.com

Abstract

Continuous glucose monitoring (CGM) provides a dense view of daily metabolic physiology, yet existing generic time-series and CGM-specific foundation models often encode glucose traces as entangled single-stream sequences, leaving their multiscale temporal structure only implicitly modeled. We present **GlucoFM**, a lightweight CGM foundation model that aligns irregular recordings to a 24-hour chronological grid, preserves observation masks, and decomposes glucose dynamics into slow-varying glycemic trend and short-term deviation streams. GlucoFM is pretrained on 109,066 hours of unlabeled CGM recordings from 477 subjects with masked contextual latent prediction over fused daily representations and temporal dynamics prediction over the two streams. Across four diverse cohorts and seven phenotype-classification tasks, GlucoFM achieves the strongest subject-disjoint linear-probing performance among evaluated baselines, improving average PR-AUC by 4.1 points over the best CGM-specific foundation model, while also supporting strong cross-dataset transfer and few-shot adaptation. Beyond phenotyping, when combined with recent CGM, meal nutrition, and subject context, the same frozen encoder achieves the lowest two-hour postprandial glycemic response errors among evaluated methods for trajectory, incremental area under the curve, peak rise, and peak timing. Together, these results demonstrate that GlucoFM learns reusable frozen CGM representations spanning metabolic phenotyping and context-conditioned glucose-response prediction.

1 Introduction

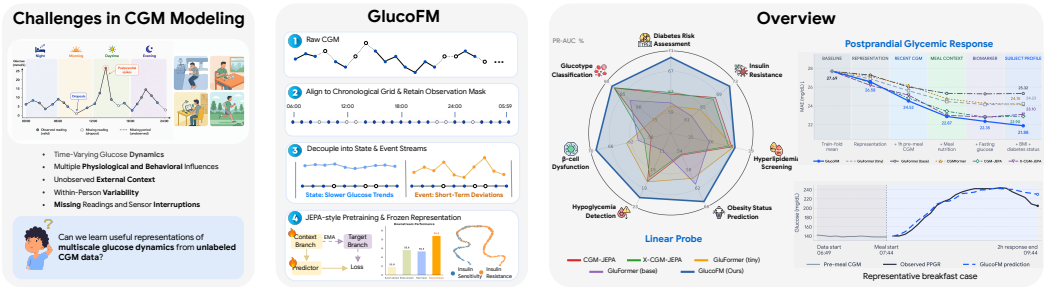


Figure 1: Overview of GlucoFM, a lightweight foundation model for continuous glucose monitoring.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Conceptual comparison between GlucoFM and representative TS/CGM foundation models. Checkmarks indicate components explicitly designed in the original model. *Decomp.* denotes decomposition, and *Aug.* denotes augmentation.

Model	Pretraining Objective	CGM-Specific	Daily Modeling	Signal Decomp.	CGM-aware Aug.
Chronos [11]	Quantized autoregressive forecasting	✗	✗	✗	✗
MOMENT [12]	Masked time-series modeling	✗	✗	✗	✗
Mantis [13]	Contrastive representation learning	✗	✗	✗	✗
GluFormer [7]	Next-token prediction	✓	✗	✗	✗
CGM-LSM [8]	Next-token prediction	✓	✗	✗	✗
CGMformer [9]	Masked token prediction	✓	✓	✗	✗
CGM-JEPA [10]	Masked latent prediction	✓	✓	✗	✗
GlucoFM **(ours)**	**Contextual JEPA + Dynamics**	✓	✓	✓	✓

Continuous glucose monitoring (CGM) provides a dense window into daily metabolic physiology, reflecting fasting baselines, nocturnal patterns, postprandial excursions, and fluctuations influenced by behavior, physiology, and sensor conditions. Its use has expanded from type 1 diabetes management, including hypoglycemia prevention, time-in-range monitoring, and closed-loop insulin delivery, to type 2 diabetes, prediabetes, and normoglycemic cohorts, where it helps characterize glycemic variability, treatment response, early metabolic dysfunction, and heterogeneous glucose phenotypes [1–5]. These signals are increasingly used for metabolic phenotyping, including diabetes risk, insulin resistance, β-cell dysfunction, hypoglycemia, obesity, hyperlipidemia, and glucotypes [6].

Despite this promise, high-quality clinical labels remain expensive, laborious to obtain, sparse, and cohort-specific, limiting fully supervised modeling [1]. This motivates self-supervised CGM foundation models that learn reusable representations from unlabeled recordings and transfer them across metabolic prediction tasks. Beyond phenotype classification, a useful CGM representation should also compose with temporally available event and subject context to predict future glucose responses without fine-tuning the encoder. Recent time-series and CGM-specific foundation models have advanced representation learning through forecasting, autoregressive generation, masked modeling, and latent prediction [7–10]. However, many existing approaches still encode CGM as a single sequence representation, leaving physiological and sensing structure only implicitly modeled. CGM dynamics are inherently multi-scale: slow glycemic trends coexist with short-term deviations potentially influenced by meals, activity, stress, or sensor artifacts. Meanwhile, heterogeneous sampling densities and missingness patterns from device protocols motivate sensing-aware modeling beyond naive interpolation and raw-value reconstruction. These challenges call for a CGM foundation model that explicitly incorporates multiscale decomposition and sensing-aware design.

Motivated by the multi-scale dynamics of CGM, we propose **GlucoFM**, a lightweight self-supervised foundation model for CGM representation learning. GlucoFM is a research prototype designed solely for retrospective physiological representation learning. GlucoFM aligns irregular recordings to a 24-hour chronological grid while preserving observation masks, and uses a dual-stream state–event encoder to separately model slow glycemic trends and transient deviations before fusion. Rather than reconstructing raw glucose values, GlucoFM is pretrained with JEPA-style latent objectives [14] that combine masked contextual representation learning and temporal dynamics modeling. It further incorporates CGM-aware augmentations that simulate value perturbations, heterogeneous sampling rates, and sensor dropouts, encouraging robust daily representations that capture both global glycemic context and local temporal changes. As summarized in Table 1, GlucoFM differs from prior TS/CGM foundation models through explicit state–event decomposition and CGM-aware augmentation.

We pretrain GlucoFM on 109,066 hours of unlabeled CGM recordings from 477 dataset-defined subjects and evaluate its frozen representations across four downstream cohorts and seven phenotype-classification tasks. For the controlled linear-probing comparison in Figure 1, CGM-specific baselines without public checkpoints are re-pretrained on the same unlabeled CGM corpus, enabling controlled comparison of architectural and objective-level differences. GlucoFM achieves the strongest average subject-disjoint linear-probing performance among evaluated baselines and further demonstrates strong cross-dataset transfer and few-shot adaptation. Pre-fusion probes show that the two streams retain distinct slow–fast temporal content. Beyond phenotyping, the same frozen encoder, when combined with recent CGM, meal nutrition, and subject context, achieves the lowest two-hour postprandial glycemic response (PPGR) errors among evaluated methods for trajectory, incremental area under the curve (iAUC), peak rise, and peak timing.

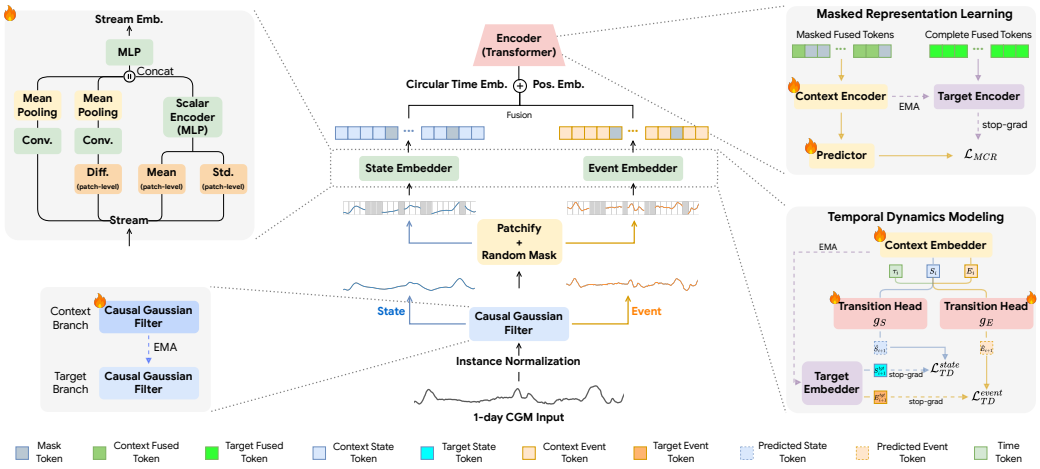


Figure 2: Model framework and pretraining objectives of `GlucoFM`.

Our contributions are threefold:

- **Lightweight daily CGM foundation model.** GlucoFM learns daily glucose representations from unlabeled CGM while preserving observation masks and absolute time-of-day structure.
- **Multiscale and sensing-aware representation learning.** GlucoFM combines state–event dual-stream modeling, CGM-aware augmentation, masked contextual latent prediction, and temporal dynamics modeling to encode complementary glycemic dynamics.
- **Broad downstream utility.** Across four CGM cohorts and seven phenotype tasks, GlucoFM shows strong linear probing, cross-dataset transfer, few-shot adaptation, and multiday prediction. Combined with recent CGM, meal nutrition, and subject context, the same frozen encoder further supports two-hour PPGR prediction.

2 Methodology

2.1 Modeling Input CGM Data Streams

Chronological Grid Alignment. Given CGM readings $X = \{x_i\}_{i=1}^{N}$ with timestamps $T = \{t_i\}_{i=1}^{N}$, our goal is to learn representations that preserve both glucose dynamics and absolute time-of-day structure. Since CGM recordings can be irregular due to heterogeneous sampling rates and sensor dropouts, we align each recording segment to a fixed 24-hour chronological grid with $\Delta t = 5$ minutes and $L = 288$ positions. The first timestamp defines the absolute circadian start index, $s = \lfloor (60 \cdot \mathrm{hour}(t_1) + \mathrm{minute}(t_1)) / \Delta t \rfloor$. The aligned input consists of a glucose sequence $\hat{X} \in \mathbb{R}^L$ and an observation mask $M \in \{0,1\}^L$, where $M_j = 1$ only for physically observed measurements. Missing positions are filled only for tensor construction, while M is preserved throughout representation learning. This separates chronological alignment from interpolation and allows GlucoFM to handle different sampling rates and sensor dropouts on a common daily grid.

CGM-aware Data Augmentation. We use two families of on-the-fly augmentations during pretraining. *Value perturbations* simulate realistic CGM signal variation, including low-frequency baseline wander and short compression-like drops, while preserving the observation mask. *Structural sparsification* alters the observation pattern itself by randomly decimating dense 5-minute recordings into 15-minute-like sampling and masking short contiguous disconnection blocks. These augmentations expose GlucoFM to sensor drift, compression artifacts, heterogeneous sampling rates, and physical dropouts, while randomized ordering and decayed co-occurrence probabilities prevent over-corrupting individual sequences.

2.2 State–Event Dual-Stream Encoder

Mask-aware Signal Decomposition. CGM signals naturally contain both low-frequency metabolic states and high-frequency transient events. Low-frequency components reflect stable glycemic

baselines and slower regulatory patterns, while high-frequency components represent short-term deviations from the local trend, which may arise from transient physiology and behavior. To avoid encoding these heterogeneous dynamics as a single entangled sequence, GlucoFM decouples the aligned signal into state and event components before masked representation learning, as shown in Figure 2.

Given the aligned sequence $\hat{X}$ and observation mask M, we first apply mask-aware normalization and compute patch-level statistics, including glucose summaries and rate-of-change features. We then estimate the low-frequency state using a learnable causal Gaussian filter. The filter bandwidth σ is initialized at $\sigma = 6.0$ and constrained within $[2.0, 12.0]$ grid steps, corresponding to approximately 10–60 minutes on the 5-minute grid. This range allows the filter to separate short-term fluctuations from slower glycemic trends while adapting its smoothing scale during pretraining. Causality is enforced by using only current and past observations when estimating the local trend, preventing future glucose values from leaking into the representation. The filter is also mask-aware, normalizing only over valid observations to avoid dropout-induced boundary artifacts.

The filtered trend defines the state component, while the residual captures high-frequency event-level deviations:

$$\hat{X}_{\text{state}} = \text{Filter}(\hat{X}, M), \qquad \hat{X}_{\text{event}} = (\hat{X} - \hat{X}_{\text{state}}) \odot M. \tag{1}$$

State and event signals are encoded by two parallel streams and fused into unified patch tokens.

Patch Tokenization and Circadian Encoding. After state–event decomposition, GlucoFM tokenizes each sequence into 24 non-overlapping temporal patches, each covering 12 grid steps, corresponding to one hour. The state stream encodes the filtered trend together with mask-aware state statistics, while the event stream encodes residual deviations together with rate-of-change statistics. The two stream outputs are then fused into unified physiological patch tokens. To preserve absolute time-of-day information, each token is augmented with circular time features, $\tau(i) = [\sin(2\pi i/L), \cos(2\pi i/L)]$, where i is the absolute grid index and $L = 288$. This encoding allows the context encoder to model glucose dynamics with awareness of circadian phase.

2.3 Latent Predictive Pretraining

GlucoFM uses two complementary JEPA-style objectives to predict latent representations rather than reconstruct potentially interpolated or noisy glucose values.

Masked Contextual Representation Learning. In our first pretraining objective, we randomly mask temporal patches in the online branch, with the masking ratio sampled from $[0.5, 0.6]$. In the online branch, each selected patch is replaced by a learnable mask token, and the resulting full patch grid is encoded by a context encoder. In parallel, the full physically observed sequence is encoded by an EMA target branch to provide stable latent targets. The target branch is initialized from the online branch and updated as $\theta_{\text{tgt}} \leftarrow m\theta_{\text{tgt}} + (1-m)\theta_{\text{online}}$, with m cosine-scheduled from 0.997 toward 1.0. A lightweight predictor maps the context-branch tokens to the target latent space and predicts the representations of masked patches. Let Z_i^{pred} and Z_i^{tgt} denote the predicted and target latent tokens at patch i. The masked contextual representation loss is:

$$\mathcal{L}_{\text{MCR}} = \frac{\sum_{i=1}^{P} w_i m_i^{\text{mask}} \,\text{SmoothL1}\left(Z_i^{\text{pred}}, Z_i^{\text{tgt}}\right)}{\sum_{i=1}^{P} w_i m_i^{\text{mask}} + \epsilon}, \tag{2}$$

where P is the number of temporal patches, m_i^{mask} indicates whether patch i is masked, and w_i is the fraction of physically observed measurements within the patch.

Temporal Dynamics Modeling. Masked contextual prediction learns daily contextual representations, but CGM phenotypes also depend on how glucose dynamics evolve over time. We therefore introduce a second JEPA-style objective that predicts next-patch state and event representations. Given context-branch state and event tokens (S_i, E_i) at patch i, two lightweight transition heads g_S and g_E predict the next-patch state and event targets:

$$\hat{S}_{i+1} = S_i + g_S([S_i, E_i, \tau_i]), \qquad \hat{E}_{i+1} = E_i + g_E([E_i, S_i, \tau_i]), \tag{3}$$

where τ_i is the temporal position embedding formed from patch position and gated circular time-of-day features. The residual update form encourages the heads to model temporal changes rather

than directly copying the current representation. The targets $(S_{i+1}^{\text{tgt}}, E_{i+1}^{\text{tgt}})$ are taken from the EMA target branch before global Transformer self-attention, so the next-patch targets do not already encode information from later patches. The temporal dynamics loss is:

$$\mathcal{L}_{\text{TD}} = \frac{1}{2}\left(\frac{\sum_{i=1}^{P-1} q_i \,\text{SmoothL1}\left(\hat{S}_{i+1}, S_{i+1}^{\text{tgt}}\right)}{\sum_{i=1}^{P-1} q_i + \epsilon} + \frac{\sum_{i=1}^{P-1} q_i \,\text{SmoothL1}\left(\hat{E}_{i+1}, E_{i+1}^{\text{tgt}}\right)}{\sum_{i=1}^{P-1} q_i + \epsilon}\right), \quad (4)$$

where q_i weights each transition by the observed support of adjacent patches and excludes patches hidden from the context branch. This objective encourages the state stream to encode gradual baseline shifts and the event stream to encode transient deviations, providing an explicit temporal constraint on the decoupled representations.

Overall Objective. The final pretraining loss is:

$$\mathcal{L} = \lambda_{\text{MCR}}\mathcal{L}_{\text{MCR}} + \lambda_{\text{TD}}\mathcal{L}_{\text{TD}}. \quad (5)$$

We set $\lambda_{\text{MCR}} = \lambda_{\text{TD}} = 1.0$. GlucoFM uses compact 3-layer Transformer context and EMA target encoders with hidden dimension 128, 4 attention heads, and feed-forward dimension 256, together with a lightweight 1-layer Transformer predictor. During downstream evaluation, we discard the EMA target branch and retain only the frozen online encoder. GlucoFM has 0.72M trainable parameters and 1.18M total parameters during pretraining, with the difference mainly due to the EMA target branch. Additional implementation details are provided in Appendix D.

3 Experiments

3.1 Experimental Setup

Datasets. Table 2 summarizes the datasets used in this study. ShanghaiT2DM counts denote recording-visit subject-entries. We use dataset-defined group separation between pretraining and downstream evaluation, ensuring no overlap between pretraining subjects and downstream test groups. For self-supervised pretraining, we aggregate five CGM cohorts spanning 5-minute and 15-minute sampling rates, totaling 477 dataset-defined subjects and 109,066 hours of recordings. This mixture exposes GlucoFM to diverse glucose dynamics, sensor densities, and population characteristics. For downstream evaluation, we use four CGM cohorts with 203 dataset-defined participants and 71,669 hours of CGM data, covering seven phenotype-classification tasks: diabetes risk assessment, insulin resistance, β-cell dysfunction, glucotype, hyperlipidemia, hypoglycemia, and obesity classification. All downstream evaluations use subject-disjoint splits to prevent overlap between training and test subjects, enabling robust assessment across heterogeneous CGM cohorts. Additional dataset details are provided in Appendix B.

Table 2: Composition of `GlucoFM` datasets.

Dataset	Sample Rate	# Subjects	Duration (h)
Pre-train Datasets			
`Wear-CGM`	5	192	75,330
`ShanghaiT2DM` [15]	15	44	12,414
`Stanford` [1]	5	19	8,761
`BIG IDEAs` [16]	5	16	3,017
`Colas` [17]	5	206	9,544
Downstream Datasets			
`CGMacros` [18]	5 | 15	45	10,376 | 10,998
`ShanghaiT2DM` [15]	15	65	15,634
`Stanford` [1]	5	37	27,571
`Hall` [6]	5	56	7,090

Baselines. We compare GlucoFM with *three groups of baselines*: general-purpose time-series foundation models, a CGM-specific open-weight model pretrained on a separate cohort, and CGM-specific models retrained on our pretraining corpus. Implementation details are provided in Appendix C. General-purpose baselines include Chronos-2 [19], MOMENT [12], Mantis [13], and MantisV2 [20]. Controlled CGM-specific comparisons use study-retrained implementations of CGM-JEPA and X-CGM-JEPA [10], and GluFormer [7], because official pretrained checkpoints were unavailable at the time of evaluation. These implementations follow the published methods and are trained on the same corpus as GlucoFM. We use GluFormer as the representative autoregressive CGM baseline because CGM-LSM [8] has a closely related architecture and next-token objective. We also evaluate CGMformer [9], an open-weight CGM-specific model pretrained on an external 964-participant cohort with non-public raw data.

Setup. We report ROC-AUC, PR-AUC, and Macro-F1. ROC-AUC measures global discriminative ability, PR-AUC emphasizes performance under class imbalance, and Macro-F1 evaluates class-balanced decision quality across phenotype classes. Detailed model training setup is provided in Appendix D.

Table 3: Linear-probe performance. The best result is shown in **bold**, and the second-best result is underlined. All values report mean performance over 10 iterations of 5-fold subject-grouped cross-validation. Core denotes key CGM-based metabolic phenotyping tasks.

Method	Params	Metrics	CGMacros				ShanghaiT2DM			Stanford			Hall				Avg
			Diabetes	IR	Hyperlip.	Obesity	IR	Hyperlip.	Hypogly.	Diabetes	β-cell Dys.	IR	Diabetes	IR	Hyperlip.	Glucotype	
General Time-series Foundation Models with Large-scale Pretraining																	
Chronos-2 (small)	28M	PR	51.6	82.7	30.5	56.0	65.0	34.8	16.0	69.2	58.6	62.5	51.7	44.9	24.2	54.2	50.1
		AUC	68.0	67.5	52.1	53.7	55.0	50.6	49.1	65.4	57.3	64.2	64.9	57.7	**61.3**	62.6	59.2
		F1	49.4	61.7	50.1	52.7	52.6	51.5	50.5	61.3	54.8	59.6	60.9	55.0	52.5	58.6	55.1
Chronos-2	120M	PR	54.3	84.1	34.0	58.1	59.8	34.7	18.9	68.5	58.5	59.2	48.6	51.3	20.6	54.4	50.4
		AUC	69.7	69.9	**55.8**	56.6	49.5	49.7	53.0	65.6	57.8	60.2	60.4	62.7	50.3	63.0	58.9
		F1	50.2	63.3	**53.7**	54.2	49.0	50.4	52.8	61.9	55.5	56.6	58.4	58.2	50.9	56.2	55.1
MOMENT (small)	40M	PR	54.0	83.8	30.3	60.9	58.3	38.3	16.3	70.4	57.2	63.0	59.2	53.1	19.4	55.6	51.4
		AUC	69.1	70.6	51.5	58.7	49.0	55.9	52.9	67.6	55.5	65.1	69.7	62.1	48.7	61.7	59.9
		F1	49.8	63.6	50.9	55.7	49.3	52.9	51.4	62.5	53.8	60.1	64.5	59.0	50.4	59.3	55.9
MOMENT (large)	385M	PR	50.3	87.4	32.0	61.8	59.6	**42.4**	14.3	71.9	61.8	58.1	52.1	44.9	18.4	55.8	50.8
		AUC	66.1	75.2	53.2	59.9	47.8	**60.8**	49.6	69.1	58.9	59.4	65.0	56.6	51.4	63.2	59.7
		F1	46.8	67.2	52.5	56.3	48.6	**57.2**	49.7	63.7	55.7	56.1	60.5	53.3	47.7	57.8	55.2
Mantis	8M	PR	61.8	90.9	30.1	63.4	59.2	30.6	**22.9**	75.1	63.5	67.1	54.2	55.7	24.8	77.5	55.5
		AUC	75.5	80.3	49.8	60.8	48.2	46.0	58.6	71.4	61.5	68.2	67.1	66.2	57.9	82.6	63.9
		F1	56.0	**71.6**	49.8	57.2	48.4	47.5	**53.8**	65.9	58.5	62.3	59.6	59.8	**53.7**	73.3	58.4
MantisV2	4.2M	PR	63.6	91.2	30.3	62.8	64.1	33.3	22.0	75.2	65.7	64.6	57.7	59.3	**25.0**	81.5	56.9
		AUC	76.7	80.3	49.1	60.3	52.7	48.5	**59.6**	71.2	64.4	65.5	67.5	66.7	58.8	84.7	64.7
		F1	56.2	68.9	49.2	55.9	51.2	48.1	53.3	65.7	58.6	60.4	61.1	60.8	52.6	76.7	58.5
Externally Pretrained CGM Foundation Model																	
CGMformer	0.85M	PR	63.3	89.9	31.8	**68.1**	54.8	40.6	11.9	75.9	62.2	61.8	51.1	48.0	18.0	79.6	54.1
		AUC	77.1	78.0	54.2	**66.0**	43.6	58.4	42.1	71.4	59.2	63.2	66.6	58.2	46.5	84.3	62.1
		F1	57.2	67.6	52.1	**60.5**	45.5	55.1	45.3	**66.2**	55.7	59.0	59.3	56.3	48.4	77.1	57.5
CGM Foundation Models Retrained on Our Pretraining Corpus																	
CGM-JEPA	0.52M	PR	63.0	86.2	28.7	55.5	**69.1**	37.4	17.8	66.4	58.7	61.5	59.9	56.8	17.3	87.6	54.7
		AUC	75.9	73.8	47.6	53.8	**60.8**	53.6	56.8	61.2	55.4	61.6	73.5	68.1	40.5	**90.7**	62.4
		F1	55.4	66.3	48.2	53.2	**57.2**	51.3	50.4	58.7	53.9	58.1	**65.0**	59.4	42.8	82.2	57.3
X-CGM-JEPA	0.52M	PR	63.6	86.6	29.8	55.3	66.9	35.5	16.9	67.4	60.0	61.9	59.3	56.2	17.5	87.7	54.6
		AUC	76.6	73.6	48.0	53.2	58.4	52.1	55.7	61.8	56.4	62.0	73.0	67.7	40.0	90.6	62.1
		F1	56.5	65.1	48.2	52.5	55.4	49.9	49.0	59.1	54.4	58.1	64.8	59.2	42.8	82.1	56.9
GluFormer (tiny)	0.65M	PR	59.4	86.1	28.2	60.2	58.1	33.9	17.9	74.3	63.3	64.5	48.9	50.9	21.6	75.4	53.0
		AUC	73.9	72.5	47.9	57.9	47.6	51.1	53.3	68.9	61.9	65.4	63.3	63.6	58.0	81.7	61.9
		F1	51.7	65.4	48.0	55.4	47.8	50.3	49.1	63.0	58.5	61.7	57.9	59.4	53.1	72.7	56.7
GluFormer (base)	135M	PR	62.9	87.2	31.4	61.4	56.5	31.5	11.5	69.8	61.1	56.1	51.9	51.6	16.4	82.7	52.3
		AUC	76.7	74.8	53.1	60.5	44.9	46.7	43.9	63.1	58.0	56.3	66.0	61.6	46.9	87.9	60.0
		F1	55.7	67.8	52.5	57.2	46.3	48.1	44.8	59.2	55.4	54.8	58.7	58.3	47.7	80.7	56.2
GlucoFM	0.72M	PR	**65.9**	**91.9**	**36.1**	64.9	67.0	33.5	21.1	**77.3**	**69.0**	**67.6**	**66.2**	**60.2**	14.4	**88.3**	**58.8**
		AUC	**78.7**	**81.2**	54.7	62.6	57.8	50.5	59.2	**72.8**	**68.7**	**69.1**	**75.9**	**70.7**	41.6	**90.7**	**66.7**
		F1	**58.3**	69.6	50.2	59.4	55.4	49.1	50.7	**66.2**	**63.3**	**64.0**	64.5	**62.0**	43.1	**82.4**	**59.9**

3.2 Experimental Results

We evaluate GlucoFM along four primary downstream axes that reflect different clinically relevant utility scenarios: subject-disjoint linear probing, postprandial glycemic response forecasting, few-shot adaptation, and cross-dataset transferability. Additional setup details and full results with fold-level variance are provided in Appendix E. Same-window GMI comparisons and frozen representation visualizations are provided in Appendices E.6 and E.8.

Subject-Disjoint Linear Probing. We first evaluate whether GlucoFM learns phenotype-informative 24-hour representations that generalize to unseen subjects while capturing day-to-day variability. For each method, we freeze the encoder, extract embeddings from non-overlapping 24-hour windows, and train the same logistic regression classifier on the extracted representations. We use subject-grouped cross-validation, ensuring that all windows from the same subject appear only in either the training or test fold. Each held-out window is predicted independently, with evaluation performed over all held-out test windows. As shown in Table 3, GlucoFM achieves the strongest overall performance, obtaining the best task-averaged PR-AUC, ROC-AUC, and Macro-F1, while ranking within the top two on 11/14, 11/14, and 9/14 task–dataset evaluations, respectively. Across the 14 task–dataset evaluations, the paired task-averaged gains over CGM-JEPA are $+4.11$ points for PR-AUC (95% CI $[2.40, 5.81]$), $+4.34$ for ROC-AUC ($[2.26, 6.41]$), and $+2.57$ for Macro-F1 ($[0.91, 4.23]$). Against CGMformer, the corresponding gains are $+4.74$ ($[2.47, 7.02]$), $+4.67$ ($[2.02, 7.32]$), and $+2.34$

Table 4: PPGR across cumulative context, averaged over Dexcom and Libre evaluations (MAE↓).

Context	Endpoint	GluFormer (tiny)	GluFormer (base)	CGMformer	CGM-JEPA	X-CGM-JEPA	GlucoFM
Frozen representation	Trajectory (mg/dL)	26.88	27.28	27.13	26.50	**26.19**	26.58
	Positive iAUC (mg/dL·h)	37.20	37.82	37.81	36.89	**35.78**	37.08
	Peak rise (mg/dL)	32.12	31.97	32.51	31.36	**31.24**	32.62
	Peak time (min)	32.13	33.70	32.28	32.27	32.19	**31.27**
+ 1 h pre-meal CGM	Trajectory (mg/dL)	25.61	26.01	26.17	24.82	25.14	**24.52**
	Positive iAUC (mg/dL·h)	35.34	36.10	36.30	**33.73**	34.09	34.01
	Peak rise (mg/dL)	30.75	30.79	31.20	**29.15**	29.96	30.28
	Peak time (min)	31.06	32.96	32.12	32.66	32.85	**30.14**
+ Meal nutrition	Trajectory (mg/dL)	24.45	25.33	24.75	23.42	23.00	**22.87**
	Positive iAUC (mg/dL·h)	33.13	34.83	33.54	31.11	**30.22**	30.76
	Peak rise (mg/dL)	28.66	29.80	28.81	27.04	**26.64**	27.14
	Peak time (min)	31.86	32.88	32.95	32.01	32.50	**30.55**
+ Fasting glucose	Trajectory (mg/dL)	24.14	25.27	24.27	22.80	22.80	**22.35**
	Positive iAUC (mg/dL·h)	32.64	34.69	32.76	30.36	30.13	**29.87**
	Peak rise (mg/dL)	28.30	29.80	28.08	**26.40**	26.58	26.47
	Peak time (min)	31.86	33.00	32.46	31.07	31.36	**30.12**
+ BMI + diabetes status	Trajectory (mg/dL)	24.23	25.32	24.15	22.90	23.10	**21.88**
	Positive iAUC (mg/dL·h)	32.56	34.68	32.17	30.24	30.56	**28.50**
	Peak rise (mg/dL)	28.10	29.70	27.93	25.97	26.59	**25.40**
	Peak time (min)	31.81	32.62	32.06	31.38	31.47	**30.31**

points ($[0.55, 4.14]$); all six Nadeau–Bengio-corrected CIs [21] exclude zero and remain significant after Holm adjustment (adjusted $p \leq 0.012$; Appendix E.1).

The gains are most evident on core metabolic phenotyping outcomes. In PR-AUC, GlucoFM leads all diabetes-risk and β-cell dysfunction evaluations, and 3/4 insulin-resistance evaluations, suggesting that its frozen representations capture clinically relevant glycemic structure rather than only generic temporal patterns. GlucoFM also outperforms CGMformer in PR-AUC on 11/14 evaluations despite CGMformer being pretrained on a larger external 964-participant cohort, and exceeds CGM-specific baselines re-pretrained on the same corpus on 11/14 evaluations. This suggests that the improvements are not solely explained by pretraining scale, but are consistent with the benefit of CGM-aware pretraining objectives and temporal inductive biases. The main exception is ShanghaiT2DM, which may reflect a device- and sampling-rate shift: it is Libre-derived with 15-minute sampling, while our pretraining corpus contains relatively fewer real Libre/15-minute recordings. Overall, GlucoFM provides the strongest frozen representation across the benchmark, with particularly clear advantages on clinically central metabolic tasks.

Postprandial Glycemic Response Forecasting. We next evaluate whether frozen CGM representations support prediction of glucose dynamics following a meal. We use paired Dexcom and Libre recordings from CGMacros, retaining 874 matched meal events per sensor from 34 participants with no additional logged meal during the two-hour prediction window. For each event, a strictly causal 24-hour pre-meal CGM window is encoded into a frozen representation, and the task predicts the subsequent two-hour glucose change relative to the observed meal-start glucose. Dexcom and Libre are modeled separately using the same MLP with two hidden layers at their native target cadences: 24 five-minute outputs for Dexcom and 8 fifteen-minute outputs for Libre. The representation is combined cumulatively with one hour of pre-meal CGM, meal nutrition, fasting glucose, and BMI and diabetes status. All encoders remain frozen and are evaluated with five subject-disjoint folds and 10 head initializations. We report MAE for the trajectory, positive incremental area under the curve (iAUC), peak rise, and peak time, computing each endpoint separately by sensor and then averaging Dexcom and Libre with equal weight.

Table 4 shows how each source of context changes PPGR performance. Once recent CGM is included, GlucoFM achieves the lowest trajectory error at every enriched stage, decreasing from 26.58 to 21.88 mg/dL across the full progression. With complete context, GlucoFM achieves the lowest error on all four endpoints for both Dexcom and Libre. In the equally weighted two-sensor mean, GlucoFM reduces trajectory, positive-iAUC, peak-rise, and peak-time MAE by 4.5%, 5.8%, 2.2%, and 3.4%, respectively, relative to CGM-JEPA, the strongest comparator across all four endpoints. These gains show that GlucoFM's frozen representation composes effectively with recent glucose, meal nutrition, and subject information. Sensor-stratified results are provided in Appendix E.2.

Table 5: Cross-dataset transfer results. "+" denotes GlucoFM's gain over the second-best method; "-" denotes its gap to the best baseline.

	Stanford → Hall				Hall → Stanford			
	Diabetes Risk Assessment		*Insulin Resistance*		*Diabetes Risk Assessment*		*Insulin Resistance*	
Metrics	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑
CGM-JEPA	53.0	73.4	55.0	67.9	68.1	64.3	61.3	61.8
X-CGM-JEPA	55.4	73.7	57.8	68.2	68.2	64.7	61.7	62.4
GluFormer (tiny)	58.6	70.4	50.8	62.2	67.6	65.6	59.7	58.3
GlucoFM	**61.6** (+3.0)	**74.7** (+1.0)	**61.6** (+3.8)	**72.1** (+3.9)	**73.4** (+5.2)	**69.7** (+4.1)	**67.7** (+6.0)	**69.2** (+6.8)
	CGMacros → Hall				Hall → CGMacros			
	Diabetes Risk Assessment		*Insulin Resistance*		*Diabetes Risk Assessment*		*Insulin Resistance*	
Metrics	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑
CGM-JEPA	50.6	69.6	53.4	67.1	87.5	78.1	89.1	77.0
X-CGM-JEPA	49.5	69.1	53.6	67.2	87.3	77.8	89.2	77.1
GluFormer (tiny)	55.9	70.2	50.0	65.1	83.9	74.1	84.8	71.0
GlucoFM	**63.3** (+7.4)	**78.2** (+8.0)	**61.7** (+8.1)	**72.5** (+5.3)	**88.8** (+1.3)	**81.3** (+3.2)	**90.0** (+0.8)	**78.8** (+1.7)
	CGMacros → Stanford				Stanford → CGMacros			
	Diabetes Risk Assessment		*Insulin Resistance*		*Diabetes Risk Assessment*		*Insulin Resistance*	
Metrics	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑	PR-AUC↑	ROC-AUC↑
CGM-JEPA	64.1	60.9	59.8	60.3	87.8	78.4	85.9	74.0
X-CGM-JEPA	65.4	62.3	59.7	60.2	**88.3**	78.5	83.6	70.7
GluFormer (tiny)	68.5	64.0	64.0	**65.2**	85.2	76.2	**87.9**	**76.3**
GlucoFM	**77.1** (+8.6)	**73.6** (+9.6)	**65.4** (+1.4)	64.8 (-0.4)	**88.3** (+0.5)	**79.9** (+1.4)	87.3 (-0.6)	73.3 (-3.0)

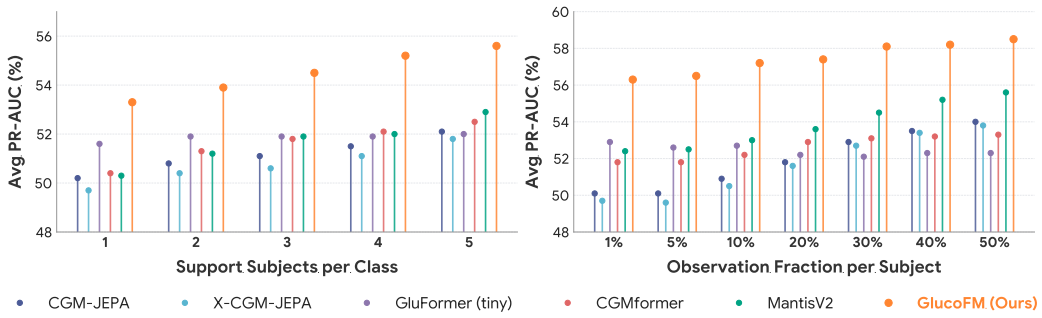


Figure 3: Few-shot adaptation under limited labeled subjects (left) and limited per-subject observations (right), reported as task-averaged PR-AUC.

Few-Shot Adaptation. We evaluate GlucoFM under two low-label adaptation settings. In *limited-subject adaptation*, we restrict the number of labeled support subjects per class within the training fold. In *limited-observation adaptation*, we retain all training subjects but use only a fraction of each subject's 24-hour windows for classifier training. We use 5-fold subject-grouped cross-validation with 10 repeated iterations and 5 random support samplings per split. Figure 3 shows that GlucoFM achieves the highest task-averaged PR-AUC at every support-subject count and observation fraction. Performance improves as additional support subjects are introduced, while the consistent lead with only one or two subjects per class demonstrates effective adaptation from very few labeled individuals. With all training subjects retained, GlucoFM also remains strongest across observation fractions from 1% to 50% and improves as additional observations are included, demonstrating effective adaptation even with sparse per-subject data. Together, these results establish GlucoFM as a label-efficient representation under both limited-subject and limited-observation supervision.

Cross-Dataset Generalization. We evaluate cohort transfer by training a linear probe on one source dataset and testing it directly on another using frozen representations. We focus on diabetes risk assessment and insulin resistance, which are shared across CGMacros, Stanford, and Hall. To isolate representation transferability, we compare with CGM-specific baselines pretrained on the same CGM corpus and use the target dataset only for final evaluation. As shown in Table 5, GlucoFM achieves the best overall transfer performance, ranking first on 21 of 24 PR-AUC and ROC-AUC evaluations. The gains are strongest for transfers involving Hall and for CGMacros→Stanford diabetes-risk assessment, indicating that GlucoFM captures metabolic structure that transfers across cohorts and devices. The

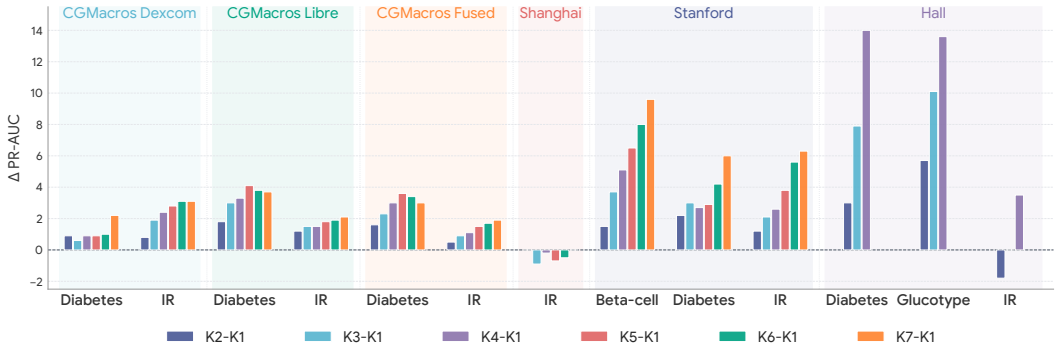


Figure 4: **Effect of multiday CGM observation.** Paired PR-AUC change from the one-day subject representation to K-day representations, $\Delta_K = \text{PR-AUC}(K) - \text{PR-AUC}(1)$. Positive values indicate improvement over $K = 1$. `Hall` uses $K \leq 4$; other cohorts use $K \leq 7$.

few non-leading cases mainly occur for Stanford→CGMacros insulin resistance, where GlucoFM remains close to the best baseline.

Multiday Representation Observation. We evaluate whether longer CGM observation improves subject-level prediction using frozen GlucoFM representations. For each subject, we select one fixed eligible $K_{\max}$-day CGM anchor episode and enumerate adjacent K-day subwindows within the same anchor for $K = 1, \ldots, K_{\max}$. Each day is independently encoded by the frozen GlucoFM encoder, and embeddings within each subwindow are mean-pooled into one subject-level representation. For each K and start position, every subject contributes one representation and each test subject receives one prediction, so subwindows from the same subject are not treated as independent test samples. Metrics are computed by start position and averaged within each repeated evaluation. We train linear probes with 10 iterations of 5-fold subject-level cross-validation and report the paired PR-AUC change. Figure 4 shows that longer CGM observation often improves subject-level prediction. The gains are strongest on Stanford and Hall, suggesting that multiple daily CGM profiles provide a more stable estimate of subject-level metabolic phenotypes. CGMacros further provides a paired-sensor setting, where Dexcom and Libre are recorded from the same subjects over the same period. We evaluate each sensor separately and include a matched fused setting, where same-subject same-day Dexcom and Libre embeddings are averaged before multiday aggregation. The mostly positive gains across these settings suggest that frozen GlucoFM embeddings capture subject-level signals that are reproducible across sensor views. Additional ShanghaiT2DM pooling comparisons are provided in Appendix E.5. Overall, these results show that GlucoFM can support multiday subject-level prediction, while the best aggregation strategy may depend on the cohort, sensor type, and clinical label.

Pretraining Scalability Analysis. We evaluate how GlucoFM scales with unlabeled data by sampling 20–100% of subjects from each pretraining cohort using five random seeds, while keeping the model architecture and downstream protocol fixed. As shown in Figure 5, average PR-AUC increases steadily as more pretraining data are used. Notably, GlucoFM trained with only 20% of the corpus already matches CGM-specific foundation model baselines trained on the full corpus, suggesting strong data efficiency, consistent with the inductive biases introduced by state–event modeling and JEPA-style pretraining. Performance continues to improve with additional unlabeled data, indicating that GlucoFM is not saturated at the current scale. The narrow shaded region shows that the trend is stable across subject subsampling seeds, suggesting that the improvements are not driven by a particular subset of pretraining subjects.

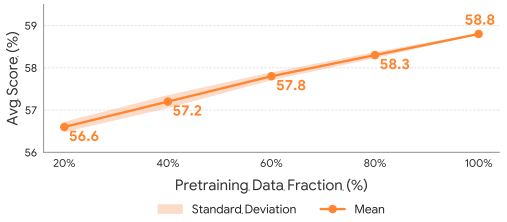


Figure 5: Pretraining data scaling analysis.

4 Ablation

We ablate key components of GlucoFM under identical pretraining and evaluation protocols. The main text reports task-averaged results, with protocols and additional architecture results in Appendix F.

Dual Stream vs. Single Stream. We compare a raw-input tokenizer, state-only and event-only single-stream encoders, and the proposed dual-stream state–event encoder. For fairness, all variants receive the corresponding patch-level statistics and temporal-difference features; they differ only in how glucose signals and auxiliary features are organized. As shown in Table 6, the dual-stream encoder achieves the strongest task-averaged performance across all three downstream metrics, supporting the value of organizing slow-varying glucose trends and short-term deviations in separate streams before fusion.

Table 6: Dual-stream ablation and probes.

Downstream utility (%)			
Encoder	**PR-AUC**	**ROC-AUC**	**Macro-F1**
Raw input	55.3	63.7	58.3
State only	55.6	63.7	57.9
Event only	51.8	60.1	55.6
Dual stream	**58.8**	**66.7**	**59.9**
Pre-fusion content probes (R^2)			
Target	**State**	**Event**	**Gap (95% CI)**
Glucose level	**0.962**	0.401	S−E: 0.561 [0.518, 0.630]
5-min abs. change	0.832	**0.888**	E−S: 0.056 [0.040, 0.073]

We further probe frozen pre-fusion state and event tokens on 5-min Dexcom recordings from CGMacros, enabling interpolation-free measurement of short-term change. State tokens retain substantially more information about hourly glucose level, whereas event tokens retain more information about 5-min absolute change. Paired subject-cluster bootstrap intervals exclude zero for both directional gaps. Together, the downstream ablation and content probes show that the two branches preserve distinct temporal emphasis and that their integration improves downstream representation quality. Protocol details are provided in Appendix F.2.

CGM-aware augmentation also improves task-averaged downstream performance across all metrics (Appendix F.4). Removing the temporal dynamics objective ($\lambda_{\mathrm{TD}} = 0$) reduces task-averaged performance across all metrics (Appendix F.3). Dense interpolation, with or without structural sparsification, underperforms the default mask-aware setting (Appendix F.5).

5 Discussion

GlucoFM provides reusable daily CGM representations for both metabolic phenotyping and event-conditioned PPGR. The pre-fusion probes and encoder ablations further connect this downstream utility to complementary slow-trend and short-term-deviation representations.

Limitations. Despite these findings, GlucoFM is limited by the scale and diversity of available CGM data. The pretraining population remains modest and may not fully capture demographic, device, lifestyle, and disease heterogeneity. Our evaluation is restricted to retrospective phenotype classification and meal-aligned PPGR prediction and does not assess longitudinal outcomes, treatment response, or prospective clinical deployment. Moreover, although multiday embedding aggregation improves subject-level prediction, the encoder itself still models 24-hour windows independently, leaving longer-range temporal modeling to simple downstream aggregation. GlucoFM is a research prototype, has not been cleared or approved by any regulatory authority, and is not intended to diagnose, treat, cure, or prevent any disease, nor should it be used as a substitute for professional medical advice.

Future Work. Future work should scale pretraining to larger and more diverse CGM cohorts, including broader device coverage and more real 15-minute recordings. Extending GlucoFM from post-hoc multiday aggregation to native multi-day context modeling may better capture longitudinal metabolic state, treatment response, and slower changes over weeks or months. Another important direction is real-time CGM modeling under evolving missingness, device changes, and behavioral context. The PPGR results also motivate free-living prediction under uncertain meal timing and evolving activity, insulin, sleep, and sensor context.

Conclusion. We introduced GlucoFM, a lightweight CGM foundation model that preserves observation patterns while modeling slow-varying trends and short-term deviations. Across multiple cohorts and seven phenotype-classification tasks, GlucoFM achieves strong frozen linear probing, low-label adaptation, cross-dataset transfer, and multiday subject-level prediction. The same frozen representation achieves the lowest full-context two-hour PPGR error across both sensors and all four response endpoints. These results suggest that CGM foundation models benefit from preserving sensing structure and explicitly modeling the multiscale nature of glucose dynamics. By learning transferable representations from unlabeled CGM, GlucoFM may reduce reliance on large labeled clinical cohorts for metabolic phenotyping; however, it should be validated within the intended population and clinical workflow before real-world use.

References

[1] Ahmed A. Metwally, Dalia Perelman, Heyjun Park, Yue Wu, Alokkumar Jha, Seth Sharp, Alessandra Celli, Ekrem Ayhan, Fahim Abbasi, Anna L. Gloyn, Tracey McLaughlin, and Michael P. Snyder. Prediction of metabolic subphenotypes of type 2 diabetes via continuous glucose monitoring and machine learning. *Nature Biomedical Engineering*, 9(8):1222–1239, Aug 2025.

[2] Ahmed A. Metwally, Heyjun Park, Yue Wu, Tracey McLaughlin, and Michael P. Snyder. Use of continuous glucose monitoring with machine learning to identify metabolic subphenotypes and inform precision lifestyle changes. *Journal of Diabetes Science and Technology*, 20(3), 2026.

[3] Yue Wu, Ben Ehlert, Ahmed A. Metwally, Dalia Perelman, Heyjun Park, Andrew Wallace Brooks, Fahim Abbasi, Basil Michael, Alessandra Celli, Caroline Bejikian, Ekrem Ayhan, Yingzhou Lu, Samuel M. Lancaster, Daniel Hornburg, Lucia Ramirez, David Bogumil, Sarah Pollock, Frank Wong, Denver Bradley, Georg Gutjahr, Ekanath Srihari Rangan, Tao Wang, Lettie McGuire, P. Venkat Rangan, Helge Ræder, Zohar Shipony, Doron Lipson, Tracey McLaughlin, and Michael P. Snyder. Individual variations in glycemic responses to carbohydrates and underlying metabolic physiology. *Nature Medicine*, 31(7):2232–2243, 2025.

[4] Heyjun Park, Ahmed A. Metwally, Alireza Delfarah, Yue Wu, Dalia Perelman, Caleb Mayer, Curtis McGinity, Majid Rodgar, Alessandra Celli, Tracey McLaughlin, Emmanuel Mignot, and Michael Snyder. High-resolution lifestyle profiling and metabolic subphenotypes of type 2 diabetes. *npj Digital Medicine*, 8(1):352, 2025.

[5] David C Klonoff, Richard M Bergenstal, Eda Cengiz, Mark A Clements, Daniel Espes, Juan Espinoza, David Kerr, Boris Kovatchev, David M Maahs, Julia K Mader, Nestoras Mathioudakis, Ahmed A Metwally, Shahid N Shah, Bin Sheng, Michael P Snyder, Guillermo Umpierrez, Mandy M Shao, Agatha F Scheideman, Alessandra T Ayers, Cindy N Ho, and Elizabeth Healey. Continuous glucose monitoring data analysis 2.0: functional data pattern recognition and artificial intelligence applications. *Journal of Diabetes Science and Technology*, 19(6):1515–1527, 2025.

[6] Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS biology*, 16(7):e2005143, 2018.

[7] Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R. Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, Eli Meirom, Eric P. Xing, Gal Chechik, Hagai Rossman, and Eran Segal. A foundation model for continuous glucose monitoring data. *Nature*, 650(8103):978–986, 2026.

[8] Junjie Luo, Abhimanyu Kumbara, Mansur Shomali, Rui Han, Anand Iyer, Grazia Aleppo, Ritu Agarwal, and Gordon Gao. A large sensor foundation model pretrained on continuous glucose monitor data for diabetes management. *npj Health Systems*, 2025.

[9] Yurun Lu, Dan Liu, Zhongming Liang, Rui Liu, Pei Chen, Yitong Liu, Jiachen Li, Zhanying Feng, Lei M. Li, Bin Sheng, Weiping Jia, Luonan Chen, Huating Li, and Yong Wang. A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data. *National Science Review*, 12(5):nwaf039, May 2025.

[10] Hada Melino Muhammad, Zechen Li, Flora Salim, and Ahmed A. Metwally. CGM-JEPA: Learning consistent continuous glucose monitor representations via predictive self-supervised pretraining. *arXiv preprint arXiv:2605.00933*, 2026.

[11] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

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

[13] Vasilii Feofanov, Marius Alonso, Songkang Wen, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

[14] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[15] Qinpei Zhao, Jinhao Zhu, Xuan Shen, Chuwen Lin, Yinjia Zhang, Yuxiang Liang, Baige Cao, Jiangfeng Li, Xiang Liu, Weixiong Rao, and Congrong Wang. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10, 2023.

[16] Peter Cho, Juseong Kim, Brinnae Bent, and Jessilyn Dunn. BIG IDEAs Lab Glycemic Variability and Wearable Device Data. *PhysioNet*, September 2023. Version 1.1.2.

[17] Ana Colás, Luis Vigil, Borja Vargas, David Cuesta-Frau, and Manuel Varela. Detrended fluctuation analysis in the prediction of type 2 diabetes mellitus in patients at risk: Model optimization and comparison with other metrics. *PLOS ONE*, 14(12):e0225817, 2019.

[18] Anurag Das, David Kerr, Namino Glantz, Wendy Bevier, Rony Santiago, Ricardo Gutierrez-Osuna, and Bobak J. Mortazavi. CGMacros: a pilot scientific dataset for personalized nutrition and diet monitoring. *Scientific Data*, 12:1557, 2025.

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

[20] Vasilii Feofanov, Songkang Wen, Jianfeng Zhang, Lujia Pan, and Ievgen Redko. MantisV2: Closing the zero-shot gap in time series classification with synthetic data and test-time strategies. *arXiv preprint arXiv:2602.17868*, 2026.

[21] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[22] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

[23] Dani Kiyasseh, Tingting Zhu, and David A Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[24] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. SimPer: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations (ICLR)*, 2023.

[25] Zhihao Yu, Xu Chu, Yujie Jin, Yasha Wang, and Junfeng Zhao. SMART: Towards pre-trained missing-aware model for patient health status prediction. In *NeurIPS 2024*, 2024.

[26] Yamin Li, Ruidong Zhang, Anran Wang, Eduardo Fonseca, Dan Ellis, Zechen Li, Simon A. Lee, Tao Pang, Jake Garrison, Maxwell A. Xu, Yicheng Wang, Jitian Zhang, Weizhi Zhang, Haichao Zhang, Catie Chang, Xin Liu, Anupam Pathak, Daniel McDuff, Shwetak Patel, Jake Sunshine, Yuzhe Yang, and Xiaoran Fan. AudioFM: Towards general purpose audio representation learning. *arXiv preprint arXiv:2610.34183*, 2026.

[27] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[28] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2024.

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

[30] Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[31] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

[32] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021.

[33] Isotta Landi, Benjamin S. Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T. Dudley, Cesare Furlanello, and Riccardo Miotto. Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digital Medicine*, 3:96, 2020.

[34] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

[35] Zilin Jing, Vincent Jeanselme, Yuta Kobayashi, Simon A Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xinzhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured EHR foundation models. *arXiv preprint arXiv:2602.00541*, 2026.

[36] Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S Aysola, Rajesh Kumar, and Yuzhe Yang. SleepLM: Natural-language intelligence for human sleep. *arXiv preprint arXiv:2602.23605*, 2026.

[37] Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On pre-training and scaling of sleep foundation models. *arXiv preprint arXiv:2603.00190*, 2026.

[38] 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. *arXiv preprint arXiv:2506.05321*, 2025.

[39] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[41] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel

McDuff. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Simon A. Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md Sazzad Hissain Khan, Keum San Chun, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, Viswam Nathan, Md Mahbubur Rahman, Li Zhu, Subramaniam Venkatraman, and Sharanya Arcot Desai. Hi-MAE: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *The Fourteenth International Conference on Learning Representations*, 2026.

[43] Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. *arXiv preprint arXiv:2605.00973*, 2026.

[44] Zechen Li, Baiyu Chen, Hao Xue, and Flora D. Salim. ZARA: Training-free motion time-series reasoning via evidence-grounded LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

[45] Baiyu Chen, Zechen Li, Wilson Wongso, Lihuan Li, Xiachong Lin, Hao Xue, Benjamin Tag, and Flora Salim. AnyMo: Geometry-aware setup-agnostic modeling of human motion in the wild. *arXiv preprint arXiv:2605.22715*, 2026.

[46] Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. COMODO: Cross-modal video-to-IMU distillation for efficient egocentric human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2026.

[47] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 69–78, New York, NY, USA, 2021. Association for Computing Machinery.

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

[49] Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. HEARTS: Benchmarking LLM reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.

[50] Julie Wagner, Howard Tennen, and Howard Wolpert. Continuous glucose monitoring: a review for behavioral researchers. *Psychosomatic Medicine*, 74(4):356–365, 2012.

[51] Ahmed A. Metwally, A. Ali Heydari, Daniel McDuff, Alexandru Solot, Zeinab Esmaeilpour, Anthony Z. Faranesh, Menglian Zhou, Girish Narayanswamy, Maxwell A. Xu, Xin Liu, Yuzhe Yang, David B. Savage, Mark Malhotra, Conor Heneghan, Shwetak Patel, Cathy Speed, and Javier L. Prieto. Insulin resistance prediction from wearables and routine blood biomarkers. *Nature*, 652:451–461, 2026.

[52] Marcos Matabuena, Rahul Ghosal, Javier Enrique Aguilar, Ayya Keshet, Robert Wagner, Carmen Fernández Merino, Juan Sánchez Castro, Vadim Zipunnikov, Jukka-Pekka Onnela, and Francisco Gude. Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics. *Scientific Reports*, 2025.

[53] David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, Jotham Suez, Jemal Ali Mahdi, Elad Matot, Gal Malka, Noa Kosower, Michal Rein, Gili Zilberman-Schapira, Lenka Dohnalová, Meirav Pevsner-Fischer, Rony Bikovsky, Zamir Halpern, Eran Elinav, and Eran Segal. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, 2015.

[54] Sarah E. Berry, Ana M. Valdes, David A. Drew, Francesco Asnicar, Mohsen Mazidi, Jonathan Wolf, Joan Capdevila, George Hadjigeorgiou, Richard Davies, Haya Al Khatib, Christopher Bonnett, Sajaysurya Ganesh, Elco Bakker, Deborah Hart, Massimo Mangino, Jordi Merino, Inbar Linenberg, Patrick Wyatt, Jose M. Ordovas, Christopher D. Gardner, Linda M. Delahanty, Andrew T. Chan, Nicola Segata, Paul W. Franks, and Tim D. Spector. Human postprandial responses to food and potential for precision nutrition. *Nature Medicine*, 26(6):964–973, 2020.

[55] Ayya Keshet, Smadar Shilo, Anastasia Godneva, Yeela Talmor-Barkan, Yaron Aviv, Eran Segal, and Hagai Rossman. CGMap: characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell metabolism*, 35(5):758–769.e3, 2023.

[56] Hikaru Sugimoto, Gal Sapir, Ayya Keshet, and Shinya Kuroda. Use of continuous glucose monitoring to stratify individuals without diabetes. *Communications Medicine*, 2026.

[57] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[58] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[59] Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

[60] Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

[61] Richard M Bergenstal, Roy W Beck, Kelly L Close, George Grunberger, David B Sacks, Aaron Kowalski, Adam S Brown, Lutz Heinemann, Grazia Aleppo, Donna B Ryan, Tonya D Riddlesworth, and William T Cefalu. Glucose management indicator (GMI): A new term for estimating A1C from continuous glucose monitoring. *Diabetes Care*, 41(11):2275–2280, Nov 2018.

[62] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2020.

[63] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W. H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 2019.

[64] F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 2019.

A Related Work

Self-Supervised and Foundation Models for Time Series. Self-supervised time-series learning commonly uses contrastive learning, masked modeling, forecasting, or latent prediction [22–26]. Recent time-series foundation models scale these ideas to heterogeneous corpora, including masked encoders such as MOMENT [12], tokenized or decoder-only forecasters such as Chronos-2 [11, 19] and TimesFM [27], probabilistic forecasters such as Lag-Llama [28], and multi-task models such as Moirai [29] and UniTS [30]. While these models provide strong general-purpose backbones, they typically treat time series as generic numerical sequences and are optimized for forecasting, reconstruction, or broad task transfer. GlucoFM instead targets CGM-specific daily representation learning, where temporal alignment, sensing irregularity, and glycemic structure are central.

Clinical and Physiological Representation Learning. Clinical representation learning has produced pretrained models for longitudinal EHRs [31–35], sleep physiology [36, 37], wearable sensing [38–45], and broader physiological signals [46–49]. These works demonstrate the value of domain-specific pretraining under label scarcity, patient heterogeneity, and noisy real-world measurement. However, CGM has distinct temporal structure: glucose traces combine slow basal regulation, circadian variation, and short-term excursions potentially influenced by meals, insulin, stress, activity, or sensor artifacts over hours to days [50]. GlucoFM focuses on this glucose-specific structure by learning daily representations directly from CGM rather than relying on generic EHR, sleep, cardiovascular, or wearable encoders.

CGM Representation Learning for Metabolic Phenotyping. CGM has been used to study metabolic heterogeneity through postprandial response prediction, glucotype discovery, cohort profiling, and prediction of insulin resistance, β-cell dysfunction, and metabolic risk [1, 5, 6, 51–56]. Much of this work relies on handcrafted metrics, clustering, supervised predictors, or study-specific protocols, limiting reusable representation learning across cohorts and tasks. Recent CGM foundation models, such as CGMformer [9], CGM-LSM [8], CGM-JEPA [10], and GluFormer [7], advance masked modeling, autoregressive forecasting, finetuning, or cohort-level prediction. In contrast, GlucoFM learns a compact frozen CGM encoder for subject-disjoint clinical prediction, cross-dataset transfer, and low-label adaptation.

Missingness-Aware CGM Modeling. Missingness is central in clinical time-series modeling and is commonly addressed through imputation, mask-aware architectures, or continuous-time methods [57–60]. In CGM, gaps can arise from sensor warm-up, removal, dropout, calibration issues, or nonwear, and naive imputation may obscure meaningful glucose dynamics. GlucoFM therefore uses a mask-aware daily grid that preserves chronological structure while distinguishing observed measurements from missing positions, supporting representation learning under real-world CGM irregularity.

B Datasets

B.1 CGM Segmentation and Missingness Handling

Before extracting 24-hour windows, we split each subject's raw CGM trace into continuous recording segments based on timestamp gaps. Gaps of at most 1 hour are treated as short interruptions within the same segment. After aligning the trace to the target 5-minute grid, these missing positions are retained as unobserved entries with NaN glucose values and mask value 0, rather than being treated as real measurements. In contrast, gaps longer than 1 hour are treated as segment boundaries, and the trace is split into separate segments.

As a result, each 24-hour training window is extracted from a continuous recording segment and does not contain any single real-data gap longer than 1 hour. This preserves realistic short missingness while avoiding windows that span long non-wear or sensor-disconnection periods. For fair comparison, all applicable baselines use the same pretraining inputs and identical downstream fold splits as GlucoFM.

B.2 Pretraining Datasets

To increase the diversity of unlabeled pretraining data, we extract 24-hour windows from each continuous CGM segment using a fixed-seed random overlapping strategy. For each segment, we randomly sample overlapping 24-hour windows with a coverage ratio between 20% and 80%, rather than using only fixed non-overlapping windows. This exposes the model to diverse daily start times

and temporal contexts while keeping sampling reproducible. Since pretraining is self-supervised and label-free, overlapping windows improve temporal coverage without introducing label leakage; downstream evaluation instead uses non-overlapping windows with subject-grouped splits.

Wear-CGM. The Wear-CGM dataset is a non-public multimodal CGM dataset collected as part of Google/Fitbit research studies across two sequential research phases with non-overlapping participants. The protocols for the Wear-CGM datasets were approved by Advarra (Institutional Review Board (IRB) no. Pro00059582 and Pro00069880). All participants provided written informed consent prior to data collection, which included permission for the use of de-identified data in secondary research and algorithm development. It contains CGM recordings from healthy, non-diabetic adults in the United States, collected using Dexcom G6 Pro sensors at 5-minute intervals. Phase 1 included 105 participants in an observational protocol lasting approximately 4 weeks, with 2–3 sensors used per participant. Phase 2 included 87 participants over up to 15 days and incorporated standardized meal challenges designed to elicit metabolic responses. Phase 2 also collected clinical measurements, including fasting glucose and insulin, comprehensive metabolic panels, lipid panels, HbA1c, C-reactive protein, and up to two 2-hour oral glucose tolerance tests after overnight fasting. Together, the two phases provide 75,330 hours of CGM recordings. Although the dataset also contains wearable, nutrition, and blood-pressure measurements, only CGM data are used in this work.

BIG IDEAs. The BIG IDEAs dataset [16] contains CGM recordings from 16 subjects, collected with Dexcom sensors at 5-minute intervals. After preprocessing, it provides 3,017 monitoring hours, corresponding to an average duration of approximately 7.9 days per subject.

Colas. The Colas dataset [17] contains CGM recordings from 206 subjects after preprocessing, collected with iPro sensors at 5-minute intervals. It provides 9,544 monitoring hours, corresponding to an average duration of approximately 1.9 days per subject.

Stanford. The Stanford dataset [1] contains CGM recordings from 56 subjects, collected with Dexcom sensors at 5-minute intervals. Among them, 19 subjects lack the clinical metadata required for downstream label construction and are therefore used only for unlabeled pretraining. This pretraining subset contains 8,761 monitoring hours, corresponding to an average duration of approximately 19.2 days per subject. The remaining 37 subjects have complete clinical profiles and are used for downstream evaluation.

ShanghaiT2DM. The ShanghaiT2DM dataset [15] contains CGM recordings from patients with type 2 diabetes, collected with FreeStyle Libre sensors at 15-minute intervals. Most participants have one recording session, while seven participants provide 2–3 separate sessions. Because clinical measurements may vary across visits separated by long intervals, we follow the baseline protocol and treat each recording visit as a distinct subject-entry. Among these subject-entries, 44 lack sufficient clinical metadata for downstream label construction and are used only for unlabeled pretraining. An identity audit found one biological participant with an earlier unlabeled visit in this pretraining subset and a later labeled visit in the downstream subset. The ShanghaiT2DM pretraining/downstream separation is therefore defined at the subject-entry (recording-visit) level. This pretraining subset contains 12,414 monitoring hours, corresponding to an average duration of approximately 11.8 days per subject-entry.

B.3 Downstream Datasets

The clinical thresholds below are used to define downstream prediction labels consistently across cohorts and are not intended as standalone diagnostic criteria. For all downstream datasets, CGM traces are divided into non-overlapping 1-day segments, which avoids duplicated temporal evidence across samples and reduces leakage or over-counting from highly correlated overlapping windows.

CGMacros. The CGMacros dataset [18] contains multimodal physiological data from 45 subjects, each wearing both Dexcom and FreeStyle Libre sensors for approximately 10 days. The Dexcom and Libre recordings provide 10,376 and 10,998 monitoring hours, respectively. To avoid subject leakage, recordings from both sensor brands for the same subject are always assigned to the same training or testing split. We define four downstream tasks: (1) diabetes risk, with three classes: normoglycemic, prediabetes, and type 2 diabetes; (2) insulin resistance, where a subject is labeled positive if $\text{HOMA-IR} = (\text{Insulin}_{\mu U/mL} \times \text{Fasting Glucose}_{mg/dL})/405.0 > 2.9$; (3) obesity, defined as $\text{BMI} = \text{Weight}_{kg}/(\text{Height}_{m})^2 \geq 30$; and (4) hyperlipidemia, where a positive label is assigned if total cholesterol ≥ 240, LDL ≥ 160, or triglycerides ≥ 200 mg/dL.

Table 7: Label distributions for downstream clinical prediction tasks across four cohorts; `ShanghaiT2DM` counts denote labeled recording sessions.

`CGMacros` **(N=45)**

Task	Class	Count
Diabetes Risk	Normoglycemic	16
	Prediabetes	14
	Type 2 Diabetes	15
Insulin Resistance	Sensitive	13
	Resistant	32
Obesity	Non-obese	22
	Obese	23
Hyperlipidemia	Normal	33
	Hyperlipidemia	12

`Hall` **(N=56)**

Task	Class	Count
Diabetes	Normoglycemic	37
	Abnormal	19
Glucotype	Normal	34
	Severe	22
Insulin Resistance	Sensitive	35
	Resistant	21
Hyperlipidemia	Normal	48
	Hyperlipidemia	8

`Stanford` **(N=37)**

Task	Class	Count
Diabetes	Normoglycemic	17
	Abnormal	20
β-cell Dysfunction	Normal	17
	Dysfunction	20
Insulin Resistance	Sensitive	17
	Resistant	20

`ShanghaiT2DM` **(N=65)**

Task	Class	Count
Hypoglycemia	No	56
	Yes	9
Insulin Resistance	Sensitive	24
	Resistant	41
Hyperlipidemia	Normal	45
	Hyperlipidemia	20

Hall. The Hall dataset [6] originally contains 57 subjects. We retain 56 subjects with sufficient clinical information for constructing diabetes risk, insulin resistance, hyperlipidemia, and glucotype labels. The processed CGM recordings are collected with Dexcom sensors at 5-minute intervals and contain 7,090 monitoring hours, with an average duration of approximately 5.3 days per subject. We define four binary downstream tasks: (1) diabetes risk, where prediabetes or diabetes is grouped as abnormal glucose regulation and compared against normoglycemic status; (2) glucotype, distinguishing severe glucose fluctuation from normal profiles, where low and moderate glucotype categories are grouped as non-severe; (3) insulin resistance, determined by steady-state plasma glucose (SSPG), where a subject is labeled positive if $\mathrm{SSPG} > 120$, or, when SSPG is unavailable, by $\text{HOMA-IR} = (\mathrm{Insulin}_{\mu U/mL} \times \mathrm{FBG}_{mg/dL})/405.0 > 2.9$; and (4) hyperlipidemia, where a positive label is assigned if total cholesterol ≥ 240, LDL ≥ 160, or triglycerides ≥ 200 mg/dL.

Stanford. The downstream Stanford subset [1] contains 37 subjects with complete clinical profiles, providing 27,571 monitoring hours and an average duration of approximately 31 days per subject. We define three binary downstream tasks: (1) insulin resistance, mapping insulin-sensitive subjects to 0 and insulin-resistant subjects to 1 based on SSPG-derived classes; (2) β-cell dysfunction, where subjects are categorized as dysfunction versus normal based on the median disposition index (DI); and (3) diabetes risk, where subjects are labeled as abnormal glucose regulation if $\mathrm{HbA1c} \geq 5.7\%$ and normoglycemic otherwise.

ShanghaiT2DM. The downstream ShanghaiT2DM subset [15] contains 65 labeled sessions from 58 biological participants, providing 15,634 monitoring hours and an average duration of approximately 10 days per session. Each session is evaluated as one subject-entry, and all 24-hour windows from that entry remain in the same cross-validation fold. We define three binary downstream tasks: (1) hypoglycemia, mapped directly from clinical records, where "yes" indicates the positive class and "no" indicates the negative class; (2) insulin resistance, computed using HOMA-IR after converting fasting insulin from pmol/L to μU/mL by dividing by 6.945, where a session is labeled positive if $\text{HOMA-IR} = (\mathrm{Insulin}_{\mu U/mL} \times \mathrm{Glucose}_{mg/dL})/405.0 > 2.9$; and (3) hyperlipidemia, where lipid profiles are converted from mmol/L to mg/dL and a positive label is assigned if total cholesterol ≥ 240 ($\mathrm{TC}_{mmol/L} \times 38.67$), LDL ≥ 160 ($\mathrm{LDL}_{mmol/L} \times 38.67$), or triglycerides ≥ 200 ($\mathrm{TG}_{mmol/L} \times 88.57$).

C Baselines

We compare GlucoFM with three groups of baselines: general-purpose time-series foundation models, an open-weight CGM-specific foundation model pretrained on an external cohort, and CGM-specific foundation models retrained on our pretraining corpus. We use author-released checkpoints for the general-purpose models and CGMformer. For CGM-JEPA, X-CGM-JEPA, and GluFormer, official pretrained checkpoints were unavailable at the time of evaluation, so we reproduced the published preprocessing, architecture, and objectives where applicable and retrained the models on our pretraining corpus. GluFormer-base follows the published model scale, whereas GluFormer-tiny is a parameter-matched variant introduced in this study. For all representation-based evaluations, encoders are frozen and the same downstream logistic regression classifier is trained unless otherwise stated.

C.1 General-purpose Time-series Foundation Models

Chronos-2. Chronos-2 [19] is a general-purpose time-series forecasting foundation model for zero-shot univariate, multivariate, and covariate-informed forecasting. It extends Chronos with group attention to share information across related series, variables, targets, and covariates. We evaluate both Chronos-2 and Chronos-2-small as general-purpose foundation model baselines for frozen representation extraction. To obtain sequence-level embeddings, we feed each CGM sequence into the model and extract the output hidden states. We aggregate valid hidden states with mean pooling while excluding the EOS token, which performed best in our preliminary pooling comparison.

MOMENT. MOMENT [12] is an open time-series foundation model pretrained on the Time Series Pile with masked time-series modeling. Following the official pipeline, we remove null values, align each sequence to 288 points using linear interpolation, and extract frozen representations for downstream evaluation. We include MOMENT to assess whether large-scale general-purpose time-series pretraining transfers to CGM-based clinical prediction.

Mantis and MantisV2. Mantis [13] is a lightweight time-series foundation model for classification, using a Vision Transformer-style architecture and contrastive pretraining. MantisV2 [20] improves Mantis for zero-shot time-series classification through stronger synthetic-data pretraining, architectural refinements, and test-time representation strategies. Following the official preprocessing for both models, we linearly interpolate each input sequence to 512 points and extract frozen encoder representations for downstream evaluation.

C.2 Externally Pretrained CGM Foundation Model

CGMformer. CGMformer [9] is a CGM-specific Transformer pretrained with masked learning on daily CGM profiles. The released checkpoint is pretrained on an external cohort of 1,917 CGM-days from 964 participants and evaluated in the original work on tasks such as diabetes screening, metabolic subtyping, and dietary recommendation. We use it as an externally pretrained CGM-specific baseline. Following the official preprocessing scheme, each CGM window is represented on a 288-point 5-minute grid; missing bins are mapped to CGMformer's `<pad>` token; glucose values are discretized using the released vocabulary; and a `<cls>` token is prepended to form a 289-token input. We keep the checkpoint frozen and use mean-pooled hidden states as window-level embeddings, following the official embedding extraction strategy; in our implementation, the mean is weighted by the attention mask to exclude `<pad>` positions.

C.3 CGM Foundation Models Retrained on Our Pretraining Corpus

CGM-JEPA. CGM-JEPA [10] is a CGM-specific predictive self-supervised model that predicts masked latent targets instead of reconstructing raw glucose values. We retrain it on the same unlabeled CGM corpus as GlucoFM using the official configuration: inputs are aligned to a 5-minute grid, linearly interpolated to 288-point daily sequences, normalized, and encoded by a 3-layer Transformer with hidden dimension 96 and 6 attention heads. It is trained with batch size 128, learning rate 10^{-4}, mask ratio 0.25, and 101 epochs. Downstream embeddings are extracted from the frozen encoder before the projection head and mean-pooled over sequence tokens.

X-CGM-JEPA. X-CGM-JEPA [10] uses the same encoder, preprocessing, and optimization configuration as CGM-JEPA, but adds an auxiliary Glucodensity-based cross-view objective. Following the official setting, Glucodensity views are constructed as KDE-based 2D joint density grids over three variable pairs: glucose–speed, glucose–acceleration, and speed–acceleration, where speed and acceleration denote first- and second-order glucose differences. Each density grid has size 32×32, is divided into non-overlapping spatial patches, and is partially masked during pretraining. We retrain X-CGM-JEPA on the same corpus as GlucoFM and extract downstream embeddings from the frozen encoder using the same mean-pooling protocol as CGM-JEPA.

GluFormer. GluFormer [7] is a generative CGM foundation model trained with self-supervised autoregressive glucose-token prediction. We retrain two variants on the same pretraining corpus as GlucoFM. GluFormer (base) follows the original setting: 5-minute alignment, imputation, clipping to $[40, 500]$, discretization into shifted glucose tokens, and a 16-layer Transformer with hidden dimension 1024, 16 attention heads, and feed-forward dimension 2048, trained with Adam at learning rate 5×10^{-5} for 76 epochs. GluFormer (tiny) uses the same preprocessing and objective but a smaller 4-layer Transformer with hidden dimension 128, 4 attention heads, and feed-forward dimension 256, chosen to provide a parameter scale comparable to GlucoFM and CGM-JEPA. It is trained with Adam at learning rate 10^{-4}, batch size 128, for 100 epochs. Downstream embeddings are extracted from frozen hidden states over valid non-padding tokens using max pooling by default.

D Additional Implementation Details

D.1 CGM Window Construction and Chronological Grid Alignment

All inputs to GlucoFM are pre-segmented 24-hour CGM windows. For datasets originally containing multi-day recordings, the segmentation strategy is described in Appendix B. Given a 24-hour segment with readings $X = \{x_i\}_{i=1}^{N}$ and timestamps $T = \{t_i\}_{i=1}^{N}$, we align it to a fixed 24-hour grid with 5-minute resolution, yielding $L = 288$ grid positions. The first timestamp t_1 defines the circadian start index

$$s = \left\lfloor \frac{60 \cdot \mathrm{hour}(t_1) + \mathrm{minute}(t_1)}{5} \right\rfloor, \tag{6}$$

and the absolute time-of-day index of grid position j is $a_j = (s + j) \bmod L$.

Each reading is assigned to a grid index based on its elapsed time from the window start:

$$u_i = \frac{t_i - t_1}{\Delta t}, \qquad j_i = \mathcal{B}(u_i), \tag{7}$$

where $\Delta t = 5$ minutes and $\mathcal{B}(\cdot)$ denotes the binning rule used during preprocessing. We use floor-based assignment for most datasets and nearest-index rounding only when it better matches the dataset's timestamp convention. Readings outside the valid range $[0, L-1]$ are excluded, and multiple readings assigned to the same grid position are averaged.

The aligned glucose sequence is denoted as $\hat{X} \in \mathbb{R}^L$, and the physical observation mask $M \in \{0,1\}^L$ indicates whether each grid position contains at least one real CGM reading. Missing positions are filled only for tensor construction; they are distinguished by the observation mask and are never treated as observed measurements. Thus, chronological grid alignment is separated from interpolation or imputation.

D.2 Mask-aware Physiological Statistics

All feature construction is mask-aware: unobserved grid positions do not contribute to statistics or loss terms. GlucoFM divides each 24-hour window into $P = 24$ non-overlapping one-hour patches, each containing $K = 12$ grid positions. For patch $\mathcal{P}_i$, the physical observation support is

$$d_i = \frac{1}{K} \sum_{j \in \mathcal{P}_i} M_j. \tag{8}$$

This patch density is reused in mask-aware loss weighting.

For the state stream, GlucoFM computes patch-level glucose statistics using only observed entries:

$$\mu_i = \frac{\sum_{j\in\mathcal{P}_i} M_j \hat{X}_j}{\sum_{j\in\mathcal{P}_i} M_j + \epsilon}, \qquad \sigma_i = \left(\frac{\sum_{j\in\mathcal{P}_i} M_j (\hat{X}_j - \mu_i)^2}{\sum_{j\in\mathcal{P}_i} M_j + \epsilon} \right)^{1/2}. \tag{9}$$

Empty patches are zeroed by the validity mask.

For the event stream, rate-of-change features are computed only from valid observed positions. The implementation searches backward up to 9 grid steps and uses the closest previous observed value:

$$r_j = \frac{\hat{X}_j - \hat{X}_{j-b}}{b}, \tag{10}$$

where j and $j-b$ are both observed and no closer valid pair exists. If no valid previous observation is found, the rate-of-change entry is set to zero and marked invalid. Patch-level event mean and standard deviation are then computed from valid rate-of-change entries only.

D.3 Causal Gaussian State–Event Decoupling

GlucoFM separates each aligned CGM sequence into a low-frequency state component and a high-frequency event component using a causal, mask-aware Gaussian filter. The filter is applied after mask-aware normalization. Let $\tilde{X}$ denote the normalized aligned glucose sequence and let M denote the corresponding observation mask used by the branch. For each grid position j, the state component is estimated from current and past observed values:

$$\tilde{X}^{\text{state}}_j = \frac{\sum_{r=0}^{R} K_\sigma(r)\, M_{j-r}\, \tilde{X}_{j-r}}{\sum_{r=0}^{R} K_\sigma(r)\, M_{j-r} + \epsilon}, \tag{11}$$

where invalid indices are ignored. The one-sided Gaussian kernel is

$$K_\sigma(r) = \frac{\exp\left(-r^2/(2\sigma^2)\right)}{\sum_{u=0}^{R} \exp\left(-u^2/(2\sigma^2)\right)}, \qquad r = 0, \ldots, R. \tag{12}$$

Using only $r \geq 0$ makes the filter causal, so future glucose values are not used to estimate the current state. In the implementation, the maximum lag is determined by the maximum allowed bandwidth and a truncation factor of 3, giving $R = \lceil 3\sigma_{\max} \rceil = 36$ grid steps.

The bandwidth σ is learnable and constrained as

$$\sigma = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot \text{sigmoid}(\rho), \tag{13}$$

where ρ is an unconstrained learnable parameter, $\sigma_{\min} = 2$, and $\sigma_{\max} = 12$. On the 5-minute grid, this corresponds to a learnable Gaussian scale of approximately 10–60 minutes, initialized at $\sigma = 6$.

The event component is defined as the observed residual after removing the state trend:

$$\tilde{X}^{\text{event}} = (\tilde{X} - \tilde{X}^{\text{state}}) \odot M. \tag{14}$$

The state stream therefore receives smoothed baseline dynamics, while the event stream receives short-term deviations supported by real observations.

D.4 Architecture and Dimensionality

Following the above patching scheme, the state stream encodes each patch using a 64-dimensional waveform feature, a 16-dimensional intra-patch trend-difference feature, and a 48-dimensional projected statistics feature from patch mean and standard deviation. The event stream encodes each patch using a 48-dimensional residual waveform feature, a 48-dimensional rate-of-change feature, and a 32-dimensional projected statistics feature from valid rate-of-change mean and standard deviation. These components are projected into 64-dimensional state and event tokens, respectively.

The state and event tokens are fused into a unified physiological patch token of dimension $D = 128$. Circular time-of-day features from the absolute grid index are projected to the same dimension and combined with patch positional embeddings through a learnable gate. The context and EMA target encoders share a 3-layer Transformer architecture with 4 attention heads and feed-forward dimension 256. The masked-context predictor is a lightweight 1-layer Transformer, and the temporal dynamics objective uses two lightweight transition heads. During downstream evaluation, only the frozen online branch is retained. GlucoFM has 0.72M trainable parameters and 1.18M total parameters during pretraining, mainly due to the additional EMA target branch.

D.5 Pretraining Configuration

GlucoFM is pretrained for 120 epochs on a single NVIDIA H100 GPU with batch size 128, learning rate 10^{-4}, weight decay 10^{-2}, and a separate learning rate of 10^{-3} for the learnable Gaussian bandwidth.

D.6 Pretraining Masking and Transition Loss

During masked contextual representation learning, GlucoFM applies random patch-level masking to the online branch, with the masking ratio sampled uniformly from $[0.5, 0.6]$. Selected patches are hidden from the visible signal used for mask-aware statistics and filtering, and their patch tokens are replaced by a learnable mask token before the context encoder. The EMA target branch uses the full physical observation mask and encodes the complete sequence to provide latent targets. The masked contextual loss is applied only at masked patch positions and is weighted by each patch's physical observation density:

$$w_i = d_i = \frac{1}{K} \sum_{j \in \mathcal{P}_i} M_j. \tag{15}$$

For temporal dynamics modeling, GlucoFM predicts next-patch state and event targets from the current online state and event tokens before Transformer self-attention. Let $m_i^{\text{mask}} \in \{0, 1\}$ indicate whether patch i is masked, and let d_i denote its physical observation density. The transition weight is

$$q_i = (1 - m_i^{\text{mask}})\, d_i\, d_{i+1}, \qquad i = 1, \ldots, P - 1. \tag{16}$$

This excludes transitions starting from masked context patches and down-weights transitions involving sparsely observed adjacent patches. The EMA next-patch targets are also taken before Transformer self-attention, preventing future contextual information from leaking into the transition objective. In the main experiments, GlucoFM optimizes the masked contextual representation loss and temporal dynamics loss with both loss weights set to 1.0.

D.7 Downstream Feature Extraction

For downstream evaluation, we freeze the pretrained online branch of GlucoFM and discard the EMA target branch and pretraining-only prediction heads. Each one-day CGM window is passed through the online state–event preprocessing modules, unified embedder, and context encoder to obtain patch representations $\{z_i\}_{i=1}^{P}$, where $P = 24$. We use global average pooling over patches to obtain a fixed-length window representation.

D.8 Data Augmentation Details

During pretraining, GlucoFM uses two augmentation families: *value perturbations* and *structural sparsification*. Value perturbations preserve the observation mask while modifying observed glucose values. Baseline wander is applied with probability 0.25 by adding a sinusoidal perturbation with amplitude sampled from $[5, 15]$ mg/dL and frequency from $[0.5, 2.0]$ cycles per window. Compression-like drops are applied with probability 0.10 by attenuating a contiguous 6–12-step segment using a V-shaped curve with minimum multiplier sampled from $[0.4, 0.7]$.

Structural sparsification alters the observation mask itself. Decimation is applied with probability 0.40 to dense windows with more than 200 observed positions by keeping every third observation from a random 5-minute offset, producing a 15-minute-like sampling pattern. Disconnection blocks are applied with probability 0.05 by removing 1–3 contiguous blocks of 2–12 grid steps. Candidate augmentations are evaluated in random order; after one is applied, subsequent probabilities are multiplied by 0.25 to avoid excessive corruption.

E Task Setup and Additional Results

E.1 Subject-Disjoint Linear Probing Details

We provide additional details for the subject-grouped linear probing protocol used in Table 3. For all methods, the pretrained encoder is frozen and only a linear classifier is trained on the extracted

Table 8: Full subject-disjoint linear-probe performance. All reported values represent the mean ± std evaluated via 10 iterations of 5-fold subject-grouped cross-validation.

Method	Metrics	CGMacros				ShanghaiT2DM			Stanford			Hall			
		Diabetes	IR	Hyperlip.	Obesity	IR	Hyperlip.	Hypogly.	Diabetes	β-cell Dys.	IR	Diabetes	IR	Hyperlip.	Glucotype
General Time-series Foundation Models with Large-scale Pretraining															
Chronos-2 (small)	PR	51.6 ± 4.5	82.7 ± 6.5	30.5 ± 6.5	56.0 ± 4.8	65.0 ± 5.4	34.8 ± 6.0	16.0 ± 4.3	69.2 ± 6.6	58.6 ± 5.1	62.5 ± 5.0	51.7 ± 10.3	44.9 ± 7.0	24.2 ± 6.8	54.2 ± 8.2
	AUC	68.0 ± 3.7	67.5 ± 10.4	52.1 ± 5.9	53.7 ± 5.1	55.0 ± 6.0	50.6 ± 6.6	49.1 ± 8.8	65.4 ± 8.0	57.3 ± 6.3	64.2 ± 5.3	64.9 ± 9.7	57.7 ± 7.7	**61.3** ± 9.3	62.6 ± 6.9
	F1	49.4 ± 4.8	61.7 ± 7.7	50.1 ± 4.1	52.7 ± 4.3	52.6 ± 4.3	51.5 ± 4.6	50.5 ± 3.8	61.3 ± 6.2	54.8 ± 4.8	59.6 ± 3.8	60.9 ± 7.1	55.0 ± 6.1	52.5 ± 6.5	58.6 ± 6.5
Chronos-2	PR	54.3 ± 4.7	84.1 ± 5.8	34.0 ± 9.1	58.1 ± 5.4	59.8 ± 5.3	34.7 ± 5.6	18.9 ± 7.8	68.5 ± 6.8	58.5 ± 4.3	59.2 ± 4.2	48.6 ± 9.2	51.3 ± 9.5	20.6 ± 8.5	54.4 ± 9.3
	AUC	69.7 ± 3.9	69.9 ± 9.9	**55.8** ± 7.9	56.6 ± 6.9	49.5 ± 7.2	49.7 ± 7.8	53.0 ± 10.8	65.6 ± 7.8	57.8 ± 5.5	60.2 ± 5.3	60.4 ± 9.2	62.7 ± 8.2	50.3 ± 10.7	63.0 ± 8.6
	F1	50.2 ± 4.4	63.3 ± 7.5	**53.7** ± 4.6	54.2 ± 5.9	49.0 ± 5.6	50.4 ± 5.4	52.8 ± 5.6	61.9 ± 5.6	55.5 ± 4.2	56.6 ± 4.2	58.4 ± 7.9	58.2 ± 6.3	50.9 ± 4.9	56.2 ± 7.8
MOMENT (small)	PR	54.0 ± 4.8	83.8 ± 5.8	30.3 ± 8.5	60.9 ± 6.3	58.3 ± 5.3	38.3 ± 7.4	16.3 ± 5.3	70.4 ± 7.4	57.2 ± 5.2	63.0 ± 5.6	59.2 ± 9.7	53.1 ± 8.2	19.4 ± 6.7	55.6 ± 10.0
	AUC	69.1 ± 4.3	70.6 ± 9.6	51.5 ± 10.3	58.7 ± 6.5	49.0 ± 6.9	55.9 ± 7.6	52.9 ± 9.8	67.6 ± 8.8	55.5 ± 6.5	65.1 ± 6.3	69.7 ± 8.3	62.1 ± 8.8	48.7 ± 10.3	61.7 ± 8.7
	F1	49.8 ± 4.7	63.6 ± 7.7	50.9 ± 6.1	55.7 ± 5.5	49.3 ± 4.8	52.9 ± 5.8	51.4 ± 4.6	62.5 ± 7.5	53.8 ± 5.0	60.1 ± 4.5	64.5 ± 6.9	59.0 ± 6.8	50.4 ± 6.5	59.3 ± 7.7
MOMENT (large)	PR	50.3 ± 5.0	87.4 ± 5.1	32.0 ± 7.9	61.8 ± 5.7	59.6 ± 5.9	**42.4** ± 6.7	14.3 ± 3.4	71.9 ± 5.3	61.8 ± 5.2	58.1 ± 4.3	52.1 ± 9.4	44.9 ± 10.1	18.4 ± 5.7	55.8 ± 7.5
	AUC	66.1 ± 4.6	75.2 ± 9.7	53.2 ± 8.3	59.9 ± 6.8	47.8 ± 7.5	**60.8** ± 6.3	49.6 ± 8.1	69.1 ± 6.4	58.9 ± 5.4	59.4 ± 4.5	65.0 ± 8.0	56.6 ± 8.8	51.4 ± 12.2	63.2 ± 7.6
	F1	46.8 ± 5.4	67.2 ± 8.2	52.5 ± 4.9	56.3 ± 4.8	48.6 ± 5.5	**57.2** ± 5.3	49.7 ± 3.5	63.7 ± 5.0	55.7 ± 4.1	56.1 ± 3.0	60.5 ± 6.5	53.3 ± 7.5	47.7 ± 5.9	57.8 ± 6.4
Mantis	PR	61.8 ± 5.9	90.9 ± 5.5	30.1 ± 10.0	63.4 ± 8.1	59.2 ± 5.5	30.6 ± 4.5	**22.9** ± 12.4	75.1 ± 5.4	63.5 ± 9.0	67.1 ± 7.3	54.2 ± 11.5	55.7 ± 12.4	24.8 ± 9.7	77.5 ± 7.6
	AUC	75.5 ± 5.1	80.3 ± 10.9	49.8 ± 13.6	60.8 ± 9.8	48.2 ± 7.3	46.0 ± 7.6	58.6 ± 17.5	71.4 ± 5.2	61.5 ± 9.3	68.2 ± 5.9	67.1 ± 10.1	66.2 ± 10.4	57.9 ± 10.5	82.6 ± 5.7
	F1	56.0 ± 6.6	**71.6** ± 8.3	49.8 ± 6.9	57.2 ± 6.9	48.4 ± 5.3	47.5 ± 5.5	**53.8** ± 7.7	65.9 ± 4.8	58.5 ± 6.3	62.3 ± 4.5	59.6 ± 8.9	59.8 ± 8.9	**53.7** ± 6.7	73.3 ± 6.1
MantisV2	PR	63.6 ± 6.2	91.2 ± 4.2	30.3 ± 10.0	62.8 ± 7.1	64.1 ± 5.9	33.3 ± 6.6	22.0 ± 12.8	75.2 ± 6.7	65.7 ± 7.2	64.6 ± 9.1	57.7 ± 14.4	59.3 ± 13.6	**25.0** ± 9.9	81.5 ± 7.8
	AUC	76.7 ± 5.2	80.3 ± 9.0	49.1 ± 13.6	60.3 ± 9.4	52.7 ± 6.8	48.5 ± 8.6	**59.6** ± 20.9	71.2 ± 7.9	64.4 ± 7.2	65.5 ± 9.3	67.5 ± 11.8	66.7 ± 11.1	58.8 ± 9.5	84.7 ± 6.0
	F1	56.2 ± 7.0	68.9 ± 9.0	49.2 ± 7.6	55.9 ± 7.2	51.2 ± 5.6	48.1 ± 6.4	53.3 ± 10.1	65.7 ± 7.7	58.6 ± 5.5	60.4 ± 7.5	61.1 ± 9.3	60.8 ± 8.9	52.6 ± 8.0	76.7 ± 7.3
Externally Pretrained CGM Foundation Model															
CGMformer	PR	63.3 ± 6.4	89.9 ± 4.6	31.8 ± 8.6	**68.1** ± 7.3	54.8 ± 3.9	40.6 ± 9.2	11.9 ± 3.0	75.9 ± 5.9	62.2 ± 8.6	61.8 ± 5.6	51.1 ± 11.4	48.0 ± 10.7	18.0 ± 7.9	79.6 ± 7.5
	AUC	77.1 ± 4.7	78.0 ± 9.5	54.2 ± 10.8	**66.0** ± 8.2	43.6 ± 6.3	58.4 ± 10.3	42.1 ± 10.4	71.4 ± 7.2	59.2 ± 8.5	63.2 ± 4.3	66.6 ± 10.1	58.2 ± 10.2	46.5 ± 20.7	84.3 ± 5.3
	F1	57.2 ± 5.4	67.6 ± 7.6	52.1 ± 6.4	**60.5** ± 6.4	45.5 ± 5.1	55.1 ± 6.3	45.3 ± 4.0	**66.2** ± 5.8	55.7 ± 6.6	59.0 ± 3.0	59.3 ± 8.8	56.3 ± 7.9	48.4 ± 8.1	77.1 ± 5.6
CGM Foundation Models Retrained on Our Pretraining Corpus															
CGM-JEPA	PR	63.0 ± 7.5	86.2 ± 7.5	28.7 ± 9.6	55.5 ± 7.4	**69.1** ± 7.5	37.4 ± 8.2	17.8 ± 5.5	66.4 ± 4.1	58.7 ± 5.9	61.5 ± 7.6	59.9 ± 12.1	56.8 ± 14.2	17.3 ± 9.8	87.6 ± 5.9
	AUC	75.9 ± 6.0	73.8 ± 12.2	47.6 ± 13.4	53.8 ± 10.0	**60.8** ± 8.9	53.6 ± 10.0	56.8 ± 11.6	61.2 ± 6.9	55.4 ± 6.7	61.6 ± 7.5	73.5 ± 9.4	68.1 ± 12.5	40.5 ± 10.5	**90.7** ± 4.5
	F1	55.4 ± 7.5	66.3 ± 8.1	48.2 ± 7.9	53.2 ± 6.9	**57.2** ± 6.9	51.3 ± 6.4	50.4 ± 6.2	58.7 ± 5.4	53.9 ± 5.5	58.1 ± 5.4	**65.0** ± 7.9	59.4 ± 10.8	42.8 ± 5.9	82.2 ± 5.9
X-CGM-JEPA	PR	63.6 ± 7.2	86.6 ± 6.9	29.8 ± 10.3	55.3 ± 7.4	66.9 ± 7.4	35.5 ± 8.1	16.9 ± 6.2	67.4 ± 3.8	60.0 ± 6.5	61.9 ± 7.7	59.3 ± 11.9	56.2 ± 14.2	17.5 ± 9.4	87.7 ± 5.9
	AUC	76.6 ± 5.8	73.6 ± 11.5	48.0 ± 13.2	53.2 ± 10.1	58.4 ± 9.0	52.1 ± 10.9	55.7 ± 13.7	61.8 ± 6.1	56.4 ± 7.1	62.0 ± 7.0	73.0 ± 9.2	67.7 ± 12.4	40.0 ± 9.4	90.6 ± 4.6
	F1	56.5 ± 7.6	65.1 ± 8.0	48.2 ± 7.9	52.5 ± 7.1	55.4 ± 7.0	49.9 ± 7.4	49.0 ± 6.7	59.1 ± 5.0	54.4 ± 5.4	58.1 ± 5.3	64.8 ± 7.5	59.2 ± 10.4	42.8 ± 5.2	82.1 ± 5.3
GluFormer (tiny)	PR	59.4 ± 5.5	86.1 ± 6.7	28.2 ± 8.8	60.2 ± 7.6	58.1 ± 3.8	33.9 ± 5.0	17.9 ± 7.5	74.3 ± 6.1	63.3 ± 6.0	64.5 ± 5.0	48.9 ± 11.0	50.9 ± 9.4	21.6 ± 7.2	75.4 ± 7.0
	AUC	73.9 ± 4.6	72.5 ± 11.8	47.9 ± 11.4	57.9 ± 8.3	47.6 ± 4.7	51.1 ± 7.7	53.3 ± 13.3	68.9 ± 6.3	61.9 ± 7.2	65.4 ± 5.0	63.3 ± 9.4	63.6 ± 8.2	58.0 ± 12.4	81.7 ± 4.8
	F1	51.7 ± 5.7	65.4 ± 8.1	48.0 ± 6.2	55.4 ± 6.1	47.8 ± 4.2	50.3 ± 5.9	49.1 ± 6.0	63.0 ± 5.4	58.5 ± 5.9	61.7 ± 3.8	57.9 ± 7.6	59.4 ± 7.3	53.1 ± 6.6	72.7 ± 5.1
GluFormer (base)	PR	62.9 ± 4.6	87.2 ± 6.2	31.4 ± 8.4	61.4 ± 6.5	56.5 ± 4.1	31.5 ± 5.4	11.5 ± 1.9	69.8 ± 6.1	61.1 ± 5.1	56.1 ± 4.4	51.9 ± 12.2	51.6 ± 10.5	16.4 ± 6.0	82.7 ± 7.8
	AUC	76.7 ± 3.7	74.8 ± 11.3	53.1 ± 10.0	60.5 ± 7.9	44.9 ± 5.8	46.7 ± 8.8	43.9 ± 10.9	63.1 ± 6.7	58.0 ± 5.0	56.3 ± 4.5	66.0 ± 9.9	61.6 ± 9.0	46.9 ± 9.9	87.9 ± 5.1
	F1	55.7 ± 5.4	67.8 ± 7.3	52.5 ± 6.0	57.2 ± 6.2	46.3 ± 4.7	48.1 ± 5.8	44.8 ± 2.6	59.2 ± 5.2	55.4 ± 3.9	54.8 ± 3.8	58.7 ± 8.7	58.3 ± 7.7	47.7 ± 4.9	80.7 ± 5.6
GlucoFM	PR	**65.9** ± 7.5	**91.9** ± 5.3	**36.1** ± 11.2	64.9 ± 11.7	67.0 ± 7.7	33.5 ± 7.8	21.1 ± 10.0	**77.3** ± 7.5	**69.0** ± 9.6	**67.6** ± 8.1	**66.2** ± 13.0	**60.2** ± 15.1	14.4 ± 5.2	**88.3** ± 5.7
	AUC	**78.7** ± 6.1	**81.2** ± 12.7	54.7 ± 15.9	62.6 ± 13.2	57.8 ± 8.2	50.5 ± 13.4	59.2 ± 15.9	**72.8** ± 7.4	**68.7** ± 8.9	**69.1** ± 6.6	**75.9** ± 10.2	**70.7** ± 12.2	41.6 ± 11.6	**90.7** ± 4.3
	F1	**58.3** ± 8.5	69.6 ± 11.0	50.2 ± 9.7	59.4 ± 9.9	55.4 ± 6.4	49.1 ± 9.1	50.7 ± 6.9	**66.2** ± 6.6	**63.3** ± 6.9	**64.0** ± 5.8	64.5 ± 9.2	**62.0** ± 10.4	43.1 ± 4.9	**82.4** ± 5.6

24-hour window representations. We use the same scikit-learn logistic regression classifier with L2 regularization for all models, using the `lbfgs` solver, a maximum of 1000 iterations, and a fixed random seed. This ensures that performance differences mainly reflect the quality of the frozen representations rather than downstream classifier capacity or optimization choices.

For each dataset and task, we perform 5-fold subject-grouped cross-validation and repeat the procedure for 10 iterations with different fold assignments. All windows from the same subject are assigned to the same fold, ensuring that training and test subjects never overlap. The same subject splits are used for all methods within each task, enabling a paired comparison across representations. All preprocessing statistics and downstream classifiers are fit only on the training subjects in each fold, and evaluation metrics are computed over held-out test windows. We report mean performance in the main paper for readability, and provide the full mean ± standard deviation across repeated folds in Table 8.

For CGMacros linear probing, Dexcom and Libre 24-hour windows are pooled into one window-level benchmark; each window is encoded independently, and all windows from the same participant, across both sensors, remain in the same fold.

Paired statistical analysis. We quantify uncertainty in the task-averaged linear-probing comparisons against CGM-JEPA and CGMformer using unrounded fold-level scores. For each metric and repeat/fold, we macro-average the 14 task–dataset scores and form the paired difference between GlucoFM and the comparator, yielding $n = 50$ paired differences from 10 repeats of five-fold cross-validation. We use the two-sided Nadeau–Bengio corrected repeated-cross-validation paired t test [21], with $n_{\text{test}}/n_{\text{train}} = 1/4$ and 95% CIs. Holm adjustment is applied across six comparisons: two comparators by three metrics.

All six corrected CIs exclude zero, and all six tests remain significant after Holm adjustment. The effects are also directionally broad: GlucoFM improves over CGM-JEPA on 11/14, 12/14, and 11/14 evaluations and over CGMformer on 11/14, 11/14, and 9/14 evaluations for PR-AUC, ROC-AUC, and Macro-F1, respectively.

E.2 Meal-Aligned PPGR Details

We construct the two-hour postprandial glycemic response (PPGR) task from paired Dexcom and Libre recordings in CGMacros. After requiring a strictly causal 24-hour input window and no

Table 9: Paired task-averaged linear-probing comparisons. Differences are GlucoFM minus the comparator in percentage points. CIs and p-values use the Nadeau–Bengio correction; Holm adjustment is applied across the six model–metric comparisons.

Comparator	Metric	Task average (GlucoFM / comparator)	Δ [95% CI]	Holm-p
CGM-JEPA	PR-AUC	58.8 / 54.7	$+4.11$ [2.40, 5.81]	< 0.001
	ROC-AUC	66.7 / 62.4	$+4.34$ [2.26, 6.41]	< 0.001
	Macro-F1	59.9 / 57.3	$+2.57$ [0.91, 4.23]	0.006
CGMformer	PR-AUC	58.8 / 54.1	$+4.74$ [2.47, 7.02]	< 0.001
	ROC-AUC	66.7 / 62.1	$+4.67$ [2.02, 7.32]	0.003
	Macro-F1	59.9 / 57.5	$+2.34$ [0.55, 4.14]	0.012

Table 10: Two-hour PPGR with complete context at each sensor's native cadence (MAE↓).

Sensor	Model	Trajectory (mg/dL)	Positive iAUC (mg/dL·h)	Peak rise (mg/dL)	Peak time (min)
Dexcom	GluFormer (tiny)	25.38	34.98	30.15	31.37
	GluFormer (base)	26.51	36.89	31.65	32.71
	CGMformer	24.71	33.11	28.86	31.70
	CGM-JEPA	23.60	31.27	27.76	30.74
	X-CGM-JEPA	24.74	33.14	29.11	31.59
	GlucoFM	**22.63**	**30.09**	**26.92**	**30.17**
Libre	GluFormer (tiny)	23.08	30.14	26.04	32.25
	GluFormer (base)	24.12	32.46	27.74	32.54
	CGMformer	23.59	31.22	26.99	32.43
	CGM-JEPA	22.21	29.21	24.19	32.01
	X-CGM-JEPA	21.46	27.99	24.06	31.34
	GlucoFM	**21.12**	**26.91**	**23.87**	**30.46**

additional logged meal in $(0, 120]$ minutes after meal onset, each sensor contributes the same 874 meals from 34 participants. Meal timestamps and nutrition come from CGMacros_Meal, but its interpolated CGM columns are not used as model inputs or prediction targets. Instead, pre-meal CGM and response targets are constructed from the original raw Dexcom and Libre exports. The meal-start baseline is the latest raw observation at or before meal onset, with a maximum age of 5 minutes for Dexcom and 15 minutes for Libre. Post-meal targets are nearest raw observations within 2.5 minutes of each Dexcom horizon and 7.5 minutes of each Libre horizon, without target interpolation.

Dexcom and Libre are trained and evaluated separately at their native target cadences. The same MLP with two hidden layers predicts 24 glucose changes at 5-minute intervals from 5 to 120 minutes for Dexcom and 8 changes at 15-minute intervals from 15 to 120 minutes for Libre. Each frozen representation is combined cumulatively with the last hour of pre-meal CGM, five nutrition variables, fasting glucose, and BMI and diabetes status. We use five dataset-defined subject-group outer folds and 10 head initializations. One repetition concatenates held-out predictions from all five folds before evaluation; the reported two-sensor results average the sensor-specific endpoint errors with equal weight.

Trajectory MAE is computed over all meals and native horizons within each sensor. Positive iAUC is computed per meal as

$$\mathrm{iAUC}^{+} = \int_{0}^{2\mathrm{h}} \max(\Delta g(t), 0)\ dt, \tag{17}$$

using meal-start glucose as the baseline and trapezoidal integration after prepending $\Delta g(0) = 0$ to each sensor's native grid; its MAE is reported in mg/dL·h. Peak-rise MAE compares the maximum predicted and observed glucose changes on the corresponding native grid. Peak-time MAE compares their first argmax times. We use trajectory, positive iAUC, and peak rise as the main endpoints; peak time is reported as a complementary timing measure.

At the complete-context setting, GlucoFM ranks first on all four endpoints for both Dexcom and Libre (Table 10).

Table 11: Task-averaged few-shot performance. Each cell reports PR-AUC / ROC-AUC / Macro-F1 (%).

Setting	Budget	CGM-JEPA	GluFormer (tiny)	CGMformer	GlucoFM
Limited subjects	$K = 1$	50.2 / 57.0 / 49.3	51.6 / 58.8 / 50.2	50.4 / 57.5 / 49.7	**53.3 / 59.6 / 50.4**
	$K = 2$	50.8 / 58.3 / 51.8	51.9 / 59.5 / 52.4	51.3 / 58.8 / 52.2	**53.9 / 60.7 / 53.3**
	$K = 3$	51.1 / 58.7 / 52.7	51.9 / 59.7 / 53.1	51.8 / 59.2 / 53.1	**54.5 / 61.5 / 54.4**
	$K = 4$	51.5 / 59.1 / 53.2	51.9 / 60.0 / 53.6	52.1 / 59.7 / 53.5	**55.2 / 62.2 / 55.2**
	$K = 5$	52.1 / 59.8 / 53.9	52.0 / 60.3 / 54.0	52.5 / 60.2 / 54.2	**55.6 / 62.8 / 56.0**
Limited observations	1%	50.1 / 58.4 / 54.2	52.9 / 61.6 / 56.5	51.8 / 59.8 / 55.6	**56.3 / 64.0 / 57.9**
	5%	50.1 / 58.3 / 54.4	52.6 / 61.2 / 56.2	51.8 / 59.8 / 55.6	**56.5 / 64.2 / 58.1**
	10%	50.9 / 59.2 / 54.9	52.7 / 61.2 / 56.2	52.2 / 60.3 / 55.9	**57.2 / 65.0 / 58.6**
	20%	51.8 / 60.0 / 55.6	52.2 / 60.7 / 56.0	52.9 / 60.9 / 56.4	**57.4 / 65.3 / 59.0**
	30%	52.9 / 60.7 / 56.1	52.1 / 60.7 / 55.9	53.1 / 61.1 / 56.5	**58.1 / 65.8 / 59.5**
	40%	53.5 / 61.2 / 56.4	52.3 / 61.0 / 56.0	53.2 / 61.3 / 56.8	**58.2 / 66.0 / 59.5**
	50%	54.0 / 61.7 / 56.7	52.3 / 61.0 / 56.1	53.3 / 61.3 / 56.8	**58.5 / 66.3 / 59.6**

E.3 Few-shot Adaptation Details

We provide additional details for the few-shot adaptation protocol used in Figure 3. For each pretrained model, we freeze the encoder and extract representations for all labeled 24-hour windows in each dataset. We perform 5-fold subject-grouped cross-validation with 10 repeated iterations, using the same subject splits across all methods. All windows from the same dataset-defined subject are assigned to the same fold, ensuring that training and test grouping units never overlap.

In the *limited-subject* setting, we sample exactly $K \in \{1, 2, 3, 4, 5\}$ labeled support subjects per class from the training fold. The same logistic regression classifier is then trained on all extracted window representations from the selected support subjects. In the *limited-observation* setting, we retain all training subjects but randomly subsample each subject's 24-hour training windows at fractions $\{1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$. For both settings, each fold and support configuration is evaluated with 5 random samplings, and metrics are computed over held-out test windows. Figure 3 reports the broader PR-AUC comparison, and Table 11 provides exact task-averaged PR-AUC, ROC-AUC, and Macro-F1 values for representative CGM-specific models.

Across all evaluated budgets, GlucoFM achieves the strongest task-averaged performance on all three metrics among the models in Table 11.

E.4 Cross-Dataset Generalization Details

We provide additional details for the cross-dataset generalization protocol used in Table 5. Unlike subject-disjoint linear probing, this experiment does not use cross-validation because the goal is to evaluate direct transfer across distinct cohorts. For each transfer direction, we freeze the pretrained encoder, extract representations from all labeled 24-hour windows in the source and target datasets, train a logistic regression classifier on the full source dataset, and evaluate it directly on the target dataset. The target dataset is used only for final evaluation; no target labels are used for training, validation, model selection, or threshold tuning. The same source–target splits, frozen representations, classifier type, and preprocessing protocol are used for all compared methods.

We evaluate diabetes risk assessment and insulin resistance because they are consistently available across CGMacros, Stanford, and Hall. To ensure label compatibility, we harmonize task definitions before transfer. For diabetes risk assessment, CGMacros originally includes normoglycemic, prediabetes, and type 2 diabetes categories; we convert it into a binary risk label by grouping prediabetes and type 2 diabetes as the positive class and normoglycemic status as the negative class, matching the binary diabetes-risk labels in Stanford and Hall. For insulin resistance, we use the dataset-provided binary labels while preserving the same positive/negative semantics across cohorts. This protocol evaluates whether frozen representations support transferable clinical decision boundaries across cohorts rather than relying only on dataset-specific label patterns.

E.5 Multiday Representation Observation Details

We provide additional details for the multiday representation observation analysis in Figure 4. The goal is to evaluate whether observing more days of CGM improves subject-level prediction using

Table 12: Subject-level multiday PR-AUC across observation lengths. Values are mean PR-AUC (%) with 95% confidence intervals over repeated subject-level cross-validation. Mean+Max denotes concat(mean, max) pooling. The best value in each row is bolded. `Hall` is evaluated only for $K = 1$–4.

Dataset	Task	Pooling	K1	K2	K3	K4	K5	K6	K7
CGMacros-Dexcom	Diabetes	Mean	57.6 [57.0, 58.2]	58.5 [57.8, 59.4]	58.2 [57.4, 59.0]	58.5 [57.6, 59.4]	58.5 [57.5, 59.4]	58.6 [57.3, 59.7]	**59.8** [58.5, 60.9]
CGMacros-Dexcom	Diabetes	Mean+Max	57.0 [56.3, 57.7]	57.9 [57.0, 58.7]	57.1 [55.7, 58.4]	58.0 [56.2, 59.8]	57.3 [54.9, 59.6]	57.3 [55.1, 59.3]	**59.5** [57.5, 61.7]
CGMacros-Dexcom	IR	Mean	92.5 [91.7, 93.2]	93.3 [92.3, 94.0]	94.4 [93.3, 95.2]	94.9 [93.8, 95.7]	95.3 [94.2, 96.1]	**95.6** [94.5, 96.5]	**95.6** [94.7, 96.3]
CGMacros-Dexcom	IR	Mean+Max	92.3 [91.4, 92.9]	92.5 [91.5, 93.2]	93.3 [92.0, 94.3]	93.8 [92.3, 94.9]	94.6 [93.1, 95.7]	95.5 [94.1, 96.6]	**97.8** [97.2, 98.3]
CGMacros-Libre	Diabetes	Mean	63.7 [62.1, 65.1]	65.5 [63.5, 67.2]	66.7 [64.5, 68.3]	67.0 [64.7, 68.8]	**67.8** [65.5, 69.6]	67.5 [65.0, 69.5]	67.4 [64.9, 69.6]
CGMacros-Libre	Diabetes	Mean+Max	62.9 [61.1, 64.4]	63.9 [61.7, 65.6]	63.3 [61.1, 64.9]	63.3 [60.8, 65.0]	**65.7** [62.9, 68.0]	63.3 [60.3, 65.9]	64.5 [61.2, 67.5]
CGMacros-Libre	IR	Mean	90.6 [90.1, 91.1]	91.8 [91.5, 92.2]	92.1 [91.8, 92.5]	92.1 [91.8, 92.5]	92.4 [92.0, 92.8]	92.5 [92.1, 92.9]	**92.7** [92.3, 93.1]
CGMacros-Libre	IR	Mean+Max	89.7 [88.9, 90.3]	91.6 [91.2, 92.0]	91.5 [91.0, 91.9]	91.5 [91.1, 92.1]	92.5 [92.1, 93.0]	**92.7** [92.3, 93.1]	92.2 [91.8, 92.7]
CGMacros-Fused	Diabetes	Mean	64.0 [63.0, 65.1]	65.6 [64.4, 66.7]	66.3 [65.0, 67.5]	67.0 [65.6, 68.3]	**67.6** [66.1, 68.9]	67.4 [65.7, 69.1]	67.0 [65.0, 68.8]
CGMacros-Fused	Diabetes	Mean+Max	63.2 [62.2, 64.3]	64.0 [62.8, 65.3]	64.2 [62.6, 65.7]	65.7 [63.7, 67.6]	**66.1** [63.9, 68.3]	63.9 [61.6, 66.1]	63.6 [61.6, 65.5]
CGMacros-Fused	IR	Mean	92.7 [92.4, 93.1]	93.2 [92.7, 93.8]	93.6 [92.9, 94.2]	93.8 [93.1, 94.6]	94.2 [93.5, 94.9]	94.4 [93.7, 95.1]	**94.6** [93.9, 95.2]
CGMacros-Fused	IR	Mean+Max	92.3 [91.9, 92.7]	93.5 [92.9, 94.0]	93.5 [92.9, 94.0]	93.7 [93.1, 94.2]	93.9 [93.2, 94.6]	94.4 [93.7, 95.0]	**95.0** [93.7, 95.8]
ShanghaiT2DM	IR	Mean	**61.9** [60.7, 63.2]	**61.9** [59.9, 63.7]	61.0 [58.4, 63.4]	61.7 [58.9, 64.4]	61.2 [58.3, 64.1]	61.4 [58.4, 64.4]	**61.9** [59.5, 64.3]
ShanghaiT2DM	IR	Mean+Max	63.1 [61.8, 64.4]	62.5 [60.6, 64.3]	60.8 [58.4, 63.3]	62.1 [59.6, 64.6]	67.8 [65.3, 70.3]	**75.2** [72.7, 77.6]	**75.2** [73.6, 76.7]
Stanford	Beta-cell	Mean	60.3 [59.2, 61.3]	61.8 [60.5, 63.0]	64.0 [62.5, 65.4]	65.4 [63.8, 66.9]	66.8 [65.0, 68.5]	68.3 [65.7, 70.7]	**69.9** [67.1, 72.4]
Stanford	Beta-cell	Mean+Max	59.7 [58.8, 60.5]	63.0 [62.2, 63.9]	63.7 [62.4, 65.0]	64.8 [63.7, 66.0]	63.6 [62.0, 65.1]	63.5 [61.5, 65.5]	**65.8** [63.2, 68.3]
Stanford	Diabetes	Mean	67.6 [66.1, 69.1]	69.8 [68.3, 71.5]	70.6 [69.0, 72.3]	70.3 [68.2, 72.5]	70.5 [67.9, 73.4]	71.8 [68.7, 74.9]	**73.6** [71.2, 75.9]
Stanford	Diabetes	Mean+Max	67.9 [66.3, 69.6]	67.7 [66.1, 69.4]	68.3 [66.5, 70.1]	66.9 [64.4, 69.6]	68.0 [65.1, 71.1]	69.5 [66.6, 72.5]	**71.7** [69.5, 74.2]
Stanford	IR	Mean	60.3 [58.7, 61.9]	61.5 [59.4, 63.5]	62.4 [60.5, 64.2]	62.9 [60.9, 64.9]	64.1 [62.1, 66.1]	65.9 [64.0, 68.0]	**66.6** [64.6, 68.9]
Stanford	IR	Mean+Max	60.1 [58.6, 61.4]	61.8 [60.0, 63.7]	63.3 [61.4, 65.0]	**66.2** [64.3, 68.1]	63.9 [62.6, 65.3]	64.0 [62.9, 65.1]	65.3 [63.6, 66.8]
Hall	Diabetes	Mean	59.8 [58.7, 60.9]	62.8 [60.0, 65.5]	67.7 [64.9, 70.5]	**73.8** [71.6, 75.9]	–	–	–
Hall	Diabetes	Mean+Max	57.8 [56.8, 58.7]	65.4 [62.7, 67.5]	63.6 [60.0, 67.1]	**70.9** [67.8, 74.1]	–	–	–
Hall	Glucotype	Mean	85.9 [84.9, 86.8]	91.6 [90.9, 92.3]	96.0 [95.4, 96.5]	**99.5** [98.9, 99.9]	–	–	–
Hall	Glucotype	Mean+Max	85.0 [83.9, 86.0]	90.9 [90.0, 91.9]	93.7 [92.5, 94.8]	**98.8** [98.0, 99.4]	–	–	–
Hall	IR	Mean	46.3 [44.2, 48.4]	44.5 [42.3, 46.5]	46.3 [42.8, 49.5]	**49.8** [46.0, 53.0]	–	–	–
Hall	IR	Mean+Max	45.0 [42.9, 47.2]	44.9 [42.8, 46.8]	41.5 [38.9, 44.2]	**45.1** [42.1, 48.0]	–	–	–

frozen GlucoFM representations. For each subject, CGM recordings are divided into non-overlapping 24-hour windows and aligned to the GlucoFM input format, a 288-point 5-minute chronological grid. A day is included only if it satisfies the preprocessing quality criteria, including a maximum consecutive missing interval of less than one hour. For each valid day, we extract one embedding using the frozen pretrained GlucoFM encoder. We then select one fixed eligible $K_{\max}$-day anchor episode for each subject. Stanford, CGMacros, and ShanghaiT2DM use $K_{\max} = 7$, while Hall uses $K_{\max} = 4$ due to the available near-continuous recordings. Within each fixed anchor, we enumerate adjacent K-day subwindows for $K = 1, \ldots, K_{\max}$. For each K-day subwindow, we aggregate the frozen daily embeddings into one subject-level representation. We evaluate two aggregation variants. The first uses mean pooling:

$$\mathbf{z}_{i,K,s}^{\text{mean}} = \frac{1}{K} \sum_{d=s}^{s+K-1} \mathbf{h}_{i,d}, \tag{18}$$

where $\mathbf{h}_{i,d}$ is the frozen daily embedding for subject i on day d, and s is the subwindow start position. The second uses concat(mean, max) pooling:

$$\mathbf{z}_{i,K,s}^{\text{mean+max}} = \left[\text{mean}_{d=s}^{s+K-1}(\mathbf{h}_{i,d}); \max_{d=s}^{s+K-1}(\mathbf{h}_{i,d})\right], \tag{19}$$

which preserves both the average daily pattern and salient embedding dimensions across the observation window.

For each K and start position, every subject contributes exactly one representation, and each test subject receives exactly one prediction. Thus, adjacent subwindows from the same subject are not treated as independent test samples. We train linear probes with 10 repeated iterations of 5-fold stratified subject-level cross-validation. Metrics are computed separately for each start position and averaged within each repeated evaluation. In CGMacros, Dexcom, Libre, and Fused are all evaluated on the same 42-subject fused-overlap 7-day anchor cohort. Dexcom and Libre are evaluated separately, while the matched fused setting averages same-subject same-day embeddings from both

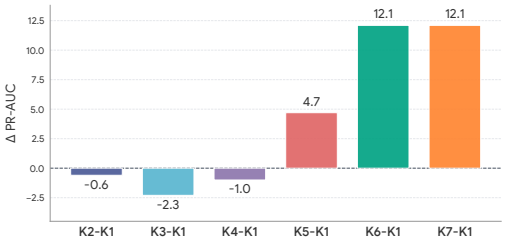


Figure 6: 7-day `ShanghaiT2DM` IR under concat(mean, max) aggregation.

sensors before multiday aggregation. We report the paired PR-AUC change relative to the one-day representation: $\Delta_K = \text{PR-AUC}(K) - \text{PR-AUC}(1)$. Deltas are computed within the same repeated evaluation, and confidence intervals are estimated over repeat-level paired deltas.

Table 12 reports the full multiday PR-AUC results for both aggregation variants; longer observation windows improve subject-level prediction in most settings.

ShanghaiT2DM pooling comparison. ShanghaiT2DM IR is an exception under simple mean pooling, but concat(mean, max) pooling (Figure 6) yields clear gains at longer horizons, especially at $K = 6$ and 7.

E.6 Same-window GMI Comparison

We compare GlucoFM with a same-window 7-day threshold baseline based on a glucose management indicator (GMI)-equivalent value [61] for diabetes classification. The GMI-equivalent value is computed from the mean glucose over the same selected seven CGM days used by GlucoFM: $\text{GMI} = 3.31 + 0.02392 \times \text{mean glucose}_{\text{mg/dL}}$. For Stanford, diabetes is treated as a binary task, where subjects with $\text{GMI} \geq 5.7$ are assigned to the risk-positive class and those with $\text{GMI} < 5.7$ to the negative class. For CGMacros, diabetes is treated as a three-class task using thresholds at 5.7 and 6.4: $\text{GMI} < 5.7$, $5.7 \leq \text{GMI} < 6.4$, and $\text{GMI} \geq 6.4$. In the fused CGMacros setting, GMI is computed from combined Dexcom and Libre readings within the matched seven-day window. Since this baseline is deterministic, we use macro-F1 as the primary metric.

As shown in Table 13, GlucoFM consistently outperforms the same-window 7-day GMI-equivalent threshold baseline across Stanford and CGMacros sensor settings. The gains are largest on CGMacros-Dexcom and the fused setting, indicating stronger diabetes category classification than a rule based only on seven-day mean glucose. Since the GMI baseline reduces each seven-day window to a single average-glucose-derived value, these results show that GlucoFM retains predictive CGM information beyond seven-day mean glucose. The smaller but positive gain on CGMacros-Libre further shows that the benefit of representation learning depends on the sensor setting and baseline strength.

Table 13: GMI-equivalent comparison on Macro-F1.

Dataset	GMI Rule	GlucoFM	Δ
Stanford	59.6	**67.0**	**+07.4**
CGMacros-Dexcom	36.3	**53.7**	**+17.4**
CGMacros-Libre	63.7	**65.6**	**+01.9**
CGMacros-Fused	56.9	**65.9**	**+09.0**

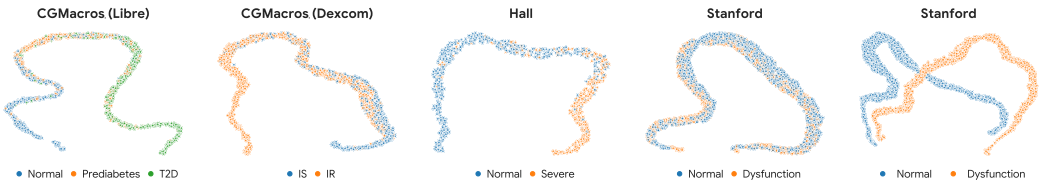
E.7 Pretraining Data Scalability Details

We provide additional details for the pretraining data scalability analysis in Figure 5. To assess how GlucoFM benefits from unlabeled CGM data, we construct reduced pretraining corpora by randomly sampling subjects from each pretraining cohort at ratios of 20%, 40%, 60%, and 80%. Subject-level sampling is performed within each cohort to preserve the overall cohort composition while varying the amount of available unlabeled data. For each ratio, we repeat the sampling with five random seeds and pretrain a separate GlucoFM model from scratch using the same architecture, optimization settings, masking strategy, and downstream evaluation protocol as the full-data model. This isolates the effect of pretraining data scale from changes in model capacity or downstream classifier settings.

Figure 5 shows that task-averaged performance improves as the amount of unlabeled pretraining data increases.

E.8 Frozen Representation Visualization

We use UMAP [62] to qualitatively inspect task-related structure in frozen GlucoFM embeddings before downstream adaptation. As shown in Figure 7(a–d), embeddings form continuous manifolds rather than clearly separated clusters, consistent with the gradual nature of glucose regulation and metabolic dysfunction [63, 64]. Even without label supervision, several labels are enriched in different regions, especially for diabetes risk, insulin resistance, and glucotype patterns, suggesting that GlucoFM captures physiologically relevant variation. More subtle outcomes such as β-cell dysfunction remain more entangled in the unsupervised view, while the supervised UMAP in Fig-



(a) Diabetes (b) IR (c) Glucotype (d) β-cell (e) β-cell (sup.)

Figure 7: UMAP visualizations of frozen `GlucoFM` embeddings. Panels (a)–(d) show unsupervised UMAP projections; panel (e) shows a supervised UMAP for β-cell dysfunction.

Table 14: Performance comparison of different encoder designs. All reported values represent the mean $\pm$ std evaluated via a 10-iteration 5-fold cross-validation.

Encoder Design	Metrics	CGMacros				ShanghaiT2DM			Stanford			Hall			
		Diabetes	IR	Hyperlip.	Obesity	IR	Hyperlip.	Hypogly.	Diabetes	β-cell Dys.	IR	Diabetes	IR	Hyperlip.	Glucotype
Raw Input	PR	61.4 ± 6.0	89.4 ± 7.0	31.1 ± 9.8	**66.0** ± 8.9	60.9 ± 6.8	34.5 ± 7.0	**23.3** ± 11.1	**77.3** ± 6.7	64.4 ± 8.1	63.8 ± 7.7	55.7 ± 12.9	53.2 ± 10.6	16.1 ± 4.9	77.0 ± 9.9
	AUC	75.4 ± 4.9	79.0 ± 13.8	49.9 ± 14.6	**64.5** ± 9.4	50.8 ± 8.4	50.1 ± 10.3	**63.2** ± 16.0	**73.0** ± 7.4	64.0 ± 8.4	65.0 ± 6.7	68.1 ± 10.4	63.1 ± 10.6	43.3 ± 11.6	82.1 ± 7.0
	F1	55.3 ± 5.9	70.6 ± 10.3	48.8 ± 7.6	**60.5** ± 7.1	50.5 ± 6.4	**49.7** ± 6.5	**56.3** ± 8.7	**66.4** ± 6.6	59.7 ± 6.8	60.1 ± 4.5	60.1 ± 7.9	58.0 ± 7.3	46.6 ± 5.8	73.4 ± 6.4
State-stream Only	PR	62.8 ± 6.7	91.0 ± 5.6	34.5 ± 12.9	62.6 ± 9.1	59.2 ± 6.1	33.2 ± 8.4	18.2 ± 7.1	74.5 ± 4.7	62.7 ± 7.9	62.2 ± 6.9	54.1 ± 11.3	58.9 ± 12.0	18.5 ± 5.1	86.2 ± 6.4
	AUC	76.0 ± 5.6	80.0 ± 12.7	51.0 ± 16.7	60.0 ± 11.0	47.8 ± 7.8	48.5 ± 13.2	55.1 ± 12.6	69.5 ± 6.7	60.8 ± 8.9	62.7 ± 5.1	68.5 ± 8.7	69.0 ± 11.0	**54.3** ± 13.5	89.2 ± 4.4
	F1	54.1 ± 7.0	**70.7** ± 10.7	50.0 ± 8.3	56.6 ± 8.0	48.4 ± 6.6	48.7 ± 8.5	50.5 ± 6.6	63.9 ± 5.7	57.2 ± 7.0	59.5 ± 4.0	61.3 ± 7.4	**62.7** ± 9.0	47.3 ± 6.7	80.2 ± 6.0
Event-stream Only	PR	60.1 ± 4.8	84.7 ± 9.7	28.6 ± 8.2	60.5 ± 7.4	62.2 ± 6.5	**34.7** ± 5.9	13.9 ± 5.8	76.9 ± 5.9	62.0 ± 8.1	57.2 ± 6.5	53.1 ± 13.2	52.0 ± 11.5	**19.4** ± 7.7	59.4 ± 9.7
	AUC	74.5 ± 4.4	72.0 ± 16.1	46.9 ± 11.1	58.7 ± 9.9	53.5 ± 8.2	**51.3** ± 9.3	46.3 ± 9.8	72.3 ± 7.0	61.9 ± 7.8	60.5 ± 7.2	62.3 ± 12.4	62.6 ± 9.4	52.6 ± 11.4	65.9 ± 9.8
	F1	54.5 ± 5.9	65.8 ± 10.0	46.8 ± 6.6	56.2 ± 8.0	53.0 ± 5.6	49.2 ± 6.6	46.4 ± 4.7	65.3 ± 6.4	56.4 ± 6.6	57.3 ± 4.6	57.4 ± 7.5	57.7 ± 6.9	**49.4** ± 7.4	62.6 ± 7.5
Dual-stream	PR	**65.9** ± 7.5	**91.9** ± 5.3	**36.1** ± 11.2	64.9 ± 11.7	**67.0** ± 7.7	33.5 ± 7.8	21.1 ± 10.0	**77.3** ± 7.5	**69.0** ± 9.6	**67.6** ± 8.1	**66.2** ± 13.0	**60.2** ± 15.1	14.4 ± 5.2	**88.3** ± 5.7
	AUC	**78.7** ± 6.1	**81.2** ± 12.7	**54.7** ± 15.9	62.6 ± 13.2	**57.8** ± 8.2	50.5 ± 13.4	59.2 ± 15.9	72.8 ± 7.4	**68.7** ± 8.9	**69.1** ± 6.6	**75.9** ± 10.2	**70.7** ± 12.2	41.6 ± 11.6	**90.7** ± 4.3
	F1	**58.3** ± 8.5	69.6 ± 11.0	**50.2** ± 9.7	59.4 ± 9.9	**55.4** ± 6.4	49.1 ± 9.1	50.7 ± 6.9	66.2 ± 6.6	**63.3** ± 6.9	**64.0** ± 5.8	**64.5** ± 9.2	62.0 ± 10.4	43.1 ± 4.9	**82.4** ± 5.6

ure 7(e) shows that label-guided projections can reveal more separable structure from the same frozen embeddings. These visualizations provide qualitative support for task-related organization in GlucoFM representations.

F Ablation Design Details

F.1 Encoder Design Ablation Details

Table 14 provides task-wise results for the encoder design ablation summarized in Table 6. All variants use the same pretraining data, optimization settings, masking strategy, and downstream evaluation protocol. To ensure a fair comparison, we do not remove auxiliary patch-level information from the ablated encoders. For each patch, we compute mask-aware physiological statistics consisting of state statistics, i.e., glucose mean and standard deviation, and event statistics, i.e., mean rate-of-change and rate-of-change standard deviation. The rate-of-change features are computed with a mask-aware dynamic backoff and normalized to the 5-minute grid.

For fairness, all variants are provided with the corresponding patch-level statistics and temporal-difference features. The variants differ only in how the waveform and auxiliary patch features are organized. The *Raw Input* variant directly tokenizes the aligned glucose sequence and receives both patch-internal differences and the full set of patch statistics, including glucose mean/std and rate-of-change mean/std. The *State-stream Only* variant uses the filtered low-frequency trend with trend differences and state statistics. The *Event-stream Only* variant uses the residual component with rate-of-change features and event statistics. The full *Dual-stream* model encodes the state and event streams separately before fusing them into a unified daily representation.

As shown in Table 14, the dual-stream encoder achieves the strongest task-averaged performance, supporting the benefit of integrating the two temporal streams.

F.2 Pre-fusion Stream Content Probe Details

We analyze the frozen 64-dimensional state and event tokens immediately before cross-stream fusion from the same pretrained checkpoint, without updating the encoder. The analysis uses 7,435 high-coverage hourly Dexcom patches from 35 CGMacros participants. Identical Ridge probes predict either mask-aware hourly mean glucose, which measures retained glucose-level content, or mean absolute adjacent 5-minute change, which measures retained local-dynamics content. Both streams use the same five repeated subject-disjoint 5-fold splits, and point estimates are means of repeat-wise

Table 15: Task-averaged results for supporting ablations.

Ablation	Setting	PR-AUC	ROC-AUC	Macro-F1
Temporal dynamics weight	$\lambda_{\mathrm{TD}} = 0.0$	55.5	63.2	57.7
	$\lambda_{\mathrm{TD}} = 0.2$	57.1	65.7	59.2
	$\lambda_{\mathrm{TD}} = 0.4$	58.2	66.5	59.5
	$\lambda_{\mathrm{TD}} = 0.6$	58.9	67.0	59.9
	$\lambda_{\mathrm{TD}} = 0.8$	58.9	67.0	60.0
	$\lambda_{\mathrm{TD}} = 1.0$ (default)	58.8	66.7	59.9
	$\lambda_{\mathrm{TD}} = 1.2$	58.6	66.5	59.7
	$\lambda_{\mathrm{TD}} = 1.4$	58.4	66.4	59.6
	$\lambda_{\mathrm{TD}} = 1.6$	58.3	66.2	59.4
	$\lambda_{\mathrm{TD}} = 1.8$	58.1	66.1	59.2
	$\lambda_{\mathrm{TD}} = 2.0$	57.9	66.0	59.2
Data augmentation	None	56.8	65.0	58.6
	Value perturbation	57.4	65.6	59.0
	Structural sparsification	58.4	66.5	59.7
	Full (default)	58.8	66.7	59.9
Interpolation	Dense interpolation	56.7	64.4	58.2
	Dense + sparsification	57.9	66.2	59.2
	Mask-aware (default)	58.8	66.7	59.9

out-of-fold R^2. The paired directional gaps reported in Table 6 use 2,000 subject-cluster bootstrap resamples while preserving the repeated evaluation structure.

F.3 Temporal Dynamics Weight Ablation Details

In the temporal dynamics weight ablation in Figure 8, we keep the architecture, pretraining data, masking strategy, augmentation pipeline, and downstream evaluation protocol fixed, and vary only the weight λ_{TD} of the temporal dynamics objective. The total pretraining loss is $\mathcal{L} = \mathcal{L}_{\mathrm{MCR}} + \lambda_{\mathrm{TD}}\mathcal{L}_{\mathrm{TD}}$, where $\mathcal{L}_{\mathrm{MCR}}$ denotes masked contextual latent prediction and $\mathcal{L}_{\mathrm{TD}}$ denotes temporal dynamics modeling. We sweep λ_{TD} from 0 to 2.0, with $\lambda_{\mathrm{TD}} = 0$ corresponding to masked contextual prediction alone.

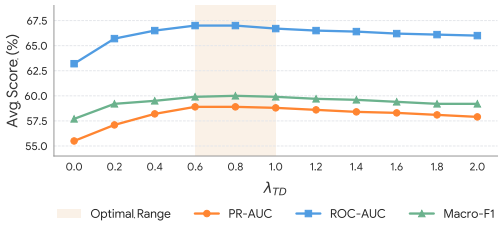


Figure 8: Temporal dynamics weight ablation.

Task-averaged performance is strongest across a broad range around $\lambda_{\mathrm{TD}} = 0.6$–$1.0$ (Table 15).

Figure 8 evaluates the weight of the temporal dynamics objective. Removing this objective ($\lambda_{\mathrm{TD}} = 0$) reduces task-averaged performance across all metrics, indicating that masked contextual latent prediction alone does not fully capture clinically useful glucose evolution. Performance improves as λ_{TD} increases and reaches a broad optimum around 0.6–1.0, then gradually declines when the dynamics term becomes too dominant. This supports a balanced pretraining objective: contextual prediction captures global daily structure, while temporal dynamics modeling provides complementary constraints on local state–event transitions.

F.4 Data Augmentation Ablation Details

For the data augmentation ablation in Figure 9, we keep the architecture, pretraining objectives, masking strategy, and downstream evaluation protocol fixed, and vary only the training-time augmentation pipeline. The *No Aug.* setting removes all augmentations. The *Value Perturb.* setting includes value-level perturbations, such as low-frequency baseline wander and compression-like transient drops. The *Struct. Spars.* setting includes structural sparsification, such as decimation and sensor-disconnection blocks. The full setting combines both augmentation families and achieves the strongest task-averaged performance (Table 15).

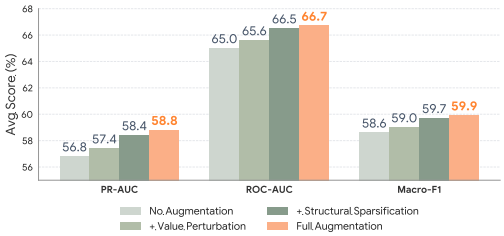


Figure 9: Data augmentation ablation.

We evaluate whether GlucoFM benefits from the proposed CGM-aware augmentation pipeline. As shown in Figure 9, augmentation improves task-averaged downstream performance across all metrics. Value-level perturbations provide modest gains by exposing the model to amplitude shifts and transient artifacts, while structural sparsification contributes more by simulating realistic missingness and heterogeneous sampling rates. This suggests that robustness to observation-pattern variation is particularly important for CGM representation learning. Overall, augmentation improves downstream representation quality but complements rather than replaces the representation design.

F.5 Dense Interpolation Ablation Details

For the dense interpolation ablation in Figure 10, we compare three preprocessing designs while keeping the architecture, pretraining objectives, and downstream protocol fixed. The *Dense Interp.* variant linearly interpolates missing grid positions during both pretraining and downstream representation extraction, and treats the resulting sequence as densely observed. The *Dense Interp. + Aug.* variant uses the same dense interpolation protocol but further applies structural sparsification during pretraining to reintroduce missingness-like perturbations. The default *No Interp.* setting preserves the original observation mask and does not treat interpolated values as real measurements, and achieves the strongest task-averaged performance (Table 15).

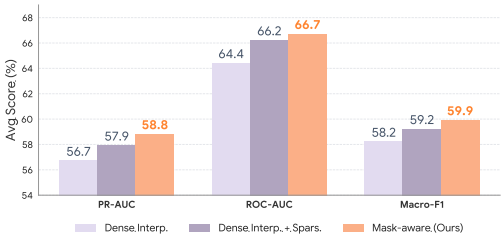


Figure 10: Dense interpolation ablation.

We test whether GlucoFM requires the dense interpolation preprocessing commonly used by CGM-specific baselines. We compare dense interpolation alone with dense interpolation plus structural sparsification. As shown in Figure 10, both interpolated variants underperform the default mask-aware setting, although they remain competitive with CGM-specific foundation model baselines on average. Adding structural sparsification narrows the gap, suggesting that exposure to structured missingness can partially reduce interpolation-induced shortcuts. Overall, preserving the observation mask is more effective for GlucoFM than treating imputed values as fully observed measurements, especially when CGM recordings contain heterogeneous sampling rates and sensor dropouts.

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 abstract and Section 1 describe GlucoFM's daily CGM representation learning, evaluation across four cohorts and seven phenotype-classification tasks, and context-conditioned PPGR prediction using the same frozen encoder. These claims are supported by the phenotyping, transfer, few-shot, multiday, and PPGR experiments in Section 3, the mechanism evidence in Section 4, and the scope discussed in Section 5.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Section 5 discusses pretraining scale and cohort/device coverage, retrospective phenotype classification and meal-aligned PPGR evaluation, and the distinction between multiday embedding aggregation and native long-range modeling. It also identifies future work on more diverse cohorts, native multiday context, and real-time CGM modeling.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper does not present theoretical results, theorems, or formal proofs; it focuses on model design and empirical evaluation.

   Guidelines:

   - The answer [N/A] 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: Section 2 specifies the model architecture and objectives, and Section 3 describes the datasets, baselines, metrics, and evaluation designs. Appendices B, C, D, and E provide preprocessing, pretraining configuration, frozen-feature extraction, dataset-defined group splits, and task-specific phenotyping, PPGR, and multiday protocols.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: Code and reproducibility scripts are planned for public release but are not yet publicly available. Appendix B documents the dataset sources and the non-public Wear-CGM component, which is not redistributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 specifies held-out window-level phenotype evaluation under subject-grouped splits, sensor-specific PPGR evaluation, and multiday subject-level prediction. Appendix D.5 reports pretraining epochs, batch size, learning rates, weight decay, and GPU type; Appendices C and E detail baseline configurations, downstream classifiers, train-only preprocessing, context inputs, and task-specific splits and repetitions.

   Guidelines:

   - The answer [N/A] 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: Section 3 and Appendix E.1 report repeated subject-grouped linear probing with Nadeau–Bengio-corrected paired 95% confidence intervals and Holm-adjusted p-values across six task-averaged comparisons; Table 8 reports fold-level standard deviations. Appendices F.2 and E.5 describe subject-cluster bootstrap intervals and intervals over repeat-level paired PR-AUC changes, respectively, and the scalability analysis reports variability across five subject-subsampling seeds.

Guidelines:

- The answer [N/A] 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: [No]

   Justification: Appendix D.5 reports pretraining on a single NVIDIA H100 GPU together with the number of epochs, batch size, learning rates, and weight decay, and Appendix D.4 reports model size. GPU memory, per-run execution time, and total compute across experiments are not reported.

   Guidelines:

   - The answer [N/A] 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 study evaluates existing CGM datasets for retrospective phenotype classification and meal-aligned PPGR research; Appendix B.2 documents the original Wear-CGM IRB approvals and consent for de-identified secondary research. Section 5 states the research-prototype scope and clinical-use boundaries, and Section 3 and Appendix E describe the evaluation splits and protocols.

Guidelines:

- The answer [N/A] 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: Sections 1 and 5 discuss the potential value of reusable CGM representations for metabolic phenotyping and context-conditioned glucose-response prediction. Section 5 also discusses cohort/device heterogeneity, retrospective evaluation, and the need for validation in the intended population and clinical workflow.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: GlucoFM is a CGM representation model for retrospective phenotyping and PPGR research, rather than a high-risk generative model, and this work does not introduce web-scraped datasets. Wear-CGM remains non-public, as described in Appendix B.2.

    Guidelines:

    - The answer [N/A] 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: [No]

    Justification: Section 3 and Appendices B and C credit the original dataset and baseline-model sources, and Appendix B.2 describes Wear-CGM's non-public access and consent conditions. The paper does not enumerate the license names and versions for all existing assets.

    Guidelines:

    - The answer [N/A] 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: [No]

    Justification: GlucoFM's architecture, training configuration, evaluation protocols, and limitations are documented in Sections 2–5 and the appendix. Code and reproducibility scripts are planned for release; a public asset package with accompanying documentation is not yet provided.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: This work performs secondary analysis of existing CGM datasets and does not conduct crowdsourcing, new participant recruitment, or participant-facing experiments. No new participant instructions or compensation procedures were introduced; the original Wear-CGM study protocols and consent for secondary research are described in Appendix B.2.

Guidelines:

- The answer [N/A] 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: [Yes]

Justification: Appendix B.2 reports that the original Wear-CGM protocols were approved by Advarra (IRB nos. Pro00059582 and Pro00069880), and that participants provided written informed consent permitting de-identified secondary research and algorithm development. The present work analyzes existing recordings and does not introduce new participant recruitment or interventions.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of GlucoFM's core method. The model, pretraining objectives, and downstream predictors described in Sections 2 and 3 do not use LLMs.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.