Time-Aware Cross-Attention for Multi-Modal Sensor-Based Blood Glucose Forecasting

Aashritha Machiraju[2], Ebrahim Farahmand[1,2], Shovito Barua Soumma[1,2], Asiful Arefeen[1,2], Carol Johnston[1], and Hassan Ghasemzadeh[1]

***Abstract*—Accurate blood glucose forecasting enables proactive management of metabolic health, particularly when leveraging data from wearable sensors that capture data about physiological and behavioral health. However, existing models struggle with integrating multimodal time-series data with inconsistent sampling rates. This paper proposes a novel forecasting framework that incorporates a time-aware cross-attention mechanism with an LSTM architecture to predict blood glucose levels using continuous glucose monitoring (CGM) data alongside physiological and behavioral signals, such as heart rate (HR), electrodermal activity, accelerometry, and dietary intake. The proposed method dynamically encodes temporal features without the need for preprocessing and employs gated multi-head cross-attention layers to fuse sensor modalities effectively. We evaluate our approach on a newly constructed dataset involving 12 participants. Our method outperforms the baseline and state-of-the-art GlySim models across multiple prediction horizons ranging from 5 minutes to 90 minutes, achieving up to 17.8% improvement in Root Mean Squared Error (RMSE) values.**

***Index Terms*—Prediabetes, attention mechanism, wearable sensors, digital health, deep learning, metabolic health**

I. Introduction

Maintaining normal blood glucose levels (BGL) and minimizing out-of-range excursions are critical to overall health, and a substantial body of prior research has demonstrated the health benefits of glucose control in healthy individuals, in people with prediabetes, and in those with diabetes. Wearable body sensors such as activity trackers and Continuous Glucose Monitor (CGM) devices are commercially available technologies employed in diabetes care to measure physiological, behavioral, and glucose level signals [1]. CGM devices provide contemporaneous glucose values every 5 minutes, capturing glucose variability over time. The high temporal resolution of CGM data enables the identification of trends in glucose levels, and support informed therapeutic decision-making by patients and healthcare providers. Traditionally, CGMs were primarily prescribed for individuals with diabetes; however, the recent availability of this technology over the counter has made it accessible to the broader population. This widespread availability presents a valuable opportunity for the early detection, prevention, and management of diabetes by tracking trends in BGL [2]. Moreover, since physiological and behavioral factors, such as physical activity, meal intake, heart rate, and stress levels, influence fluctuations in blood glucose levels, there is the opportunity to develop analytical tools capable of processing CGM and other sensor data in real-time and delivering actionable feedback to support effective glucose regulation [3].

In recent years, deep learning (DL) models have been increasingly employed for time-series forecasting tasks, including the prediction of BGL. Modern DL architectures, such as Convolutional and LSTM networks, have been proposed to overcome the limitations of traditional BGL forecasting methods. GlySim [4] proposed a stacked multimodal Convolutional and LSTM to predict BGL. However, this model struggles to capture the long-term dependencies present in time-series data, which limits their effectiveness for long-term forecasting. Moreover, DL-based BGL forecasting models such as Gluformer [5] leverage a multi-head attention mechanism that effectively models both short-term and long-term temporal dependencies. Nevertheless, Gluformer demonstrates limited performance when integrating irregularly sampled time-series data. Despite the advantages offered by the multi-head attention architecture [6] and its improved studies, these models still face challenges in capturing the relative importance of heterogeneous time-series features, particularly those with irregular sampling rates.

Key Limitations and Associated Challenges: The primary limitations of current state-of-the-art methods and major challenges in accurate blood glucose prediction include:(1) limited ability to deliver accurate long-term blood glucose forecasts; (2) data sensor sources such as CGM readings, physiological signals, and behavioral variables exhibit mismatched temporal resolutions, known as irregular sampling rates; and (3) scarcity of real-world datasets, particularly for healthy populations.

Novel Contributions: To address these limitations, we propose a novel glucose forecasting model based on multi-modal attention and LSTM architectures. This model integrates CGM data and macronutrient data, such as information about fat, carbohydrates, and protein, with various body sensor measurements, including accelerometer, heart rate (HR), and electrodermal activity (EDA), using a cross-attention mechanism without requiring a preprocessing step. To enhance the long-term forecasting accuracy of the model, we propose a time-aware cross-attention mechanism capable of capturing temporal characteristics of different signals and their impact on blood glucose levels. The key contributions of our work

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA. Email: {efarahma, shovito, aarefeen, carol.johnston, hghasemz}@asu.edu.

[2]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA. Email: amachira@asu.edu

This work was supported in part by the National Science Foundation (NSF) under grant IIS-2402650. A. Arefeen was supported in part by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under award T32DK137525. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF and NIH.

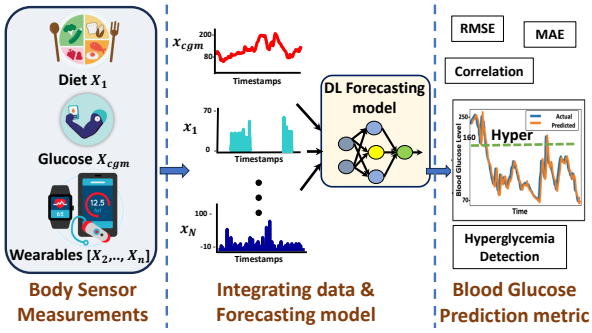

Fig. 1. Overview of our proposed framework.

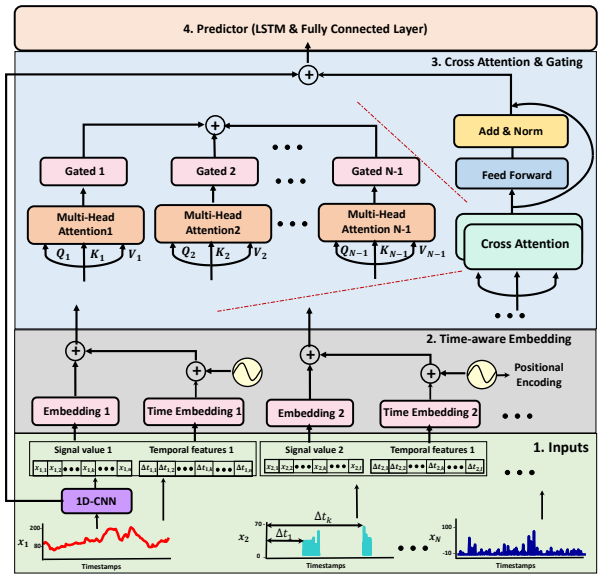

Fig. 2. Proposed Multimodal Time-aware Cross-Attention Model Architecture. It consists of (1) Inputs, (2) Time-aware Embedding, (3) Cross-Attention & Gating, and (4) Prediction steps.

are summarized as follows.

- We developed a cross-attention mechanism to integrate multimodal body sensor data, including heart rate, meal intake, EDA, and accelerometer signals, which often have irregular sampling rates.
- We introduce a time-aware cross-attention mechanism to model temporal dependencies across different physiological and behavioral signals.
- We evaluate our forecasting model on a new dataset collected from healthy individuals.

II. Proposed Model

An overview of our proposed framework is shown in Fig. 2. Our proposed framework consists of the following main modules. 1. Body Sensor Measurements: This module measures physiological and behavioral signals from wearable sensors from healthy individuals. 2. Integrating data & Forecasting model: This module integrates irregularly sampled body sensor data using a time-aware cross-attention mechanism and performs long-term BGL prediction using a 1D CNN stacked with LSTM layers. 3. Blood Glucose Prediction metric: This module generates BGL predictions and evaluates performance using established error metrics.

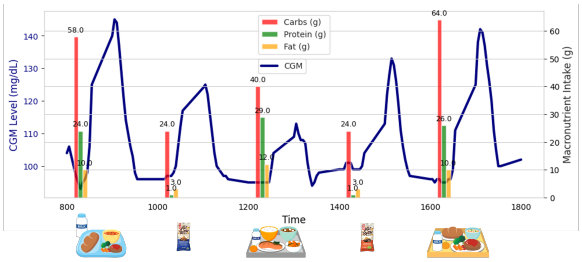

Fig. 3. An example of a CGM signal from a subject as a response to the standardized food items fed at different times.

Our forecasting model is specifically designed for predicting BGL of healthy individuals and incorporates CGM data along with various physiological and behavioral signals, including physical activity, HR, and EDA. The attention mechanism within the Transformer architecture enables the effective fusion of multi-modal time-series signals recorded at varying sampling rates and supports the modeling of long-term temporal dependencies. To evaluate the effectiveness of the proposed model, we conduct experiments using multimodal data from 12 subjects. The following sections provide further details on the forecasting problem formulation and the Integrating Data & Forecasting Model module. The dataset description and Blood Glucose Prediction Metrics module are presented in Section III.

A. Problem Formulation

The multimodal BGL forecasting task is mathematically formulated as a multivariate time-series sequential prediction downstream task. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ represent a set of n body sensor measurements module. Each sensor's data stream is represented by $\mathbf{x}_k = [x_{k,1}, \ldots, x_{k,t}]^\top$, where t indicates the duration of the observed time. Given that $\mathbf{x}_{\text{cgm}}$ corresponds to the CGM signal. Thus, the objective is to predict the future values of CGM over a defined horizon can be mathematically formulated by Eq. 1.

$$\hat{\mathbf{x}}_{cgm} = [x_{cgm,t+1}, \ldots, x_{cgm,t+ph}]^\top = \mathbf{F}(\mathbf{X}; \mathbf{\Theta}) \quad (1)$$

where $\mathbf{F}$ represents the forecasting model, which is developed using a time-aware cross attention and LSTM architecture in this paper, parameterized by Θ, which is learned during the training process.

B. Multimodal Time-aware Cross-Attention Model

The proposed forecasting model employs a multimodal time-aware cross-attention LSTM-based architecture, as shown in Fig.2. This architecture integrates a cross-attention mechanism that enables the effective fusion of multivariate time-series inputs with heterogeneous sampling rates [7]. The attention mechanism contributes to improved performance in predicting blood glucose levels. The model receives as input the historical target data (e.g., CGM measurements), denoted as $\mathbf{x}_{cgm} = [x_{cgm,ST}, \ldots, x_{cgm,ET}] \in \mathbb{R}^{T_h}$, along with other body sensor measurement signals $[\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{T_h}$. Here, T_h represents the size of the historical input window, while ST and ET indicate the start and end times of this historical period in the target signal (CGM), respectively. The

same start and end times (ST and ET) are applied to the historical windows of the other body sensor measurements. The proposed architecture consists of four main steps: (1) Inputs, (2) Time-aware Embedding, (3) Cross-Attention & Gating, and (4) Prediction.

In the Inputs step, we extract two feature vectors from each body sensor and CGM signal without applying any preprocessing, such as down-sampling or up-sampling. For each body sensor signal, the time difference between each sample point and the start time of the historical window is then computed. Therefore, for each body sensor signal k, two feature vectors are generated: (1) a signal value vector containing sample points between ST and ET, denoted as $\mathbf{x}_k = [x_{k,1}, \ldots, x_{k,l}]^\top$) where, l is the number of sample points for the kth body sensor measurement, (2) a temporal feature vector $\mathbf{\Delta T}_k = [\Delta t_{k,1}, \ldots, \Delta t_{k,l}]^\top$), where each $\Delta t_{k,1}$ is calculated as the time difference between the ith sample point and ST, i.e. $\Delta t_{k,1} = Time(x_{k,1}) - ST$. Note that the CGM data is fed into a 1D CNN to extract informative features, which are also used as residual connections to mitigate the issue of exploding gradients.

In the Time-aware Embedding step, each body sensor measurement's temporal feature is then passed through an embedding layer ($f_{T\_embed}(\mathbf{\Delta T}_k)$), and a positional encoding layer ($f_{T\_pos}(\mathbf{\Delta T}_k)$). The outputs of these two layers are added together to obtain the transformed representation of the temporal feature vector. Furthermore, body sensor measurement's signal value vector is just passed through an embedding layer ($f_{embed}(\mathbf{x}_k)$) to produce the transformed signal values feature representations. The outputs of the transformed temporal and signal values are added and fed to the Cross-Attention and gating step. The vanilla multi-head attention mechanism [6] operates by weighting the value matrix ($\mathbf{V} \in \mathbb{R}^{t \times d_{\text{model}}}$) according to the relationships between the query ($\mathbf{Q} \in \mathbb{R}^{t \times d_{\text{model}}}$) and key ($\mathbf{K} \in \mathbb{R}^{t \times d_{\text{model}}}$) matrices. The mathematical formulation of the attention mechanism is provided in Eq. 2.

$$\text{Atten}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{\text{model}}}}\right)\mathbf{V} \tag{2}$$

In the Cross-Attention and Gating step, we incorporate a cross-attention mechanism, which has demonstrated notable success in integrating information from multiple modalities across a range of domains for downstream tasks [8]. This mechanism enables different sensor-derived signals with various sample rates to exchange information and compute correlations between various body sensor measurements and the CGM data. Thus, we design an n-branch cross-attention layer in which all branches share $\mathbf{X}_{\text{cgm}}$ as the query input. In the i-the branch, the key and value matrices are obtained from $\mathbf{X}_i$. The cross-attention (CA) of the i-th branch is computed using Eqs. 3 and 4.

$$\text{CA}\left(\mathbf{X}_{\text{cgm}}, \mathbf{X}_i, \mathbf{X}_i\right) = [\mathbf{H}_1, \ldots, \mathbf{H}_{m_H}]\mathbf{W}_H^{\text{CA}} \tag{3}$$

$$\mathbf{H}_h = \text{Atten}(\mathbf{X}_{\text{cgm}}\mathbf{W}_{\mathbf{Q}}^{\text{CA}}, \mathbf{X}_i\mathbf{W}_{\mathbf{K}}^{\text{CA}}, \mathbf{X}_i\mathbf{W}_{\mathbf{V}}^{\text{CA}}) \tag{4}$$

Here, $\mathbf{W}_{\mathbf{Q}}^{\text{CA}}$, $\mathbf{W}_{\mathbf{K}}^{\text{CA}}$, and $\mathbf{W}_{\mathbf{V}}^{\text{CA}}$ are weight matrices specific to the attention head and belong to $\mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$. Moreover, $\mathbf{W}_H^{\text{CA}} \in \mathbb{R}^{(m_H \cdot d_{\text{model}}) \times d_{\text{model}}}$ is the final projection matrix that maps the concatenated outputs from all attention heads back to the original model dimension. The attention mechanisms for the remaining $n-1$ branches are computed independently using the same procedure. The attention outputs are filtered using Gated Linear Units (GLUs), which selectively retain relevant information. These are then processed by a feedforward network and an Add & Norm layer. The outputs are summed with a residual connection from the 1D-CNN and passed to a prediction module, where an LSTM followed by a fully connected layer forecasts BGLs over the prediction horizon.

TABLE I

MODEL-LEVEL PERFORMANCE. POSITIVE Δ INDICATES LOWER ERROR (RMSE/MAE) OR HIGHER CORRELATION (CORR).Δ_B: Δ VS BASELINE, Δ_G VS GLYSIM. THE BEST IMPROVEMENTS ARE HIGHLIGHTED IN BOLD.

Horizon	Metric	Baseline	GlySim	Proposed	Δ_B	Δ_G
5 min	RMSE	7.60 ± 2.85	9.92 ± 2.75	7.99 ± 2.37	-5.2%	$+19.4\%$
	MAE	5.58 ± 2.07	6.91 ± 2.31	5.90 ± 2.00	-5.7%	$+14.5\%$
	Corr	0.84 ± 0.18	0.83 ± 0.13	0.85 ± 0.11	$+1.7\%$	$+3.2\%$
30 min	RMSE	14.46 ± 3.88	13.81 ± 4.66	12.64 ± 3.88	$+12.6\%$	$+8.5\%$
	MAE	10.75 ± 3.41	10.40 ± 3.76	9.86 ± 3.55	$+8.3\%$	$+5.2\%$
	Corr	0.56 ± 0.19	0.61 ± 0.18	0.59 ± 0.18	$+5.3\%$	-3.4%
60 min	RMSE	18.17 ± 5.02	16.09 ± 5.05	15.00 ± 4.52	$+\mathbf{17.5}\%$	$+6.8\%$
	MAE	13.78 ± 4.65	11.78 ± 3.99	11.67 ± 4.11	$+15.4\%$	$+1.0\%$
	Corr	0.44 ± 0.19	0.51 ± 0.16	0.49 ± 0.19	$+11.6\%$	-3.7%
90 min	RMSE	20.13 ± 5.69	17.53 ± 5.54	16.54 ± 5.08	$+\mathbf{17.8}\%$	$+5.6\%$
	MAE	14.81 ± 4.75	12.82 ± 4.24	12.59 ± 4.37	$+15.0\%$	$+1.8\%$
	Corr	0.30 ± 0.21	0.40 ± 0.18	0.45 ± 0.21	$+\mathbf{50.3}\%$	$+\mathbf{11.2}\%$

III. RESULT

In this section, we first introduce the dataset and perform experiments on this dataset. We then present the effectiveness of our proposed model by comparing to baseline and to state-of-the-art forecasting models such as GlySim [4]. The comparison is carried out using standard evaluation metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), correlation analysis, and accuracy of hyperglycemia event detection. In this study, we consider 160 mg/dl as the threshold for a hyperglycemia event. The experimental details will be discussed in Section III-B.

A. Dataset

Data were collected as part of an ongoing study involving 12 healthy college students, aged 23 to 31 years, with BMIs ranging from 20.08 to 31.84. Participants were monitored for device-recorded responses following the consumption of standardized meals, which contained $19-126$ grams of carbohydrates, $1-54$ grams of protein, and $0-30.2$ grams of fat. Individuals recruited in the study had no diagnosed chronic metabolic or thyroid disorders, did not use recreational drugs, consumed no more than two servings of alcohol per day, and were not engaged in competitive sports or resistance training. All participants provided written informed consent. The study was approved by the Institutional Review Board at Arizona State University (IRB #15102).

The study included three non-consecutive 10-hour in-house laboratory sessions (8:00 AM–6:00 PM). Upon enrollment, participants were fitted with a Dexcom G6 CGM and an

Empatica E4 wristband worn on the dominant arm. The CGM recorded blood glucose every 5 minutes, while the E4 captured acceleration (Acc), EDA, HR, blood volume pulse (BVP), and skin temperature (TEMP) at sampling rates between 4–64 Hz. A trained staff instructed participants on proper device use.

Each session began after a 12-hour fast. Participants received meals tailored to their energy needs, calculated using the Mifflin-St Jeor equation [9], and classified as hyper-, eu-, or hypocaloric. Meals followed standard macronutrient ratios (20% protein, 55% carbs, 25% fat) and were served at 8:30 AM, 12:30 PM, and 4:30 PM, with snacks at 10:30 AM and 2:30 PM. Mealtimes were recorded, and participants responded to smartphone prompts every 30 minutes to monitor activity. Figure 3 shows the meal events and the resulting glycemic response of a subject from a random study day.

B. Baseline and Experimental Details

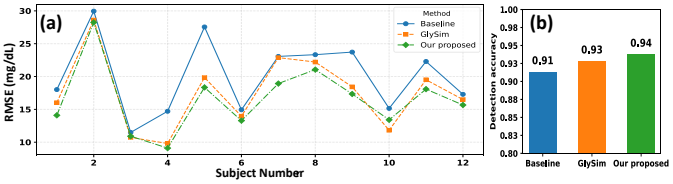

Fig. 4. Comparison of our proposed method with baseline and GlySim for 90 minutes PH in terms of (a) RMSE for each subject (b) Hyperglycemia Detection Accuracy mean across 12 subjects.

The baseline model consists of stacked 1D CNN and FC layers. It begins with a Conv1D layer projecting the input to 64 channels, followed by MaxPooling and a second Conv1D layer with 128 channels. The output is passed through two FC layers with a size of 64, and the second maps it to the desired number of outputs for prediction horizons (PH). The desired PHs in this study are 5, 30, 60, and 90 minutes. Moreover, we consider 6 hours (72 samples) as historical data for feeding to the forecasting model.

Table I presents the mean ± SD of evaluation metrics across the 12 test participants, along with the percentage change of the proposed model relative to the baseline and GlySim models. The proposed forecasting model consistently achieves the lowest RMSE and MAE, as well as higher correlation coefficients at all PHs, with the most substantial improvements observed at longer PH times. For instance, at the 90-minute PH, our forecasting model demonstrates a 17.8% improvement over the baseline and a 6.8% improvement over GlySim.

Furthermore, Figure 4a presents a comparison of 90-minute BGL forecasting for each subject using our proposed model, the baseline, and GlySim models in terms of RMSE, to highlight subject-to-subject consistency. While the absolute error varies widely across individuals, the proposed model remains lower than both baselines for nearly all twelve subjects. In fact, our forecasting model achieves the lowest RMSE for every subject when compared to the other models. In addition, we assessed each model's hyperglycemia-detection accuracy, a clinically important metric for early warning. Figure 4b shows the mean hyperglycemia-detection accuracy across all participants. The proposed model achieves the highest score (0.94), outperforming GlySim by one percentage point and the 1-D CNN baseline by three. These improved RMSE and MAE values, as reported in Table I, contribute to a reduced number of missed hyperglycemic events. Overall, the results indicate that incorporating time-aware cross-attention plays a significant role in enhancing model performance.

IV. Conclusion & Discussion

In this study, we proposed a novel glucose forecasting model utilizing a time-aware cross-attention mechanism integrated with LSTM layers, effectively addressing key limitations of combining various body sensor measurements and CGM data with different sample rates. By integrating multimodal physiological and behavioral data streams with irregular sampling rates, such as CGM readings, dietary information, heart rate, accelerometer, and EDA signals, the proposed architecture significantly improves long-term blood glucose forecasting. The cross-attention mechanism facilitates efficient fusion of heterogeneous data sources without preprocessing, while the temporal embeddings adeptly capture intricate temporal dependencies among diverse signals. Our experimental evaluation on a newly collected dataset from healthy individuals demonstrates the model's superiority over baseline methods and existing state-of-the-art frameworks in a longer prediction horizons.

References

[1] M. M. H. Shuvo and S. K. Islam, "Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1612–1623, 2023.

[2] S. D. Imrisek, M. Lee, D. Goldner, H. Nagra, L. M. Lavaysse, J. Hoy-Rosas, J. Dachis, and L. E. Sears, "Effects of a novel blood glucose forecasting feature on glycemic management and logging in adults with type 2 diabetes using one drop: retrospective cohort study," *JMIR diabetes*, vol. 7, no. 2, p. e34624, 2022.

[3] E. Farahmand, S. B. Soumma, N. T. Chatrudi, and H. Ghasemzadeh, "Hybrid attention model using feature decomposition and knowledge distillation for glucose forecasting," *arXiv preprint arXiv:2411.10703*, 2024.

[4] A. Arefeen and H. Ghasemzadeh, "Glysim: Modeling and simulating glycemic response for behavioral lifestyle interventions," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2023, pp. 1–5.

[5] R. Sergazinov, M. Armandpour, and I. Gaynanova, "Gluformer: Transformer-based personalized glucose forecasting with uncertainty quantification," in *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] E. Farahmand, R. R. Azghan, N. T. Chatrudi, E. Kim, G. K. Gudur, E. Thomaz, G. Pedrielli, P. Turaga, and H. Ghasemzadeh, "Attengluco: Multimodal transformer-based blood glucose forecasting on ai-readi dataset," *arXiv preprint arXiv:2502.09919*, 2025.

[8] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in neural information processing systems*, vol. 34, pp. 14 200–14 213, 2021.

[9] M. D. Mifflin, S. T. S. Jeor, L. A. Hill, B. J. Scott, S. A. Daugherty, and Y. O. Koh, "A new predictive equation for resting energy expenditure in healthy individuals." *The American journal of clinical nutrition*, vol. 51 2, pp. 241–7, 1990.