

Article

GluKDnet: A Lightweight Blood Glucose Prediction Model Based on Heterogeneous Knowledge Distillation

Aowei Teng, Xiaoyu Sun *, Hongru Li and Xia Yu

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; tengaowei@mails.neu.edu.cn (A.T.); lihongru@ise.neu.edu.cn (H.L.); yuxia@ise.neu.edu.cn (X.Y.)

* Correspondence: sunxiaoyu1@ise.neu.edu.cn

Abstract

Accurate blood glucose prediction is essential for glycemic management in people with diabetes, but the size of many high-performing models complicates execution on resource-constrained artificial pancreas controllers. We propose GluKDnet, a lightweight glucose-forecasting model for prospective Android-smartphone-based mobile edge controllers. GluKDnet transfers the representational capacity of a time-series foundation model to a compact causal CNN through heterogeneous knowledge distillation. The teacher model, MOMENT, is adapted to continuous glucose monitoring (CGM) data through risk-event-aware masking, which prioritizes abnormal glucose levels, rapid glucose fluctuations, and CGM-defined dawn phenomenon and Somogyi effect patterns during masked reconstruction. A transient-state and steady-state distillation module jointly aligns ordered patch-level dynamics and day-level summaries between teacher and student. Using DLCP3 for teacher pretraining and leave-one-patient-out evaluation on OhioT1DM, GluKDnet achieves RMSE values of 20.04, 32.04, and 45.33 mg/dL for 30, 60, and 120 min prediction, respectively, with about 53K parameters. Auxiliary evaluation on T1D-UoM shows a similar offline accuracy-parameter count pattern. On a vivo V2072A Android smartphone, the 30 min model achieved a mean API inference latency of 0.470 ms (P95: 0.855 ms), a maximum sampled process proportional-set-size memory of 47.06 MiB, and a median incremental device energy estimate of 0.277 mJ per inference. These device measurements characterize the exported student model under one hardware and software configuration; insulin dosing and prospective closed-loop clinical evaluation remain outside the scope of this study.

Keywords: continuous glucose monitoring; blood glucose prediction; artificial pancreas; knowledge distillation; time-series foundation model; mobile edge controller



Academic Editor: Carson K. Leung

Received: 23 June 2026

Revised: 21 July 2026

Accepted: 30 July 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\) license](#).

1. Introduction

Diabetes is a severe chronic disease where the blood glucose levels (BGLs) are consistently high because the body cannot produce enough insulin or cannot use insulin effectively. According to the International Diabetes Federation, there are approximately 537 million people living with diabetes worldwide [1]. Diabetes and its long-term complications substantially increase mortality and disability burdens. Peripheral neuropathy, macro- and micro-vascular disease, like renal and kidney diseases, are the major complications of diabetes, leading to a reduction in life quality and increasing medical costs [2]. A substantial body of evidence shows that effective glycemic control can reduce the harm caused by diabetes and its complications [3]. More broadly, the analysis and mining of medical and physiological signal data have become important for health management, supporting tasks

such as EEG-based cognitive assessment and EEG-assisted early disease detection [4,5]. With the rapid development of continuous glucose monitoring (CGM) technology [6] and the emergence of artificial pancreas systems [7], conventional clinical glycemic management is being reshaped. Artificial pancreas systems automatically adjust insulin infusion based on the real-time feedback of BGLs provided by the CGM system. Despite artificial pancreas systems having been shown to reduce hyperglycemia and hypoglycemia, and improve glycemic control, two major challenges are impeding the efficiency of the systems, including the time delay caused by the CGM system and the slow absorption of subcutaneously infused insulin [8]. To overcome the issues caused by time delays, accurate BGL prediction becomes essential for the artificial pancreas system. The prospective deployment scenario considered in this study is an Android smartphone-based mobile edge controller between the CGM and the insulin delivery workflow. In that architecture, the forecasting module provides glucose trajectories to a separate dosing and safety control layer rather than directly issuing insulin delivery commands. We evaluate retrospective predictive performance and the exported student's execution on one Android smartphone; controller integration and closed-loop clinical operation are outside the present scope.

With the accumulation of CGM data, data-driven models, including machine learning and deep learning techniques, have been increasingly used for BGL prediction. In linear models, the relationship between future BGLs, history of CGM data, and other variables is described by autoregressive, autoregressive moving average, autoregressive moving average with exogenous variables, latent variable regression, and state space models. For capturing the nonlinearity in glucose metabolism, the support vector regression and Gaussian process models project the variables into the Hilbert space before establishing the relationship between variables [9–11]. Due to the simple model structures, the ability of machine learning techniques in modeling nonlinearity is limited.

Deep learning techniques have been widely utilized in modeling complex nonlinear processes and outperform conventional machine learning methods on many tasks. Nonlinear models based on deep neural networks have been extensively investigated for forecasting future BGLs, ranging from common architectures, including feedforward neural networks, convolutional neural networks [12–14], recurrent neural networks [15–18], Transformers and their variants [19–21], to advanced architectures, like encoder–decoder, generative adversarial networks, diffusion models, and reinforcement learning techniques. The activation function enables deep neural network models to learn nonlinear relationships, and the model's flexibility is determined by the number of trainable parameters.

Large pretrained models have recently been developed for CGM, including CGMformer [21] and GluFormer [22]. They demonstrate the value of pretraining on large glucose corpora, but their scale motivates compact task-specific students when prediction must run on an edge device. General masked reconstruction also does not explicitly differentiate clinically important glycemic states. Compared with euglycemia, a forecasting teacher should be more sensitive to abnormal BGL, rapid fluctuations, and recurring nocturnal-to-morning risk patterns.

Motivated by the above considerations, this study proposed a risk-aware lightweight BGL prediction model based on heterogeneous knowledge distillation, where a risk-aware foundation model is developed as the teacher by paying attention to abnormal glucose events, and a CNN-based lightweight student model is trained to inherit useful glucose dynamics by learning from the teacher. The major contributions of this paper are as follows:

- (1) We propose a risk-event-aware masking (REAM) strategy for glucose prediction. Higher risk scores are assigned to abnormal glycemic states and then used to select masking segments during self-supervised reconstruction, encouraging the teacher model to learn risk-sensitive representations.

- (2) We decompose the feature structure into two complementary levels. Ordered patch-level knowledge represents short-term trends and event patterns, whereas day-level knowledge provides a global constraint on the overall glucose state. Their joint use transfers local transient dynamics and global daily morphology without claiming that temporal ordering is retained by global pooling alone.
- (3) We build a heterogeneous distillation framework for lightweight glucose prediction. Compact projectors map Transformer–teacher and CNN–student representations into shared alignment spaces, where supervised response, first-order-difference, patch-level, and day-level objectives are optimized jointly. The resulting 53K-parameter student is additionally benchmarked on an Android smartphone for latency, process memory usage, and incremental device energy.

The remainder of this paper is organized as follows: Section 2 summarizes related BGL forecasting models and knowledge distillation techniques. The proposed method is presented in Section 3. Prediction accuracy, component contributions, and computational measurements are investigated in Section 4, and the results are further discussed in Section 5. Conclusions are drawn in Section 6.

2. Related Work

2.1. Blood Glucose Prediction

Blood glucose prediction models can be broadly divided into physiological models, data-driven models, and hybrid models that combine both approaches [23]. Data-driven models, which learn glucose dynamics directly from historical data without requiring complex physiological parameters, have become a major research direction because they can exploit large-scale CGM data. Early studies used CGM sequences to build autoregressive and other statistical models for short-horizon glucose prediction and showed that clinically meaningful accuracy can be obtained from recent glucose history alone [24–26]. Their relatively small parameterizations and capacity for online updating motivate their consideration in subsequent embedded or portable device studies. However, purely linear assumptions often cannot fully represent the nonlinear dynamics caused by postprandial fluctuations, sensor noise, and other factors.

To capture nonlinear patterns and external disturbances, researchers have increasingly adopted neural network models. Online artificial neural networks have been used to learn short-term evolution from historical CGM sequences and perform rolling real-time prediction [27]. With the growing importance of multimodal inputs, insulin, meals, and lifestyle information have also been incorporated to improve prediction stability and clinical usability in free-living conditions [28]. Carbohydrate intake information has been systematically introduced into short-term prediction frameworks, and related work has shown that such event information can improve postprandial prediction [29].

More recently, deep sequence models and hybrid structures have been widely used to provide stronger representation capacity and model longer dependencies. Traditional and shallow learning methods, including support vector regression, wearable-processor-oriented short-term forecasting, and Gaussian processes, remain competitive in some data regimes [9–11]. Recurrent neural networks, especially LSTMs [30], have become common in glucose prediction because of their ability to model temporal dependence. For example, Rabby et al. proposed stacked LSTMs combined with Kalman smoothing to improve robustness under sensor errors [17]. CNN–LSTM hybrid structures extract local fluctuation patterns through convolutions and then model temporal dependence through recurrent layers, showing potential for longer prediction horizons [31]. Lag fusion and ensemble forecasting methods have also been proposed to address individual differences in look-back window selection for blood glucose time series [32].

With the development of attention mechanisms, Transformers have been introduced for glucose prediction to model long-range dependencies and key segments. Lee et al. used a Transformer to perform both glucose sequence prediction and hyperglycemia/hypoglycemia event prediction [20]. Transformer-based CGM prediction has also been examined on OhioT1DM under imputation and smoothing settings, showing the relevance of self-attention models for short-term blood glucose forecasting [33]. Bian et al. proposed a Transformer–LSTM hybrid model to combine global dependency modeling and local dynamics [34]. To alleviate attention cost and representation bottlenecks in long sequences, the broader time-series forecasting literature has proposed variants such as PatchTST, Crossformer, and iTransformer, which use patching, cross-dimension attention, and inverted modeling to reduce complexity and improve representation of long-range dependencies and multivariate interactions [35–37]. Broader forecasting studies have also compared linear and Transformer-style architectures, reinforcing the idea that architecture choice depends on sequence characteristics and the forecasting horizon [38]. Lu et al. followed the foundation model paradigm from natural language processing and computer vision: they pretrained CGMformer on large-scale CGM data using masked reconstruction and then transferred the learned representations to blood glucose prediction and other clinical assessment tasks through downstream fine-tuning [21].

Recent work further clarifies the scale–deployment spectrum. GluFormer establishes large-scale CGM pretraining as a general representation-learning paradigm [22], while PaPaGei develops open foundation models for optical physiological signals [39]. At the compact end, neural architecture search has been used to construct ultra-low-memory glucose forecasters for edge hardware [40], and GluLLM has examined on-device language-model-based glucose prediction [41]. These studies target either broad pretrained representations or deployment-oriented architectures. GluKDnet instead studies how risk-adapted CGM knowledge can be transferred across architectures into a causal CNN and then measures the exported student’s execution on a smartphone.

Although these advanced models perform well, they often have large architectures. In clinical and wearable settings, high parameter counts motivate the development of more compact models for subsequent device-specific evaluation. In addition, most existing models mainly optimize overall prediction accuracy and pay insufficient attention to clinically important risk regions, such as abnormal glucose levels and rapid glucose fluctuations. Parameter-compact glucose prediction models that emphasize risk-relevant regions are therefore an important step toward future evaluation in resource-constrained environments.

2.2. Knowledge Distillation

Deep learning models have achieved substantial progress in many fields, but their large scale motivates model compression before evaluation under resource constraints. Knowledge distillation is an important model compression technique. Since Bucilua et al. introduced model compression in 2006 [42], the idea has been widely studied. Hinton et al. further developed the concept by training a small “student” network under the guidance of a larger “teacher” network [43]. Recent compression work has further combined distillation with pruning to improve controllable compactness and parameter efficiency [44]. Pruning methods based on feature map similarity scores have also been proposed to identify redundant representations and reduce architectural complexity [45].

Most knowledge distillation methods can be classified by their information source into logit-based distillation [46–49], feature-based distillation [50–53], and similarity-based distillation [54–57]. Logit-based distillation encourages the student to mimic the teacher’s output distribution [43]. Feature-based distillation, as introduced in FitNets by Romero et al. [50], aligns intermediate representations rather than only logits. Some

methods further encourage the student to learn higher-order relationships among samples, channels, or classes. Similarity-based distillation transfers knowledge by preserving relative relationships, such as pairwise feature similarities [56].

Distillation can also be categorized by training paradigm. In offline distillation, the teacher is pretrained and frozen during student training; this is common when using large pretrained models [58,59]. In online distillation, teacher and student models are trained simultaneously and exchange information during training [60,61]. In self-distillation, no external teacher is required; the student learns from its own predictions or from outputs of different stages or branches [62,63]. Some studies combine logit-based and feature-based distillation. For example, Sun et al. proposed multilayer learning from a large teacher to improve training efficiency and task performance [64]. Other frameworks introduce teacher assistants to bridge the capacity gap between teacher and student models [65].

Cross-architecture distillation has recently received more explicit attention. One-for-All uses projectors to bridge heterogeneous teachers and students [66], and redundancy suppression distillation reduces incompatible information transferred between architectures [67]. In glucose forecasting, future-aware sequence-to-sequence distillation transfers future-disturbance-aware forecasting knowledge to a deployment-time student that uses historical inputs only [68]. Modular medical-AI frameworks such as HEAL-LLMF also emphasize separating large fusion or reasoning components from task-facing healthcare modules, although they address a broader multimodal decision setting rather than CGM distillation [69]. Relative to these studies, GluKDnet combines risk-aware adaptation of a time-series teacher with temporally aligned local/global feature transfer to a causal CNN; the contribution lies in this glucose-specific integration and its component-level and device-level evaluation, not in claiming that foundation models, projectors, or CNNs are individually new.

Most existing distillation methods focus on homogeneous teacher–student architectures. In these settings, both teacher and student usually share the same or similar neural structures. However, high-performing modern models often use architectures that differ substantially from compact student models. For instance, a teacher may be Transformer-based, while the student is CNN-based. Under such heterogeneous conditions, traditional homogeneous distillation may be limited or may even harm student performance because the teacher and student encode information in different feature spaces.

For this reason, we identify transient-state and steady-state knowledge in CGM signals and propose a heterogeneous knowledge distillation framework that transfers blood glucose dynamics effectively between a time-series foundation model and a lightweight CNN.

3. Methods

3.1. Overall Framework

Figure 1 shows the overall framework of GluKDnet. At each prediction time, the model uses the most recent rolling 24 h CGM history, comprising 288 readings at a five-minute sampling interval, and predicts the next H glucose values. Its prospective deployment setting is an Android smartphone-based mobile edge controller within an artificial pancreas system. In this setting, the GluKDnet student receives historical CGM readings and provides predicted glucose trajectories to a separate dosing and safety control layer, which may subsequently determine pump actions. GluKDnet is therefore a forecasting component rather than a standalone insulin-dosing controller. This study evaluates retrospective predictions and exported student inference on one Android device, but does not integrate a live CGM stream, pump controller, or clinical closed loop. The framework consists of three parts. The top branch is the teacher model, which uses the MOMENT

time-series foundation model. To make the teacher more suitable for glucose dynamics, we introduce a risk-event-aware mechanism into masked reconstruction self-supervised pre-training. For REAM annotation, CGM samples with $g_t < 70$ mg/dL and $g_t > 180$ mg/dL are labeled as hypoglycemic and hyperglycemic risk points, respectively, following the international consensus on time in range [70]. This mechanism encourages the teacher to focus on these risk-relevant regions and provides stable supervisory signals and intermediate representations for distillation. The bottom branch is the student model, a parameter-compact dilated causal CNN residual backbone. Under strict causal constraints, dilated convolutions expand the receptive field within a shallow architecture, enabling the model to represent multiscale glucose dynamics while limiting architectural complexity and trainable parameter count. Because the model uses CGM alone, the effects of meals, insulin, activity, and sleep are represented only indirectly through their signatures in the observed glucose trajectory rather than through explicit exogenous inputs. Patch-aligned embedding aligns the student's temporal coverage with the teacher's patch tokens. The middle module is hierarchical distillation: using the same historical input, the teacher and student generate both multi-step predictions and intermediate features. Patch-level representations describe local glucose fluctuation patterns, and day-level representations provide a global summary of the whole-day glucose state. During distillation, the teacher is frozen, and only the student network and its projectors are optimized. The following subsections describe data processing, the student architecture, risk-event-aware masking, and the distillation objective.

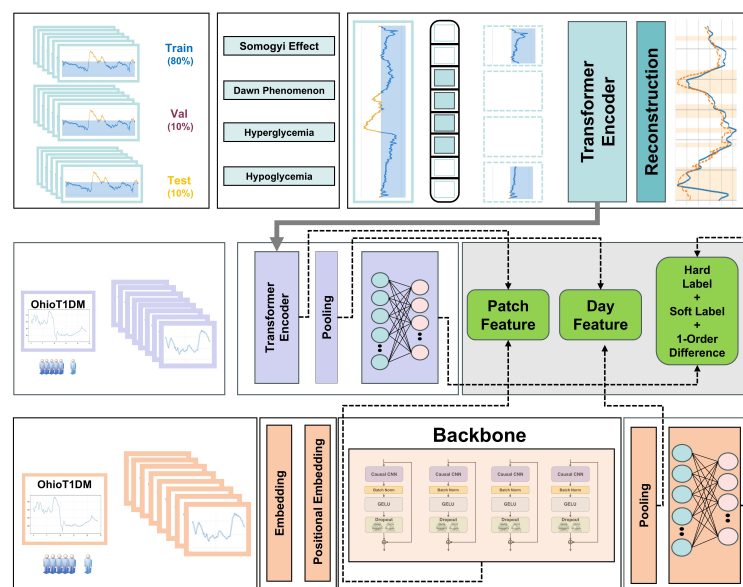


Figure 1. Overall framework of GluKDnet. The teacher branch adapts MOMENT to CGM data through risk-event-aware masked reconstruction, while the lightweight causal CNN student receives output, patch-level, and day-level supervision for multi-step prediction.

3.2. Data Preprocessing

The data used in this study are single-channel CGM glucose sequences. Readings below 40 mg/dL or above 400 mg/dL are treated as invalid and flagged as missing [71]. The reported retrospective benchmarks use the quality-controlled records of the respective public datasets; they were not designed as an online missingness experiment. For future causal operation, we specify an explicit short-gap extrapolation and timeout policy that never uses future values. If at most three consecutive five-minute samples are unavailable (a gap of at most 15 min) and at least two recent valid historical observations exist, the missing values are extrapolated linearly from the recent historical slope and clipped

to 40–400 mg/dL. If the gap exceeds 15 min, or the historical evidence is insufficient, the forecaster abstains instead of returning a potentially unsafe estimate. Prediction resumes only after valid sampling has recovered and the input buffer again satisfies the quality policy. Thus, no cubic-spline or bidirectional interpolation is assumed for online inference.

At every forecast time, the input is the most recent contiguous 24 h of CGM data available up to that time; it is therefore a rolling causal history rather than a fixed midnight-to-midnight calendar day. On the five-minute grid used in both evaluation datasets, this gives an input length of $L = 288$. We selected this evidence-informed observation unit for three related reasons. First, it contains one complete day–night cycle, so every prediction has both nocturnal-to-morning and daytime context. A shorter history is phase-dependent and may omit one of these portions of the preceding daily pattern. Second, this scale has direct precedent in glucose forecasting. CGM-LSM likewise used the preceding 24 h, comprising 288 CGM readings, to predict the subsequent 2 h and selected this window to capture daily patterns and a comprehensive view of a patient’s typical day [72]. On OhioT1DM, Ghimire et al. found strong autocorrelation between current glucose and glucose values from the preceding day and consequently adopted a 24 h rather than a 2 h history [73]; complete 288-point daily profiles have also been used as a meaningful representation unit in CGMformer [21]. Third, the full-day context provides natural temporal support for the day-level distillation objective and for REAM’s nocturnal-to-morning risk patterns. In the intended future artificial pancreas workflow, the same fixed and bounded 288-point buffer can be updated whenever a new CGM value arrives, without requiring continual accumulation of an increasingly long history. We therefore regard 24 h as a physiologically and methodologically motivated design choice, rather than as a universally optimal input length.

3.3. Risk-Event-Aware Masked Reconstruction

We use MOMENT as the teacher model in GluKDnet because MOMENT is a pretrained foundation model for general time-series analysis, and its representations can be transferred to downstream prediction tasks in different domains [74]. MOMENT first divides the input sequence into patches and uses patch tokens as the units for attention modeling. This reduces the number of tokens, lowers attention cost and memory usage, and supports effective modeling of longer historical windows. The patch embeddings are then fed into a Transformer encoder to capture long-range dependencies. During pretraining, the model randomly masks some patches and learns temporal representations through masked reconstruction.

To transfer MOMENT effectively from general time-series data to the glucose domain, we propose REAM, as shown in Figure 2. Unlike uniform random masking, REAM does not mask all daily time points uniformly under a fixed masking ratio. Instead, it preferentially masks segments that contain glucose risk events. This forces the model to learn contextual reconstruction patterns for clinically important segments and thereby obtain more meaningful domain representations. Given a rolling 24 h CGM sequence, REAM computes a risk score for each segment. The base score includes the intensity of hypoglycemia (CGM < 70 mg/dL), hyperglycemia (CGM > 180 mg/dL), and rapid glucose changes. If the sequence contains a CGM-defined pattern consistent with the dawn phenomenon or Somogyi effect, additional scores are assigned to the corresponding time windows. Segments are then ranked by risk score.

Following the international consensus on time in range [70], we use fixed CGM cutoffs throughout REAM: a sample with $g_t < 70$ mg/dL is labeled hypoglycemic, whereas a sample with $g_t > 180$ mg/dL is labeled hyperglycemic. For segment k , $r_{\text{hypo}}(k)$ and $r_{\text{hyper}}(k)$ are the proportions of samples satisfying these respective conditions. These labels are used

to prioritize masked segments during teacher pretraining; the downstream forecasting task remains continuous glucose regression. Rapid changes are used to capture steep rises or drops between adjacent samples. If the absolute change between two consecutive readings reaches or exceeds 15 mg/dL, the latter time point is marked as a rapid-change point, and the segment intensity is computed as the proportion of such points [75]. We further introduce two day-level CGM pattern rules to emphasize important morning patterns. A pattern consistent with the dawn phenomenon is defined as elevated fasting glucose in the morning without nocturnal hypoglycemia [76]. A pattern consistent with the Somogyi effect is defined as morning hyperglycemia after nocturnal hypoglycemia [77]. REAM converts these CGM-defined criteria into segment-level risk scores and selects masking positions according to the ranked scores. If one type of risk segment is insufficient, the masking budget is filled with other high-score or random segments. In this implementation, the segment risk score is

$$s_k = w_{\text{hypo}} r_{\text{hypo}}(k) + w_{\text{hyper}} r_{\text{hyper}}(k) + w_{\text{roc}} r_{\text{roc}}(k) + w_{\text{dawn}} \phi_{\text{dawn}}(k) + w_{\text{som}} \phi_{\text{som}}(k),$$

where s_k is the comprehensive risk score for segment k . The weights w_{hypo} , w_{hyper} , w_{roc} , w_{dawn} , and w_{som} control the contribution of each risk factor; in this study, we set $w_{\text{hypo}} = 0.3$, $w_{\text{hyper}} = 0.3$, $w_{\text{roc}} = 0.2$, $w_{\text{dawn}} = 0.1$, and $w_{\text{som}} = 0.1$. The terms $r_{\text{hypo}}(k)$, $r_{\text{hyper}}(k)$, and $r_{\text{roc}}(k)$ denote the proportions of hypoglycemic, hyperglycemic, and rapid-change points in segment k . The terms $\phi_{\text{dawn}}(k)$ and $\phi_{\text{som}}(k)$ are time window enhancement terms for the CGM-defined dawn phenomenon and Somogyi effect patterns. If the corresponding pattern is not detected at the day level, the term is zero for all segments. If it is detected, positive enhancement is assigned only to the defined key time window. Under a fixed masking budget, this increases the masking priority of morning risk patterns. REAM sorts segments by s_k and generates time point masks accordingly. The reconstruction mean squared error is computed only at masked positions. In downstream prediction, the pretrained teacher weights are reused to provide future predictions and intermediate features for distillation.

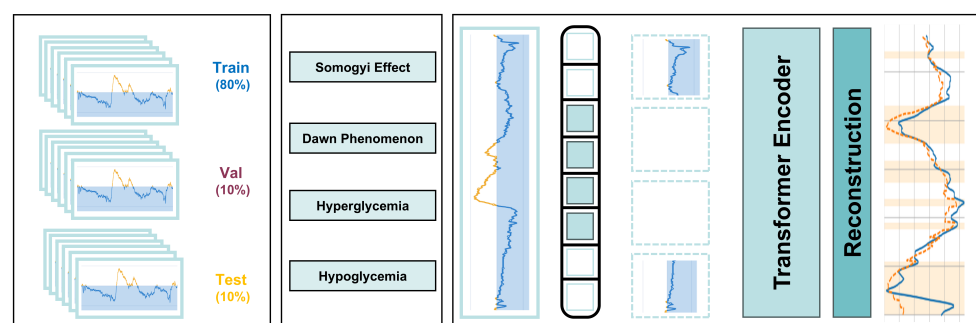


Figure 2. Risk-event-aware masking strategy for glucose domain adaptation. Daily CGM segments are scored according to hypoglycemia, hyperglycemia, rapid glucose changes, dawn phenomenon patterns, and Somogyi effect patterns before masked reconstruction in the teacher model.

REAM does not use teacher predictions or model-generated pseudo-labels to identify risk segments. The masking priorities are calculated directly from observed CGM values using the predefined glucose thresholds and trajectory rules above, while the reconstruction target at every masked position is the corresponding observed CGM value. REAM therefore changes the sampling emphasis of the masked-reconstruction task without replacing ground truth with the model's own predictions. When the predefined risk segments do not fill the masking budget, the remaining positions are drawn from other high-score or random segments, retaining exposure to non-risk background dynamics. This design avoids the

classical self-reinforcing confirmation-bias loop of pseudo-label-based self-training. It does not, however, eliminate possible prior selection bias introduced by the manually specified clinical rules and risk weights; this limitation motivates further validation of REAM under alternative rule and weighting schemes.

3.4. Student Model

The student is a lightweight patch-aligned causal CNN for multi-step CGM prediction. Given a historical input window of length L , the model predicts the next H values. As shown in Figure 3, the student contains patch embedding, an optional positional embedding, a dilated causal convolutional backbone, and a pooling plus MLP prediction head. The design learns temporal representations at the same patch granularity as the teacher while maintaining a small architectural footprint; its measured server and smartphone costs are reported in Section 4.13.

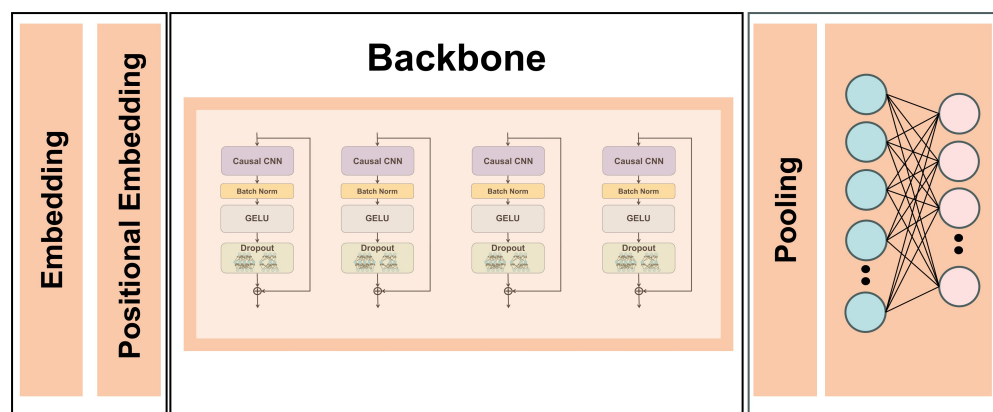


Figure 3. Patch-aligned causal CNN student model. The student embeds non-overlapping CGM patches, adds positional information, processes them with a dilated causal convolutional backbone, and pools the resulting features for multi-step glucose prediction.

The point-level CGM sequence is first divided into non-overlapping patches of length P , producing $N = L/P$ patches. A one-dimensional convolution performs patch partitioning and linear projection jointly, mapping each patch into a C -dimensional student token representation. A learnable positional embedding can be added to preserve the absolute temporal order of the patches.

The backbone is composed of stacked dilated causal convolutional residual blocks, as shown in Figure 4. Each residual block includes a main branch and a residual branch. The main branch applies causal one-dimensional convolution to extract local temporal features, followed by batch normalization, GELU activation, and dropout. Its output is added element-wise to the residual branch. Left padding enforces causality, ensuring that each output depends only on historical inputs and preventing future information leakage. Increasing dilation rates across layers enlarges the effective receptive field in a shallow network, enabling the student to capture local rapid fluctuations and medium-term trends in the observed CGM trajectory. These patterns may contain signatures consistent with meal-related rises, delayed insulin-associated declines, or circadian variation. However, meals, insulin, activity, and sleep are not model inputs, so the network infers such effects only indirectly from CGM and neither explicitly models nor causally attributes them.

The causal CNN was chosen by considering both temporal inductive bias and theoretical cost. For N patch tokens, C hidden channels, kernel size K , and B residual blocks, its approximate cost is $O(BNKC^2)$ and all temporal positions within a layer can be evaluated in parallel. A small LSTM has a comparable $O(NC^2)$ arithmetic order but retains a sequential recurrence across N steps, which can limit parallel inference. A dense Transformer requires

approximately $O(N^2C + NC^2)$, and pruning reduces parameter count without necessarily producing proportional latency gains because irregular sparsity depends on hardware and kernel support. A linear forecaster is cheaper, approximately $O(LH)$, but lacks the nonlinear hierarchical representation used here. Dilated causal convolution therefore provides a bounded state buffer, strict temporal causality, multiscale receptive fields, and regular dense operators. This comparison motivates the design; it does not imply that a CNN is universally faster than every optimized recurrent, sparse, or attention implementation.

At the output end, patch-level features generated by the backbone are pooled into a compact global representation and passed through the prediction head to produce the next H glucose values. In the knowledge transfer setting, the patch-level backbone output and pooled global representation serve as two alignment interfaces, constraining local dynamics and global morphology respectively.

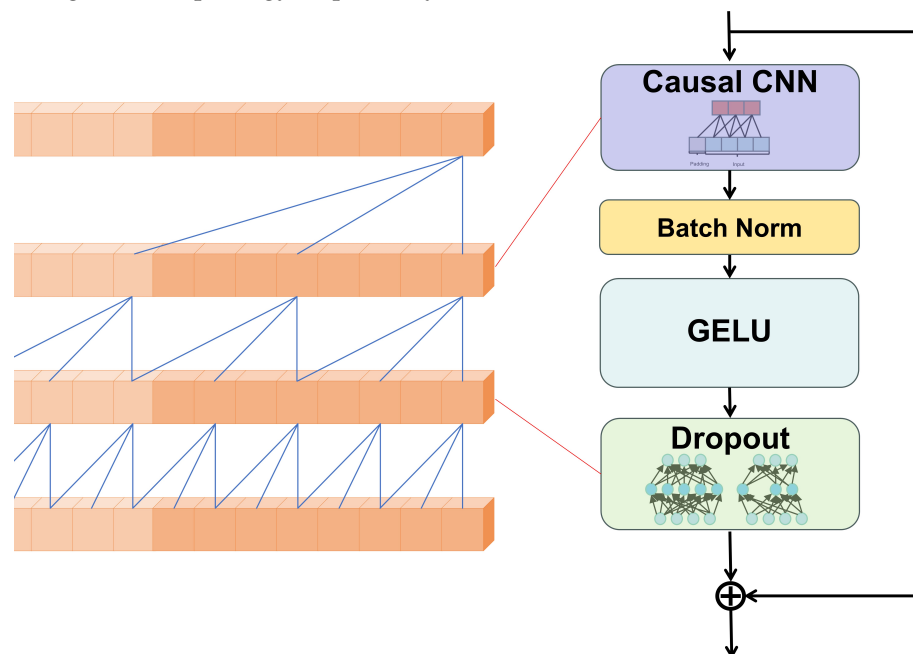


Figure 4. Dilated causal convolutional residual block. Causal convolutions preserve temporal order, dilation expands the receptive field, and the residual path combines convolution, batch normalization, GELU activation, and dropout without using future information.

3.5. Feature Selection

To represent both local dynamics and daily morphology in a rolling 24 h CGM sequence, and to provide hierarchical supervision for knowledge transfer, we design two feature levels: patch-level features and day-level features. Given an input sequence $\mathbf{x} \in \mathbb{R}^L$, the sequence is divided into $N = L/P$ continuous patches of fixed length P . A convolutional embedding maps each patch into a C -dimensional hidden space, and causal convolutions model temporal dependence across the patch sequence. This produces ordered local representations

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{C \times N},$$

where $\mathbf{z}_i$ is the contextual feature of the i -th patch and captures fine-grained patterns such as local trends, abrupt changes, and short-term fluctuations. To obtain a compact summary of the whole-day glucose curve, global average pooling is applied over the patch dimension:

$$\mathbf{g} = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_i \in \mathbb{R}^C.$$

The vector $\mathbf{g}$ is defined as the day-level feature and summarizes the overall daily glucose state. Because global averaging alone does not retain complete temporal ordering or curve shape, it is used as a global constraint rather than as a self-sufficient representation of daily morphology.

The advantage of this two-level design is that it provides complementary and alignable supervision for distillation. Patch-level features retain ordered intra-day structure and key local event information, allowing fine-grained alignment of rapid rises, drops, and high-variability regions. Day-level features provide a stable global constraint on the whole-day state. Overall morphology is therefore preserved jointly by ordered patch alignment, the day-level constraint, and first-order-difference matching. This combination avoids relying on a global average alone and reduces the instability that may arise from relying only on token-wise local alignment.

3.6. Distillation Method

The Transformer-based teacher and CNN-based student differ in network structure, feature distribution, and hidden dimensionality. The Transformer aggregates global information through attention, whereas the CNN extracts features through local convolutions and hierarchical pooling. Directly aligning hidden representations may therefore be unstable or semantically weak because the two representations lie in different feature spaces. We introduce lightweight projectors on both the teacher and student sides to map intermediate representations into a shared alignment space, enabling stable cross-architecture comparison and gradient transfer.

During distillation, all teacher parameters are frozen, and only the lightweight student and its projectors are trained. Given the same historical input $\mathbf{x} \in \mathbb{R}^L$ and future ground truth $\mathbf{y} \in \mathbb{R}^H$, the teacher and student produce multi-step predictions $\hat{\mathbf{y}}_T \in \mathbb{R}^H$ and $\hat{\mathbf{y}}_S \in \mathbb{R}^H$. The student also outputs patch-level features $\mathbf{Z}_S \in \mathbb{R}^{C \times N}$ and the day-level representation $\mathbf{g}_S \in \mathbb{R}^C$. The teacher constructs corresponding patch-level features $\mathbf{Z}_T \in \mathbb{R}^{d \times N}$ and day-level features $\mathbf{g}_T \in \mathbb{R}^d$ from Transformer encoder hidden states. Because the reported model uses only single-channel CGM sequences, the observed data contain multiscale trajectory signatures, including rapid rises, delayed declines, and recurring time-of-day patterns. Such signatures may be consistent with postprandial responses, delayed insulin action, and circadian influences, but these physiological effects are inferred only indirectly from the glucose trajectory: meal intake, insulin delivery, activity, sleep, and circadian phase are not provided as explicit inputs. Cross-patient distributions also differ substantially. A lightweight model trained only with supervised regression may drift during multi-step forecasting or produce overly smooth predictions. Therefore, we use a hierarchical distillation objective. Teacher predictions stabilize multi-step regression, and day-level and patch-level distribution alignment transfers structured representations of global states and local event segments.

Patch is used as the shared minimum alignment unit on both sides, avoiding interpolation or dynamic matching. The student uses patch-aligned embedding, and the teacher uses patch tokens, so the i -th patch covers the same time interval in both networks. This enables direct token-wise patch-level alignment, which is important for transferring representations of local glucose fluctuations and peak neighborhoods. These are trajectory-level patterns rather than explicitly annotated meal responses.

The student-side features are

$$\mathbf{Z}_S \in \mathbb{R}^{C \times N},$$

$$\mathbf{g}_S = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_{S,i} \in \mathbb{R}^C.$$

The teacher-side features are constructed from encoder hidden states:

$$\mathbf{Z}_T \in \mathbb{R}^{d \times N},$$

$$\mathbf{g}_T = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_{T,i} \in \mathbb{R}^d,$$

where $\mathbf{z}_{S,i}$ and $\mathbf{z}_{T,i}$ are the hidden representations of the i -th patch token for the student and teacher respectively.

Because the teacher and student hidden dimensions differ ($d \neq C$), and because representation scales and statistics may shift across patients, we do not force numerical feature regression. Instead, we use distribution-based representation distillation. Lightweight projectors map the teacher and student representations to the same logit-sized alignment space, and KL divergence is applied after softmax normalization. For any intermediate representation $\mathbf{F}$, the projector is defined as

$$X_{\text{pro}}(\mathbf{F}) = FC(Conv_{1 \times 1}(\mathbf{F})),$$

where $Conv_{1 \times 1}$ performs channel mapping and distribution shaping, and FC projects features to a D -dimensional logit space. The teacher and student use independent projectors $X_{\text{pro}}^T(\cdot)$ and $X_{\text{pro}}^S(\cdot)$.

In the implemented configuration, the 288-point history is divided into $N = 36$ non-overlapping patches of length $P = 8$. The teacher hidden dimension is $d = 1024$, the student hidden dimension is $C = 64$, and the projector bottleneck contains 32 channels. Four independent projectors are used: teacher-side and student-side projectors for both day-level and patch-level features. The teacher projectors use $Conv1D_{1 \times 1}(1024, 32)$ followed by $FC(32, H)$, whereas the student projectors use $Conv1D_{1 \times 1}(64, 32)$ followed by $FC(32, H)$. For day-level vectors, a singleton temporal dimension is introduced before the 1×1 convolution; for patch-level tensors, the same channel mapping and fully connected layer are applied at each of the 36 aligned patch positions. No additional nonlinear activation is used. The shared projection dimension is $D = H$, giving $D = 6, 12$, and 24 for the 30, 60, and 120 min tasks, respectively. These projectors are optimized jointly with the student during distillation but are discarded after training, so they add no parameters or operations to student-only inference.

For patch-level representations, the projected logits are

$$\mathbf{Q}_T = X_{\text{pro}}^T(\mathbf{Z}_T) \in \mathbb{R}^{D \times N},$$

$$\mathbf{Q}_S = X_{\text{pro}}^S(\mathbf{Z}_S) \in \mathbb{R}^{D \times N}.$$

Let $\mathbf{Q}^{(i)} \in \mathbb{R}^D$ denote the logits of the i -th patch token. The patch-level distillation loss is the average KL divergence across all tokens:

$$\mathcal{L}_{\text{pat}} = \frac{1}{N} \sum_{i=1}^N KL(\text{softmax}(\mathbf{Q}_T^{(i)}) \parallel \text{softmax}(\mathbf{Q}_S^{(i)})).$$

Because patches are strictly aligned, no additional alignment cost is needed. Teacher encodings of rapid glucose changes and other local dynamics can therefore be transferred directly to the student, reducing shape bias in high-variability segments. For day-level features, the projected logits are

$$\mathbf{q}_T = X_{\text{pro}}^T(\mathbf{g}_T) \in \mathbb{R}^D,$$

$$\mathbf{q}_S = X_{\text{pro}}^S(\mathbf{g}_S) \in \mathbb{R}^D.$$

The day-level distillation loss is

$$\mathcal{L}_{\text{day}} = KL(\text{softmax}(\mathbf{q}_T) \parallel \text{softmax}(\mathbf{q}_S)).$$

This term aligns the student's global temporal representation distribution with that of the teacher and improves stable modeling of long-range dynamics across patients.

In the output space, teacher predictions are used as soft supervision for multi-step regression. We also introduce first-order-difference distillation to align the rate of change of the prediction curves. This is important because CGM sequences are noisy and often contain peak-to-valley transitions. Matching only prediction values can make the student overly smooth, whereas a rate-of-change constraint encourages the student to match dynamic trends and reduces peak timing shifts and trend lag.

The supervised regression loss is

$$\mathcal{L}_{\text{sup}} = \frac{1}{H} \sum_{h=1}^H (\hat{y}_{S,h} - y_h)^2.$$

The output distillation loss uses SmoothL1:

$$\mathcal{L}_{\text{kd}} = \frac{1}{H} \sum_{h=1}^H \rho(\hat{y}_{S,h} - \hat{y}_{T,h}),$$

where $\rho(\cdot)$ is the SmoothL1 function. Defining the first-order difference as $\Delta\hat{y}_{\cdot,h} = \hat{y}_{\cdot,h+1} - \hat{y}_{\cdot,h}$, the difference distillation loss is

$$\mathcal{L}_{\text{d1}} = \frac{1}{H-1} \sum_{h=1}^{H-1} \rho(\Delta\hat{y}_{S,h} - \Delta\hat{y}_{T,h}).$$

The final training objective is

$$\mathcal{L} = \lambda_{\text{sup}} \mathcal{L}_{\text{sup}} + \lambda_{\text{kd}} \mathcal{L}_{\text{kd}} + \lambda_{\text{d1}} \mathcal{L}_{\text{d1}} + \lambda_{\text{day}} \mathcal{L}_{\text{day}} + \lambda_{\text{pat}} \mathcal{L}_{\text{pat}},$$

where λ_{sup} , λ_{kd} , λ_{d1} , λ_{day} , and λ_{pat} are the weights of the corresponding loss terms.

4. Experiments

4.1. Datasets

We used three public CGM datasets with distinct roles. DLCP3 was used only for masked-reconstruction pretraining of the teacher model. OhioT1DM was used as the primary downstream benchmark for blood glucose prediction training and evaluation. T1D-UoM was used as an auxiliary evaluation dataset to examine whether the offline accuracy-parameter count pattern observed on OhioT1DM also appears in a broader public CGM cohort. The three cohorts were collected independently, are mutually disjoint, and have no known patient overlap.

The DLCP3 dataset is from the six-month randomized multicenter iDCL trial [78]. The study enrolled 168 participants with type 1 diabetes aged 14–71 years and followed them for six months. The control group received sensor-augmented insulin pump therapy, and the primary endpoint was based on the percentage of time in range measured by CGM.

The OhioT1DM dataset [79], released by Marling and Bunescu, is widely used for blood glucose prediction. It contains eight weeks of data from six participants with type 1 diabetes (IDs 559, 563, 570, 575, 588, and 591). It provides CGM readings sampled every five minutes and multiple event streams strongly related to glucose fluctuations. Although OhioT1DM includes these additional event streams, all reported experiments

used only the five-minute CGM glucose sequence as model input; the event streams were not provided to GluKDnet or the baseline models.

The T1D-UoM dataset [80], released by the University of Manchester team, was used in this manuscript as an auxiliary evaluation dataset for type 1 diabetes management and glucose prediction. It contains approximately 12 weeks of longitudinal real-world data from 17 people with type 1 diabetes. In addition to CGM values, it includes basal and bolus insulin doses, nutrition intake, physical activity, and sleep-related information. To keep the comparison consistent with the single-channel setting used for OhioT1DM and the student model, the present experiments used only CGM glucose sequences as model inputs; the additional insulin, nutrition, activity, and sleep modalities were not used in the reported models.

4.2. Evaluation Metrics

We evaluate prediction performance using numerical error and clinical acceptability metrics. Mean absolute error (MAE) and root mean squared error (RMSE) quantify prediction error:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|,$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2},$$

where $\hat{y}_i$ is the predicted value, y_i is the corresponding ground truth, and N is the total number of prediction points.

We also use mean absolute relative difference (MARD), reported as a percentage:

$$MARD = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{y}_i - y_i|}{\max(|y_i|, \epsilon)} \times 100\%,$$

where ϵ is a small constant to avoid numerical instability when the denominator is too small.

To evaluate trend fitting, we use the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the ground truth values. A higher R^2 indicates stronger explanatory power.

For clinical acceptability, we use Clarke Error Grid (CEG) analysis [81]. The CEG divides the plane formed by predicted and true glucose values into five zones, as illustrated in Figure 5. Zone A indicates clinically accurate predictions that do not affect treatment decisions. Zone B indicates small errors that may lead only to benign or noncritical treatment adjustments. Zone C indicates predictions that may cause unnecessary or excessive treatment. Zone D indicates missed dangerous glucose events. Zone E indicates severe errors that may lead to dangerous or wrong treatment decisions. Predictions in Zones A and B are usually considered clinically acceptable.

4.3. Experimental Settings

To evaluate generalization to unseen patients, we use leave-one-patient-out (LOPO) cross-validation [82]. For a dataset containing N patients, the model is trained on $N - 1$ patients and evaluated on the remaining patient. This process is repeated N times so that

each patient is used once as the held-out test subject, providing a realistic assessment of cross-patient generalization.

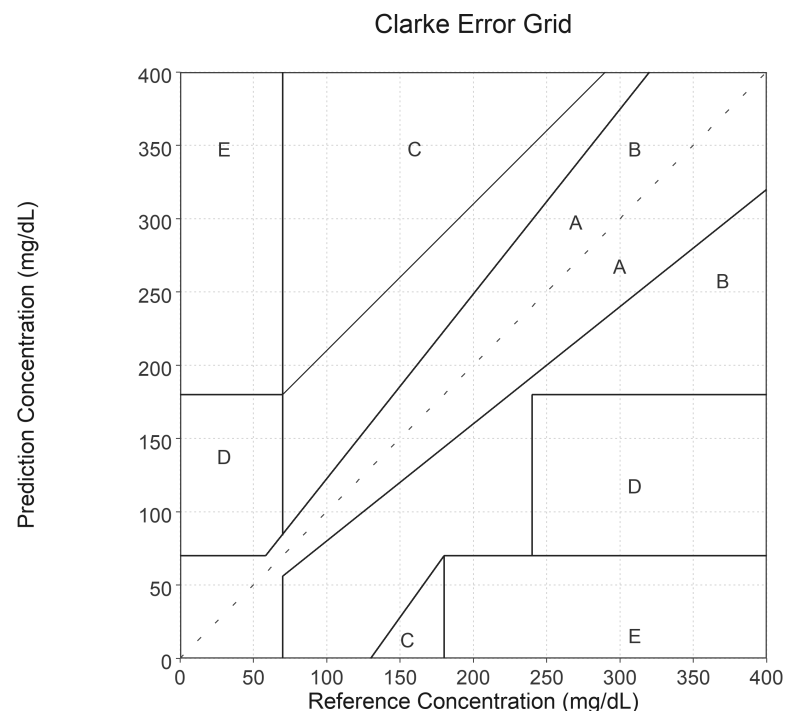


Figure 5. Clarke Error Grid zones used for clinical safety evaluation. Predicted glucose concentrations are plotted against reference concentrations and assigned to Zones A–E, which represent clinically accurate, benign, unnecessary-treatment, missed-risk, and dangerous-treatment regions.

Transfer from DLCP3 to the downstream datasets was not performed as direct zero-shot inference. In every OhioT1DM and T1D-UoM LOPO fold, the MOMENT teacher was initialized from the DLCP3 REAM masked-reconstruction checkpoint and then adapted using only the $N - 1$ non-held-out patients from the corresponding downstream dataset. The target patient was excluded from fold-specific teacher adaptation, student training, validation-based model selection, and hyperparameter selection. This two-stage procedure separates the roles of the datasets: DLCP3 provides self-supervised initialization from generic CGM trajectory structure, whereas fold-specific downstream adaptation calibrates the teacher to the sensor, sampling, and cohort characteristics observable in the training patients of OhioT1DM or T1D-UoM. The reported T1D-UoM results therefore provide auxiliary evidence that the transferred initialization remains useful after downstream adaptation; they should not be interpreted as proof of demographic or sensor-invariant zero-shot generalization.

Model development and same-server timing were performed on a shared laboratory server equipped with an Intel Xeon Platinum 8280 CPU at 2.70 GHz and an NVIDIA L40S GPU. The code used Python 3.11.13, PyTorch 2.0.1, CUDA 11.7, and scikit-learn 1.7.1. In addition, exported FP32 students were benchmarked on a vivo V2072A smartphone running Android 13 using ExecuTorch 1.3.1. Server and smartphone protocols are reported separately because their latency and memory measurements are not directly interchangeable.

In the teacher masked-reconstruction stage, the initial learning rate was 0.0001, training lasted 50 epochs, batch size was 64, and AdamW was used as the optimizer. The masking ratio was 0.30. Automatic mixed precision and gradient clipping were enabled, and the best validation checkpoint was saved for downstream prediction.

For fold-specific downstream adaptation, the MOMENT teacher was initialized from the masked-reconstruction checkpoint and fully fine-tuned for at most 30 epochs with

batch size 256. AdamW used learning rates of 2×10^{-5} for the encoder and embedder and 2×10^{-4} for the prediction head. The supervised-only causal CNN was trained for at most 50 epochs, whereas the distilled student was trained for at most 75 epochs; both used a batch size of 1024, AdamW, an initial learning rate of 10^{-3} , and validation-based checkpoint selection. During distillation, teacher parameters were frozen. The total loss combined supervised regression, output response distillation, first-order-difference distillation, day-level feature distillation, and patch-level feature distillation. The loss weights were set to $\lambda_{\text{sup}} = 1.0$, $\lambda_{\text{kd}} = 0.7$, $\lambda_{\text{d1}} = 0.02$, $\lambda_{\text{day}} = 0.5$, and $\lambda_{\text{pat}} = 0.5$.

The numerical magnitudes of these coefficients are balancing factors rather than directly comparable measures of component importance because the constituent losses differ in definition, normalization, and gradient scale. The supervised MSE loss is the primary ground-truth regression objective, and $\lambda_{\text{sup}} = 1.0$ serves as the reference. In contrast, $\mathcal{L}_{\text{d1}}$ applies SmoothL1 to adjacent prediction differences. Differencing emphasizes local high-frequency variations and CGM noise; therefore, $\lambda_{\text{d1}} = 0.02$, which is 50-fold smaller than λ_{sup} , keeps this term as a weak trend-preserving regularizer and limits competition with absolute-glucose regression. The patch- and day-level losses use KL divergence to align softmax-normalized projected feature distributions. Setting $\lambda_{\text{pat}} = \lambda_{\text{day}} = 0.5$ provides intermediate-strength local and global representation guidance without allowing heterogeneous feature matching to dominate the supervised objective. The empirical sensitivity of these settings is evaluated in Section 4.10. A 10-epoch linear warm-up was applied to the distillation terms to improve stability. Other hyperparameters were tuned on the validation set, and the best checkpoint of each fold was used for testing.

4.4. Baseline Models

We selected representative deep learning baselines covering recurrent, convolutional, and self-attention architectures. All baselines were retrained and tested under the same data splits, preprocessing pipeline, and evaluation metrics to ensure fair comparison.

Specifically, CNN-LSTM [31] extracts local temporal features through one-dimensional convolutional layers and then uses a two-layer LSTM and fully connected layers for multi-step prediction. Transformer [20] models historical CGM sequences through multi-head self-attention and supports both glucose regression and hyperglycemia/hypoglycemia event classification. GluNet [12] is a dilated causal convolutional model that transforms glucose change ΔG into discrete classes before inferring future glucose values. Gluformer [83] uses a Transformer backbone to model future glucose trajectories as an infinite mixture distribution and outputs both predictions and uncertainty estimates. BGformer [19], which is based on Informer, introduces a feature enhancement module and an encoder–decoder structure to capture local and global temporal dependencies.

We followed the original architecture and key module designs of these baselines. Because of dataset characteristics and the objectives of this study, some experimental configurations were unified, including input dimensions, historical window length, prediction horizon, and preprocessing procedure. These changes affect only input/output settings and do not alter the core modeling mechanisms.

4.5. Experimental Results

We compared GluKDnet with five representative deep learning baselines: CNN-LSTM, Transformer, GluNet, Gluformer, and BGformer. For the main OhioT1DM evaluation, the five-minute sampling interval means that horizons of 6, 12, and 24 time steps correspond to 30, 60, and 120 min. All methods used the same most recent rolling 24 h CGM history, namely 288 time steps, as input. The T1D-UoM sequences were also represented at a five-minute sampling interval and evaluated with the same 24 h input and 6-, 12-, and 24-step horizon

definitions. Table 1 reports the primary OhioT1DM results, and Table 2 reports the auxiliary T1D-UoM evaluation under the same 30, 60, and 120 min horizon naming convention.

Overall, prediction errors increase and R^2 decreases as the prediction horizon becomes longer, reflecting the greater uncertainty of long-horizon glucose forecasting. Although different baselines achieve competitive results on some individual metrics, GluKDnet provides a favorable overall offline accuracy–parameter count trade-off. On the primary OhioT1DM benchmark, GluKDnet achieves the lowest RMSE across the 30, 60, and 120 min horizons, while maintaining competitive MAE, MARD, and R^2 values. On the auxiliary T1D-UoM dataset, GluKDnet also shows a similar accuracy–parameter count pattern, although its advantage is less uniform across all metrics and patient-level variability remains substantial.

Table 1. Primary performance on the OhioT1DM dataset. Values report mean $\pm$ standard deviation under leave-one-patient-out evaluation for 30, 60, and 120 min prediction horizons; bold values indicate the best result within each horizon and metric.

Horizon	Model	RMSE (mg/dL)	MAE (mg/dL)	MARD (%)	R^2
30 min	CNN-LSTM	22.40 $\pm$ 3.52	15.33 $\pm$ 1.37	9.73 $\pm$ 1.85	0.8414 $\pm$ 0.02
	Transformer	21.68 $\pm$ 2.52	15.74 $\pm$ 1.59	10.55 $\pm$ 2.40	0.8464 $\pm$ 0.03
	GluNet	21.59 $\pm$ 2.14	15.91 $\pm$ 1.30	10.62 $\pm$ 2.24	0.8473 $\pm$ 0.03
	Gluformer	20.70 $\pm$ 2.44	14.64 $\pm$ 1.44	9.83 $\pm$ 2.24	0.8594 $\pm$ 0.03
	BGformer	21.78 $\pm$ 2.81	15.77 $\pm$ 1.81	10.60 $\pm$ 2.58	0.8466 $\pm$ 0.03
	GluKDnet	20.04 $\pm$ 2.49	15.13 $\pm$ 1.75	10.17 $\pm$ 2.59	0.8545 $\pm$ 0.03
60 min	CNN-LSTM	35.06 $\pm$ 4.44	25.69 $\pm$ 3.01	16.55 $\pm$ 3.04	0.6101 $\pm$ 0.03
	Transformer	34.16 $\pm$ 3.19	25.77 $\pm$ 2.59	17.88 $\pm$ 4.39	0.6192 $\pm$ 0.08
	GluNet	34.45 $\pm$ 2.95	25.63 $\pm$ 2.08	17.09 $\pm$ 3.68	0.6105 $\pm$ 0.09
	Gluformer	33.52 $\pm$ 3.26	25.18 $\pm$ 2.33	17.30 $\pm$ 3.98	0.6352 $\pm$ 0.07
	BGformer	33.96 $\pm$ 3.91	25.54 $\pm$ 3.07	17.75 $\pm$ 4.47	0.6231 $\pm$ 0.08
	GluKDnet	32.04 $\pm$ 3.07	23.88 $\pm$ 2.19	16.24 $\pm$ 3.74	0.6630 $\pm$ 0.07
120 min	CNN-LSTM	50.54 $\pm$ 7.14	38.99 $\pm$ 6.51	25.32 $\pm$ 4.63	0.1860 $\pm$ 0.10
	Transformer	49.59 $\pm$ 4.45	38.74 $\pm$ 3.47	27.10 $\pm$ 5.73	0.2052 $\pm$ 0.11
	GluNet	51.84 $\pm$ 3.93	39.60 $\pm$ 2.96	27.06 $\pm$ 5.68	0.1208 $\pm$ 0.17
	Gluformer	48.71 $\pm$ 4.47	38.05 $\pm$ 3.77	26.38 $\pm$ 6.25	0.2358 $\pm$ 0.09
	BGformer	48.88 $\pm$ 4.18	38.38 $\pm$ 3.49	27.04 $\pm$ 5.85	0.2274 $\pm$ 0.11
	GluKDnet	45.33 $\pm$ 3.92	35.68 $\pm$ 3.13	25.16 $\pm$ 5.73	0.3350 $\pm$ 0.10

To further analyze the RMSE-oriented trade-off between model complexity and prediction performance, we plotted model parameter counts against RMSE values, as shown in Figure 6. The horizontal axis represents parameter count on a logarithmic scale, and the vertical axis represents mean RMSE and standard deviation under the 30, 60, and 120 min horizons. On this RMSE-focused comparison, GluKDnet lies in the region with the smallest parameter count and the lowest RMSE across the three OhioT1DM horizons. With only about 53K parameters, GluKDnet achieves the lowest RMSE on the primary OhioT1DM benchmark at all three horizons, even compared with much larger models such as Transformer (about 1.09M parameters) and BGformer (about 663K parameters). This result supports the main RMSE advantage, but it should be interpreted together with the MAE, MARD, and R^2 results, where several baselines remain competitive. Through heterogeneous knowledge distillation, GluKDnet compresses the model substantially while retaining key temporal representations. Parameter count remains an architectural measure rather than a substitute for execution evidence; Section 4.13 therefore reports separate server and Android measurements.

4.6. Patient-Specific Evaluation

To examine patient-level behavior beyond average metrics, we report the 30 min results for all six held-out OhioT1DM participants in Table 3 and Figure 7, followed by the

best and most challenging cases. This fine-grained analysis is important because cohort averages hide substantial inter-patient variability.

Table 2. Auxiliary performance on the T1D-UoM dataset. Values report mean $\pm$ standard deviation for RMSE, MAE, MARD, and R^2 across 30, 60, and 120 min prediction horizons; bold values indicate the best result within each horizon and metric.

Horizon	Model	RMSE (mg/dL)	MAE (mg/dL)	MARD (%)	R^2
30 min	CNN-LSTM	29.54 $\pm$ 12.12	21.99 $\pm$ 9.56	16.59 $\pm$ 7.55	0.6085 $\pm$ 0.251
	Transformer	30.22 $\pm$ 12.30	22.34 $\pm$ 9.32	16.89 $\pm$ 7.13	0.5919 $\pm$ 0.260
	GluNet	31.76 $\pm$ 12.77	23.51 $\pm$ 9.50	17.45 $\pm$ 7.06	0.5471 $\pm$ 0.286
	Gluformer	29.40 $\pm$ 12.25	21.95 $\pm$ 9.67	16.71 $\pm$ 7.65	0.6100 $\pm$ 0.248
	BGformer	31.65 $\pm$ 11.99	23.39 $\pm$ 8.67	17.64 $\pm$ 6.47	0.5547 $\pm$ 0.279
	GluKDnet	29.25 $\pm$ 12.33	21.99 $\pm$ 9.79	16.54 $\pm$ 7.47	0.6119 $\pm$ 0.243
60 min	CNN-LSTM	40.94 $\pm$ 13.87	31.29 $\pm$ 11.22	23.83 $\pm$ 8.81	0.3044 $\pm$ 0.308
	Transformer	41.79 $\pm$ 13.79	31.79 $\pm$ 10.84	24.14 $\pm$ 8.34	0.2783 $\pm$ 0.320
	GluNet	45.09 $\pm$ 15.91	33.99 $\pm$ 12.20	25.36 $\pm$ 9.07	0.1464 $\pm$ 0.409
	Gluformer	40.88 $\pm$ 13.54	30.99 $\pm$ 10.88	23.83 $\pm$ 8.52	0.3189 $\pm$ 0.296
	BGformer	42.34 $\pm$ 13.77	32.61 $\pm$ 10.72	24.89 $\pm$ 8.01	0.2576 $\pm$ 0.327
	GluKDnet	40.10 $\pm$ 13.98	31.15 $\pm$ 11.45	23.70 $\pm$ 8.83	0.3127 $\pm$ 0.304
120 min	CNN-LSTM	48.99 $\pm$ 13.17	38.24 $\pm$ 10.68	29.46 $\pm$ 8.40	0.0657 $\pm$ 0.230
	Transformer	49.56 $\pm$ 13.63	38.67 $\pm$ 10.75	29.73 $\pm$ 8.32	0.0476 $\pm$ 0.236
	GluNet	57.40 $\pm$ 17.60	43.61 $\pm$ 13.29	32.64 $\pm$ 9.77	0.0256 $\pm$ 0.421
	Gluformer	48.41 $\pm$ 12.86	37.83 $\pm$ 10.38	29.14 $\pm$ 7.78	0.0877 $\pm$ 0.218
	BGformer	49.46 $\pm$ 13.44	38.55 $\pm$ 10.54	29.45 $\pm$ 7.78	0.0506 $\pm$ 0.227
	GluKDnet	47.70 $\pm$ 13.37	38.02 $\pm$ 11.02	29.03 $\pm$ 8.29	0.0774 $\pm$ 0.229

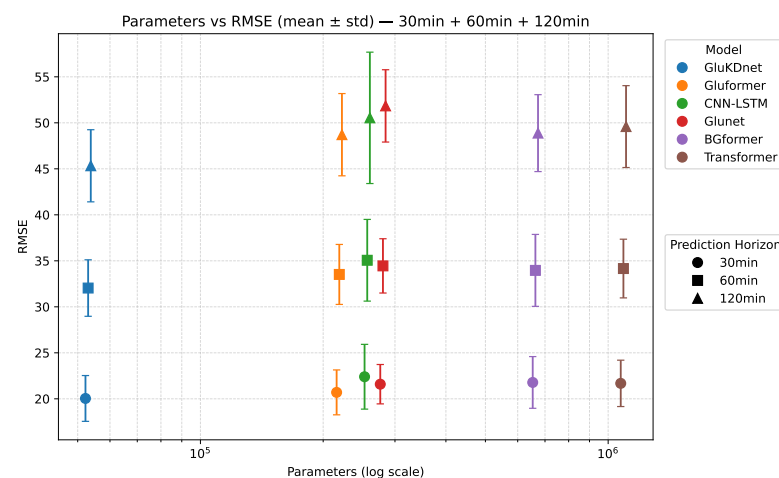


Figure 6. Trade-off between parameter count and RMSE. The horizontal axis shows model size on a logarithmic scale, the vertical axis reports RMSE, marker shapes denote prediction horizons, and error bars indicate standard deviations across patients.

Table 3. GluKDnet performance for every held-out OhioT1DM participant in the 30 min task.

PID	RMSE (mg/dL)	MAE (mg/dL)	MARD (%)	CEG Zone D (%)
559	20.67	16.15	11.11	1.85
563	18.27	14.22	9.09	0.92
570	17.92	12.87	6.54	0.36
575	25.38	18.38	14.60	5.69
588	19.13	14.22	8.27	0.00
591	18.86	14.97	11.44	2.70

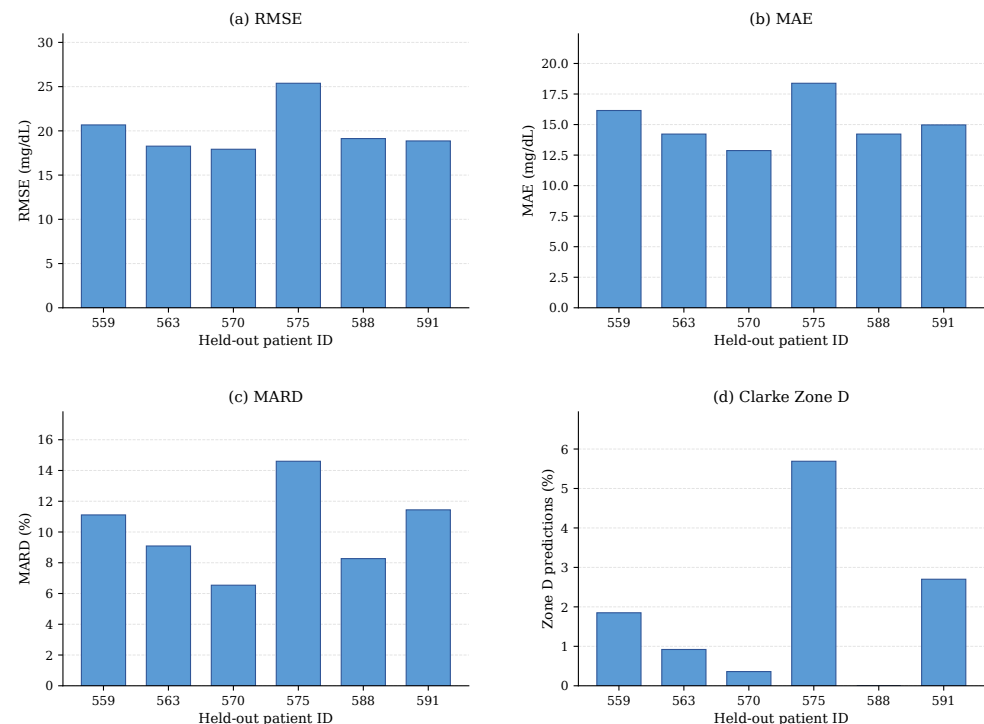


Figure 7. Patient-level numerical performance of GluKDnet for all six OhioT1DM participants in the 30 min task.

The patient-specific analysis further shows that the advantage of GluKDnet is not uniform across all individuals. In the best-case sequence, as shown in Figure 8 and Table 4, GluKDnet achieves the strongest performance across all four metrics, with an RMSE of 17.92 mg/dL. The curve comparison shows that GluKDnet closely follows the main intra-day trends and maintains smaller amplitude errors and time lags in rapidly changing intervals. This suggests that the distilled lightweight student can closely follow intra-day glucose dynamics when the held-out patient distribution is well aligned with the learned representations.

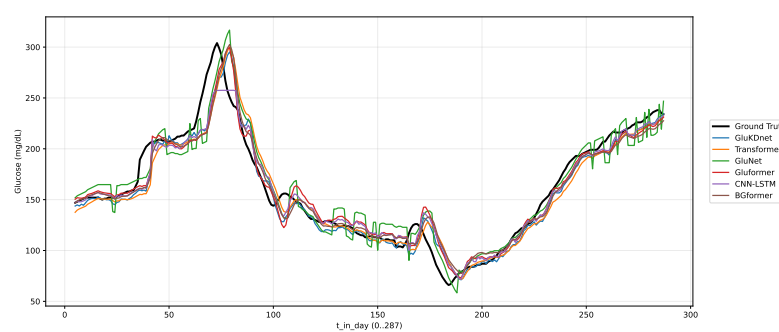
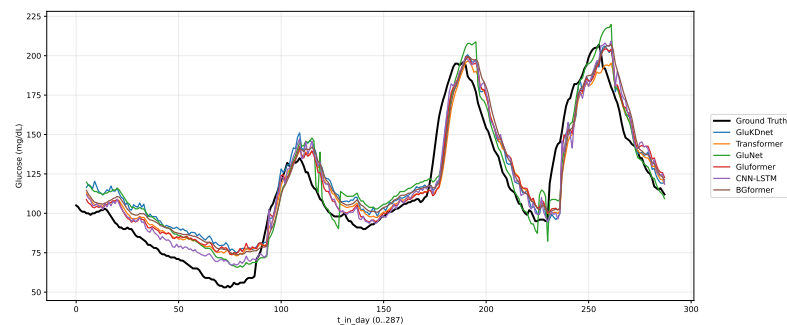


Figure 8. Best-case patient-level prediction comparison. Ground truth CGM values are compared with predictions from GluKDnet and five baselines across one daily sequence, illustrating the case in which GluKDnet most closely follows intra-day glucose dynamics.

However, in the challenging case shown in Figure 9 and Table 5, CNN-LSTM obtains a lower RMSE than GluKDnet, and several models show larger deviations in low-glucose and rapidly changing intervals. GluKDnet performs at a similar level to several baselines, with an RMSE of 25.38 mg/dL, whereas CNN-LSTM achieves a lower RMSE of 23.57 mg/dL. These results indicate that GluKDnet improves the overall offline accuracy-parameter count trade-off, but its cross-patient robustness remains limited for difficult individuals with stronger physiological or behavioral variability.

Table 4. Best-case patient-level performance. The table reports RMSE, MAE, MARD, and R^2 for the same representative best-case sequence; bold values denote the best metric values.

Model	RMSE (mg/dL)	MAE (mg/dL)	MARD (%)	R^2
CNN-LSTM	23.25	15.99	7.36	0.8691
Transformer	18.84	13.68	6.97	0.9140
GluNet	19.05	14.31	7.28	0.9121
Gluformer	18.33	13.16	6.70	0.9072
BGformer	19.38	14.15	7.17	0.9090
GluKDnet	17.92	12.87	6.54	0.9222

**Figure 9.** Challenging patient-level prediction comparison. Ground truth CGM values are compared with predictions from GluKDnet and five baselines for a difficult daily sequence with larger deviations in low-glucose and rapidly changing intervals.**Table 5.** Challenging patient-level performance. The table reports RMSE, MAE, MARD, and R^2 for the corresponding difficult daily sequence; bold values denote the best metric values.

Model	RMSE (mg/dL)	MAE (mg/dL)	MARD (%)	R^2
CNN-LSTM	23.57	16.34	12.20	0.8504
Transformer	25.93	18.45	14.05	0.8190
GluNet	25.33	18.29	13.92	0.8272
Gluformer	24.50	16.84	12.94	0.8384
BGformer	25.92	18.61	14.42	0.8191
GluKDnet	25.38	18.38	14.60	0.8266

Overall, under strict cross-patient evaluation, performance is strongly affected by individual differences. GluKDnet shows a favorable accuracy–complexity trade-off in the best case, but its cross-patient generalization still needs improvement for challenging individuals.

4.7. Clinical Safety Evaluation

We used Clarke Error Grid analysis to evaluate the clinical acceptability of GluKDnet predictions. Patient-level analysis is important for understanding stability and risk distribution. Table 6 and Figure 10 show the CEG distribution for six patients in the 30 min task, while Figures 11 and 12 extend the analysis to 60 and 120 min.

Table 6. Patient-level Clarke Error Grid distribution for the 30 min prediction task. Values are percentages of predictions in each clinical zone for each OhioT1DM participant; Zones A and B are generally considered clinically acceptable.

PID	Zone A (%)	Zone B (%)	Zone C (%)	Zone D (%)	Zone E (%)
559	87.56	10.32	0.26	1.85	0
563	90.65	8.44	0	0.92	0

Table 6. Cont.

PID	Zone A (%)	Zone B (%)	Zone C (%)	Zone D (%)	Zone E (%)
570	95.42	4.22	0	0.36	0
575	74.09	20.22	0	5.69	0
588	91.51	8.49	0	0	0
591	83.98	13.32	0	2.70	0

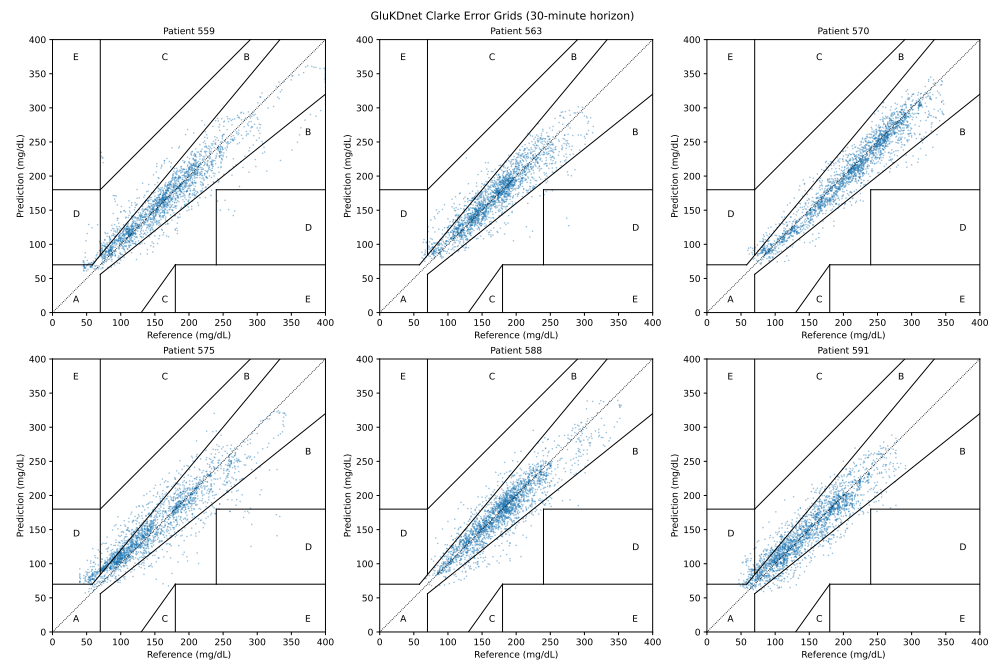


Figure 10. Patient-level Clarke Error Grid analysis for the 30 min prediction task. Each panel shows predicted versus reference glucose values for one OhioT1DM participant. The dashed diagonal indicates perfect agreement; Zones A–E denote the clinical error regions.

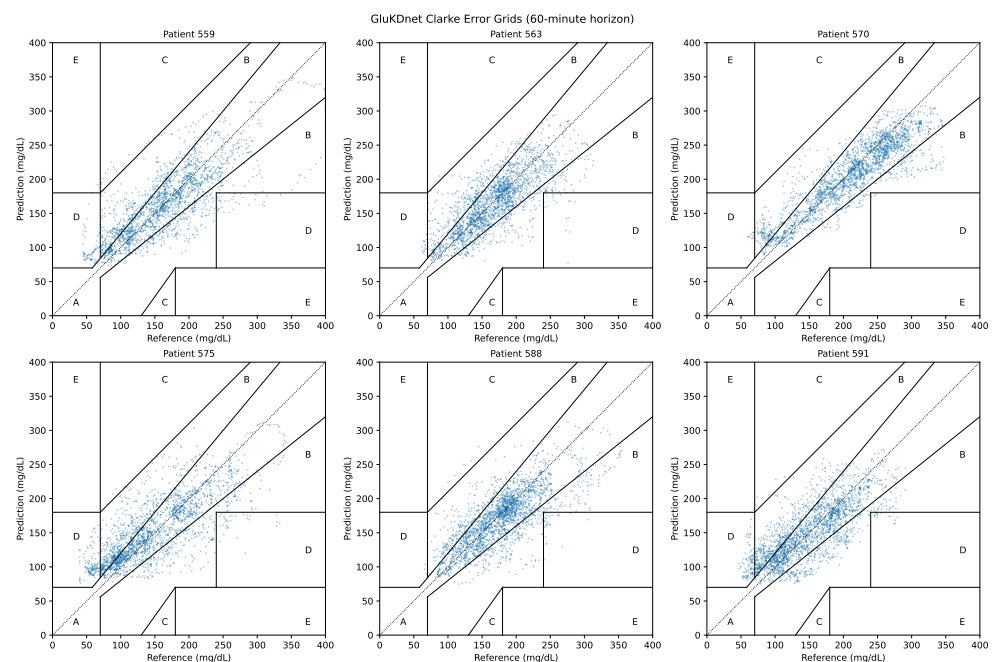


Figure 11. Patient-level Clarke Error Grid analysis for the 60 min prediction task. The dashed diagonal indicates perfect agreement; Zones A–E denote the clinical error regions.

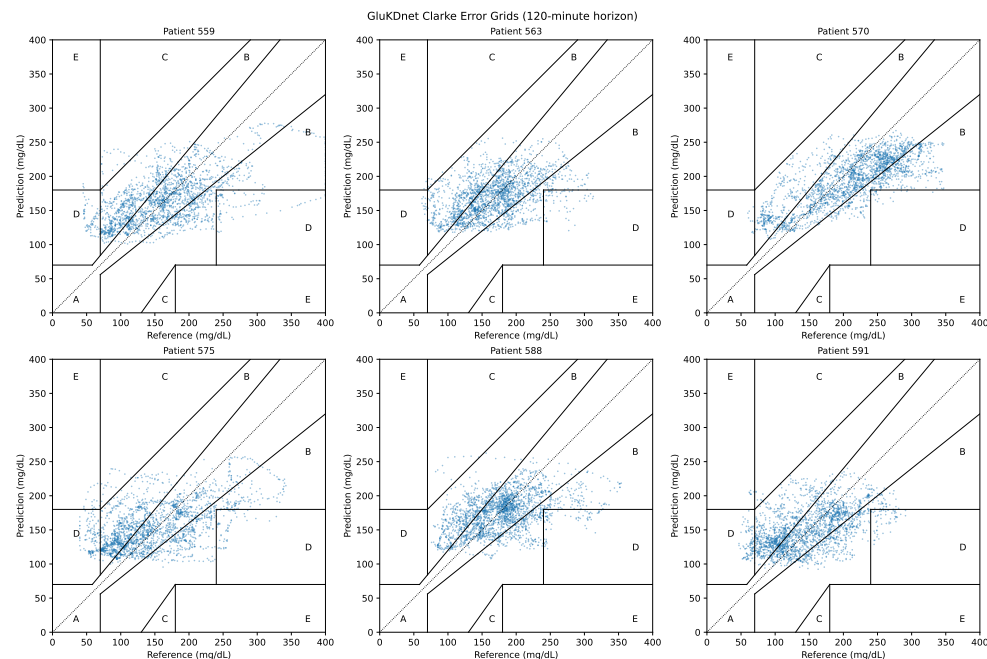


Figure 12. Patient-level Clarke Error Grid analysis for the 120 min prediction task. The dashed diagonal indicates perfect agreement; Zones A–E denote the clinical error regions.

The results show substantial differences in clinical prediction accuracy across patients. At 30 min, patient 570 has 95.42% of predictions in Zone A, 4.22% in Zone B, only 0.36% in Zone D, and none in Zone E. Patients 575 and 591 are more challenging: their Zone D proportions are 5.69% and 2.70%, respectively. As the horizon increases, the plots show a visible migration from Zone A toward Zones B and D, consistent with the larger numerical errors at 60 and 120 min. Thus, clinical acceptability is horizon- and patient-dependent, and the nonzero Zone D fractions preclude a claim of uniformly safe forecasting.

4.8. Ablation Study

The preceding experiments show the overall performance of GluKDnet. To isolate the contribution of each training component, we tested the complete model; removal of REAM, response distillation $\mathcal{L}_{kd}$, first-order-difference distillation $\mathcal{L}_{d1}$, patch-level distillation $\mathcal{L}_{pat}$, day-level distillation $\mathcal{L}_{day}$, and selected combinations; and a supervised-only PatchCNN. All configurations used the same splits and evaluation protocol. Table 7 reports RMSE at all three horizons.

The complete model achieves the lowest RMSE at all three horizons. Removing $\mathcal{L}_{pat}$ increases 30 and 60 min RMSE by 0.21 and 0.77 mg/dL, whereas removing $\mathcal{L}_{day}$ increases 120 min RMSE from 45.33 to 46.89 mg/dL, supporting the role of the global constraint at longer horizons. Removing $\mathcal{L}_{kd}$ gives 21.48, 33.92, and 47.35 mg/dL; removing $\mathcal{L}_{d1}$ causes smaller but consistent increases. The supervised-only PatchCNN is worst at every horizon (23.76, 36.42, and 50.61 mg/dL), showing that the gains cannot be attributed to the student architecture alone. Removing REAM also degrades performance, and the combined removals show complementary contributions rather than a single dominant module.

Table 7. Component ablation on OhioT1DM. RMSE is reported for 30-, 60-, and 120 min prediction under the same LOPO protocol; bold values indicate the lowest RMSE at each horizon.

Configuration	RMSE 30 min	RMSE 60 min	RMSE 120 min
Full GluKDnet	20.04	32.04	45.33
w/o $\mathcal{L}_{pat}$	20.25	32.81	45.56
w/o $\mathcal{L}_{day}$	20.39	32.23	46.89

Table 7. Cont.

Configuration	RMSE 30 min	RMSE 60 min	RMSE 120 min
w/o $\mathcal{L}_{\text{pat}}$ and $\mathcal{L}_{\text{day}}$	22.16	34.95	48.77
w/o REAM	20.54	33.38	45.96
w/o REAM, $\mathcal{L}_{\text{pat}}$, and $\mathcal{L}_{\text{day}}$	22.95	35.40	49.38
w/o $\mathcal{L}_{\text{kd}}$	21.48	33.92	47.35
w/o $\mathcal{L}_{\text{d1}}$	20.18	32.29	45.51
Supervised-only PatchCNN	23.76	36.42	50.61

Figure 13 provides an implementation illustration of REAM, showing how risk-focused masked regions and reconstructed trajectories are generated for representative CGM sequences.

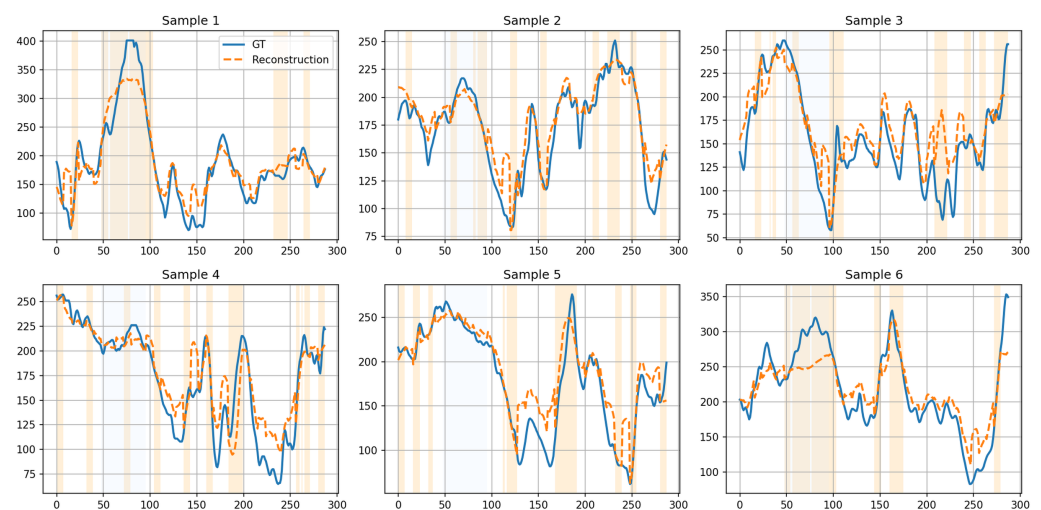


Figure 13. Implementation illustration of the risk-event-aware masking strategy. Example CGM sequences show risk-focused masked regions and reconstructed trajectories, illustrating how REAM prioritizes clinically informative segments during teacher adaptation.

4.9. Teacher Model Selection

Table 8 compares three candidate teacher models: MOMENT [74], CGMformer [21], and Moirai [84]. This selection experiment was conducted on OhioT1DM for the 30 min prediction horizon, using the same LOPO split, preprocessing pipeline, and evaluation protocol as the main experiments. The candidate teacher models were evaluated without downstream fine-tuning. MOMENT achieves the lowest RMSE in this selection setting. Its pretrained representation capacity can also provide higher-quality feature information for the student. Therefore, MOMENT is selected as the teacher model.

Table 8. Teacher model selection results on OhioT1DM for 30 min prediction. Candidate teacher models are compared using the same split and preprocessing protocol; bold indicates the lowest RMSE.

Model	RMSE (mg/dL)
MOMENT	22.04
CGMformer	25.67
Moirai [84]	23.39

4.10. Loss Weight Sensitivity Analysis

We conducted a one-factor-at-a-time sensitivity analysis on OhioT1DM for the 30 min prediction horizon under the same six-fold LOPO protocol as the main evaluation. In each experiment, only one distillation weight was varied while all remaining weights were fixed at their adopted values; $\lambda_{\text{sup}} = 1.0$ was fixed throughout. RMSE was used as

the focused evaluation metric because it is the primary numerical error criterion in this study and penalizes large glucose prediction errors more strongly. Figure 14 reports the LOPO-averaged RMSE for every tested setting.

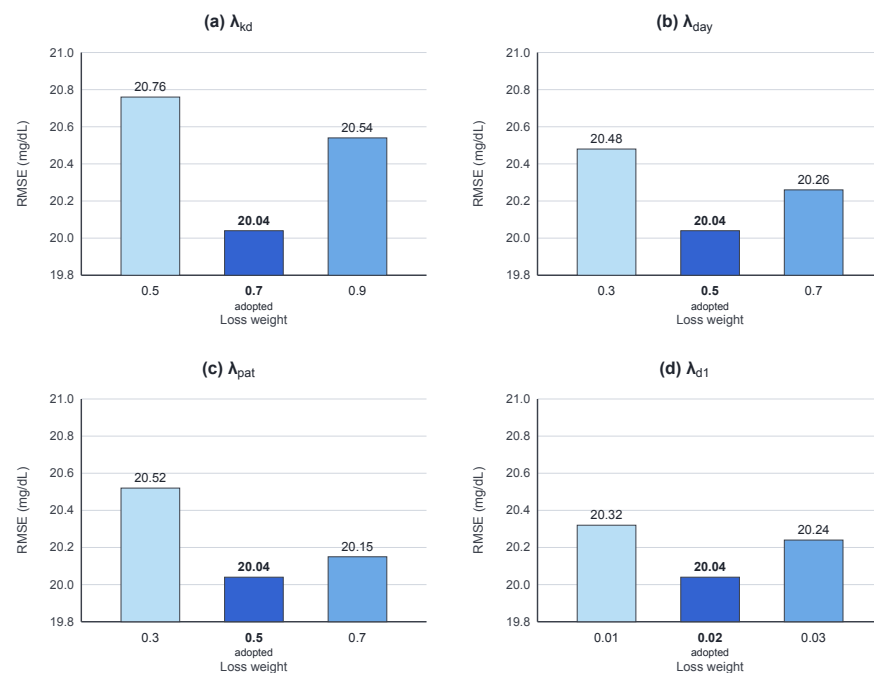


Figure 14. One-factor-at-a-time sensitivity analysis of the four distillation loss weights on OhioT1DM for 30 min glucose prediction. Only the indicated weight is varied in each panel, while the remaining weights are fixed at the adopted values. The adopted values are shown in dark blue. Lower RMSE indicates better performance.

For $\lambda_{kd} = 0.5, 0.7$, and 0.9 , the RMSE values were $20.76, 20.04$, and 20.54 mg/dL, respectively. For $\lambda_{day} = 0.3, 0.5$, and 0.7 , the corresponding values were $20.48, 20.04$, and 20.26 mg/dL. For $\lambda_{pat} = 0.3, 0.5$, and 0.7 , the values were $20.52, 20.04$, and 20.15 mg/dL. Finally, $\lambda_{d1} = 0.01, 0.02$, and 0.03 produced RMSE values of $20.32, 20.04$, and 20.24 mg/dL. Thus, the adopted configuration achieved the lowest observed RMSE in every examined discrete candidate range. In particular, the small first-order-difference weight of 0.02 outperformed both 0.01 and 0.03 , while the feature-distillation weight of 0.5 outperformed both 0.3 and 0.7 for the day- and patch-level terms. Across all alternatives, the RMSE increase relative to the adopted configuration ranged from 0.11 to 0.72 mg/dL, indicating limited but observable local sensitivity. These results support the adopted settings within the tested grids; they do not establish a global optimum over the continuous coefficient space or statistically significant differences between closely spaced settings.

4.11. REAM Weight Sensitivity Analysis

We additionally varied each REAM risk score weight one factor at a time on OhioT1DM for the 30 min horizon, while retaining all other weights at the adopted values. The weights scale a ranking score and are not mixture probabilities; therefore, a perturbed setting is not renormalized to sum to one. Table 9 shows that the adopted setting has the lowest RMSE among the tested discrete candidates. The largest increase is 0.19 mg/dL and changes to the dawn and Somogyi terms alter RMSE by only 0.03 – 0.05 mg/dL, so this is evidence of local robustness rather than proof of a unique or global optimum.

Table 9. One-factor-at-a-time sensitivity of the REAM risk score weights on OhioT1DM for 30 min prediction. Bold indicates the lowest RMSE.

Configuration	w_{hypo}	w_{hyper}	w_{roc}	w_{dawn}	w_{som}	RMSE (mg/dL)	Δ RMSE
Adopted	0.30	0.30	0.20	0.10	0.10	20.04	–
Hypo—Low	0.20	0.30	0.20	0.10	0.10	20.23	+0.19
Hypo—High	0.40	0.30	0.20	0.10	0.10	20.16	+0.12
Hyper—Low	0.30	0.20	0.20	0.10	0.10	20.18	+0.14
Hyper—High	0.30	0.40	0.20	0.10	0.10	20.12	+0.08
ROC—Low	0.30	0.30	0.10	0.10	0.10	20.14	+0.10
ROC—High	0.30	0.30	0.30	0.10	0.10	20.10	+0.06
Dawn—Low	0.30	0.30	0.20	0.05	0.10	20.08	+0.04
Dawn—High	0.30	0.30	0.20	0.15	0.10	20.07	+0.03
Somogyi—Low	0.30	0.30	0.20	0.10	0.05	20.09	+0.05
Somogyi—High	0.30	0.30	0.20	0.10	0.15	20.08	+0.04

4.12. Distillation Loss and Teacher Error Audit

To examine whether the student simply reproduces teacher errors, we conducted a post hoc six-fold audit for the 30 min task using the 75-epoch KD configuration. Table 10 reports fold means and sample standard deviations of the raw losses and their weighted contributions. The output response term is the largest distillation contribution, whereas the first-order and feature terms act as small regularizers after weighting; the combined distillation contribution remains much smaller than the supervised term.

Table 10. Six-fold loss audit for the 30 min KD model. Values are mean $\pm$ sample standard deviation.

Term	Raw Loss	Weighted Contribution
$\mathcal{L}_{\text{sup}}$	0.830389 $\pm$ 0.263866	0.830389 $\pm$ 0.263866
$\mathcal{L}_{\text{kd}}$	0.097690 $\pm$ 0.029459	0.068383 $\pm$ 0.020622
$\mathcal{L}_{\text{d1}}$	0.009882 $\pm$ 0.003103	0.000198 $\pm$ 0.000062
$\mathcal{L}_{\text{day}}$	0.000407 $\pm$ 0.000151	0.000204 $\pm$ 0.000076
$\mathcal{L}_{\text{pat}}$	0.000676 $\pm$ 0.000374	0.000338 $\pm$ 0.000187
Combined distillation	–	0.069122 $\pm$ 0.020755
Total objective	–	0.899511 $\pm$ 0.283206

Within each held-out fold, we then selected the 20% of samples with the largest teacher absolute errors; selection was performed only for this diagnostic and did not affect training. On this difficult subset, the teacher, supervised-only student, and KD student had RMSE values of 29.181 ± 4.209 , 30.975 ± 5.024 , and 29.020 ± 5.108 mg/dL, respectively. The KD student improved over the supervised-only student in all six folds and over the teacher at the fold-aggregate level in three of six folds. The teacher error correction fraction—the proportion of selected samples for which the KD student’s absolute error was lower than the teacher’s—was $60.00\% \pm 3.85\%$. These observations do not prove that teacher error transfer or overfitting is impossible, but they provide no evidence of simple systematic copying under the audited protocol.

4.13. Computational and On-Device Evaluation

We first compared the frozen teacher, distilled student, and supervised-only student on the same L40S server. Table 11 reports six-fold mean $\pm$ sample standard deviation. The distilled and supervised students have nearly identical inference costs, while the teacher is approximately 27 times slower by mean latency and uses about 202 times more peak GPU memory. The projectors and teacher are training-only components and are discarded for student inference.

Table 11. Teacher–student execution comparison on the same L40S server. Values are mean $\pm$ sample standard deviation across six folds.

Model	Mean Latency (ms)	P95 Latency (ms)	Peak GPU Memory (MiB)
Student (KD)	0.889 $\pm$ 0.093	0.925 $\pm$ 0.095	9.960 $\pm$ 0.003
Student (supervised only)	0.885 $\pm$ 0.093	0.920 $\pm$ 0.092	9.960 $\pm$ 0.003
MOMENT teacher	24.227 $\pm$ 2.354	25.324 $\pm$ 3.046	2013.781 $\pm$ 0.000

Observed wall-clock training times on the shared laboratory server were approximately 6 h for teacher adaptation, approximately 1 h for the complete six-fold supervised student run (50 epochs), and 5721.77 s (95.36 min, or 1.59 h) for the complete six-fold KD student rerun (75 epochs; 15.89 min per fold on average). The latter implies about 0.59 h additional observed wall time relative to the supervised run, but it includes 25 more allowed epochs and was not a controlled matched-budget speed experiment. Moreover, the server was shared by students and staff, so contention makes these times descriptive rather than reproducible hardware benchmarks. Distillation and teacher adaptation are one-time offline costs; neither the teacher nor the projectors is required after the student is exported.

For mobile evaluation, three FP32 student models for the 30, 60, and 120 min horizons were exported to ExecuTorch 1.3.1 and executed with XNNPACK/portable backends and one CPU thread on a vivo V2072A (MediaTek MT6891Z/CZA, Android 13/API 33, 11.27 GB physical RAM). A real 288-point window from OhioT1DM participant 563 was used as input. After 50 warm-up calls, 1000 API forward calls were timed for each model. Table 12 reports the latency distributions and maximum sampled process proportional set size (PSS).

Table 12. Android student model measurements on a vivo V2072A. PSS includes the application and runtime, not only model tensors.

Horizon	Mean (ms)	P50 (ms)	P95 (ms)	P99 (ms)	Max Sampled PSS (MiB)
30 min ($H = 6$)	0.470	0.431	0.855	0.980	47.06
60 min ($H = 12$)	0.155	0.149	0.174	0.211	46.71
120 min ($H = 24$)	0.158	0.146	0.170	0.335	48.45

The non-monotonic horizon timings should not be interpreted as a model capacity trend: the three models share the same backbone and differ only in a small output head, while Android scheduling and runtime execution vary between benchmark runs. In a separate saturated continuous-throughput run, 1,081,600 executions of the 30 min model completed in 150 s, corresponding to $150 \times 1000 / 1,081,600 \approx 0.139$ ms per inference after overhead amortization. We retain the 0.470 ms API mean as the primary single-call latency and report 0.139 ms only as a throughput-derived equivalent.

Energy was measured for the 30 min model in five paired, alternating 30-s idle/active stages using Android BatteryManager charge counter and voltage readings, with the idle baseline subtracted. The median incremental device energy estimate was 0.276721 mJ per inference (IQR: 0.275641–0.277427 mJ; five of five valid pairs). This is a device-level incremental estimate and cannot isolate the model from all operating system or application activity. Only the exported students were executed on Android; the much larger MOMENT teacher was retained on the server as an offline training component.

5. Discussion

GluKDnet combines three established ideas—foundation model adaptation, cross-architecture knowledge distillation, and a causal CNN student—for a specific

glucose-forecasting objective. Its contribution is the integrated design: REAM introduces glucose risk priorities during teacher adaptation; temporally aligned patch and day representations transfer complementary local and global information; and the final student is evaluated through component ablations, error propagation diagnostics, and device measurements. This positioning differs from large CGM foundation models that prioritize representation breadth [21,22], direct edge architecture search [40], and recent forecast output distillation [68]. The individual primitives are not claimed as new.

Under strict LOPO validation on OhioT1DM, GluKDnet achieves the lowest RMSE among the compared baselines at 30, 60, and 120 min with about 53K parameters. Its advantage is not uniform across every MAE, MARD, or R^2 comparison, so the result is best interpreted as a favorable overall accuracy–size trade-off. The expanded ablation supports complementary roles for all training components. In particular, removing the day-level constraint increases 120 min RMSE by 1.56 mg/dL, while the much larger deterioration of the supervised-only PatchCNN shows that performance is not explained by the student backbone alone. The loss and REAM weight studies identify the adopted values as the best within their tested discrete grids, but the small differences for some candidates do not establish statistical separation or a continuous global optimum.

The execution results provide direct, configuration-specific evidence beyond parameter count. On the same L40S server, the KD student and supervised student have nearly identical costs, whereas the MOMENT teacher is substantially slower and more memory intensive. On the vivo V2072A, all three exported students execute with sub-millisecond mean API latency and maximum sampled process PSS below 49 MiB; the 30 min model has a median baseline-corrected device energy estimate of 0.277 mJ per inference. These measurements support feasibility on the tested phone and software stack, but they do not automatically transfer to other processors, runtime versions, thread counts, precisions, background workloads, or battery states.

Patient-level results remain heterogeneous. Participant 570 has an RMSE of 17.92 mg/dL and 0.36% of 30 min predictions in CEG Zone D, whereas participant 575 has an RMSE of 25.38 mg/dL and 5.69% in Zone D. The 60 and 120 min CEG plots also show worsening clinical acceptability with horizon. Thus, low average RMSE and fast execution do not imply uniform patient-level reliability. The teacher-high error audit similarly supports only a bounded conclusion: the KD student improves over the supervised-only student in all six audited folds and corrects a majority of teacher-difficult samples, but it does not prove that teacher error propagation or overfitting can never occur.

The training time observations show that distillation adds offline cost—the 75-epoch six-fold KD rerun took 1.59 h versus approximately 1 h for the 50-epoch supervised run—in addition to teacher adaptation. This is not a matched-epoch speed comparison and was collected on a shared server. The extra cost is nevertheless separated from inference because the teacher and projectors are discarded after training. This separation is aligned with the intended role of a large offline teacher and a small task-facing student, but future studies should use controlled, matched-budget measurements before making a stronger training efficiency claim.

Limitations and Future Work

This study has several limitations. First, the downstream cohorts remain small: OhioT1DM contains six participants and T1D-UoM contains 17. Although the three datasets are mutually disjoint and LOPO prevents held-out-patient leakage, generalization to larger, more diverse populations, sensor brands, and acquisition protocols remains uncertain. Second, the model uses CGM alone. T1D-UoM also contains insulin, meals, activity, and sleep, but these modalities were intentionally excluded to preserve a common compact input.

Their physiological effects are inferred only indirectly from the glucose trajectory and cannot be interpreted as explicit or causal modeling.

Third, the 24 h history was selected a priori to provide one complete day–night cycle and support day-level distillation and nocturnal-to-morning REAM patterns. We did not rerun the model with 4, 8, or 12 h histories, so the performance loss caused by shortening the window has not been quantified and 24 h is not established as universally optimal. Fourth, the Android results cover one phone, one runtime, FP32 inference, one thread, retrospective input windows, and short benchmark sessions. PSS includes application/runtime memory, and BatteryManager provides a baseline-corrected device-level estimate rather than model-exclusive energy. Prospective live-CGM streaming, communication failures, sustained thermal behavior, multiple devices, and other precision/runtime configurations require further study.

Fifth, the short-gap extrapolation and timeout rule specifies how future online inference should respond to missing samples, but its effect on predictive accuracy and abstention frequency has not yet been evaluated prospectively. Sixth, the model has not been connected to an insulin dosing and safety controller. Any artificial pancreas use would require independent dosing logic, fault handling, prospective validation, and regulatory evaluation. Finally, REAM's fixed thresholds, pattern rules, and risk weights encode clinical priors. The local sensitivity analysis reduces concern about modest weight perturbations but does not remove possible selection bias or establish validity in every population.

6. Conclusions

We propose GluKDnet, a risk-aware heterogeneous distillation framework that transfers MOMENT representations into a 53K-parameter causal CNN. On OhioT1DM, the student achieves RMSE values of 20.04, 32.04, and 45.33 mg/dL at 30, 60, and 120 min, and auxiliary T1D-UoM results show a similar accuracy–size pattern. Component ablations support the complementary contributions of REAM and the four distillation terms. On a vivo V2072A, the 30 min student achieves 0.470 ms mean API latency, 47.06 MiB maximum sampled process PSS, and 0.277 mJ median incremental device energy per inference. These findings support GluKDnet as a compact forecasting candidate under the tested configuration. Broader cohorts, shorter-window comparisons, multimodal inputs, multi-device and sustained online tests, and integration with an independently validated dosing and safety layer remain necessary.

Author Contributions: A.T. developed the method and the draft, H.L. revised the manuscript and supervised the project, X.S. revised the manuscript and analyzed the results, X.Y. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The National Natural Science Foundation of China (Grant 62403114), and the Fundamental Research Funds for the Central Universities (Grant N2404030 and N25ZLH009).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the following repositories: the DCLP3 dataset is available from the Jaeb Center for Health Research at <https://public.jaeb.org/dataset/573> (accessed on 15 April 2026; see Reference [78]); the OhioT1DM dataset is available from Ohio University at <https://webpages.charlotte.edu/rbunescu/ohiot1dm.html> (accessed on 15 April 2026; see Reference [79]); and the T1D-UoM dataset is available from Zenodo at <https://doi.org/10.5281/zenodo.15806142> (accessed on 15 April 2026; see Reference [80]).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BGL	Blood glucose level
CGM	Continuous glucose monitoring
CEG	Clarke error grid
CNN	Convolutional neural network
KD	Knowledge distillation
LOPO	Leave-one-patient-out
MAE	Mean absolute error
MARD	Mean absolute relative difference
REAM	Risk-event-aware masking
RMSE	Root mean squared error
T1D	Type 1 diabetes

References

1. Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.N.; Mbanya, J.C.; et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119. <https://doi.org/10.1016/j.diabres.2021.109119>.
2. Merid, F.; Getahun, F.; Esubalew, H.; Gezahegn, T. Diabetic microvascular complications and associated factors in patients with type 2 diabetes in Southern Ethiopia. *Front. Endocrinol.* **2024**, *15*, 1342680. <https://doi.org/10.3389/fendo.2024.1342680>.
3. Khunti, K.; Zaccardi, F.; Amod, A.; Aroda, V.R.; Aschner, P.; Colagiuri, S.; Mohan, V.; Chan, J.C.N. Glycaemic control is still central in the hierarchy of priorities in type 2 diabetes management. *Diabetologia* **2025**, *68*, 17–28. <https://doi.org/10.1007/s00125-024-06254-w>.
4. Gramouseni, F.; Tzimourta, K.D.; Angelidis, P.; Giannakeas, N.; Tsipouras, M.G. Cognitive Assessment Based on Electroencephalography Analysis in Virtual and Augmented Reality Environments, Using Head Mounted Displays: A Systematic Review. *Big Data Cogn. Comput.* **2023**, *7*, 163. <https://doi.org/10.3390/bdcc7040163>.
5. Santarrosa-López, I.; Alor-Hernández, G.; Bustos-López, M.; Hernández-Capistrán, J.; Sánchez-Morales, L.N.; Sánchez-Cervantes, J.L.; Marín-Vega, H. DETEC-ADHD: A Data-Driven Web App for Early ADHD Detection Using Machine Learning and Electroencephalography. *Big Data Cogn. Comput.* **2025**, *9*, 3. <https://doi.org/10.3390/bdcc9010003>.
6. Kwon, S.Y.; Moon, J.S. Advances in Continuous Glucose Monitoring: Clinical Applications. *Endocrinol. Metab.* **2025**, *40*, 161–173. <https://doi.org/10.3803/enm.2025.2370>.
7. de Visser, H.S.; Waraich, S.; Chhabra, M.; Yamamoto, J.; Zenlea, I.; Askin, N.; Rabbani, R.; McGavock, J.; Greenberg, D.; Greenberg, M.; et al. Automated Insulin Delivery Systems and Glucose Management in Children and Adolescents with Type 1 Diabetes. *JAMA Pediatr.* **2025**, *179*, 1162. <https://doi.org/10.1001/jamapediatrics.2025.2740>.
8. Govindaraju, D.; Subbian, S. Event-triggered smart dual hormone artificial pancreas for patient-specific drug delivery. *Sci. Rep.* **2025**, *15*, 32045. <https://doi.org/10.1038/s41598-025-18085-9>.
9. Bunesco, R.; Struble, N.; Marling, C.; Shubrook, J.; Schwartz, F. Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression. In Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 4–7 December 2013; pp. 135–140. <https://doi.org/10.1109/ICMLA.2013.30>.
10. Rodriguez-Rodriguez, I.; Chatzigiannakis, I.; Rodriguez, J.-V.; Maranghi, M.; Gentili, M.; Zamora-Izquierdo, M.-A. Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques. *Sensors* **2019**, *19*, 4482. <https://doi.org/10.3390/s19204482>.
11. Valletta, J.J.; Chipperfield, A.J.; Byrne, C.D. Gaussian Process modelling of blood glucose response to free-living physical activity data in people with type 1 diabetes. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Minneapolis, MN, USA, 3–6 September 2009; pp. 4913–4916. <https://doi.org/10.1109/IEMBS.2009.5332466>.
12. Li, K.; Liu, C.; Zhu, T.; Herrero, P.; Georgiou, P. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 414–423. <https://doi.org/10.1109/JBHI.2019.2931842>.
13. Seo, W.; Park, S.-W.; Kim, N.; Jin, S.-M.; Park, S.-M. A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. *Comput. Methods Programs Biomed.* **2021**, *211*, 106424. <https://doi.org/10.1016/j.cmpb.2021.106424>.
14. Zaidi, S.M.A.; Chandola, V.; Ibrahim, M.; Romanski, B.; Mastrandrea, L.D.; Singh, T. Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients. *Sci. Rep.* **2021**, *11*, 24332. <https://doi.org/10.1038/s41598-021-03341-5>.
15. Zhu, T.; Li, K.; Chen, J.; Herrero, P.; Georgiou, P. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J. Healthc. Inform. Res.* **2020**, *4*, 308–324. <https://doi.org/10.1007/s41666-020-00068-2>.

16. Martinsson, J.; Schliep, A.; Eliasson, B.; Mogren, O. Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res.* **2020**, *4*, 1–18. <https://doi.org/10.1007/s41666-019-00059-y>.
17. Rabby, M.F.; Tu, Y.; Hossen, M.I.; Lee, I.; Maida, A.S.; Hei, X. Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 101. <https://doi.org/10.1186/s12911-021-01462-5>.
18. van Doorn, W.P.T.M.; Foreman, Y.D.; Schaper, N.C.; Savelberg, H.H.C.M.; Koster, A.; van der Kallen, C.J.H.; Wesselius, A.; Schram, M.T.; Henry, R.M.A.; Dagnelie, P.C.; et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE* **2021**, *16*, e0253125. <https://doi.org/10.1371/journal.pone.0253125>.
19. Xue, Y.; Guan, S.; Jia, W. BGformer: An improved Informer model to enhance blood glucose prediction. *J. Biomed. Inform.* **2024**, *157*, 104715. <https://doi.org/10.1016/j.jbi.2024.104715>.
20. Lee, S.-M.; Kim, D.-Y.; Woo, J. Glucose Transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 1600–1611. <https://doi.org/10.1109/JBHI.2023.3236822>.
21. Lu, Y.; Liu, D.; Liang, Z.; Liu, R.; Chen, P.; Liu, Y.; Li, J.; Feng, Z.; Li, L.M.; Sheng, B.; et al. A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data. *Natl. Sci. Rev.* **2025**, *12*, nwaf039. <https://doi.org/10.1093/nsr/nwaf039>.
22. Lutsker, G.; Sapir, G.; Shilo, S.; Merino, J.; Godneva, A.; Greenfield, J.R.; Samocha-Bonet, D.; Dhir, R.; Gude, F.; Mannor, S.; et al. A foundation model for continuous glucose monitoring data. *Nature* **2026**, *650*, 978–986. <https://doi.org/10.1038/s41586-025-09925-9>.
23. Oviedo, S.; Vehi, J.; Calm, R.; Armengol, J.; Llobet, R. A review of personalized blood glucose prediction strategies for T1DM patients. *Int. J. Numer. Method Biomed. Eng.* **2017**, *33*, e2833. <https://doi.org/10.1002/cnm.2833>.
24. Sparacino, G.; Zanderigo, F.; Corazza, S.; Maran, A.; Facchinetti, A.; Cobelli, C. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 931–937. <https://doi.org/10.1109/TBME.2006.889774>.
25. Reifman, J.; Rajaraman, S.; Gribok, A.; Ward, W.K. Predictive monitoring for improved management of glucose levels. *J. Diabetes Sci. Technol.* **2007**, *1*, 478–486. <https://doi.org/10.1177/193229680700100405>.
26. Eren-Oruklu, M.; Cinar, A.; Quinn, L.; Smith, D. Estimation of future glucose concentrations with subject-specific recursive linear models. *Diabetes Technol. Ther.* **2009**, *11*, 243–253. <https://doi.org/10.1089/dia.2008.0065>.
27. Perez-Gandia, C.; Facchinetti, A.; Sparacino, G.; Cobelli, C.; Gomez, E.J.; Rigla, M.; de Leiva, A.; Hernando, M.E. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol. Ther.* **2010**, *12*, 81–88. <https://doi.org/10.1089/dia.2009.0076>.
28. Pappada, S.M.; Cameron, B.D.; Rosman, P.M.; Bourey, R.E.; Papadimos, T.J.; Olorunto, W.; Borst, M.J. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes Technol. Ther.* **2011**, *13*, 135–141. <https://doi.org/10.1089/dia.2010.0104>.
29. Zecchin, C.; Facchinetti, A.; Sparacino, G.; De Nicolao, G.; Cobelli, C. Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 1550–1560. <https://doi.org/10.1109/TBME.2012.2188893>.
30. Mirshekarian, S.; Bunesco, R.; Marling, C.; Schwartz, F. Using LSTMs to learn physiological models of blood glucose behavior. In Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, Republic of Korea, 11–15 July 2017; pp. 2887–2891. <https://doi.org/10.1109/EMBC.2017.8037460>.
31. Jaloli, M.; Cescon, M. Long-Term Prediction of Blood Glucose Levels in Type 1 Diabetes Using a CNN-LSTM-Based Deep Neural Network. *J. Diabetes Sci. Technol.* **2023**, *17*, 1590–1601. <https://doi.org/10.1177/19322968221092785>.
32. Khadem, H.; Nemat, H.; Elliott, J.; Benaissa, M. Blood Glucose Level Time Series Forecasting: Nested Deep Ensemble Learning Lag Fusion. *Bioengineering* **2023**, *10*, 487. <https://doi.org/10.3390/bioengineering10040487>.
33. Acuna, E.; Aparicio, R.; Palomino, V. Analyzing the Performance of Transformers for the Prediction of the Blood Glucose Level Considering Imputation and Smoothing. *Big Data Cogn. Comput.* **2023**, *7*, 41. <https://doi.org/10.3390/bdcc7010041>.
34. Bian, Q.; As'arry, A.; Cong, X.; Rezali, K.A.b.M.; Raja Ahmad, R.M.K.b. A hybrid Transformer-LSTM model apply to glucose prediction. *PLoS ONE* **2024**, *19*, e0310084. <https://doi.org/10.1371/journal.pone.0310084>.
35. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
36. Zhang, Y.; Yan, J. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
37. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
38. Alharthi, M.; Mahmood, A. Enhanced Linear and Vision Transformer-Based Architectures for Time Series Forecasting. *Big Data Cogn. Comput.* **2024**, *8*, 48. <https://doi.org/10.3390/bdcc8050048>.

39. Pillai, A.; Spathis, D.; Kawsar, F.; Malekzadeh, M. PaPaGei: Open Foundation Models for Optical Physiological Signals. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025. Available online: <https://openreview.net/forum?id=kYwTmlq6Vn> (accessed on 15 April 2026).
40. Al Zein, H.; van de Waterlaet, N.; Alkanat, T. Neural Architecture Search for Ultra-low Memory Blood Glucose Forecasting on the Edge. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; p. 10890864. <https://doi.org/10.1109/ICASSP49660.2025.10890864>.
41. Zhu, T.; Howson, J.; Nevado-Holgado, A. Empowering digital health management with on-device large language models for glucose prediction: A model development and validation study. *eBioMedicine* **2026**, *129*, 106343. <https://doi.org/10.1016/j.ebiom.2026.106343>.
42. Bucilua, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541. <https://doi.org/10.1145/1150402.1150464>.
43. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
44. Malihi, L.; Heidemann, G. Efficient and Controllable Model Compression through Sequential Knowledge Distillation and Pruning. *Big Data Cogn. Comput.* **2023**, *7*, 154. <https://doi.org/10.3390/bdcc7030154>.
45. Cui, J.; Wang, Z.; Yang, Z.; Guan, X. A Pruning Method Based on Feature Map Similarity Score. *Big Data Cogn. Comput.* **2023**, *7*, 159. <https://doi.org/10.3390/bdcc7040159>.
46. Ba, J.B.; Caruana, R. Do Deep Nets Really Need to be Deep? In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; Advances in Neural Information Processing Systems; pp. 2654–2662.
47. Guo, J.; Chen, M.; Hu, Y.; Zhu, C.; He, X.; Cai, D. Reducing the Teacher-Student Gap via Spherical Knowledge Distillation. *arXiv* **2020**, arXiv:2010.07485.
48. Cho, J.H.; Hariharan, B. On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4794–4802. <https://doi.org/10.1109/ICCV.2019.00489>.
49. Tian, Y.; Krishnan, D.; Isola, P. Contrastive representation distillation. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 26 April–1 May 2020.
50. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
51. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
52. Huang, Z.; Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv* **2017**, arXiv:1707.01219.
53. Heo, B.; Lee, M.; Yun, S.; Choi, J.Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019. <https://doi.org/10.1609/aaai.v33i01.33013779>.
54. Yim, J.-H.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7130–7138. <https://doi.org/10.1109/CVPR.2017.754>.
55. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976. <https://doi.org/10.1109/CVPR.2019.00409>.
56. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1365–1374. <https://doi.org/10.1109/ICCV.2019.00145>.
57. Zhu, J.; Tang, S.; Chen, D.; Yu, S.; Liu, Y.; Rong, M.; Yang, A.; Wang, X. Complementary Relation Contrastive Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9260–9269. <https://doi.org/10.1109/CVPR46437.2021.00914>.
58. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
59. Chen, X.; He, B.; Hui, K.; Sun, L.; Sun, Y. Simplified TinyBERT: Knowledge Distillation for Document Retrieval. In *Advances in Information Retrieval*; Springer: Cham, Switzerland, 2021; Volume 12657, pp. 241–248. https://doi.org/10.1007/978-3-030-72240-1_21.
60. Lan, X.; Zhu, X.; Gong, S. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems, Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 7517–7527.
61. Zhang, X.; Lu, S.; Gong, H.; Luo, Z.; Liu, M. AMLN: Adversarial-based mutual learning network for online knowledge distillation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020. https://doi.org/10.1007/978-3-030-58610-2_10.

62. Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; Anandkumar, A. Born again neural networks. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 1607–1616.
63. Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; Li, C. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021. <https://doi.org/10.1109/ICCV48922.2021.01523>.
64. Sun, S.; Cheng, Y.; Gan, Z.; Liu, J. Patient knowledge distillation for BERT model compression. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4323–4332. <https://doi.org/10.18653/v1/D19-1441>.
65. Mirzadeh, S.I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 5191–5198. <https://doi.org/10.1609/aaai.v34i04.5963>.
66. Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; Xu, C. One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/fb8e5f198c7a5dcd48860354e38c0edc-Abstract-Conference.html (accessed on 15 April 2026).
67. Zhang, W.; Liu, Y.; Ran, W.; Ma, C. Cross-Architecture Distillation Made Simple with Redundancy Suppression. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 23256–23266. Available online: https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Cross-Architecture_Distillation_Made_Simple_with_Redundancy_Suppression_ICCV_2025_paper.html (accessed on 15 April 2026).
68. Sun, X.; Li, H.; Yu, X. Future-aware blood glucose forecasting using knowledge distillation with transformer-based sequence-to-sequence models. *Sci. Rep.* **2026**, *16*, 11404. <https://doi.org/10.1038/s41598-026-41787-7>.
69. Haq, M.A.; Mohamed, A.A.; Mahendran, R.K. HEAL-LLMF: A hierarchical explainable AI framework using LLM fusion for multi-modal and personalized healthcare decision making. *Inf. Syst.* **2026**, *140*, 102696. <https://doi.org/10.1016/j.is.2026.102696>.
70. Battelino, T.; Danne, T.; Bergenstal, R.M.; Amiel, S.A.; Beck, R.; Biester, T.; Bosi, E.; Buckingham, B.A.; Cefalu, W.T.; Close, K.L.; et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range. *Diabetes Care* **2019**, *42*, 1593–1603. <https://doi.org/10.2337/dci19-0028>.
71. Giammarino, F.; Senanayake, R.; Prahalad, P.; Maahs, D.M.; Scheinker, D. A Machine Learning Model for Week-Ahead Hypoglycemia Prediction From Continuous Glucose Monitoring Data. *J. Diabetes Sci. Technol.* **2025**, *19*, 1353–1361. <https://doi.org/10.1177/19322968241236208>.
72. Luo, J.; Kumbara, A.; Shomali, M.; Han, R.; Iyer, A.; Aleppo, G.; Agarwal, R.; Gao, G. A large sensor foundation model pretrained on continuous glucose monitor data for diabetes management. *Npj Health Syst.* **2025**, *2*, 35. <https://doi.org/10.1038/s44401-025-00039-y>.
73. Ghimire, S.; Celik, T.; Gerdes, M.; Omlin, C.W. Physiology-guided blood glucose predictive model using minimal blood glucose dynamics. In Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOS-TEC 2025), Porto, Portugal, 20–22 February 2025; pp. 984–990. <https://doi.org/10.5220/0013395200003911>.
74. Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; Dubrawski, A. MOMENT: A Family of Open Time-series Foundation Models. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; pp. 16115–16152.
75. Chico, A.; Aguilera, E.; Ampudia-Blasco, F.J.; Bellido, V.; Cardona-Hernandez, R.; Escalada, F.J.; Fernandez, D.; Gomez-Peralta, F.; Gonzalez Perez de Villar, N.; Gorgojo, J.J.; et al. Clinical Approach to Flash Glucose Monitoring: An Expert Recommendation. *J. Diabetes Sci. Technol.* **2020**, *14*, 155–164. <https://doi.org/10.1177/1932296819841911>.
76. Porcellati, F.; Lucidi, P.; Bolli, G.B.; Fanelli, C.G. Thirty Years of Research on the Dawn Phenomenon: Lessons to Optimize Blood Glucose Control in Diabetes. *Diabetes Care* **2013**, *36*, 3860–3862. <https://doi.org/10.2337/dc13-2088>.
77. Huang, Y.; Lou, X.; Huang, W.; Qiu, J.; Jiang, C.; Sun, J.; Tao, X. Confirmation of the Absence of Somogyi Effect in Patients With Type 2 Diabetes by Retrospective Continuous Glucose Monitoring Systems. *Int. J. Endocrinol.* **2022**, *2022*, 6599379. <https://doi.org/10.1155/2022/6599379>.
78. Brown, S.A.; Kovatchev, B.P.; Raghinaru, D.; Lum, J.W.; Buckingham, B.A.; Kudva, Y.C.; Laffel, L.M.; Levy, C.J.; Pinsky, J.E.; Wadwa, R.P.; et al. Six-Month Randomized, Multicenter Trial of Closed-Loop Control in Type 1 Diabetes. *N. Engl. J. Med.* **2019**, *381*, 1707–1717. <https://doi.org/10.1056/NEJMoa1907863>.
79. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction. In Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data, Stockholm, Sweden, 13 July 2018; pp. 60–63.
80. Alsuhaymi, A.; Bilal, A.; Gasca Garcia, D.; Kongdee, R.; Lubasinski, N.; Thabit, H.; Nutter, P.W.; Harper, S. A Longitudinal Multimodal Dataset of Type 1 Diabetes. *Sci. Data* **2025**, *12*, 1379. <https://doi.org/10.1038/s41597-025-05695-1>.
81. Clarke, W.L.; Cox, D.; Gonder-Frederick, A.; Carter, W.; Pohl, S.L. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* **1987**, *10*, 622–628. <https://doi.org/10.2337/diacare.10.5.622>.

82. Elhoushi, A.; Hiary, H.; Abusnaina, A.; Otoom, A.; Alomari, A. Personalized blood glucose prediction in type 1 diabetes using meta-learning with bidirectional long short term memory-transformer hybrid model. *Sci. Rep.* **2025**, *15*, 30636. <https://doi.org/10.1038/s41598-025-13491-5>.
83. Sergazinov, R.; Armandpour, M.; Gaynanova, I. Gluformer: Transformer-Based Personalized Glucose Forecasting with Uncertainty Quantification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10096419>.
84. Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; Sahoo, D. Unified Training of Universal Time Series Forecasting Transformers. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; pp. 53140–53164.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.