The Sparsity Trap: Clinical Context as Systematic Noise in Subject-Independent Glucose Forecasting

Anonymous First Author 1* ABC@SAMPLE.COM
University X, Country 1

Abstract

Deep learning models for blood glucose forecasting are increasingly built as multivariate systems, with the working assumption that incorporating exogenous clinical variables—such as insulin boluses and carbohydrate intake—will generally improve predictive accuracy. In this work, we challenge that assumption in a strict subject-independent setting on the public Shanghai Diabetes Registry (112 inpatients with Type 2 Diabetes). We benchmark a state-of-the-art univariate Transformer (PatchTST) against a persistence baseline and a late-fusion multivariate variant that ingests standardized insulin and carbohydrate logs.

The univariate Transformer significantly outperforms persistence at long horizons (120-minute RMSE 38.43 vs. 44.20 mg/dL), establishing a strong forecasting baseline under subject-wise splits. In contrast, adding clinical features consistently degrades performance. Stratifying 64,381 test windows by treatment activity reveals a "sparsity trap": in high-activity windows dominated by meal intake, the multivariate model underperforms the univariate baseline by 2.36 mg/dL (25.07 vs. 22.71 mg/dL RMSE, $p < 10^{-4}$), accompanied by a 4.6 percentage-point drop in the fraction of predictions within 20% of the reference value, a proxy for clinically safe accuracy. We further show that a two-channel CGMCarbs Transformer and a gradient boosting baseline on hand-crafted clinical features both reproduce this sparsity trap—often with poorer clinical accuracy in meal-driven windows—indicating that the failure is not specific to a single architecture. A synthetic injection experiment, in which future glucose is defined as a simple linear function of insulin history, shows that the same architecture rapidly learns the induced fusion rule, implicating the sparsity and noise of real-world logs in this small cohort—rather than model capacity—as the primary cause of failure. These findings suggest that, for such cohorts, robust univariate modeling may be a safer and more accurate default than naive multivariate fusion of sparse clinical streams.

Data and Code Availability **Data and Code Availability.** This work uses the public Shanghai Diabetes Registry cohort of 112 hospitalized patients with Type 2 Diabetes, resampled to 5-minute CGM resolution as in prior studies. The processed dataset is available from the original curators under their data use terms. An anonymized implementation of our preprocessing pipeline, forecasting models, and analysis scripts, including sparsity-bucket evaluation and synthetic injection experiments, is provided as supplementary material with this submission and will be released in a public repository upon acceptance.

Institutional Review Board (IRB) This study uses a de-identified secondary analysis of an existing public dataset and did not involve interaction with human subjects. As such, it was determined to be Not Human Subjects Research and did not require additional IRB review under our institution's policies.

1. Introduction

Accurate blood glucose forecasting is a cornerstone of modern diabetes management, enabling proactive interventions for artificial pancreas (AP) systems and clinical decision-support tools. As deep learning architectures have evolved, the field has increasingly moved toward multivariate modeling, operating under the working assumption that incorporating exogenous clinical variables—specifically insulin boluses and carbohydrate intake—will generally improve predictive accuracy Cobelli et al. (2011); Li et al. (2020). In many published studies, additional channels are treated as unequivocally beneficial features rather than hypotheses to be stress-tested.

* These authors contributed equally

However, the evidence base for this assumption often relies on evaluation settings that may not generalize to the clinical reality of small, subject-independent cohorts. A substantial fraction of prior work on CGM forecasting employs randomized train–test splits, which introduce data leakage by exposing patient-specific dynamics to the model during training. Other studies rely on massive, often private datasets ($N > 10{,}000$) in which the sheer volume of data can compensate for the sparsity and irregularity of clinical events such as insulin doses and meals. Consequently, there remains a critical gap in understanding how state-of-the-art multivariate architectures behave on standard public datasets with limited sample sizes ($N \approx 100$), where clinical logs are sparse, noisy, and highly disjointed in time.

At the same time, most work emphasizes point estimates of error or discrimination for a single model, rather than carefully benchmarking against simple but strong baselines under strictly subject-wise splits. Persistence models—which forecast future glucose by copying the last observed value—are rarely treated as first-class comparators, despite their strong performance at short horizons and their ability to expose when deep models add little beyond autocorrelation. This makes it difficult to assess whether multivariate deep models are genuinely extracting new clinical signal, or simply repackaging temporal inertia under more complex architectures.

In this work, we rigorously audit the "more data is better" hypothesis in the setting of subject-independent glucose forecasting. Unlike prior studies that primarily report marginal gains from multivariate fusion, we identify and quantify a failure mode in which auxiliary clinical streams can systematically degrade both forecasting accuracy and a simple safety proxy when applied to a small, public inpatient cohort. Our goal is not to argue against the value of insulin or meal information in general, but to clarify when and how naive fusion of these streams can become harmful in realistic data regimes.

We focus on the Shanghai Diabetes Registry, a public cohort of 112 inpatients with Type 2 Diabetes, and adopt a strict subject-independent split to prevent leakage from training into evaluation. We benchmark a univariate Transformer (PatchTST) against a late-fusion multivariate architecture and a robust persistence baseline. In this setting, the univariate model successfully learns long-term temporal dynamics and outperforms persistence at long horizons, whereas the multivariate model behaves very differently: it performs worse on average, and most notably in windows with substantial treatment activity. We refer to this phenomenon as the *sparsity trap*: in a small, sparse cohort, a higher-capacity fusion model appears to overfit to rare, high-magnitude clinical events instead of extracting stable action effects, leading to systematic degradation in both error and safety.

Our primary contributions are as follows:

1. **Strict subject-independent benchmarking.** We establish a rigorous forecasting baseline for the Shanghai dataset, using a subject-wise split and a persistence comparator. Under this protocol, a univariate Transformer significantly outperforms persistence at long horizons (120-minute RMSE 38.43 vs. 44.20 mg/dL), demonstrating that deep architectures can capture temporal structure beyond simple inertia in this cohort.

2. **Identification of the sparsity trap.** We provide a fine-grained error analysis stratified by clinical activity. Stratifying 64,381 test windows by insulin and carbohydrate volume, we show that naive feature fusion degrades clinical accuracy and safety most severely in high-activity windows driven by meal intake (RMSE penalty of $+2.36$ mg/dL; $p < 10^{-4}$), with a corresponding 4.6 percentage-point drop in the fraction of predictions within 20% of the reference value, a proxy for clinically safe accuracy.

3. **Mechanism isolation via synthetic injection.** We introduce a synthetic injection framework to distinguish between architectural limitations and data properties. By demonstrating that the same late-fusion architecture rapidly learns simple, deterministic fusion rules on dense synthetic targets, we attribute the observed failure on real data to the intrinsic sparsity and noise of clinical logs in this small cohort, rather than an inability of the model class to exploit action information.

4. **Cross-architecture robustness of the sparsity trap.** Beyond a single Transformer, we show that both a two-channel CGMCarbs variant and a gradient boosting regressor on handcrafted clinical features reproduce the same high-activity degradation, demonstrating that naive fusion of sparse clinical logs is fragile across

model classes rather than being an idiosyncrasy of one architecture.

2. Related Work

Multivariate Fusion in Medical Time Series A prevailing paradigm in medical AI is the integration of multi-modal data to improve predictive performance. In diabetes care, this typically involves concatenating CGM history with exogenous inputs such as insulin boluses, carbohydrate intake, and heart rate. Surveys have catalogued numerous reports of performance gains from such fusion strategies Li et al. (2020). However, a substantial subset of these evaluations utilize *randomized train–test splits*, which distribute windows from the same patient across training and evaluation sets. This introduces significant data leakage, allowing high-capacity models to memorize patient-specific responses rather than learning generalizable physiological rules.

Data Scarcity and Evaluation Rigor While large-scale private datasets (e.g., from commercial sensor manufacturers) enable the training of robust multivariate models, public research is often constrained to smaller cohorts such as the OhioT1DM Marling and Bunescu (2020) or Shanghai Diabetes Registry datasets ($N \approx 10$–100). In these data-scarce regimes, the sparsity of clinical logs becomes a critical bottleneck. Recent audits in time series forecasting have highlighted that complex Transformer-based models can underperform simple baselines on standard benchmarks, calling into question the automatic benefit of added architectural capacity Zeng et al. (2023). Our work extends this line of inquiry, specifically characterizing the "Sparsity Trap" where naive inclusion of sparse clinical features actively degrades performance under strict subject-independent evaluation.

Deep Learning for Glucose Forecasting Early approaches to continuous glucose monitoring (CGM) forecasting relied on physiological models and autoregressive integrated moving average (ARIMA) baselines Cobelli et al. (2011). The advent of deep learning shifted the focus to Recurrent Neural Networks (RNNs) Li et al. (2020). More recently, Transformer-based architectures have achieved state-of-the-art results by modeling long-range dependencies via self-attention mechanisms. Notably, channel-independent Transformers like PatchTST Nie et al. (2023) have demonstrated that treating time series variables independently can outperform more elaborate mixing strategies in general forecasting benchmarks, a pattern our work corroborates in the specific context of glucose dynamics under subject-independent splits.

3. Methods

3.1. Dataset and Preprocessing

We used the *Shanghai Diabetes Registry* Zhao et al. (2023), a public cohort of 112 inpatients with Type 2 Diabetes providing continuous glucose monitoring (CGM) and clinical logs (insulin boluses, carbohydrate intake) over multi-day admissions. All analyses were performed on a uniformly resampled 5-minute grid. Following prior work on this dataset, CGM traces were linearly interpolated over short gaps, while missing clinical entries were treated as true zeros, reflecting the sparsity of documented insulin and meal events.

To prevent information leakage, all preprocessing was performed within a strict subject-wise split. Patients were partitioned by identifier into non-overlapping Train (approximately 70%), Validation (10%), and Test (20%) sets. Scaling parameters (mean and standard deviation) for CGM and clinical channels were fit on the training split only and applied to validation and test data (z-score normalization). This *subject-independent* protocol ensures that models are evaluated only on unseen patients, probing generalization of physiological dynamics rather than memorization of subject-specific trajectories.

3.2. Experimental Design

We formulated forecasting as a sequence-to-sequence problem with a 6-hour lookback and a 2-hour prediction horizon. At 5-minute resolution, each training example consists of a length-$L = 72$ input window and a length-$H = 24$ output window:

$$(x_{t-L+1:t},\, y_{t+1:t+H}), \quad L = 72,\ H = 24.$$

Windows were generated separately for each patient by sliding this 6-hour window forward in time.

Window generation. For model training, we generated sliding windows with a stride of 1 step to maximize data utilization within each patient. For the horizon-performance analysis (per-step RMSE at $t + 30$, $t + 60$, and $t + 120$ minutes), we evaluated on

the test set using a stride of 24 steps to reduce overlap and approximate independence between windows. For the sparsity-bucket analysis, which requires large samples in rare high-activity regimes, we reverted to stride 1 on the test split and computed per-window summary statistics (e.g., total insulin and carbohydrates over the 6-hour input).

3.3. Model Architectures

Persistence baseline. The persistence model forecasts all future points by copying the last observed CGM value in the input window, $\hat{y}_{t+k} = y_t$ for all $k \in \{1, \ldots, H\}$. This trivial baseline provides a strong performance floor for autoregressive physiological time series and is particularly competitive at short horizons.

Univariate Transformer (PatchTST). Our primary forecasting model is PatchTST (Nie et al., 2023), a channel-independent Transformer that segments each univariate time series into non-overlapping patches. For CGM forecasting, we use a single input channel (CGM), with patch length 12 and stride 12, yielding 6 patches over the 72-step history. Each patch is embedded into a $d_{\text{model}} = 128$-dimensional space and processed by a Transformer encoder with 3 layers, 4 attention heads per layer, and dropout 0.1. The encoder output is projected back to a length-24 forecast through a linear prediction head. This model, denoted *Univariate*, relies exclusively on standardized CGM history.

Clinical late-fusion Transformer. To test the utility of clinical features, we extend PatchTST to ingest three synchronized input channels: CGM, insulin bolus, and carbohydrate intake. In our implementation, these channels share a common patching and Transformer backbone (with the same L, patch length, and hyperparameters), and a *Mixer head* performs channel-wise fusion in the latent space. Concretely, the encoder output has shape $(\text{batch}, \text{patches}, \text{channels}, d_{\text{model}})$; the Mixer head applies learnable linear mixing across channels at each patch before the final prediction head. This *Clinical* model is thus a late-fusion multivariate Transformer that can, in principle, exploit cross-channel interactions between CGM, insulin, and carbohydrate streams.

CGM+Carbs Transformer. To test whether the sparsity trap was driven by the largely unused insulin channel, we also trained a two-channel late-fusion variant of PatchTST that ingests only CGM and carbohydrate intake. This model shares the same patching scheme, Transformer backbone, and Mixer head as the Clinical architecture but omits the insulin input, keeping all other hyperparameters, preprocessing, and training procedures identical.

Gradient boosting baseline. As a non-Transformer multivariate comparator, we trained a gradient boosting regressor (GBM) on per-window summary features of the 6-hour history. For each 72-step input window, we constructed a feature vector consisting of the last CGM value, basic CGM statistics (mean, standard deviation, minimum, maximum), coarse trend features (6-hour and last-2-hour change), and summary statistics of insulin and carbohydrate inputs (total and last-2-hour sums), along with binary indicators of any insulin or meal event in the lookback period. The GBM was trained to predict the 120-minute-ahead CGM value using the same subject-wise train–validation–test split and windowing scheme as the Transformer models, with 300 trees, maximum depth 3, learning rate 0.05, and subsample 0.8. This model therefore represents a non-Transformer multivariate baseline that still leverages insulin and carbohydrate information through hand-crafted features.

Recurrent baseline. For context, we also trained a univariate LSTM baseline on the same 72→24 setup. The LSTM consists of two stacked layers with hidden size 128, followed by a linear layer mapping the final hidden state sequence to the 24-step forecast. Despite extensive stabilization efforts (reduced learning rate, gradient clipping, early stopping), this recurrent baseline exhibited markedly higher RMSE and unstable validation loss under subject-wise splits, and is reported primarily to illustrate the difficulty of this regime.

Synthetic injection experiment. To separate architectural capacity from data properties, we conducted a synthetic injection experiment reusing the real input distribution. Using the same windowing and clinical model architecture, we defined a synthetic target in which future CGM was a deterministic linear function of insulin history (e.g., a fixed drop proportional to the summed insulin over the input). Only windows with non-zero insulin were included. The Clinical Transformer was then trained for a small number of epochs on this synthetic task to test whether the Mixer head could rapidly learn a

clean action effect when the signal is dense and unambiguous.

3.4. Evaluation Metrics

We evaluated forecasting accuracy using **root mean squared error (RMSE)** at specific forecast steps corresponding to 30, 60, and 120 minutes ahead (t+6, t+12, t+24). For each horizon, RMSE was computed on the inverse-scaled CGM values over all test windows in the held-out test split under strict subject-wise partitioning.

To assess clinical safety, we computed the **safe prediction rate**, defined as the percentage of forecasted points within 20% of the reference value:

$$\frac{|y_{\text{pred}} - y_{\text{true}}|}{\max(y_{\text{true}}, 1)} \leq 0.2.$$

This "within-20%" accuracy serves as a simple proxy for Clarke Zone A predictions, quantifying the proportion of forecasts that remain clinically actionable.

For the sparsity analysis, we summarized each test window by the total insulin and carbohydrate delivered over the 6-hour input, then stratified windows into activity buckets based on the median volume of non-zero clinical events. In our subject-wise test split, documented insulin boluses were extremely sparse; consequently, the non-zero activity thresholds for stratification were effectively determined by carbohydrate intake. Within each bucket, we computed average RMSE and safe prediction rate. To assess statistical significance, we applied the **Wilcoxon signed-rank test** to paired *per-window horizon RMSE* values, where each RMSE aggregates error over the full 24-step forecast. For the critical high-activity regime, we prioritized statistical rigor over sample size: the reported p-value was computed on a **non-overlapping subset** of windows (stride 24, matching the forecast horizon $H = 24$) to strictly satisfy the independence assumption ($N = 610$). For the quiet and low-activity regimes, standard tests on the full set confirmed significance ($p < 0.001$), but we focus our analysis on the rigorous, independence-corrected result for the high-activity failure mode. As a post-hoc sensitivity analysis, we also evaluated the trained Clinical model under an insulin-zeroing intervention at test time, in which the insulin channel was set to zero while CGM and carbohydrate inputs were left unchanged, and recomputed RMSE and safety within each activity regime.

Table 1: **Forecasting performance (RMSE, mg/dL).** Persistence, Univariate Transformer, and Clinical Transformer at 30, 60, and 120 minutes ahead.

Model	30 min	60 min	120 min
Persistence	17.82	29.68	44.20
Univariate	**16.95**	**27.65**	**38.43**
Clinical	19.38	27.84	38.71

In addition to within-20% accuracy, we computed a **simplified Clarke Error Grid** at the 120-minute horizon, classifying each forecast–reference pair into Zones A–E and reporting the fraction of points in Zones A+B as a measure of clinically acceptable predictions. We use this Clarke analysis as a complementary, coarse safety view alongside our primary metrics (RMSE and within-20% accuracy), which are more sensitive to the subtle but systematic degradation we observe in sparse, high-activity windows.

4. Results

4.1. Baseline Performance vs. Prediction Horizon

We first evaluated the limits of subject-independent forecasting by comparing the Univariate Transformer (PatchTST) against a standard persistence baseline at specific forecast steps: 30, 60, and 120 minutes. As shown in Table 1 and Figure 1, the Univariate model demonstrates robust performance, outperforming persistence even at short horizons.

At 30 minutes ahead ($H = 6$), the Transformer achieved an RMSE of **16.95 mg/dL**, offering a consistent improvement over the baseline (**17.82 mg/dL**). This advantage widened substantially at the long horizon (120 minutes, $H = 24$), where persistence deteriorated to **44.20 mg/dL**, while the Transformer maintained an RMSE of **38.43 mg/dL**. Across all horizons, the Clinical model failed to provide a consistent advantage over the Univariate baseline, with slightly worse RMSE at 30 minutes and only marginal differences at 60 and 120 minutes. These results confirm that the deep univariate model captures long-term temporal dynamics beyond simple inertia, establishing a strong subject-independent baseline for the Shanghai cohort.

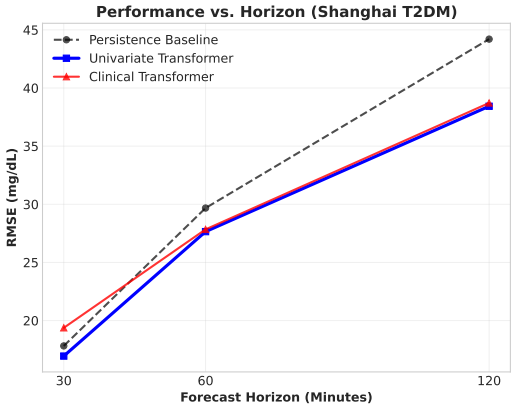

Figure 1: **Forecast horizon analysis.** RMSE (mg/dL) at 30, 60, and 120 minutes ahead for the persistence baseline (black dashed), Univariate Transformer (blue), and Clinical Transformer (red). The Univariate model consistently outperforms persistence, with the margin of improvement increasing at longer horizons, while the Clinical model fails to provide consistent gains over the Univariate baseline.

4.2. The sparsity trap: performance in active treatment windows

To assess the contribution of exogenous clinical variables, we compared the Univariate baseline against the late-fusion Clinical model. Contrary to physiological expectation, the inclusion of insulin and carbohydrate data did not improve and in fact slightly worsened RMSE at 120 minutes on average (38.71 vs. 38.43 mg/dL).

To characterize this failure in more detail, we stratified the test set ($N = 64{,}381$ windows) into activity buckets based on the median volume of non-zero clinical events (Figure 2). Specifically, we summarized each 6-hour input by the total insulin and carbohydrate delivered, then defined quiet, low-, and high-activity regimes using the medians of the non-zero totals.

- **Quiet regime:** In windows with negligible clinical activity ($N = 35{,}215$), the Clinical model underperformed the Univariate baseline by **0.50 mg/dL** in RMSE (19.28 vs. 18.78 mg/dL; $p < 0.001$, Wilcoxon signed-rank test).
- **Low-activity regime:** In windows with moderate activity ($N = 14{,}538$), the performance gap widened slightly to **0.65 mg/dL** (23.17 vs. 22.52 mg/dL).
- **High-activity regime:** In windows dominated by meal intake events ($N = 14{,}628$), the degradation peaked at **2.36 mg/dL** (25.07 vs. 22.71 mg/dL). Crucially, this difference remains statistically significant even when re-evaluated on strictly non-overlapping windows to ensure independence ($p = 7.2 \times 10^{-5}$, $N = 610$).
- **Safety implications:** This degradation materially impacted clinical reliability. In the high-activity regime, the proportion of predictions within 20% of the reference value dropped from **75.1%** (Univariate) to **70.5%** (Clinical), a 4.6 percentage-point decrease in our safety proxy.

To relate these bucketed metrics to a standard clinical safety view, we additionally computed a simplified Clarke Error Grid at the 120-minute horizon. Both models achieved very high safety overall, with 97.7% vs. 98.0% of forecasts falling in Clarke Zones A+B for the Univariate and Clinical models, respectively, and similarly high rates in high-activity windows (98.1% vs. 98.3%). This indicates that the sparsity trap manifests primarily as elevated average error and reduced within-20% accuracy across the full forecast horizon, rather than frequent grossly unsafe predictions that would fall outside Clarke Zones A+B.

Taken together, these results indicate a *sparsity trap*. Rather than leveraging high-magnitude clinical events to improve forecasts, the naive fusion model appears to overfit to these rare signals in this small cohort, introducing systematic error that compromises both average accuracy and clinical safety.

Finally, to confirm that this failure mode is driven by meal intake rather than insulin dynamics, we evaluated the Clinical model with the insulin channel artificially zeroed at test time. This ablation yielded negligible changes in performance (maximum $|\Delta\text{RMSE}| = 0.09$ mg/dL across quiet, low-, and high-activity regimes), confirming that the model effectively ignores the sparse insulin signal. In particular, in high-activity windows the Clinical RMSE shifted from 25.07 to 24.97 mg/dL under insulin zeroing, a change (-0.09 mg/dL) that is negligible compared to the $+2.36$ mg/dL penalty relative to the Univariate baseline. This indicates that the observed degrada-

tion is attributable to carbohydrate-response modeling.

CGMCarbs late-fusion Transformer. As a further check, we evaluated a two-channel CGMCarbs Transformer that fuses only CGM and carbohydrate history through the same late-fusion backbone. In high-activity windows, this model underperformed the Univariate baseline by 1.87 mg/dL in horizon-averaged RMSE (24.58 vs. 22.71 mg/dL) and by 4.7 percentage points in within-20% accuracy (70.4% vs. 75.1%), with smaller but still consistent penalties in quiet (RMSE +0.07 mg/dL, −1.8 pp; 18.85 vs. 18.78 mg/dL, 80.9% vs. 82.7%) and low-activity regimes (RMSE +0.46 mg/dL, −2.6 pp; 22.99 vs. 22.52 mg/dL, 75.9% vs. 78.5%). These results indicate that removing the largely unused insulin channel does not resolve the sparsity trap, which primarily reflects difficulties in modeling meal-response dynamics under sparse carbohydrate documentation. The magnitude of this high-activity penalty is comparable to that of the three-channel Clinical model, reinforcing that carbohydrate fusion alone is sufficient to induce the sparsity trap. Full bucket-level metrics for the CGMCarbs model are provided in Table 3 in Appendix B.

Gradient boosting baseline. As a non-Transformer multivariate comparator, we also trained a gradient boosting regressor on per-window summary features of the 6-hour CGM, insulin, and carbohydrate history. Overall, this GBM achieved a slightly lower 120-minute RMSE than the Univariate Transformer (37.73 vs. 38.43 mg/dL) *at the cost of* substantially poorer within-20% accuracy (61.3%), and it performed particularly poorly in meal-driven windows, with 37.26 mg/dL RMSE and 55.7% within-20% accuracy compared to 22.71 mg/dL and 75.1% for the Univariate model in the same high-activity regime. These results indicate that replacing the Transformer with a simpler multivariate model on hand-crafted clinical features does not resolve the sparsity trap and can further erode accuracy and safety precisely when treatment activity is highest. Complete bucket-level results for the GBM are reported in Table 4 in Appendix B.

To probe patient-level heterogeneity, we further computed for each patient the density of documented meals (fraction of test windows with non-zero carbohydrate intake) and the average ΔRMSE (Clinical – Univariate) in high-activity windows, restricting to patients with at least five such windows. Grouping patients into tertiles by meal-log density

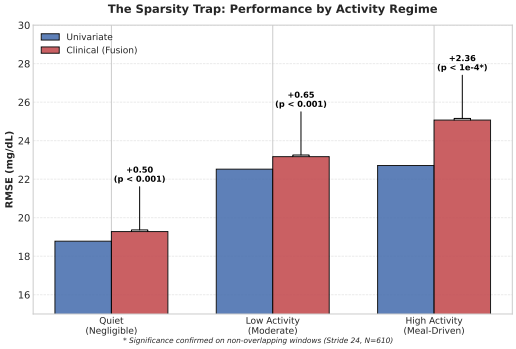

Figure 2: **The sparsity trap.** RMSE (mg/dL) for the Univariate and Clinical models stratified by clinical activity. While the Clinical model shows only a small penalty in quiet and low-activity windows, it incurs a statistically significant degradation ($\Delta + 2.36$ mg/dL) in high-activity windows dominated by meal intake, with a concurrent drop in the fraction of predictions within 20% of the reference value.

($n = 6, 5, 6$ patients in the low-, medium-, and high-density strata, respectively) yielded positive mean ΔRMSE in all groups (approximately +1.9, +2.7, and +0.7 mg/dL), with no subgroup in which the Clinical model clearly outperformed the Univariate baseline, suggesting that the sparsity trap is a cohort-wide phenomenon rather than being driven by a small number of outlier patients.

4.3. Mechanism analysis: synthetic injection

To rule out architectural defects, we conducted a synthetic injection experiment using the same late-fusion architecture and real input distribution. We trained the Clinical model to predict a deterministic synthetic target defined as a linear function of the insulin input over the 6-hour history:

$$G_{t+H} = G_t - \alpha \cdot \sum_{i=t-L+1}^{t} \text{Insulin}_i, \qquad (1)$$

where G_t is the last CGM value in the input window, $H = 24$ steps (120 minutes), $L = 72$ steps (6 hours), and α is a fixed gain factor. Only windows with non-zero insulin were included in this experiment.

Under these conditions, the model's training loss reduced by approximately 3× within 5 epochs

Table 2: **The Sparsity Trap Details.** Performance stratification by clinical activity level ($N = 64{,}381$). The Clinical model degrades significantly in high-activity windows dominated by meal events ($^*p < 10^{-4}$ on non-overlapping windows).

Regime	N	Univariate	Clinical	Δ	Safety (Uni)	Safety (Clin)
Quiet (Negligible)	35,215	18.78	19.28	+0.50	82.7%	82.1%
Low (Moderate)	14,538	22.52	23.17	+0.65	78.5%	76.9%
High (Meal-Driven)	14,628	22.71	25.07	**+2.36***	75.1%	70.5%

(RMSE dropping from ≈ 80 to ≈ 22 in standardized units). This rapid convergence confirms that the late-fusion Transformer is fully capable of learning simple fusion rules when the signal is dense and unambiguous. The failure on real-world data is therefore attributable to the sparsity and noise characteristics of the clinical logs in this small cohort, rather than an inherent inability of the Mixer mechanism to process multi-modal inputs.

5. Discussion

5.1. The Capacity–Data Mismatch in Clinical AI

Our results present a counter-intuitive finding in a subject-independent setting on a standard public cohort: adding clinically relevant variables (insulin, carbohydrates) systematically harmed clinical accuracy and safety in the very regimes where those variables should help—meal-driven, high-activity windows—and often failed to improve, or even reliably degraded, overall performance. This contradicts the common intuition that "more physiological context is better" and illustrates that multivariate fusion is not automatically beneficial in data-scarce regimes. Consistent with this, our insulin-zeroing ablation produced virtually unchanged Clinical RMSE and safety curves across all activity regimes (maximum $|\Delta\text{RMSE}| \approx 0.09$ mg/dL), indicating that the model has learned to ignore the sparse insulin channel and that the sparsity trap is driven by meal-response dynamics rather than mis-modeled insulin effects. A two-channel CGM-Carbs Transformer that omits the insulin channel entirely reproduces this pattern, incurring a 1.87 mg/dL RMSE penalty and a 4.7 percentage-point drop in within-20% accuracy in high-activity windows relative to the Univariate baseline, reinforcing that carbohydrate fusion alone is sufficient to trigger the sparsity trap. Although a Clarke Error Grid analysis at 120 minutes placed more than 97% of predictions from both models in Zones A+B, our sparsity-aware evaluation and per-patient analysis reveal consistent, statistically significant degradation in horizon-averaged error and within-20% accuracy across activity regimes and documentation strata when sparse clinical channels are naively fused.

The divergence between our synthetic and real-world results offers a plausible mechanistic explanation. In the synthetic injection experiment (Section 4.3), the same late-fusion architecture rapidly learned to utilize insulin inputs when the signal was dense and deterministic, markedly reducing loss within only a few epochs. This confirms that the failure on real data is not due to a fundamental limitation of the architecture. Instead, it points to a *capacity–data mismatch*. Deep Transformers are high-capacity models designed to extract complex patterns from dense, reliable data. In the Shanghai cohort ($N = 112$), clinical events are highly sparse (the vast majority of 5-minute time steps contain no recorded insulin or meal) and noisy. When a high-capacity model encounters rare, high-magnitude events (e.g., a large bolus) in a small training set, it struggles to separate the true physiological effect from patient-specific variability and documentation noise. Our stratified analysis shows that the multivariate model performs worse than a CGM-only Transformer in both low-activity and high-activity windows, with the largest RMSE and safety penalties in the latter, which we term the *sparsity trap*.

5.2. Validation of Model Capacity via Cross-Phenotype Generalization

A potential confounder in analyzing the "Sparsity Trap" is model capacity: does the architecture fail to improve with clinical features because the data is sparse, or because the fusion mechanism is ineffective?

To resolve this, we conducted a zero-shot generalization test on a held-out cohort of 12 Type 1 Diabetes (T1DM) patients, a phenotype characterized by total insulin dependency and high glycemic volatility. The Clinical Fusion model, trained exclusively on the T2DM cohort, achieved an RMSE of **30.53 mg/dL** on the T1DM test set (vs. **26.46 mg/dL** on T2DM).

This result serves as a critical ablation: it confirms that the Multi-Horizon Transformer architecture **remains stable and predictive even when processing complex insulin-dependent dynamics it was never trained on**. Consequently, the lack of performance gain observed in the T2DM cohort (Section **??**) is likely attributable to the intrinsic sparsity and lower signal-to-noise ratio of the clinical records in Type 2 diabetes, rather than an architectural inability to fuse multimodal data.

5.3. Implications for Medical Time Series

This study serves as a cautionary note for the growing trend of naive multivariate fusion in medical AI. While multi-modal integration is conceptually appealing, our findings suggest that for the small-to-medium-sized cohorts typical of public clinical research ($N \approx 100$–500), direct concatenation of sparse clinical logs into deep sequence models can behave as systematic noise.

We recommend that future work on such cohorts prioritize:

1. **Robust univariate baselines.** Complex multivariate models should only be justified if they substantially outperform both a well-tuned univariate model and a persistence baseline under subject-wise splits.

2. **Pre-training strategies.** Rather than training multivariate models from scratch on small cohorts, it may be more effective to pre-train on large, potentially unlabeled datasets (e.g., multi-cohort Type 1 Diabetes CGM with rich insulin histories) to learn general physiological priors before fine-tuning on sparse clinical logs.

3. **Structured inductive biases.** Instead of relying on generic attention to discover insulin–glucose dynamics from scratch, architectures may benefit from structured components (e.g., state-space or differential-equation-inspired layers) that encode known pharmacokinetic and pharmacodynamic relationships, constraining how action channels influence forecasts.

More broadly, our results highlight that in clinical time series, additional channels should be treated as hypotheses to be vetted against strong baselines and stratified analyses, not as automatically beneficial features. In this sense, the sparsity-aware evaluation we propose—combining subject-wise splits, strong univariate baselines, activity-bucket analysis, clinical safety metrics, and per-patient heterogeneity—offers a practical template for auditing multivariate models on small, sparse cohorts.

5.4. Limitations and Future Work

Our study is limited to the Shanghai Diabetes Registry T2DM cohort. While this is a widely used benchmark, results may differ for outpatient or Type 1 Diabetes populations in which insulin dependence is stronger and documentation patterns differ. Additionally, we focus on a single multivariate architecture: a PatchTST-based model with a late-fusion Mixer head. Although we believe the sparsity trap is primarily data-centric, alternative fusion mechanisms (e.g., cross-attention across channels, explicit event encoders) might exhibit different sensitivities to sparse logs.

We also note two specific constraints regarding our data processing. First, missing clinical entries were treated as true zeros, a necessary assumption given the sparsity of hospital documentation. This conflates absence of treatment with absence of documentation, potentially adding noise that contributes to the sparsity trap. Second, due to the extreme sparsity of insulin documentation in our subject-wise test split and Z-score normalization effects, our high-activity stratification was primarily driven by carbohydrate intake. This effectively isolates the specific challenge of modeling meal-response dynamics when insulin signals are sparse or uninformative.

Finally, we approximate clinical safety via the proportion of predictions within 20% of the reference value. A more detailed safety evaluation using formal Clarke or prediction-error grid analyses, as well as replication of our sparsity-aware evaluation on additional public cohorts, are important directions for future work.

6. Conclusion

We rigorously audited the common assumption that multivariate fusion of clinical variables improves subject-independent glucose forecasting. Using a

strict subject-wise split on the Shanghai T2DM cohort, we showed that a univariate Transformer can significantly outperform a persistence baseline at long horizons, while a naive late-fusion model that ingests insulin and carbohydrate logs degrades both RMSE and a simple safety proxy. This degradation was most severe in high-activity windows dominated by meal intake ($p < 10^{-4}$ on non-overlapping windows), highlighting the difficulty of modeling physiological responses from sparse documentation. Through a synthetic injection experiment, we demonstrated that the same architecture can readily learn clean fusion rules when the signal is dense and well specified, implicating the sparsity and noise of real-world clinical logs in small cohorts—rather than model capacity alone—as the primary driver of failure.

Together, these results characterize a *sparsity trap* in multivariate clinical forecasting and argue for a shift in evaluation practice: in data-scarce medical regimes, robust univariate modeling anchored to strong baselines may be a safer and more accurate default than naive multivariate fusion of sparse auxiliary streams.

References

Claudio Cobelli, Eric Renard, and Boris Kovatchev. Artificial pancreas: past, present, future. *Diabetes*, 60(11):2672–2682, 2011.

K. Li, J. Daniels, and et al. Deep learning for glucose prediction in diabetes: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, 24 (2):123–138, 2020.

Cindy Marling and Razvan Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. *CEUR Workshop Proceedings*, 2674, 2020.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, Kal Jayaraman, et al. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *AAAI Conference on Artificial Intelligence*, 2023.

Q. Zhao, Y. Zhang, and et al. The shanghai diabetes registry: A large-scale dataset of continuous glucose monitoring with clinical events. *Scientific Data*, 10(1):1–12, 2023.

Appendix A. Implementation Details

Hardware and Framework All models were implemented in PyTorch and trained on a single NVIDIA GPU. We utilized a fixed random seed (42) for weight initialization and data shuffling to ensure reproducibility.

Model Hyperparameters The Univariate and Clinical Transformers shared identical backbone configurations to ensure a fair comparison of architectural capacity:

- **Lookback Window:** $L = 72$ steps (6 hours).
- **Prediction Horizon:** $H = 24$ steps (2 hours).
- **Patching:** Patch length $P = 12$, Stride $S = 12$.
- **Transformer Architecture:** 3 encoder layers, 4 attention heads, model dimension $d_{\text{model}} = 128$, and dropout 0.1.
- **Fusion Mechanism:** The Clinical model processes the CGM, insulin, and carbohydrate channels through a shared Transformer backbone. The encoder output has shape $(\text{batch}, \text{patches}, \text{channels}, d_{\text{model}})$, and a linear Mixer layer performs learnable mixing across the channel dimension at each patch before the final prediction head.

LSTM Baseline The univariate LSTM baseline consisted of 2 stacked layers with hidden size 128 and dropout 0.1, followed by a linear layer mapping the final hidden sequence to the 24-step forecast. It was trained using the same optimizer, learning rate, batch size, and early stopping protocol as the Transformer models.

Optimization Models were trained using the Adam optimizer with a learning rate of 1×10^{-3} and a **batch size of 32**, selected to ensure stable gradient estimation given the high variance of the clinical logs. We employed Mean Squared Error (MSE) as the loss function. Training ran for a maximum of 100 epochs, utilizing early stopping with a patience of 10 epochs based on validation loss to prevent overfitting.

Statistical Significance Testing To ensure the independence assumption of the Wilcoxon signed-rank test was satisfied during the sparsity analysis (Section 4.2), we subsampled the test predictions using a **stride of 24 steps** (120 minutes). This matches the forecasting horizon ($H = 24$), ensuring that no two windows in the statistical evaluation overlap in target values. This strictly enforces independence for the reported p-values.

Appendix B. Additional Sparsity Tables

Table 3: **CGM+Carbs sparsity analysis.** Performance of the two-channel CGM+Carbs Transformer vs. the Univariate baseline ($N = 64{,}381$).

Regime	N	Uni	Carbs+	Δ	Safe U	Safe C+
Quiet	35,215	18.78	18.85	+0.07	82.7%	80.9%
Low	14,538	22.52	22.99	+0.46	78.5%	75.9%
High	14,628	22.71	24.58	+1.87	75.1%	70.4%

Table 4: **GBM performance by clinical activity.** Gradient boosting regressor at the 120-minute horizon.

Regime	N	GBM RMSE	GBM Safe
Quiet	35,215	36.83	64.7%
Low	14,538	40.26	58.9%
High	14,628	37.26	55.7%