# Extended Abstract

**Motivation** Managing Type 1 Diabetes (T1D) is a relentless challenge of balancing insulin against dynamic factors like diet and exercise, where traditional, reactive adjustments are often insufficient. This project explores the feasibility of applying offline Reinforcement Learning (RL) to create a more dynamic and personalized insulin dosing policy. Specifically, we investigate the use of a Decision Transformer (DT), a model that frames RL as a sequence modeling problem, to learn a dosing strategy directly from my own personal, multi-modal sensor data from my Continuous Glucose Monitor (CGM), insulin pump, and smartwatch.

**Method** To ensure safety, an offline RL approach was chosen, learning exclusively from historical data sampled into discrete timesteps at 5-minute intervals. These were structured into trajectories of states, actions, and rewards. The **state** vector provides a physiological snapshot at each timestep, including glycemic data (value and trend), insulin on board (IOB), multi-modal activity metrics, and cyclical temporal features. The continuous **action** of delivering an insulin bolus was discretized into 8 categories, including a "no action" category. The **reward** for each action was defined as the Time-in-Range (70-180 mg/dL) over the subsequent 3-hour window to teach the model the long-term consequences of its decisions.

**Implementation** The core of our implementation is a Decision Transformer model, which utilizes a GPT-based causal Transformer architecture Chen et al. (2021). Our model uses an embedding dimension of 128, 3 transformer layers, and 4 attention heads. The state, action, and return-to-go (RTG) inputs are projected into this embedding space and combined with learned positional embeddings. The model was trained using the AdamW optimizer and a weighted cross-entropy loss function. This weighting was essential to combat the severe class imbalance of the "no action" class and force the model to learn the rare but clinically vital bolus actions.

**Results** Hyperparameter tuning was crucial; a low learning rate ($1 \times 10^{-5}$) and large batch size (64) were required to guide the model beyond simple behavioral cloning towards a more proactive policy. Our qualitative analysis, which directly compares the model's recommended actions against historical actions for the same event, showed clear instances where the well-tuned model chose to administer a corrective bolus during a period of rising glucose while the human historically did not. This demonstrates that the model successfully learned a novel and potentially improved policy rather than just imitating the training data.

**Discussion** This project successfully demonstrated that a Decision Transformer can learn a plausible, proactive dosing policy from real-world data, moving beyond simple behavioral cloning. The primary limitation, however, is the open-loop nature of our evaluation; because the model's actions do not influence the subsequent states it sees in our analysis, we cannot reliably measure their cumulative impact or generate aggregate quantitative metrics. This underscores the fundamental difficulty of offline policy evaluation in this domain.

**Conclusion** This work serves as a successful proof-of-concept for applying Decision Transformers to personalized T1D management. We developed a complete pipeline and trained a model that learned a clinically relevant, proactive policy for correcting hyperglycemia. The critical next step for this research is to perform a closed-loop evaluation by integrating the trained agent with a validated physiological simulator. This will allow for a robust, quantitative assessment of the policy's true clinical impact on glycemic control.

# Offline RL with Decision Transformers for T1D Glucose Control

**Katherine Greatwood**
Department of Computer Science
Stanford University
kgreat@stanford.edu

## Abstract

The management of Type 1 Diabetes (T1D) is a relentless challenge where traditional, reactive strategies are often insufficient. This project investigates the feasibility of applying an offline Reinforcement Learning (RL) approach, specifically a Decision Transformer (DT), to learn a personalized insulin dosing policy from real-world, multi-modal sensor data (CGM, pump, smartwatch). We processed historical data into trajectories of states, actions, and returns-to-go and trained a DT model using a weighted cross-entropy loss function to address the severe data imbalance of infrequent insulin actions. Our results show that with careful hyperparameter tuning, the model learns a proactive policy that goes beyond simple behavioral cloning. In our qualitative analysis, the trained agent correctly recommended corrective insulin doses during periods of rising blood sugar where the human historically did not. While this work serves as a successful proof-of-concept, our open-loop evaluation prevents a quantitative assessment of the policy's cumulative impact. This highlights the limitations of purely offline evaluation and underscores that the critical next step is the integration of the learned policy with a validated physiological simulator for robust, closed-loop testing.

## 1 Introduction

The management of Type 1 Diabetes (T1D) is a significant and relentless challenge, requiring individuals to navigate a complex interplay of factors including insulin dosing, diet, physical activity, and stress to maintain glycemic control. A particularly difficult aspect of T1D management is the period following physical exercise, which can induce unpredictable and delayed glucose fluctuations, elevating the risk of both hypoglycemia and hyperglycemia. Traditional management strategies often depend on a combination of general guidelines and reactive, trial-and-error adjustments. This approach can be insufficient due to high individual variability in physiological responses.

To address the need for more dynamic and personalized treatment strategies, this project explores the application of offline Reinforcement Learning (RL). RL is a paradigm well-suited for sequential decision-making problems, such as insulin dosing, where actions have delayed and complex consequences on long-term glucose stability. To mitigate the significant safety risks associated with online RL methods that require direct patient interaction for exploration, this work employs an offline RL approach, learning exclusively from historical data.

Specifically, we utilize the Decision Transformer (DT) architecture, which reframes offline RL as a sequence modeling problem Chen et al. (2021). This is potentially advantageous for handling the noisy, multi-modal time-series data streams generated by modern diabetes technology, including Continuous Glucose Monitors (CGM), insulin pumps, and smartwatches. While our initial objective was to focus on the challenging post-exercise period, the scope was revised to develop a general insulin delivery recommendation model to leverage a larger and more diverse dataset, which is crucial for training robust sequence models. The core contribution of this work is the application and evaluation of the standard Decision Transformer methodology to generate personalized insulin dosing recommendations from my own integrated sensor data, addressing a clear gap in current research.

## 2 Related Work

The application of Reinforcement Learning to T1D glucose control has been an active area of research, though early efforts were primarily conducted within in-silico environments Yau et al. (2023). These simulation-based studies successfully demonstrated the potential for RL to optimize insulin dosing policies. However, the direct clinical applicability of these models is limited, as simulators often fail to capture the full complexity and individual variability of real-world glucose dynamics, especially concerning the nuanced effects of physical activity and stress. Furthermore, traditional online RL methods are infeasible for this domain due to the significant safety risks posed by the exploratory actions required during training. The broader challenges and potential of RL in healthcare have been surveyed extensively by Yu et al. Yu et al. (2021).

To overcome these safety concerns, research has increasingly shifted towards offline RL, which learns policies from pre-existing datasets without risky online interaction. Foundational studies by Pajech et al. Pajech et al. (2022) and Emerson et al. Emerson et al. (2023) have explored offline RL for personalized glucose control, demonstrating potential improvements in glycemic outcomes like time-in-range. While promising, this prior work has generally focused on overall glucose management or specific scenarios like meal responses, with less emphasis on the uniquely challenging post-exercise physiological state, a period that requires distinct management strategies.

Separately, sequence modeling, particularly with the Transformer architecture, has proven valuable for T1D management, most notably for glucose prediction tasks using CGM data Li et al. (2024). Building on this, the Decision Transformer (DT) model proposed by Chen et al. Chen et al. (2021) emerged as a powerful paradigm that frames offline RL as a conditional sequence modeling task. While other Transformer-based offline RL algorithms have been evaluated for glucose control Viroonluecha et al. (2023) and adaptations have been developed for general healthcare decisions Zhang et al. (2023), to our knowledge, the standard Decision Transformer architecture has not yet been specifically applied to generate insulin dose recommendations for T1D.

Therefore, a clear gap exists in applying the standard Decision Transformer methodology to the specific problem of generating insulin dose recommendations for T1D glucose control. This project addresses this gap by implementing and evaluating a Decision Transformer on a personalized, multi-modal dataset comprising integrated CGM, insulin pump, and smartwatch data to generate insulin recommendations.

## 3 Method and Experimental Setup

### 3.1 Data

The dataset for this project was constructed from my own personal, multi-modal sensor data, including time-series data from a Continuous Glucose Monitor (CGM), an insulin pump, and a Garmin smartwatch. The raw data streams were integrated, cleaned, and time-aligned to create a unified historical record. This continuous record was then sampled at regular 5-minute intervals to form discrete timesteps, which were used to build the final sequences for the model.

**State Representation** At each timestep $t$, a state vector $s_t$ was engineered to provide the model with a comprehensive view of the current physiological situation. The state vector is composed of several key components:

- **Glycemic Features:** The current blood glucose value (mg/dL) and the glucose trend, calculated as the slope of a linear regression over the preceding 30 minutes of glucose readings.
- **Insulin Features:** The current Insulin on Board (IOB), representing the amount of active insulin from previous boluses still working in the body.
- **Activity Features:** A set of metrics derived from smartwatch data to capture the context of physical activity, which significantly impacts insulin sensitivity. These include the duration and average heart rate of the last completed workout, the time elapsed since the last workout ended, a binary flag indicating if a workout is currently active, and the total exercise duration over the last 6 hours.
- **Temporal Features:** The time of day and day of the week are encoded into four continuous features using sine and cosine transformations to capture their cyclical nature.

**Action Space** The model's action space is discretized into 8 distinct categories to make the decision-making problem tractable for a classification-based model. The continuous real-world action of delivering an insulin bolus was binned into several ranges (e.g., 0.1-0.5 IU, 0.5-1.0 IU). A specific and crucial action, ID 7, was designated as the "no action" case, representing a 5-minute interval where no bolus was delivered. This is by far the most frequent action in the dataset.

**Reward and Trajectory Formulation** To train the Decision Transformer, the data was structured into trajectories of '(state, action, reward)' tuples. The reward $r_t$ for taking an action $a_t$ at state $s_t$ was defined as the percentage of time the glucose level remained within the target range (70-180 mg/dL) over the subsequent 3-hour window. This delayed reward mechanism is designed to teach the model the long-term consequences of its actions. This is a great fit for the application because both insulin and carbohydrates can take hours to fully absorb.

From these rewards, the return-to-go (RTG) $\hat{R}_t$ was calculated for each timestep, representing the cumulative sum of all future rewards in that episode ($\hat{R}_t = \sum_{t'=t}^{T} r_{t'}$). These trajectories were then segmented into fixed-length sequences of 24 timesteps, corresponding to a 2-hour context window, which serve as the direct input to the Decision Transformer.

**Data Splitting** The complete set of generated sequences was split by a date into a training set (80%) and a validation set (20%). This ensures that the model is evaluated on data that occurs later in time than all of its training data, preventing any look-ahead bias and providing a more realistic assessment of its generalization performance.

### 3.2 Model

To address the task of learning an insulin dosing policy from offline data, we employ a Decision Transformer (DT) architecture, which casts reinforcement learning as a conditional sequence modeling problem Chen et al. (2021). Unlike traditional RL methods that learn a value function or a policy, the DT learns to model the entire trajectory of states, actions, and returns. It is conditioned on a desired outcome, specified by a target return-to-go (RTG), and autoregressively predicts the sequence of actions that will achieve it. The underlying architecture is a GPT-based causal Transformer.

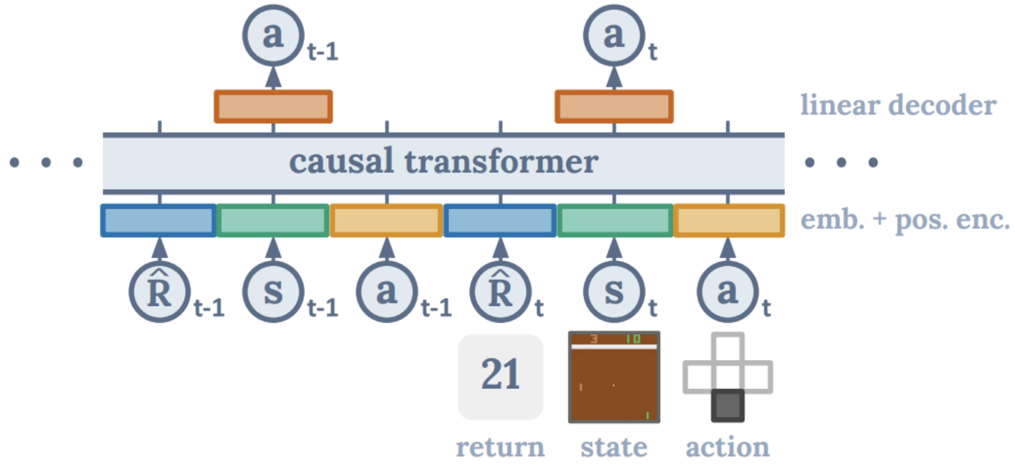

Figure 1: Decision Transformer architecture Chen et al. (2021)

Our specific implementation uses an embedding dimension ($d_{model}$) of 128, processed by 3 transformer encoder layers. Each layer contains a multi-head self-attention mechanism with 4 heads and a feedforward network with an inner dimension of 256.

**State Encoder** The state $s_t$ is a concatenated vector of multi-modal sensor data, including glycemic features, current insulin on board (IOB), activity metrics, and temporal features, as detailed in Section 3.1. To process this input, a dedicated **state encoder**, implemented as a linear layer, projects the concatenated state vector $s_t$ into the 128-dimensional model embedding space. Similarly, separate linear layers are used to encode the scalar return-to-go $\hat{R}_t$ and the discrete action $a_{t-1}$ into this same embedding space.

**Learned Positional Embeddings** The standard self-attention mechanism in a Transformer is permutation-invariant. To provide the model with crucial temporal context, we add a **learned positional embedding** to each token embedding at every timestep. An embedding layer within the model maintains a unique, trainable vector for each position index (from 1 up to the context length of 24). These positional embeddings are initialized randomly and updated via backpropagation, allowing the model to learn the significance of temporal positioning within a trajectory. The resulting sequence of summed embeddings is then passed through the Transformer blocks to predict the next action.

### 3.3 Training

The Decision Transformer model was trained using the **AdamW optimizer** Loshchilov and Hutter (2019), with learning rates explored in the range of $1 \times 10^{-5}$ to $1 \times 10^{-4}$. The training process minimized a **weighted cross-entropy loss** function. Given the nature of the dataset, where the "no action" class (ID 7) is overwhelmingly more frequent than any specific bolus action, class weighting was essential to prevent the model from overfitting to the majority class. The weights were calculated to be inversely proportional to the frequency of each action class in the training data, which significantly increases the penalty for misclassifying rare but clinically important bolus actions.

To mitigate overfitting, a dropout rate of 0.1 was applied to the attention and feedforward layers within the Transformer blocks. The model was trained using batches of 32 or 64 sequences, where each sequence had a context length of 24 timesteps, corresponding to 2 hours of historical data.

### 3.4 Evaluation

Given the significant safety risks of testing a new policy on a live subject, all model evaluation was performed offline using the held-out validation set. Our evaluation strategy combined an initial sanity check with a detailed qualitative comparison of the model's learned policy against the historical human policy.

**L1 Metric Sanity Check** In initial experiments on a smaller, over-fitted dataset, we used an L1 loss metric to measure the mean absolute difference between the model's recommended insulin dose and the historical dose. Observing a steady decrease in this metric confirmed that the model was capable of learning from the data and that the training pipeline was functioning correctly, providing a valuable sanity check before proceeding to more complex evaluations.

**Qualitative Policy Comparison** To gain an intuitive understanding of the model's learned behavior, we performed a direct policy comparison on challenging samples from the validation set. This method avoids the logical inconsistencies and error accumulation of a multi-step simulation by focusing on a single question: for a given real-world sequence of states, how would the model's recommended actions have differed from the actions actually taken?

The process is as follows: A complete historical trajectory (e.g., 2 hours, or 24 timesteps of states, actions, and returns) is selected. We feed the entire sequence of historical states into the trained Decision Transformer in a single forward pass. The model's output is a corresponding sequence of recommended actions for each timestep. The resulting visualization directly contrasts the two action sequences (historical vs. model) on a shared timeline, with the real historical glucose trajectory providing the necessary clinical context. This allows for a direct, "apples-to-apples" comparison of the two policies under the exact same conditions.

**Limitations and Future Work** While this qualitative comparison is robust and interpretable, it is an open-loop evaluation. The model's recommended actions at each timestep do not influence the future states it sees, as it only ever processes the single, fixed historical state sequence. This makes it impossible to assess the true, cumulative consequences of the model's policy. Because of this limitation, generating reliable aggregate quantitative metrics like Time-in-Range from this method is not feasible.

Therefore, robust quantitative evaluation requires future work to integrate the trained Decision Transformer policy with a validated, closed-loop physiological simulator, such as the UVA/Padova T1D simulator. This would allow the model's actions to dynamically influence the subsequent states, providing a trustworthy assessment of the policy's real-world performance potential.

# 4 Results

This section details the results from our experimental evaluations, covering both quantitative hyperparameter tuning and a qualitative analysis of the final model's learned policy.

## 4.1 Quantitative Evaluation

A series of experiments were conducted to determine the optimal hyperparameters for training the Decision Transformer on this dataset.

**Learning Rate** Learning rate tuning shows the lower learning rate converges to a much lower loss. As shown in Figure 2, a learning rate of $1 \times 10^{-4}$ (pink) led to a much more stable and lower loss profile compared to $1 \times 10^{-3}$ (orange). All subsequent experiments utilized the more stable $1 \times 10^{-4}$ or lower learning rates.

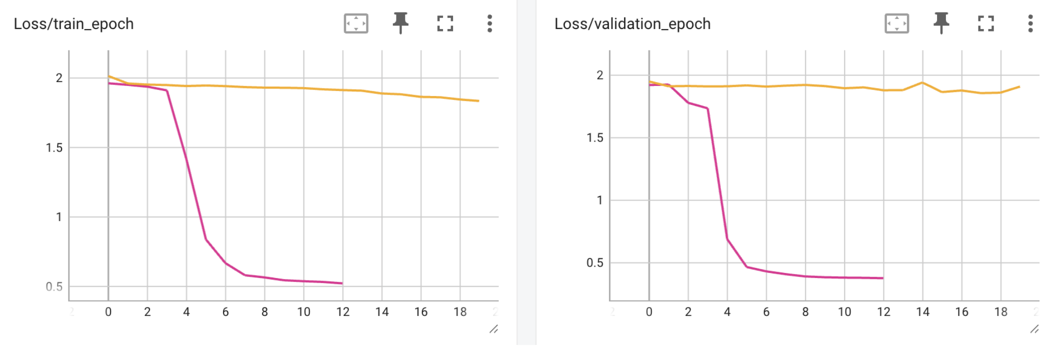

Figure 2: Loss for learning rate comparison. lr $1 \times 10^{-3}$ (orange) vs. $1 \times 10^{-4}$ (pink). Batch size 32. Context 2 hours.

**Batch Size** Batch size hyperparameter tuning (Figures 3 and 4) hows a lower batch size performs better in terms of training loss and the L1 metric, as seen in Figures 3 and 4. A batch size of 8 (black) achieved both a lower final loss and a lower L1 metric compared to a size of 32 (blue). The L1 metric, which measures the absolute difference between the model's recommended dose and the historical dose, suggests that the model trained with a smaller batch size was able to more closely mimic the patterns in the training data.

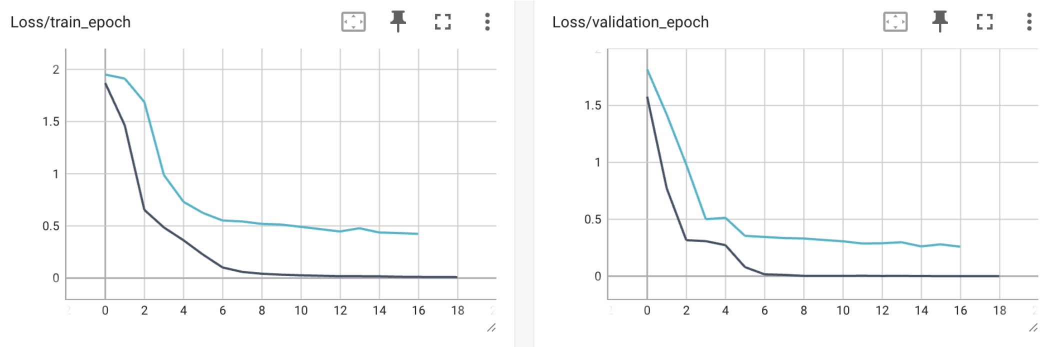

Figure 3: Loss for Batch size comparison (black=bs8, blue=bs32). Learning rate=$1 \times 10^{-4}$. Context 2 hours.

**Context Length** Context length tuning (Figure 5) showed that lengths longer than 2 hours (24 timesteps) began to reduce performance on the validation set. This may suggest that for many general-purpose decisions, more distant history introduces noise rather than valuable signal. However,

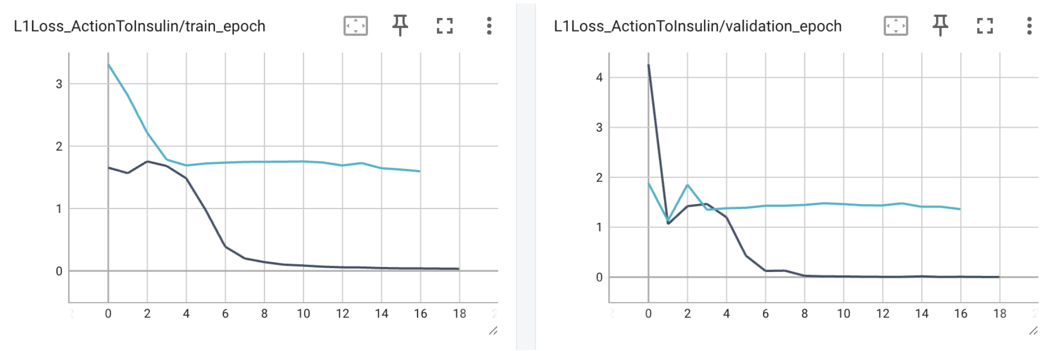

Figure 4: L1 metric compared to historical episode. For Batch size comparison (black=bs8, blue=bs32). Learning rate=$1 \times 10^{-4}$. Context 2 hours.

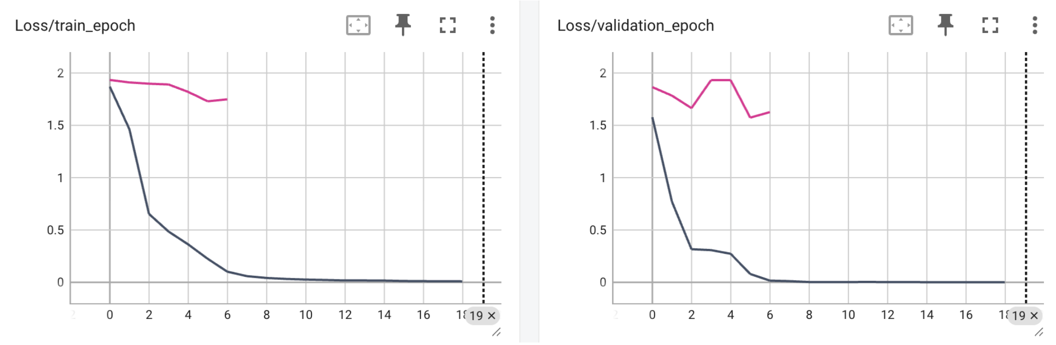

Figure 5: 6 hours context (pink) vs 2 hours.

further tuning of other hyperparameters and potentially increasing model capacity could reverse this conclusion, so we leave more extensive exploration of very long context lengths to future work.

### 4.2 Qualitative Analysis

To evaluate the clinical plausibility of the learned policy, we employed the direct policy comparison method detailed in the Method section. This analysis revealed that the model's behavior is highly sensitive to the training hyperparameters.

Initial experiments using a smaller batch size (8) and a higher learning rate ($1 \times 10^{-4}$) resulted in a policy that perfectly mimicked the historical actions, a phenomenon known as behavioral cloning. As shown in Figure 6, the model's recommended actions are identical to the historical actions. While this leads to a low training loss, it demonstrates a failure to learn a novel or potentially improved policy beyond simply copying the training data.

In contrast, after adjusting hyperparameters to use a larger batch size (64) and a lower learning rate ($1 \times 10^{-5}$), the model learned a more nuanced and proactive policy. Figure 7 shows a scenario where glucose is rising. While the human took no action, the model recommended a series of small corrective boluses. This demonstrates that with careful tuning, the Decision Transformer is capable of learning a policy that is not merely a clone of the dataset but a novel strategy that attempts to proactively correct undesirable trends.

## 5 Discussion

This project successfully demonstrated the feasibility of applying a Decision Transformer to learn a personalized insulin dosing policy from real-world, multi-modal sensor data. The key challenge in training such a model is guiding it to learn a policy that is more effective than the potentially

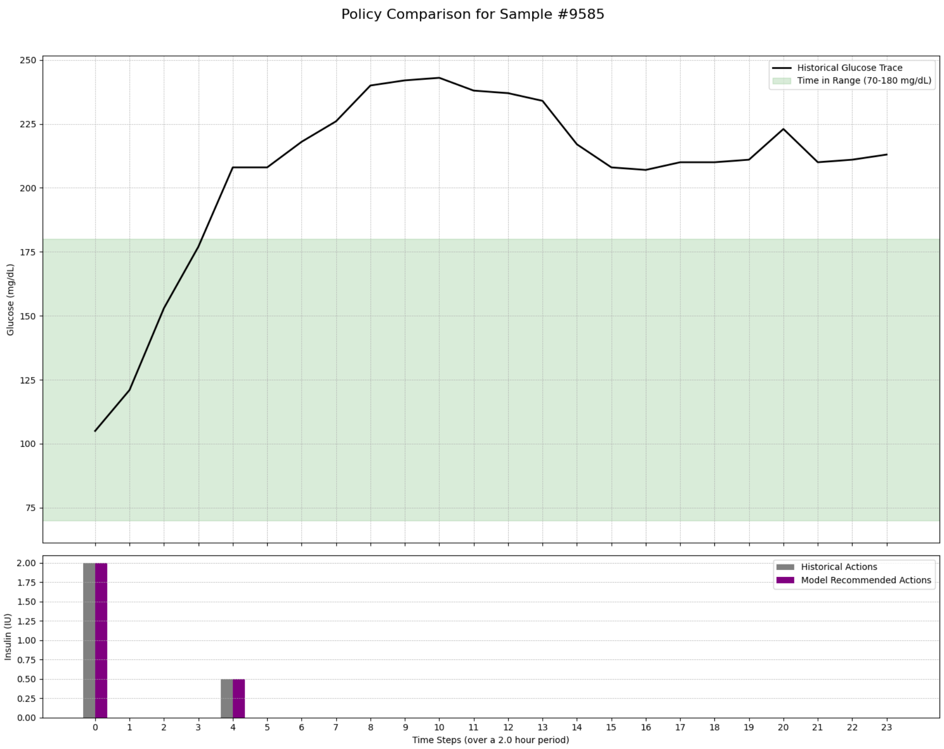

Figure 6: With a batch size of 8 and learning rate of $1 \times 10^{-4}$, the model exhibits overfitting behavior, precisely cloning the historical actions.

suboptimal behavior present in the historical data, rather than simply imitating it. Our results indicate that this is achievable, but highly sensitive to the choice of training hyperparameters.

The initial quantitative results, particularly the L1 metric, showed that certain hyperparameter settings (small batch size, high learning rate) excelled at behavioral cloning. While this minimizes training error, it fails the primary objective of finding a better policy. By adjusting the hyperparameters (larger batch size, lower learning rate), we were able to train a model that, in qualitative analysis, exhibited a more desirable, proactive policy. This suggests a trade-off between imitation and optimization that must be carefully managed. The model's ability to recommend proactive corrections, as seen in Figure 7, is a significant success and demonstrates the potential of this approach.

The primary limitation of this work remains the evaluation methodology. Our qualitative analysis is an open-loop evaluation; we can see that the model recommends a better action, but we cannot see the causal outcome of that action, as the model's decisions do not influence the future states it sees. This prevents the generation of reliable quantitative metrics, such as an improved Time-in-Range. The second limitation is the use of a single-subject dataset, which means the resulting policy is highly personalized but not generalizable.

Future work must prioritize closing the evaluation loop. The most critical next step is to integrate the promising, proactive model with a validated physiological simulator (e.g., the UVA/Padova T1D simulator). This would allow for robust, closed-loop testing of the policy's long-term effects on glycemic control and the generation of trustworthy quantitative metrics.

## 6 Conclusion

This project explored the application of Decision Transformers for personalized diabetes management, developing a complete pipeline from raw, multi-modal sensor data to a trained policy model. We demonstrated that with careful hyperparameter tuning and techniques like weighted loss functions, it is possible to train a model that goes beyond simple behavioral cloning and learns a proactive

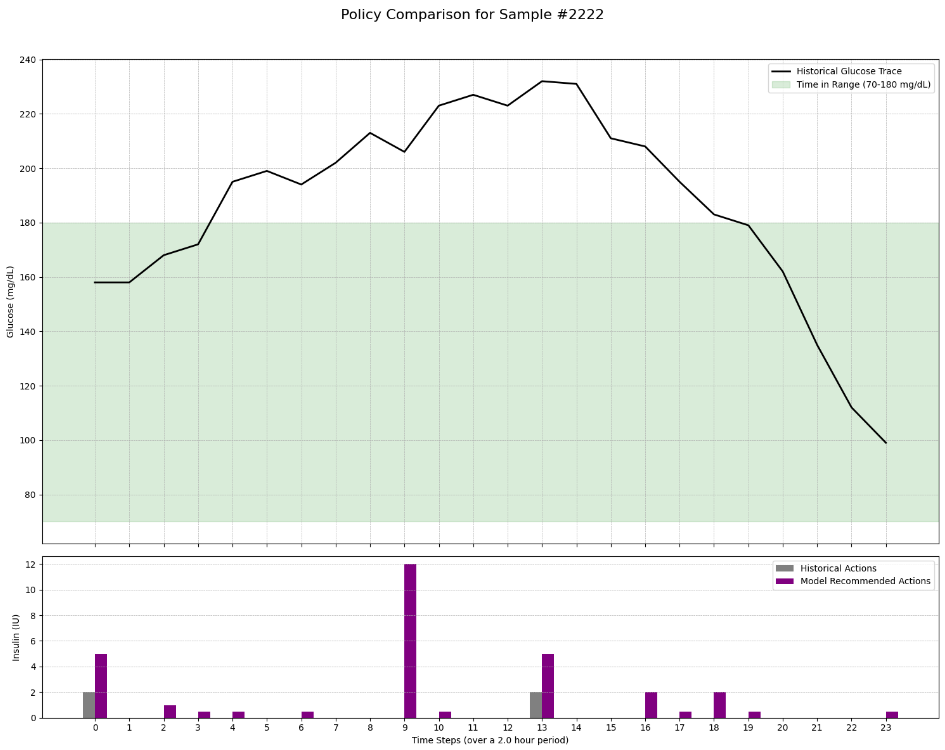

Figure 7: With a batch size of 64 and learning rate of $1 \times 10^{-5}$, the model learns a more proactive policy, recommending corrective insulin during a period of rising blood sugar where the historical user did not.

policy for recommending insulin corrections. The direct policy comparison proved to be a robust and interpretable method for qualitatively evaluating and understanding the nuanced behavior of the learned agent.

The main conclusion of this work is that Decision Transformers represent a viable and promising approach for this offline reinforcement learning problem. However, the project also underscores the profound difficulty of offline evaluation. While we have successfully trained a model that appears to make intelligent, proactive decisions, validating the real-world benefit of these decisions requires a more advanced, closed-loop simulation environment. This work serves as a successful proof-of-concept and lays the groundwork for future research integrating such learned policies with physiological simulators to truly assess their clinical potential.

**Changes from Proposal** The most significant deviation from the initial project proposal was the expansion of the project's scope. The original objective was to focus specifically on the challenging post-exercise period. However, as documented in the project milestone, this narrow focus yielded a dataset of only 159 sequences, which was insufficient for robustly training a Decision Transformer. To address this, the objective was revised to develop a general insulin delivery recommendation model, which allowed for the creation of a substantially larger and more diverse dataset from the available data. Consequently, the evaluation strategy also evolved. While the stretch goal of integrating a formal physiological simulator was not reached, this work instead focused on developing a robust qualitative analysis pipeline to directly compare the model's learned policy against the historical policy in a variety of real-world scenarios.

## References

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

Harry Emerson, Matthew Guy, and Ryan McConville. 2023. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Artificial Intelligence in Medicine* 140 (2023), 102554.

Rui Li, Jingjing Wang, Yan Li, Yifan Zhao, Xiaoyu Chen, Yi Zhang, Cheng Cheng, Xiangyu Wang, Xiaoyan Wei, and Bin Liu. 2024. Exploring the potential of deep learning models integrating transformer and LSTM in predicting blood glucose levels for T1D patients. *Computer Methods and Programs in Biomedicine* 244 (2024), 107936.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Martin Pajech, Corentin Sautier, Laurent F Gauthier, Michel P Bonnier, Dana Lewis, Pau Cruañas, Scott Smedley, Pierre-Yves Oudeyer, and Martin F Hemberg. 2022. Offline reinforcement learning for personalized blood glucose control in type 1 diabetes. *arXiv preprint arXiv:2204.03376* (2022).

Phuwadol Viroonluecha, Francisco Javier Egea Lopez, Diego Garcia-Vazquez, and Roberto Henriques. 2023. Evaluation of offline reinforcement learning for blood glucose level control in type 1 diabetes. *Plos one* 18, 3 (2023), e0274608.

Kok-Lim Alvin Yau, Yung-Wey Chong, Xiumei Fan, Celimuge Wu, Yasir Saleem, and Phei-Ching Lim. 2023. Reinforcement learning models and algorithms for diabetes management. *IEEE Reviews in Biomedical Engineering* (2023).

Chao Yu, Jiming Liu, Shamim Nemati, and Guoshuai Zhao. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–36.

Zhiyue Zhang, Hongyuan Mei, and Yanxun Xu. 2023. Continuous-time decision transformer for healthcare applications. In *Conference on Health, Inference, and Learning*. PMLR, 284–301.

## A Additional Experiments

In additional experiments, unweighted cross entropy loss demonstrated the importance of class weighting. Without the weighting, the model learns to always select no insulin (the most frequent in the dataset). See Figures 8 and 9.

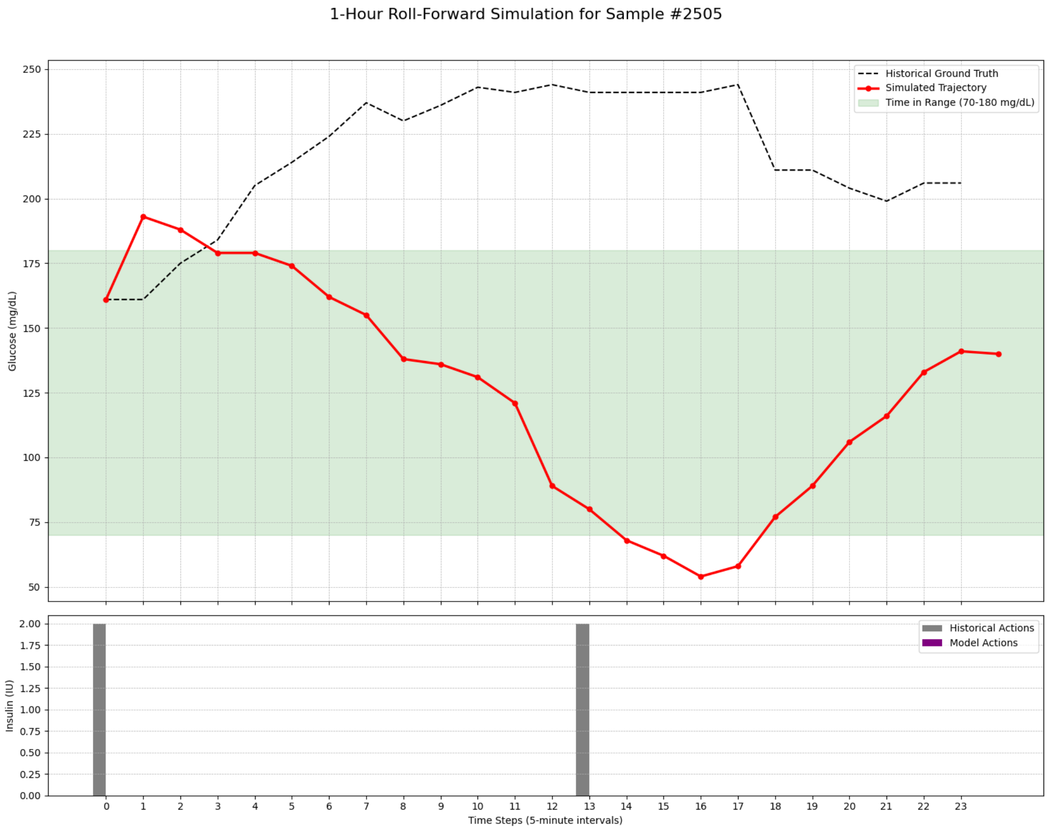

Figure 8: Example with unweighted cross-entropy loss. All of the qualitative examples are similar because the model learned to always select 0 insulin.

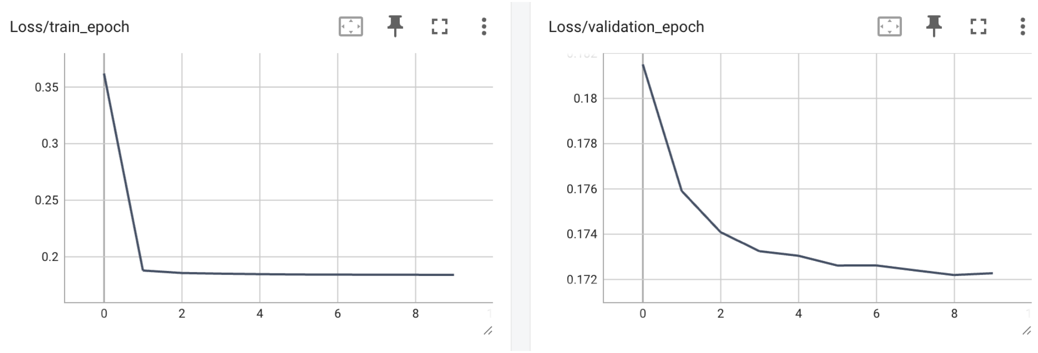

Figure 9: Loss with unweighted cross-entropy loss.