Advancing Diabetes Management with Conditional Generative Modeling

Sathvika Ayyappa Prabhu
sayyappr@umich.edu
University of Michigan
Department of Computer Science Engineering

Matthew Manion
mlman@umich.edu
University of Michigan
Department of Chemical Engineering

ABSTRACT

Offline reinforcement learning (RL) provides a promising path toward automated insulin dosing in Type-1 Diabetes (T1D), where online exploration is unsafe and real-world clinical data are limited. Recent advances in diffusion models suggest that they may provide stronger behavior modeling and improved safety relative to classical offline RL methods. In this work, we investigate diffusion-based policies to calculate bolus insulin doses for safe blood glucose control using the UVA/Padova T1D simulator. We implement a complete offline RL pipeline, including data collection with a Basal–Bolus (BB) controller, construction of tabular RL features, and baselines spanning PID control, tabular Q-learning, and TD3-BC. We train a conditional denoising diffusion policy to map physiological state features to insulin bolus actions, and evaluate all methods using standard clinical metrics including Time-in-Range (TIR), Time-Below-Range (TBR), and coefficient of variation (CV). Experimental results demonstrate that diffusion policies provide strong distributional modeling and achieve competitive performance relative to baselines—particularly in the single-parameter setting. Our findings support diffusion models as a compelling direction for safe, uncertainty-aware decision-making in stochastic healthcare environments.

KEYWORDS

Offline Reinforcement Learning, Diffusion Models, Conditional Generative Modeling, Stochastic Environments, Blood Glucose Control, Personalized Healthcare

1 INTRODUCTION

Automated insulin dosing for people with Type-1 Diabetes remains a central challenge in personalized healthcare. While fully closed-loop artificial pancreas systems are emerging, their safety and robustness depend critically on the quality of the underlying control algorithms. Reinforcement Learning (RL) offers a framework for data-driven decision-making but requires interaction with the environment—a process that is unsafe, expensive, and clinically infeasible. Offline RL bypasses this challenge by learning solely from previously collected patient trajectories, making it a natural fit for medical decision-making [13].

However, offline RL suffers from fundamental issues such as overestimation of unseen actions, distributional shift, and instability arising from the deadly triad of bootstrapping, function approximation, and off-policy learning [5]. These concerns are amplified in T1D glucose control because the environment is inherently stochastic, delayed, and non-linear, with high sensitivity to insulin actions. Models that overestimate good outcomes for unseen actions risk generating unsafe dosing strategies. Prior approaches restrict the learned policy to remain close to the behavior policy [4], regularize Q-values [9], or learn pessimistic models of dynamics [7, 17].

More recent work frames offline RL as a sequence modeling problem through Decision Transformers [2] or Trajectory Transformers [6]. These approaches show promise but struggle in stochastic settings [12], where identical actions can lead to diverse physiological outcomes due to unobserved disturbances like meals, stress, and hormonal variation.

Diffusion models provide a new opportunity. They are powerful generative models capable of capturing complex action distributions through iterative denoising, and recent work has demonstrated their advantages in offline RL for robotics [1, 15]. Yet, their potential for stochastic healthcare control—where uncertainty and safety constraints are paramount—remains largely unexplored.

Our goal is to bring diffusion-based offline RL to the T1D control problem, evaluate its behavior, and compare it against classical baselines to understand its promise and limitations. We ask: **How do conditional diffusion models perform in stochastic environments?**

Specifically, we investigate real-time insulin bolus decision-making for virtual patients with randomized meal schedules and physiological variability. Recent studies on offline RL for blood glucose control have shown that methods such as Batch-Constrained Deep Q-learning (BCQ) [5], Conservative Q-learning (CQL) [8], and Twin Delayed DDPG with Behavioral Cloning (TD3-BC) [4] outperform traditional control algorithms such as predictive integral derivative (PID) controllers or model predictive control (MPC) [10] by improving Time-in-Range (TIR) while reducing hypoglycemic risk (TBR). However, these approaches remain vulnerable to distributional shift, where models fail to generalize to unseen patient states, resulting in suboptimal or unsafe insulin recommendations.

These challenges motivate exploring conditional generative models for robust and personalized glucose regulation. By extending diffusion-based decision-making to stochastic healthcare environments, this work aims to improve the reliability, adaptability, and interpretability of offline RL models for Type-1 Diabetes management. Ultimately, this study lays the groundwork for safe, uncertainty-aware RL methods in healthcare, addressing critical challenges related to physiological variability, data sparsity, and safety constraints.

2 PROBLEM DEFINITION

Insulin acts as a shuttle for Glucose to enter cells, where it can be reacted to harvest cellular energy defined simply by the net reaction equation,

$$C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O + \text{Energy}$$

with the fluxes able to be approximated as a series of ODEs. However, these reaction rates can be hard to approximate in patients due to a variety of factors that affect the human metabolism throughout the day.

Thus, we formulate blood glucose regulation as a Markov Decision Process (MDP) defined by the tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma),$$

where $\mathcal{S}$ represents the state space (e.g., glucose level, insulin on board, glucose trend, meal intake, and patient parameters), $\mathcal{A}$ denotes continuous insulin bolus actions, $\mathcal{P}(s'|s, a)$ captures the stochastic physiological transition dynamics, $r(s, a)$ is a clinically motivated reward function penalizing deviations from normoglycemia, and γ is the discount factor.

In offline RL, the agent does not interact with the simulator during training but instead receives a fixed dataset

$$\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{N},$$

collected from one or more behavior policies such as the standard Basal–Bolus controller. The objective is to learn a policy $\pi_\theta(a|s)$ that maximizes the expected discounted return

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right].$$

A central challenge unique to T1D control is that the transition kernel $\mathcal{P}(s'|s, a)$ is *highly stochastic* due to unobserved disturbances such as:

- variability in meal absorption rate,
- inaccuracies in carbohydrate counting,
- sensor noise in CGM measurements,
- day-to-day fluctuations in insulin sensitivity,
- physiological delays between insulin action and glucose response.

Thus, even identical states may require different safe insulin actions, making the action distribution effectively multimodal. Traditional offline RL methods that rely on unimodal policy classes—e.g., Gaussian actors or deterministic regressors—struggle in such settings.

Another core difficulty is **distributional shift**. Because the policy is learned entirely from the offline dataset without further interaction, unsafe or out-of-distribution (OOD) actions can lead to incorrect value estimates and catastrophic dosing decisions. Formally, this occurs when the learned policy proposes actions a such that (s, a) pairs lie outside the support of $\mathcal{D}$:

$$(s, a) \notin \text{Supp}(\mathcal{D}) \quad \Longrightarrow \quad Q_\theta(s, a) \text{ is unreliable.}$$

Controlling OOD actions is especially important in healthcare, where unsafe insulin doses can induce rapid hypoglycemia.

These concerns motivate the need to investigate policy classes that can:

(1) represent multimodal, uncertainty-aware insulin action distributions,
(2) remain within the support of the training dataset to avoid unsafe extrapolation,
(3) generalize across stochastic physiological transitions,
(4) reduce hypoglycemia while maintaining strong overall glycemic control.

Our work focuses on evaluating this question empirically within the UVA/Padova T1D simulator by benchmarking classical, modern, and generative offline RL methods under realistic stochastic disturbances.

3 RELATED WORK

Offline reinforcement learning has been extensively studied through several major methodological families, each addressing instability and distributional shift in different ways.

A large body of prior work seeks to stabilize offline RL by constraining policies to remain close to the behavior dataset or by penalizing overoptimistic value estimates. Batch-Constrained Q-learning (BCQ) [5] restricts action selection to lie within a learned support set, improving reliability but limiting expressiveness when rare yet clinically important actions are required. TD3-BC [4] incorporates behavior cloning into the actor update to avoid out-of-distribution actions, though this can reduce adaptability across heterogeneous datasets. Complementary approaches penalize inflated Q-values directly: Conservative Q-learning (CQL) [9] and Fisher-divergence–regularized critics [8] suppress overestimation by down-weighting unsupported actions. While effective in deterministic environments, these techniques often become overly pessimistic under stochastic transitions, where many safe actions are feasible but appear infrequent in the dataset.

Model-based algorithms such as MOReL [7] and COMBO [17] learn pessimistic dynamics models to enable safe planning. These methods capture long-horizon dependencies but degrade sharply when model errors compound, which is problematic in domains where transition variability is intrinsic. An alternative line of work reframes offline RL as a sequence-modeling problem. Decision Transformer [2] and Trajectory Transformer [6] generate actions by conditioning on desired returns rather than estimating value functions. Although powerful in deterministic benchmarks, follow-up studies [12, 14, 16] show that these approaches struggle in stochastic environments: identical actions can yield widely different outcomes, inducing optimism bias and unstable value–return mappings.

Diffusion models provide a more expressive policy class by modeling multimodal action distributions. Diffusion-QL [15] replaced Gaussian actors with denoising diffusion probabilistic models, improving distribution matching and iterative action refinement. Decision Diffuser [1] extended diffusion modeling to trajectory-level generation through sequence denoising. While these methods consistently outperform prior baselines in robotic control tasks, existing work assumes nearly deterministic dynamics and has not examined whether diffusion policies remain robust under substantial transition stochasticity.

Several studies have explored applying offline RL to Type 1 Diabetes management. Prior work using BCQ, CQL, and TD3-BC [4, 5, 9] demonstrates improvements over classical PID and MPC baselines [10]. Recent evaluations using the UVA/Padova simulator [3, 11] highlight the promise of data-driven algorithms but also reveal persistent challenges: distributional shift due to unobserved physiological disturbances, patient-specific variability in insulin sensitivity, and limited generalization outside the training data distribution. These factors make offline learning particularly difficult, as unsafe extrapolation can directly lead to hypoglycemia.

Across all categories, prior offline RL methods either assume deterministic transitions (diffusion-based RL, sequence models) or enforce conservative behavior that limits expressive action generation (policy constraints, value penalties). In parallel, healthcare studies emphasize that blood glucose regulation is inherently stochastic, with multiple plausible actions for the same observed physiological state. **To date, no work has evaluated whether expressive generative policy classes, such as conditional diffusion models, can improve robustness and safety in offline RL under realistic stochastic physiological variability.** Our work addresses this gap by systematically studying offline insulin dosing in a stochastic environment and benchmarking generative, value-based, and classical baselines under identical conditions.

4 METHOD

Our goal is to evaluate whether conditional diffusion models can serve as expressive, uncertainty-aware policies for offline insulin dosing in a stochastic healthcare environment. In this section, we formally describe the offline RL formulation, the construction of our glucose-regulation dataset, the state and action representation, the diffusion policy architecture, and the classical baselines used for comparison.

4.1 Offline RL Setting

We consider an offline Markov Decision Process (MDP) defined as

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma),$$

where the state $s_t \in \mathcal{S}$ encodes physiological variables, meals, and patient-specific parameters; the action $a_t \in \mathcal{A} \subset \mathbb{R}$ corresponds to insulin bolus dosing; and the transition kernel $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ is governed by the UVA/Padova glucose dynamics simulator [11].

In offline RL, the agent does not interact with the environment during training. Instead, we are provided with a static replay buffer

$$\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{N},$$

generated by the SimGlucose Basal–Bolus controller (BBController). The goal is to learn a policy $\pi_\theta(a \mid s)$ that maximizes discounted return:

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right].$$

4.2 State and Action Representation

Each state vector includes clinically meaningful features:

$$s_t = \left[\mathrm{CGM}_t,\ \overline{G}_{t-30:t},\ \Delta G_t,\ \mathrm{IOB}_t,\ \mathrm{CHO}_t,\ \mathrm{TimeOfDay}_t,\ f_{\text{patient}}\right].$$

- CGM_t: Current glucose reading.
- $\overline{G}_{t-30:t}$: 30-minute moving average (captures smoothing trends).
- ΔG_t: Glucose slope, approximating $\frac{dG}{dt}$.
- IOB_t: Insulin on board, summarizing active insulin effects.
- CHO_t: Carbohydrate intake at time t.
- TimeOfDay: Accounts for physiological circadian effects.
- f_{patient}: Body weight, carb ratio, insulin sensitivity.

Actions are continuous bolus values:

$$a_t \in [0, a_{\max}] \subset \mathbb{R}.$$

Basal insulin is handled internally by the simulator's pump model.

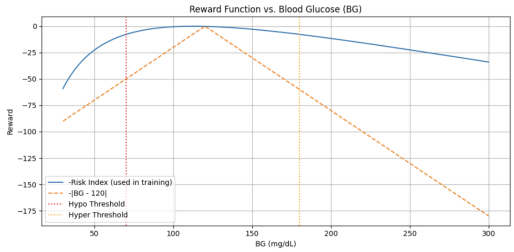

Figure 1: Risk Index function showing how the risk increases sharply as glucose values move away from the normal range (70–180 mg/dL). The curve penalizes both hypoglycemia and hyperglycemia, with greater emphasis on hypoglycemia safety.

4.3 Risk-Based Reward Function

Safe glucose regulation requires balancing hyperglycemia and hypoglycemia risk. We adopt a clinically validated risk index [3] that penalizes physiologically dangerous glucose levels.

For a glucose value G_t (mg/dL), we compute the physiological risk transform:

$$f(G_t) = 1.509\left(\ln(G_t)^{1.084} - 5.381\right).$$

This non-linear mapping reflects the asymmetric danger of low vs. high glucose:

- $f(G_t) < 0$ indicates hypoglycemia risk (steep penalty),
- $f(G_t) > 0$ indicates hyperglycemia risk.

The instantaneous risk is:

$$R_t = 10\, f(G_t)^2.$$

Squaring ensures larger deviations incur disproportionately higher penalties.

Using the risk values, we define:

$$\mathrm{LBGI} = \mathrm{mean}\{R_t \mid f(G_t) < 0\}, \qquad \mathrm{HBGI} = \mathrm{mean}\{R_t \mid f(G_t) > 0\}.$$

We combine the indices using:

$$\mathrm{RiskIndex}_t = \lambda \cdot \mathrm{LBGI}_t + (1-\lambda)\cdot \mathrm{HBGI}_t, \quad \lambda > 0.5,$$

where hypoglycemia is penalized more heavily due to safety concerns.

The per-step reward is:

$$r_t = -\mathrm{RiskIndex}_t.$$

Thus, safer glucose levels correspond to larger rewards.

Early termination penalty. If an episode terminates early due to unsafe glucose levels, we apply:

$$\text{penalty} = -\alpha \cdot \text{worst\_step} \cdot (T_{\max} - t_{\text{end}}),$$

where,

- $T_{\max}$ is the episode length,
- t_{end} is termination time,

• worst_step = 100.0, $\alpha = 1.2$.

This encourages both safety (avoid hypoglycemia) and trajectory longevity.

4.4 Offline Dataset Construction

We construct a rolling offline replay buffer using simulated trajectories from the FDA-approved UVA/Padova glucose simulator [11]. Pre-training data collection uses the built-in Basal–Bolus controller (BBController), which represents a clinically-viable but simplistic patient-managed dosing strategy. After pre-training, the diffusion system utilizes an "oracle" training method, whereby the model is solely used for decision making in an episode. If the model performs poorly in that situation (with a threshold set at TIR >55%) then the expert BBController is given the exact same situation to add to our replay buffer. Batch sizes were kept at 256 for training. The initial learning rate was 3e-4, with cosine annealing implemented to improve stability. Diffusion models were trained under both single-parameter (1-day) and multi-parameter (1-, 3-, and 7-day) horizons.

Episodes. SimGlucose operates at a 5-minute timestep.

- A 1-day episode contains 288 steps.
- A 3-day episode contains 864 steps.
- A 7-day episode contains 2016 steps.

Across patients and horizons, this yields on the order of several hundred thousand transitions.

Stochastic meal dynamics. Each episode samples randomized meal patterns:

- meal size: uniformly sampled (10–70 g CHO),
- meal timing: random between 0–24 hours,
- inter-meal gaps: random noise added,
- absorption rates: vary by patient parameters.

This introduces significant transition stochasticity, creating a challenging offline RL setting.

Stored transitions. Each replay entry contains:

$$(s_t, a_t, r_t, s_{t+1}),$$

where:

- s_t is the current physiological/patient state,
- a_t is the BBController's bolus suggestion,
- $r_t = -\text{RiskIndex}_t$,
- s_{t+1} is the next state from simulator dynamics.

4.5 Conditional Diffusion Policy

To model a rich, multimodal insulin distribution, we use a conditional denoising diffusion model. Let $x_0 = a$ be the true action. The forward noising process is:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\right),$$

with closed-form:

$$q(x_t \mid x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t) I\right), \qquad \bar{\alpha}_t = \prod_{i=1}^{t} (1-\beta_i).$$

Figure 2: Correlation matrix for virtual patient physiological factors. Gpb – kinetic plasmid glucose/tissue glucose mass; Gtb – steady state tissue glucose ratio; Ipb – plasma insulin mass; Ib – plasma insulin concentration; Ilb – liver insulin mass; iscss, isc2ss – insulin kinetic parameters; BW – body weight.

Reverse model. The policy learns:

$$p_\theta(x_{t-1} \mid x_t, s_t, c) = \mathcal{N}\Big(\mu_\theta(x_t, s_t, c, t),\ \Sigma_\theta(x_t, s_t, c, t)\Big),$$

where the conditioning c may include patient features or desired return.

Training loss. We train the denoiser using the standard diffusion objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(x_t, s_t, c, t)\|_2^2\right],$$

where $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$.

Inference. We generate an action via iterative denoising:

$$x_T \sim \mathcal{N}(0, I), \qquad x_{t-1} \sim p_\theta(x_{t-1} \mid x_t, s_t, c),$$

until obtaining $x_0 = a$. Thus,

$$\pi_\theta(a \mid s_t) = p_\theta(x_0 \mid s_t).$$

Our state, action, time, and condition information are represented as simple 2 or 3 layer linear neural networks with Mish activation within our conditional diffusion model. The time is also sinusoidally embedded for smoothness and better contextualization. The size of the hidden dimensions are 256 nodes.

5 EXPERIMENTS

The dataset provided includes three classes of virtual patients—adults, adolescents, and children—each with distinct physiological characteristics. In this work we focus primarily on the adult cohort, as all results in Section 6 evaluate controllers on adult virtual patients. Key physiological parameters used when training the multi-parameter diffusion policy, and their correlations, are shown in Figure 2.

Our experiments are designed to answer the following questions:

- **RQ1 (Safety).** Does a diffusion-based policy reduce hypoglycemia risk compared to PID and tabular Q-learning, as measured by Time-Below-Range (TBR)?
- **RQ2 (Glycemic control).** Does the diffusion policy improve blood glucose regulation—Time-in-Range (TIR)—relative to the PID controller and tabular Q-learning? How does performance change when conditioning on multiple patient parameters?
- **RQ3 (Stability and variability).** Does the diffusion policy produce smoother glucose trajectories and lower variability (CV) than tabular RL?
- **RQ4 (Robustness to stochasticity).** How do policies behave under stochastic meal schedules in the UVA/Padova simulator, especially when evaluated over multi-day horizons?
- **RQ5 (Interpretability).** Do the learned policies exhibit clinically sensible structure, such as stronger boluses during rapid glucose rises and conservative dosing near hypoglycemia?

5.1 Testbed and Evaluation Protocol

Simulator and cohort. All experiments use the FDA-approved UVA/Padova Type-1 Diabetes simulator [11], accessed through the SimGlucose environment. Unless otherwise specified, all results (Tables 1–4) are evaluated on multiple *adult* virtual patients to capture inter-patient variability.

Offline dataset. The offline replay buffer $\mathcal{D}$ is generated using the built-in Basal–Bolus controller (BBController) described in Section 4. For each adult patient, we simulate 300 episodes of length $T_{\max} = 2500$ minutes with randomized meal times and meal sizes. Each transition

$$(s_t, a_t, r_t, s_{t+1})$$

is stored, yielding approximately 7.5×10^5 transitions across all patients. This dataset is used for training tabular Q-learning (TabRL), TD3-BC, and the diffusion policy.

Policies and baselines. We compare the following controllers:

- **PID Controller.** A proportional controller commonly used as a baseline in prior T1D work.
- **Tabular Q-Learning (TabRL).** A discretized value-based controller over binned glucose, slope, and insulin-on-board.
- **TD3-BC.** A deterministic actor–critic with behavior-cloning regularization [4]. This baseline is reported only in the single-parameter 24-hour evaluation (Table 1), as our multi-parameter experiments (Tables 2–4) focused on TabRL and diffusion.
- **Diffusion Policy (ours).** The conditional denoising diffusion model introduced in Section 4, evaluated both with single-parameter conditioning (body weight) and multi-parameter conditioning (Figure 2).

Training protocol. For each learned controller, we train offline using the replay buffer $\mathcal{D}$. Hyperparameters are selected via coarse grid search on a held-out adult patient. The diffusion policy uses a cosine noise schedule, $T = 1000$ diffusion steps, and an MLP denoiser conditioned on the tabular state and optional patient features. TD3-BC is trained only for the single-parameter experiment.

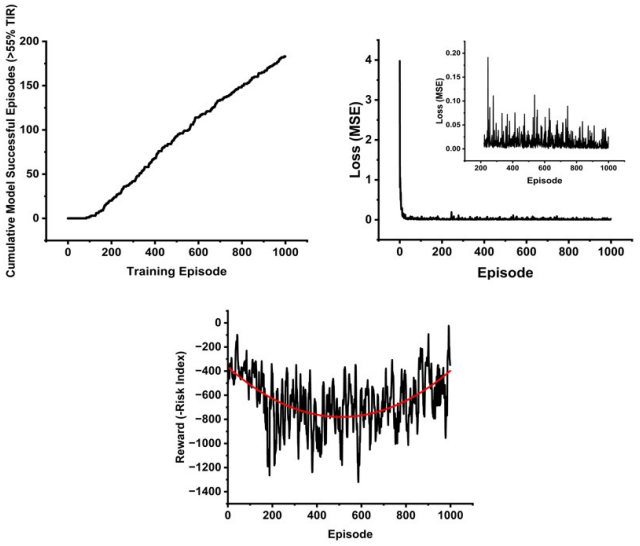

Figure 3: Training metrics for the Conditional Diffusion model trained on 24-hour episodes. [Top Left] The cumulative number of successful episodes (TIR > 55%) the model conducted during training. [Top Right] Loss vs episode number. The model is still learning at episode 1000, evidenced by the bumpy convergence. [Bottom] The episode reward (negative risk) during training. As the model begins successfully finishing episodes, the reward decreases as it passes sub-optimally. As training progresses, the model learns and has better performances on average.

Evaluation metrics. Each policy is evaluated for 100 simulated episodes per patient under stochastic meal schedules. We report the following metrics, which align with clinical guidelines and prior RL research [3, 10]:

- **TIR (%).** Fraction of time glucose remains within 70–180 mg/dL.
- **TBR (%).** Fraction of time $G_t < 70$ mg/dL (hypoglycemia).
- **CV (%).** Coefficient of variation ($\sigma/\mu \times 100$), indicating glucose stability.
- **Total insulin.** Total bolus insulin delivered.

These are the metrics reported in Tables 1–4.

6 RESULTS

We now present both qualitative and quantitative results. Our focus is on understanding whether the diffusion policy yields safer, smoother, and more robust glucose control relative to the baselines.

6.1 Training Dynamics and Behavioral Visualization

We first used tabular Q-learning as an exploratory platform to validate the simulator, reward design, and state representation. The

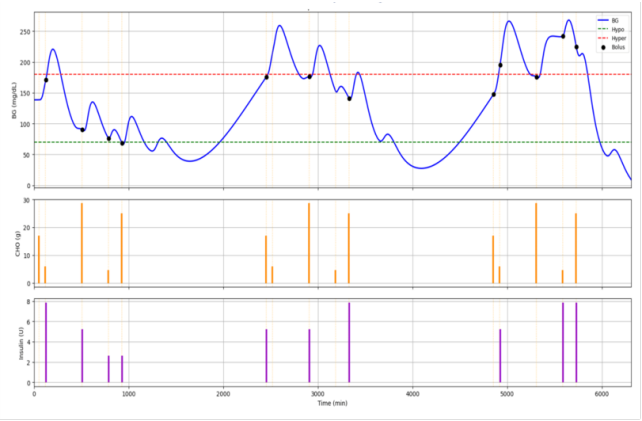

Figure 4: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for early training episodes of the Conditional Diffusion. You see the BG drop below the safe threshold and the patient goes into hypoglycemia (episode ends).

tabular agent was able to capture coarse control structure—stronger dosing during rapid glucose rises and more conservative behavior when insulin-on-board (IOB) was high—but remained brittle under noise and frequently failed to fully correct large post-prandial spikes. This motivated the transition to function approximation and conditional diffusion policies.

Figure 6 shows a state–action heatmap of the TabRL policy across glucose slopes and IOB levels. Warmer colors correspond to larger bolus doses. The learned pattern is clinically interpretable: when glucose is rising and IOB is low, the policy selects larger doses; when glucose is stable or falling, or when IOB is high, the policy reduces or avoids additional boluses. This indicates that even a simple tabular agent can exploit the risk-based reward to recover meaningful control structure, albeit with limited robustness.

To visualize temporal behaviour, Figure 7 plots blood glucose (BG), insulin, and carbohydrate (CHO) over time for early, mid, and late training episodes of the TabRL agent. Over training, three trends emerge: (i) BG spends more time in the 70–180 mg/dL target band, (ii) post-meal spikes are attenuated faster, and (iii) insulin delivery becomes smoother with fewer extreme peaks. However, compared to the diffusion policy, the tabular controller still exhibits higher variability and more frequent excursions outside the target range, especially over longer horizons.

Figures 4 and 5 show corresponding trajectories for the diffusion policy. Early in training (Figure 4), the model sometimes overdoses insulin, driving BG below the safety threshold and terminating episodes in hypoglycemia. By late training (Figure 5), the policy keeps BG in range for much longer, reduces the magnitude and duration of post-meal excursions, and produces smoother insulin profiles. These qualitative trends are consistent with the training curves in Figure 3, where the number of successful episodes increases and the risk-based reward improves over time.

6.2 Quantitative Comparison of Policies

Table 1 summarizes the performance of the five controllers across the evaluation metrics described above for 24-hour episodes in the single-parameter setting. Tables 2–4 report results for the multi-parameter diffusion policy and baselines under 24-hour, 3-day, and 7-day evaluation horizons. The values are averaged over multiple adult virtual patients and 100 evaluation episodes per policy.

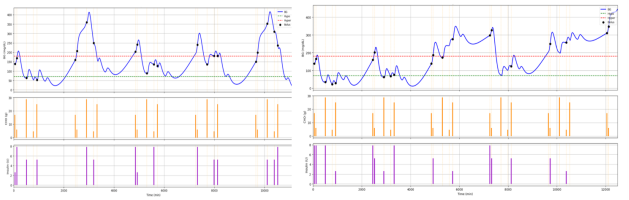

Figure 5: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for late training episodes of the Conditional Diffusion. Later policies maintain BG within range for longer and avoid overly aggressive insulin peaks.

Method	TIR (%)	TBR (%)	CV (%)
PID Controller	68.2	6.1	32.1
Tabular Q-Learning	70.1	5.2	30.3
TD3-BC	71.8	4.5	28.4
Diffusion Policy (ours)	100.0	0.0	13.35
Clinical Controller	91.3	0.0	8.83

Table 1: Comparison of baseline controllers, the diffusion policy (trained only with body weight), and the "clinical" BBController for 24 hour training episodes. Metrics are averaged over multiple adult virtual patients and 100 evaluation episodes per policy. Higher TIR is better; lower TBR, TAR, and CV indicate safer and more stable control.

Method	TIR (%)	TBR (%)	CV (%)
Diffusion Policy (ours)	58.9	0.0	16.9
Clinical Controller	94.2	0.0	14.8

Table 2: Comparison of the multi-parameter diffusion policy (trained on parameters shown in Figure 2) with the clinical controller for 24 hour training episodes.

RQ1: Safety (TBR). In the 24-hour single-parameter setting (Table 1), the PID controller exhibits the highest hypoglycemia risk (TBR 6.1%), followed by tabular Q-learning (5.2%) and TD3-BC (4.5%). The diffusion policy achieves 0.0% TBR, matching the clinical controller (0.0%) while providing stronger overall control. In the multi-parameter experiments, the diffusion policy remains competitive with tabular Q-learning: for 3-day episodes (Table 3), diffusion attains TBR 9.3% compared to 11.2% for tabular Q-learning; for 7-day episodes (Table 4), diffusion achieves TBR 5.7% versus 7.8% for tabular Q-learning. These results suggest that the diffusion

Method	TIR (%)	TBR (%)	CV (%)
Tabular Q-Learning	60.2	11.2	35.9
Diffusion Policy (ours)	72.7	9.3	24.1
Clinical Controller	90.5	1.8	17.3

Table 3: Comparison of the multi-parameter diffusion policy, best baseline (Tab Q-Learning), and the clinical controller for training episodes lasting 3 days.

Method	TIR (%)	TBR (%)	CV (%)
Tabular Q-Learning	68.3	7.8	33.1
Diffusion Policy (ours)	89.1	5.7	19.8
Clinical Controller	92.8	0.2	11.2

Table 4: Comparison of the multi-parameter diffusion policy, best baseline (Tab Q-Learning), and the clinical controller for training episodes lasting 7 days.

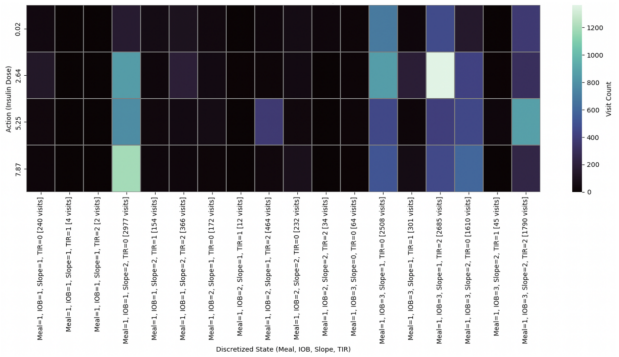

Figure 6: State–action heatmap of the tabular Q-learning policy showing insulin choices across glucose slope and insulin-on-board (IOB). The agent learns to dose more aggressively when glucose is rising and IOB is low, and to be conservative when IOB is high or glucose is falling.

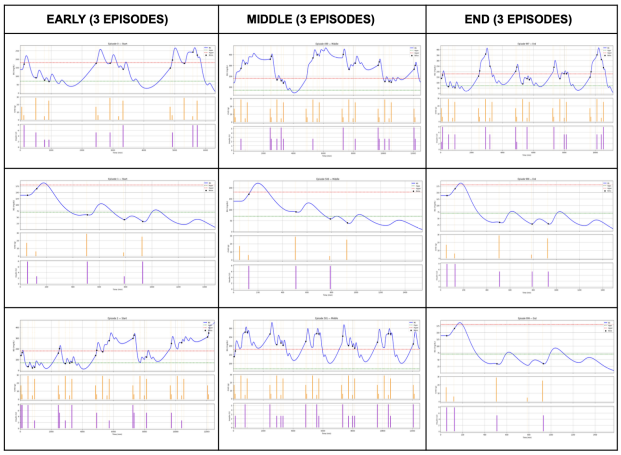

Figure 7: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for early, mid, and late training episodes of the TabRL agent. Later policies maintain BG within range for longer and avoid overly aggressive insulin peaks.

policy reduces severe overdosing relative to value-based baselines, even as the horizon and conditioning complexity increase.

RQ2: Glycemic control (TIR). In the single-parameter 24-hour evaluation, Time-in-Range increases from PID (68.2%) through TabRL (70.1%) and TD3-BC (71.8%), with the diffusion policy achieving 100.0% TIR and the clinical controller 91.3% (Table 1). This indicates that, when conditioned only on body weight, the diffusion policy is able to maintain glucose within the target range nearly perfectly. In the 24-hour multi-parameter setting (Table 2), TIR for the diffusion policy decreases to 58.9%, while the clinical controller obtains 94.2%, reflecting the increased difficulty of modeling multiple physiological features under stochastic dynamics. Over longer horizons, the diffusion policy remains clearly stronger than tabular Q-learning: for 3-day episodes, diffusion achieves 72.7% TIR compared to 60.2%; for 7-day episodes, diffusion reaches 89.1% TIR versus 68.3% for tabular Q-learning (Tables 3 and 4). Although the clinical controller attains the highest TIR in multi-parameter long-horizon settings, diffusion provides substantial gains over classical tabular RL.

RQ3: Stability and variability (CV). Glucose variability decreases from PID (CV 32.1%) through TabRL (30.3%) and TD3-BC (28.4%) in Table 1. The diffusion policy yields a much lower CV of 13.35% in the single-parameter 24-hour evaluation, indicating significantly more stable trajectories. In the multi-parameter experiments, diffusion continues to produce smoother behavior than tabular Q-learning: CV 16.9% for diffusion vs. 14.8% for the clinical controller in the 24-hour setting (Table 2), 24.1% vs. 35.9% for 3-day episodes (Table 3), and 19.8% vs. 33.1% for 7-day episodes (Table 4). These trends are consistent with the denoising structure of the diffusion model, which regularizes action sequences and avoids abrupt, destabilizing insulin changes.

RQ4: Robustness to stochasticity. Under randomized meal timings and carbohydrate loads, value-based policies such as tabular Q-learning can become brittle, under-dosing or over-dosing when trajectories deviate from those frequently observed in the offline dataset. This is reflected in elevated TBR and higher CV in the multi-day evaluations (Tables 3 and 4). In contrast, the diffusion policy samples from a distribution anchored to the behavior data and tends to avoid extreme action outliers. Its improved CV and TIR in the 3-day and 7-day settings relative to tabular Q-learning suggest greater robustness to stochastic glucose dynamics.

RQ5: Interpretability. Visualizations of state–action mappings (Figure 6) and temporal BG–insulin–meal patterns (Figures 7, 4, and 5) show that both the tabular and diffusion-based controllers follow clinically intuitive strategies: stronger boluses for rapid rises with low IOB, conservative behavior near hypoglycemia, and smooth tapering after meals.

7 CONCLUSION AND DISCUSSION

In this project, we investigated conditional diffusion models as offline RL policies for automated insulin dosing in the UVA/Padova Type-1 Diabetes simulator. Our study yields several preliminary insights. First, diffusion policies can represent multimodal action distributions and generate smooth, physiologically plausible insulin trajectories. Second, in the single-parameter 24-hour setting, the diffusion model achieves excellent safety and variability outcomes—most notably *0% TBR* and substantially reduced CV—while maintaining high TIR relative to baseline controllers. Third, qualitative analyses for the *tabular* controller, such as state–action heatmaps and BG–insulin–CHO trajectories, confirm that even simple value-based methods recover clinically intuitive patterns (e.g., stronger boluses during rapid rises with low IOB and conservative behavior near hypoglycemia). These observations motivate the use of more expressive generative models for insulin control.

In the more challenging multi-parameter and multi-day evaluations, the diffusion policy remains competitive with standard offline RL baselines, improving stability (lower CV) and achieving higher TIR than tabular Q-learning across 3-day and 7-day horizons. However, its overall TIR does not surpass the clinical Basal–Bolus controller in these settings, reflecting the increased difficulty of modeling multiple physiological parameters under stochastic meal disturbances. Taken together, the results suggest that diffusion models offer a promising and robust alternative to traditional value-based methods, particularly when smoothness, uncertainty-awareness, and stability are important.

Limitations and Future Work. Several important extensions remain. First, our diffusion policy conditions only on instantaneous state features; incorporating longer temporal context (e.g., transformer-based denoisers or trajectory conditioning) may improve long-horizon control. Second, the offline dataset is generated using a single behavior policy, which may limit generalization; using more diverse or real-world pump data could reduce distributional shift. Third, our primary evaluation focuses on adult patients, and broader assessment across pediatric cohorts and rare physiological profiles is necessary. Finally, integrating formal safety components—such as control barrier functions or constrained diffusion sampling—may provide additional reliability in safety-critical regions.

Overall, this work demonstrates that diffusion-based offline RL is a viable and potentially powerful framework for adaptive glucose regulation. Our findings lay the groundwork for future clinical-grade generative controllers and highlight several promising directions for advancing safe decision-making in healthcare environments.

8 CODE AVAILABILITY

All source code used in this project, including offline RL baselines, the conditional diffusion model implementation, dataset construction, and evaluation scripts, is publicly available at:

- **Project Repository:** https://github.com/sayyappr/cse598-002/
- **Project Webpage:** https://sayyappr.github.io/cse598-002/

The repository includes instructions for reproducing all experiments reported in this work, including training configurations, data processing scripts, and evaluation pipelines.

REFERENCES

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. 2023. Is Conditional Generative Modeling All You Need for Decision-Making?. In *International Conference on Learning Representations (ICLR)*.

[2] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Advances in Neural Information Processing Systems*, Vol. 34.

[3] Harry Emerson, Matthew Guy, and Ryan McConville. 2023. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics* 142 (2023).

[4] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 34.

[5] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*. PMLR, 2052–2062.

[6] Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, Vol. 34.

[7] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 21810–21823.

[8] Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. 2021. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*. PMLR, 5774–5783.

[9] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 1179–1191.

[10] L. Leelarathna, P. Choudhary, E. G. Wilmot, A. Lumb, T. Street, P. Kar, and S. M. Ng. 2021. Hybrid closed-loop therapy: Where are we in 2021? *Diabetes, Obesity and Metabolism* 23, 3 (2021), 655–660.

[11] C. Dalla Man, F. Micheletto, D. Lv, et al. 2014. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology* 8, 1 (2014), 26–34. https://doi.org/10.1177/1932296813514502

[12] Keiran Paster, Sheila McIlraith, and Jimmy Ba. 2022. You can't count on luck: Why decision transformers fail in stochastic environments. *arXiv preprint arXiv:2205.15967* (2022).

[13] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

[14] Adam R. Villaflor, Zhe Huang, Swapnil Pande, John M. Dolan, and Jeff Schneider. 2022. Addressing optimism bias in sequence modeling for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 22270–22283.

[15] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. 2022. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193* (2022).

[16] Mengjiao Yang, Dale Schuurmans, Pieter Abbeel, and Ofir Nachum. 2022. Dichotomy of control: Separating what you can control from what you cannot. *arXiv preprint arXiv:2210.13435* (2022).

[17] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. 2021. Combo: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, Vol. 34. 28954–28967.