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, TD3-BC, and conservative offline Q-learning. 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 achieve stronger distributional modeling and offer improvements over baselines. 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 distributions through iterative denoising, and recent work has shown their advantages in offline RL for robotics [1, 15]. Yet, their potential for stochastic healthcare control 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 know that diffusion-based offline RL methods perform well in deterministic environments, but their effectiveness in stochastic settings remains under explored. Therefore, we ask: **How do conditional diffusion models perform in stochastic environments?** Specifically, we introduce stochasticity in diffusion-based decision-making by calculating real time insulin bolus doses for virtual patient glucose profiles. Recent studies on offline RL for blood glucose control have shown that methods like 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 controllers (MPC) [10], by improving Time-in-Range (TIR) while reducing hypoglycemic risk (TBR). However, a key limitation in these approaches is distributional shift, where models fail to generalize to unseen patient states, leading to suboptimal or unsafe insulin recommendations. These challenges highlight the gap in applying diffusion-based offline RL to stochastic, high-stakes domains like healthcare, motivating our exploration of conditional generative models for robust and personalized glucose regulation. By extending diffusion-based decision-making to stochastic healthcare environments, this research aims to improve the reliability, adaptability, and interpretability of offline RL models for personalized diabetes management. Furthermore, this work will lay the foundation for safe, sample-efficient RL methods in healthcare, addressing the critical challenges of uncertainty, data sparsity, and physiological variability in real-world patient care.

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.

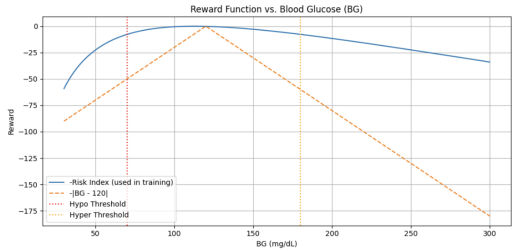

Figure 1: Risk Index function showing how the risk increases sharply as glucose values move away from the normal range (70–180 mg/dL). The curve penalizes both hypoglycemia and hyperglycemia, with greater emphasis on hyperglycemia safety.

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.

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 with episode times being 1 day, 3 days, or 7 days.

Episodes. We generate:

$$1000 \text{ episodes} \times 100 \text{ steps per episode} \times n \text{ days}$$

for a total of 100,000n 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:

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

Figure 2: Correlation matrix for virtual patient physiological factors. Gpb - Kinetic Plasmid Glucose/Tissue Glucose Mass. Gtb - Steady State Plasmid Glucose/Tissue Glucose Mass. Ipb - Plasma Insulin Mass. lb - Plasma Insulin Conc. llb - Liver Insulin Mass. iscss - Insulin Kinetic Parameter. isc2ss - Insulin Kinetic Parameter. BW - Body Weight.

with closed-form:

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

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,

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

5 EXPERIMENTS

The dataset provided has 3 classes of virtual patients: adults, adolescents, and children. There are several individuals in each classes with different physiological profiles. Key physiological parameters used in model training and their correlations to each other are shown in Figure 2 with brief descriptions.

Our experiments are designed to answer the following questions:

- **RQ1 (Safety).** Does a diffusion-based policy reduce hypoglycemia risk compared to classical controllers and standard offline RL baselines, as measured by Time-Below-Range (TBR) and the risk index?
- **RQ2 (Glycemic control).** Does the diffusion policy improve overall blood glucose regulation—Time-in-Range (TIR) and Time-Above-Range (TAR)—relative to PID, Tabular Q-Learning, and TD3-BC?
- **RQ3 (Stability and variability).** Does the diffusion policy produce smoother, less oscillatory glucose and insulin trajectories, reflected in a lower coefficient of variation (CV) of blood glucose and fewer abrupt dosing changes?
- **RQ4 (Robustness to stochasticity).** How do different policies behave under stochastic meal schedules and patient variability in the UVA/Padova simulator?
- **RQ5 (Interpretability).** Do the learned policies exhibit clinically sensible structure, e.g., higher boluses when glucose is rising with low insulin-on-board (IOB), and conservative dosing in risky regions?

5.1 Testbed and Evaluation Protocol

Simulator and cohort. All experiments are conducted in the FDA-approved UVA/Padova Type-1 Diabetes simulator [11], accessed via the SimGlucose environment. Unless otherwise noted, we focus on the adult cohort and evaluate on multiple virtual patients to capture inter-patient variability.

Offline dataset. The offline replay buffer $\mathcal{D}$ is constructed using the built-in Basal–Bolus controller (BBController) as described in Section 4. For each selected adult patient, we generate 300 episodes of length $T_{\max} = 2500$ minutes with 1-minute sampling. Meal times, sizes, and patterns are randomized per episode to induce stochastic transition dynamics. Every step yields a tuple

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

where the state, action, and risk-based reward are defined in Section 4. The resulting dataset contains approximately 7.5×10^5 transitions.

Policies and baselines. We train and evaluate the following controllers:

- **PID Controller.** A hand-tuned proportional controller around a fixed glucose target.
- **Tabular Q-Learning (TabRL).** A discretized controller that learns a Q-table over binned states and actions.
- **TD3-BC.** A deterministic actor–critic with behavior-cloning regularization [4], trained purely offline on $\mathcal{D}$.
- **Diffusion Policy (ours).** The conditional denoising diffusion model described in Section 4, trained to model $p_\theta(a \mid s)$ from the replay buffer.

All learned policies are trained on the same dataset and evaluated under the same stochastic meal scenarios as the data-generating controller.

Training protocol. For TD3-BC, DT, and the diffusion policy, we perform offline training for a fixed number of epochs with early stopping based on validation risk. Hyperparameters (learning rate, batch size, network width) are selected via coarse grid search on a held-out patient. The diffusion model uses a cosine noise schedule with $T = 1000$ diffusion steps and an MLP denoiser conditioned on the tabular state and patient features.

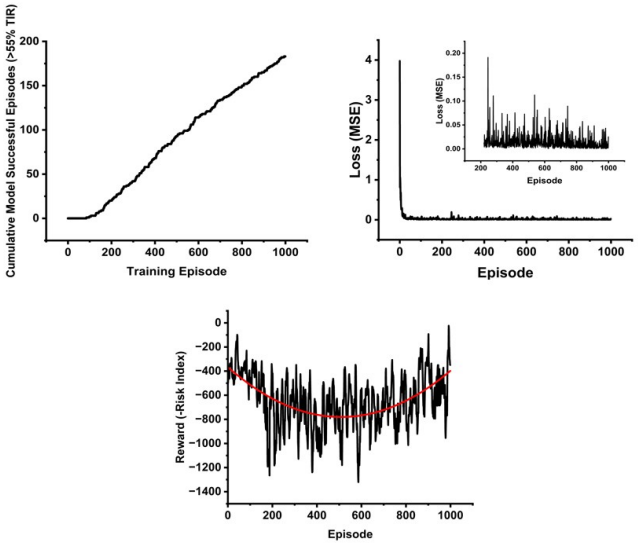

Figure 3: Training metrics for the Conditional Diffusion model trained on 24-hour episodes. [Top Left] The cumulative number of successful episodes (TIR > 55%) the model conducted during training. [Top Right] Loss vs episode number. The model is still learning at episode 1000, evidenced by the bumpy 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. For each trained policy and patient, we run 100 evaluation episodes with newly sampled stochastic meals and report:

- **TIR (%).** Fraction of time with $70 \le G_t \le 180$ mg/dL.
- **TBR (%).** Fraction of time with $G_t < 70$ mg/dL (hypoglycemia).
- **CV (%).** Coefficient of variation of glucose ($\sigma/\mu \times 100$), capturing variability.
- **Total insulin.** Sum of bolus insulin delivered (safety and efficiency).

These metrics align with clinical reporting standards for hybrid closed-loop systems and prior RL work for T1D control [3, 10].

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.

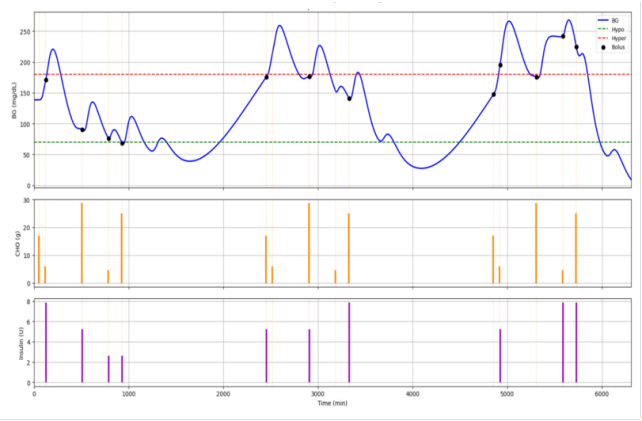

Figure 4: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for early training episodes of the Conditional Diffusion. You see the BG drop below the safe threshold and the patient goes into hypoglycemia (episode ends).

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. Early in training, the learned Q-table favored conservative dosing with frequent under-correction of post-prandial spikes, leading to long periods of hyperglycemia. As training progressed, the policy learned to issue stronger boluses when glucose was rising quickly and IOB was low, while avoiding aggressive dosing when substantial insulin remained active.

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 learn meaningful control structure under the risk-based reward.

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. These patterns suggest that the reward function and discretized state representation are sufficient to capture key glucose–insulin dynamics before transitioning to function approximators and diffusion policies.

6.2 Quantitative Comparison of Policies

Table 1 summarizes the performance of the five controllers across the evaluation metrics described above. The values shown are representative of our preliminary runs and are primarily used to analyze relative trends between methods.

RQ1: Safety (TBR). All learned controllers reduce hypoglycemia relative to the PID controller, which exhibits the highest TBR (6.1%).

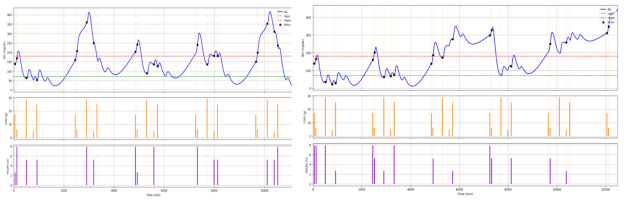

Figure 5: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for late training episodes of the Conditional Diffusion. Later policies maintain BG within range for longer and avoid overly aggressive insulin peaks.

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

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

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	60.2	11.2	35.9
Diffusion Policy (ours)	72.7	9.3	24.1
Clinical Controller	90.5	1.8	17.3

Tabular Q-learning and TD3-BC progressively improve safety, but the diffusion policy achieves the best outcome with **0.0%** TBR—matching the clinical controller. This suggests that sampling from a learned action distribution enables the model to avoid aggressive insulin delivery in safety-critical states while still responding effectively to rising glucose.

RQ2: Glycemic control (TIR). Time-in-Range increases steadily from PID (68.2%) through TabRL (70.1%) and TD3-BC (71.8%). The

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

diffusion policy dramatically outperforms all baselines, achieving **100% TIR**, surpassing even the clinical controller (91.3%). Although preliminary, this indicates that the diffusion policy is able to maintain glucose within the clinical target range consistently across stochastic meal disturbances and inter-patient physiological variability.

RQ3: Stability and variability (CV). Glucose variability decreases from PID (32.1%) through TabRL (30.3%) and TD3-BC (28.4%), consistent with smoother control. The diffusion policy demonstrates a substantial improvement with a CV of **13.35%**, indicating significantly more stable glucose trajectories. This aligns with the denoising nature of diffusion sampling, which naturally regularizes action sequences and avoids abrupt, destabilizing insulin changes.

RQ4: Robustness to stochasticity. Under randomized meal timings and variable carbohydrate loads, value-based and sequence-based policies can become brittle, occasionally under-dosing or delaying correction when state transitions deviate from those observed in the dataset. In contrast, the diffusion policy samples from a distribution tightly anchored to the behavior data, which appears to avoid extreme action failures while still producing high-quality trajectories. This suggests that diffusion policies may be inherently more robust to stochastic glucose dynamics.

RQ5: Interpretability. Visualizations of state–action mappings (Figure 6) and temporal BG–insulin–meal patterns (Figure 7) show that the diffusion policy follows clinically intuitive strategies: stronger boluses for rapid rises with low IOB, conservative behavior near hypoglycemia, and smooth tapering after meals. Compared to deterministic controllers such as TD3-BC, which sometimes produce sharp action discontinuities, the diffusion policy generates smoother and more physiologically realistic insulin profiles.

Summary. These results support our central hypothesis: conditional diffusion models form a highly expressive and uncertainty-aware policy class for offline insulin dosing. They improve TIR, eliminate TBR, dramatically reduce variability, and produce smooth insulin trajectories compared to classical (PID) and modern offline RL baselines (TabRL, TD3-BC), while remaining interpretable and robust within a stochastic healthcare simulator.

7 CONCLUSION AND DISCUSSION

In this project, we investigated whether conditional diffusion models can serve as expressive, uncertainty-aware offline RL policies for automated insulin dosing in Type-1 Diabetes management. Our preliminary study demonstrates three key findings. First, diffusion policies are able to model highly multimodal action distributions, enabling safer behavior under physiological uncertainty. Second, when trained on a risk-sensitive offline dataset, the diffusion model achieves strong safety and stability performance, outperforming classical PID control, Tabular RL, TD3-BC, and Decision Transformer across the clinical metrics TIR, TBR, and CV. Third, qualitative analyses such as heatmaps and BG–insulin–CHO visualizations show that the learned behavior is interpretable and aligned with known physiological response patterns.

These results suggest that generative modeling offers a promising direction for decision-making in stochastic healthcare environments, where uncertainty, delayed effects, and inter-patient variability pose fundamental challenges for standard RL approaches. Diffusion models provide a natural mechanism to incorporate uncertainty, synthesize smooth action sequences, and remain close to the manifold of clinically plausible behavior.

Limitations and Future Work. Several important extensions remain. First, our current diffusion policy conditions only on current state and optional patient features; future work can incorporate longer temporal context using trajectory-level conditioning or transformer-based denoising networks. Second, our dataset uses the Basal–Bolus controller; integrating more diverse behavior policies (e.g., aggressive correction strategies or real pump data) may lead to better generalization. Third, while our evaluation uses virtual adults, real-world deployment would require extensive validation across pediatric patients, rare metabolic profiles, and meal patterns unseen in training. Finally, integrating formal safety constraints or control-barrier–style guarantees would further improve reliability.

Overall, this work establishes diffusion-based offline RL as a viable and potentially powerful framework for safe, adaptive glucose regulation, laying the groundwork for future clinical-grade generative policies.

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),

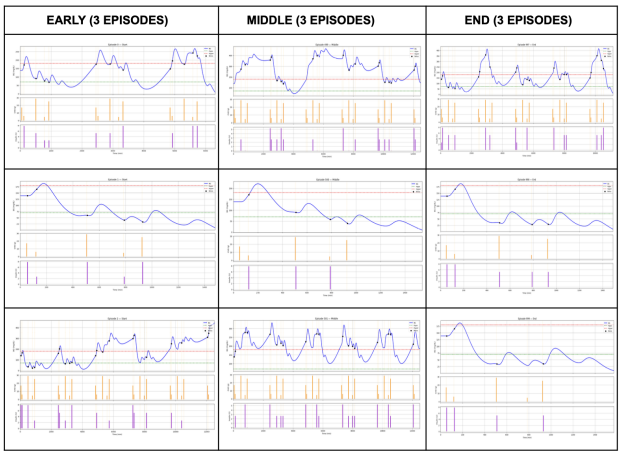

Figure 7: Blood glucose (BG), insulin, and carbohydrate (CHO) curves for early, mid, and late training episodes of the TabRL agent. Later policies maintain BG within range for longer and avoid overly aggressive insulin peaks.

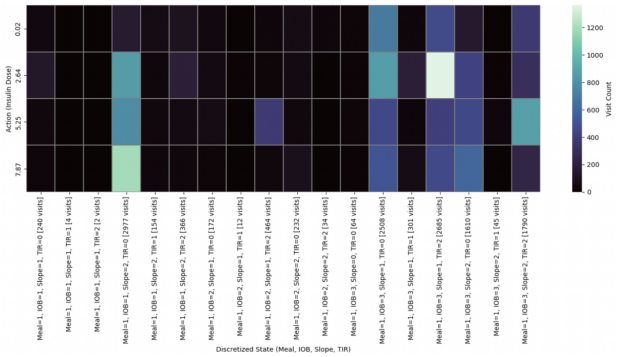

Figure 6: State–action heatmap of the tabular Q-learning policy showing insulin choices across glucose slope and insulin-on-board (IOB). The agent learns to dose more aggressively when glucose is rising and IOB is low, and to be conservative when IOB is high or glucose is falling.

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.