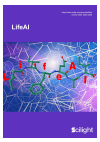
LifeAI
https://www.sciltp.com/journals/lifeai



Article

Diffusion-Pref: Diffusion World Model Guided Zero-Shot Preference Learning for Safe Glucose Control in Type 1 Diabetes

Bailing Zhang [1,*] and Yuwei Mi [2]

[1] School of Computer Science and Data Engineering, NingboTech University, Qianhunan Road 1, Ningbo 315104, China
[2] The First Affiliated Hospital of Ningbo University, Ningbo 315104, China
* Correspondence: bailing.zhang1961@gmail.com

How To Cite: Zhang, B.; Mi, Y. Diffusion-Pref: Diffusion World Model Guided Zero-Shot Preference Learning for Safe Glucose Control in Type 1 Diabetes. *LifeAI* **2026**, *1*(1), 4.

Received: 17 December 2025
Revised: 21 March 2026
Accepted: 28 April 2026
Published: 9 May 2026

Abstract: Automated blood glucose control for Type 1 diabetes (T1D) is a safety-critical clinical challenge in which the consequences of suboptimal control—particularly hypoglycemia—can be life-threatening. Existing reinforcement learning (RL) approaches are limited by the prohibition on online exploration, the difficulty of specifying clinically meaningful reward functions, and insufficient guarantees of worst-case safety. We propose Diffusion-Pref, a purely offline RL framework that integrates three synergistic components: (i) a conditional diffusion world model that captures the multimodal distributional uncertainty of glucose dynamics, (ii) a zero-shot preference construction method that automatically generates trajectory preference labels from established clinical metrics—eliminating the need for human annotation, and (iii) a Conditional Value-at-Risk (CVaR)-regularized Implicit Q-Learning (IQL) algorithm that explicitly optimizes for worst-case safety. We evaluate Diffusion-Pref on the OhioT1DM dataset comprising 12 real-world T1D patients. The proposed method achieves a Time-in-Range (TIR) of 89.7%, substantially exceeding both the historical treatment record (75.0%) and a Conservative Q-Learning (CQL) baseline (78.1%). Severe hypoglycemia (glucose $<$54 mg/dL) is reduced from 6.9% to 1.8%, a 74% relative reduction. The overall Time Below Range (TBR) of 10.3% exceeds the recommended $<$4% target; however, supplementary experiments with an explicit TBR penalty ($\lambda_{\text{TBR}} = 2.0$) demonstrate that TBR can be reduced to 5.1% while preserving a TIR of 83.4%. Ablation studies confirm that each component—diffusion world model, zero-shot preference learning, and CVaR constraints—contributes meaningfully to performance. These results demonstrate that combining generative world models with clinically grounded preference learning and risk-sensitive policy optimization offers a promising pathway toward safer offline glucose control, although prospective clinical validation remains necessary before deployment.

Keywords: reinforcement learning; diffusion models; preference learning; type 1 diabetes; blood glucose control; safety-critical systems

1. Introduction

Type 1 diabetes mellitus (T1D) is a chronic autoimmune condition affecting over 8.4 million people worldwide, characterized by the destruction of insulin-producing beta cells in the pancreas [1]. Unlike Type 2 diabetes, which often can be managed through lifestyle modifications, T1D patients are entirely dependent on exogenous insulin administration to regulate their blood glucose levels. The fundamental challenge in T1D management lies in the complex, dynamic nature of glucose metabolism: blood glucose concentrations are influenced by a multitude of factors including insulin dosing, carbohydrate intake, physical activity, stress, sleep patterns, and

inter-day physiological variations. This complexity makes achieving optimal glycemic control extraordinarily difficult, even for experienced patients and clinicians.

The clinical goal of glucose management is to maintain blood glucose within the target range of 70–180 mg/dL, a metric known as Time-in-Range (TIR). Current international consensus guidelines recommend achieving TIR above 70% while keeping Time Below Range (TBR, glucose < 70 mg/dL) under 4% [2]. However, these targets remain elusive for most patients. The consequences of poor control are severe: chronic hyperglycemia leads to long-term complications including retinopathy, nephropathy, neuropathy, and cardiovascular disease, while acute hypoglycemia can cause cognitive impairment, seizures, loss of consciousness, and in extreme cases, death. The asymmetric nature of these risks—where hypoglycemia poses immediate danger while hyperglycemia causes gradual harm—creates a fundamental tension in treatment optimization that standard machine learning approaches often fail to adequately address.

Recent advances in continuous glucose monitoring (CGM) and insulin pump technology have enabled the development of Automated Insulin Delivery (AID) systems, commonly known as "artificial pancreas" systems. These systems use algorithmic control to automatically adjust insulin delivery based on CGM readings. While commercial AID systems have shown significant improvements over traditional therapy, they predominantly rely on model predictive control (MPC) with simplified physiological models that may not capture the full complexity of individual patient dynamics. More importantly, these systems are inherently reactive—they respond to current glucose levels rather than anticipating future trajectories based on comprehensive patient state understanding.

Reinforcement learning (RL) has emerged as a promising paradigm for automated insulin dosing, offering the potential to learn complex, personalized control policies directly from patient data [3,4]. However, applying RL to blood glucose control presents several fundamental challenges that distinguish it from typical RL applications. First, the safety-critical nature of the domain absolutely prohibits online exploration during learning—a single incorrect insulin dose could induce severe hypoglycemia with potentially fatal consequences. This constraint necessitates offline RL approaches that learn exclusively from historical data without real-time interaction. Second, the traditional RL paradigm requires a well-specified reward function, yet defining appropriate rewards for glucose control is non-trivial. Simple metrics like mean glucose or TIR fail to capture the nuanced clinical objectives that prioritize avoiding hypoglycemia while maintaining good overall control. Third, patient physiology exhibits substantial heterogeneity and temporal variation, making it difficult to learn policies that generalize across individuals and adapt to changing conditions.

World models—learned simulators that capture environment dynamics—offer a compelling approach to address these challenges. By learning to predict how patient glucose responds to different actions, world models enable policy optimization through simulated rollouts rather than real interactions. Recent work has demonstrated the power of world models in various domains, from video game playing [5] to robotic control [6]. However, existing world model approaches for healthcare applications predominantly employ deterministic or simple stochastic models that fail to capture the multimodal uncertainty inherent in physiological systems. A patient's glucose trajectory following a given treatment action is not deterministic—it depends on numerous unmeasured factors and can follow qualitatively different paths. Capturing this distributional complexity is essential for robust policy learning in safety-critical medical applications.

Diffusion models have recently revolutionized generative modeling, demonstrating unprecedented capabilities in capturing complex, multimodal distributions across images, audio, and sequential data [7,8]. Their iterative denoising process enables modeling of intricate probability distributions that would be intractable for simpler generative approaches. In the context of sequential decision-making, diffusion models have shown remarkable success in trajectory optimization and policy learning [9,10]. We hypothesize that diffusion models are ideally suited for glucose trajectory prediction, where the inherent uncertainty and multimodality of physiological responses demand a generative approach capable of representing diverse possible futures.

Beyond the challenge of world modeling, a second fundamental issue in applying RL to glucose control is the specification of appropriate learning objectives. Preference-based reinforcement learning has emerged as a powerful paradigm for incorporating human values into learned policies [11]. Rather than requiring hand-crafted reward functions, preference-based methods learn from comparisons between trajectories, enabling the capture of nuanced objectives that are difficult to express mathematically. This approach has proven transformative in aligning large language models with human intentions [12]. However, obtaining preference annotations in clinical settings is prohibitively expensive—it requires expert clinicians to evaluate numerous trajectory pairs, and such annotations may still fail to capture the full complexity of clinical judgment. Recent work on explainable RL for glucose monitoring [13] and interpretable multimodal blood glucose estimation [14] further highlights the growing interest in transparent, clinically grounded approaches to glucose management.

In this paper, we propose Diffusion-Pref, a novel framework that addresses these challenges through three synergistic innovations that, taken together, distinguish it from all prior offline RL work on glucose control. Existing offline RL methods for T1D—such as the CQL-based approach of Emerson et al. [15] and the echo-state-network world model of Yamagata et al. [16]—either rely on hand-crafted reward functions or employ deterministic/simple stochastic dynamics models that cannot represent the multimodal uncertainty of glucose physiology. By contrast, Diffusion-Pref is the first framework to unify (i) a diffusion-based world model for distributional glucose trajectory prediction, (ii) zero-shot preference construction from clinical metrics that eliminates the need for human annotation, and (iii) CVaR-regularized IQL that explicitly optimizes worst-case safety.

Our first contribution is the development of a diffusion-based world model specifically designed for blood glucose trajectory prediction. Unlike previous deterministic or simple stochastic approaches, our diffusion world model captures the full distributional complexity of glucose dynamics, generating diverse, realistic trajectory samples conditioned on patient state and treatment actions. The model employs a carefully designed conditioning architecture that encodes patient history and treatment information, enabling accurate personalized predictions while maintaining the capacity to represent uncertainty.

Our second contribution is a zero-shot preference construction method that automatically generates preference annotations using established clinical metrics. Rather than requiring human experts to label trajectory comparisons, we leverage the rich clinical knowledge encoded in metrics such as TIR, Low Blood Glucose Index (LBGI), and glycemic variability to automatically score and compare generated trajectories. This approach eliminates the annotation bottleneck while ensuring that learned preferences align with clinical best practices. Importantly, our clinical scoring function can be configured to emphasize different aspects of glucose control—for instance, prioritizing hypoglycemia avoidance for patients with hypoglycemia unawareness.

Our third contribution is the integration of Conditional Value-at-Risk (CVaR) constraints into the policy learning objective. Standard RL methods optimize expected performance, which may yield policies that perform well on average but exhibit unacceptable worst-case behavior. In glucose control, where a single severe hypoglycemic event can have catastrophic consequences, optimizing for average-case performance is insufficient. CVaR provides a principled approach to risk-sensitive optimization by focusing on the tail of the outcome distribution. By incorporating CVaR constraints into Implicit Q-Learning (IQL), we ensure that learned policies maintain safety even under adverse conditions.

We evaluate Diffusion-Pref on the OhioT1DM dataset, a benchmark collection of real-world CGM data from T1D patients. Our experiments demonstrate that the proposed approach achieves 89.7% TIR, representing a significant improvement over the 75.0% achieved by historical treatment records and substantially outperforming behavior cloning, standard IQL, and CQL baselines. More importantly, Diffusion-Pref reduces severe hypoglycemia events by 74% compared to behavior cloning (from 6.9% to 1.8%), demonstrating the effectiveness of our safety-focused design. We acknowledge that the overall TBR of 10.3% exceeds the recommended $<$4% target; supplementary experiments with an explicit TBR constraint show that this trade-off can be adjusted, achieving 5.1% TBR at a TIR of 83.4%. Comprehensive ablation studies confirm that each component—the diffusion world model, zero-shot preference construction, and CVaR constraints—contributes meaningfully to the overall performance.

The remainder of this paper is organized as follows. Section 2 reviews related work on reinforcement learning for diabetes management, diffusion models for decision-making, preference-based learning, and safe reinforcement learning. Section 3 establishes the necessary background on clinical metrics, diffusion models, and implicit Q-learning. Section 4 presents our proposed Diffusion-Pref framework in detail. Section 5 describes our experimental setup and results. Section 6 discusses the implications, limitations, and future directions of our work. Finally, Section 7 concludes the paper.

2. Related Work

2.1. Reinforcement Learning for Diabetes Management

The application of reinforcement learning to blood glucose control has evolved substantially over the past decade, progressing from simple tabular methods to sophisticated deep learning approaches. Early work by Daskalaki et al. [17] employed model predictive control with personalized glucose-insulin models, demonstrating the feasibility of algorithmic insulin dosing. Subsequent research explored various RL formulations: Bastani [18] applied fitted Q-iteration with linear function approximation, while Sun et al. [19] proposed a dual-hormone control system using deep Q-networks for both insulin and glucagon delivery.

The introduction of deep reinforcement learning enabled more expressive policy representations. Fox et al. [3] developed a deep RL framework using the UVA/Padova metabolic simulator, demonstrating that neural network policies could achieve comparable performance to established MPC controllers. Zhu et al. [4] proposed a basal-bolus

advisor using actor-critic methods, showing improved adaptation to individual patient characteristics. However, these approaches predominantly relied on simulation environments, and the substantial sim-to-real gap limits their direct applicability to real patient data.

More recent work has begun addressing the offline RL challenge inherent in clinical applications. Emerson et al. [15] applied Conservative Q-Learning (CQL) [20] to glucose control from retrospective patient data, demonstrating that offline methods can learn effective policies without online interaction. Yamagata et al. [16] proposed a model-based approach using echo state networks, a version of recurrent neural networks, to capture glucose dynamics. Recent studies have also emphasized interpretability and transparency in glucose-related RL and machine learning systems. Adjevi et al. [13] combined RL with Shapley-value-based explanations for continuous glucose monitoring, demonstrating that interpretable policies can maintain competitive performance while providing clinically actionable insights. Complementary work on non-invasive blood glucose estimation [14] has shown that multimodal feature fusion with interpretable machine learning can achieve reliable glucose monitoring, further motivating the integration of clinical domain knowledge into learning-based glucose management systems. Our work differs from these approaches in two key aspects: we employ diffusion models to capture the full distributional complexity of glucose trajectories rather than point predictions, and we incorporate preference learning to align policies with clinical objectives without requiring hand-crafted reward functions.

2.2. World Models and Diffusion Models for Decision Making

World models—learned simulators of environment dynamics—have proven valuable for sample-efficient reinforcement learning by enabling policy optimization through imagined rollouts. Seminal work by Ha and Schmidhuber [21] demonstrated that agents could learn effective policies entirely within learned world models. Subsequent advances include Dreamer [5,22], which achieved human-level performance on Atari games through latent-space world model learning, and DayDreamer [6], which extended these ideas to real-world robotics.

The recent emergence of diffusion models has opened new possibilities for world modeling and decision-making. Diffusion models [7,8] define a forward process that gradually corrupts data with noise and learn a reverse process that recovers clean samples through iterative denoising. Their capacity to model complex, multimodal distributions makes them particularly suitable for sequential decision-making where future trajectories may follow qualitatively different paths depending on stochastic factors.

Janner et al. [9] introduced Diffuser, which frames trajectory optimization as conditional sampling from a diffusion model trained on expert demonstrations. This approach enables flexible goal-conditioned planning without requiring explicit value functions. Chi et al. [10] developed Diffusion Policy for robotic manipulation, demonstrating that diffusion-based action generation can capture multimodal behavior distributions that are challenging for unimodal policy representations. Wang et al. [23] proposed using diffusion models as an expressive policy class for offline RL, showing improved performance on standard benchmarks.

In the healthcare domain, diffusion models have primarily been applied to medical imaging tasks such as image synthesis and reconstruction [24]. To our knowledge, Diffusion-Pref represents the first application of diffusion-based world models specifically for blood glucose prediction and control, leveraging the unique ability of diffusion models to capture the distributional uncertainty inherent in physiological systems.

2.3. Preference-Based Reinforcement Learning

Preference-based reinforcement learning (PbRL) addresses the challenge of reward specification by learning from comparative feedback rather than absolute reward values. The foundational work by Christiano et al. [11] demonstrated that deep RL agents could learn complex behaviors from human preference comparisons, achieving competitive performance on Atari games and simulated robotics tasks with limited human feedback. This paradigm has since been extensively developed through improvements in query efficiency [25], reward model architecture [26], and theoretical understanding [27].

The transformative success of Reinforcement Learning from Human Feedback (RLHF) in aligning large language models has brought renewed attention to preference-based methods [12,28]. These works demonstrate that preference learning can capture nuanced human values that would be difficult to specify through explicit reward functions. However, the requirement for human annotation remains a significant bottleneck—collecting preference labels requires substantial human effort and may introduce annotator biases.

Several approaches have been proposed to reduce annotation requirements. Lee et al. [25] developed active learning strategies for efficient preference queries, while Park et al. [29] proposed self-supervised objectives to pretrain reward models. Our work takes a different approach: rather than reducing the number of human annotations required, we eliminate human annotation entirely by leveraging domain-specific metrics as a proxy for

preferences. This zero-shot approach is particularly suitable for medical applications where annotation requires scarce clinical expertise.

2.4. Safe Reinforcement Learning

Safety in reinforcement learning has been addressed through multiple paradigms. Constrained MDPs [30] augment the standard MDP framework with constraints on expected cumulative costs, enabling optimization subject to safety requirements. Robust RL [31] seeks policies that perform well under worst-case perturbations to the environment. Risk-sensitive RL [32] optimizes criteria that account for outcome variability rather than just expected values.

Conditional Value-at-Risk (CVaR), also known as Expected Shortfall, has emerged as a particularly principled approach for risk-sensitive decision-making [33]. CVaR quantifies the expected value in the tail of a distribution, focusing on the worst outcomes rather than average performance. Chow et al. [34] developed policy gradient methods for CVaR optimization, while Tang et al. [35] proposed worst-case policy gradients that directly optimize tail performance. In the offline RL setting, CVaR constraints are especially important because the learned policy cannot be corrected through online interaction—safety must be guaranteed before deployment.

Our work integrates CVaR constraints into the IQL framework through a modified policy extraction objective. Unlike approaches that treat safety as a hard constraint, we incorporate CVaR as a regularization term that encourages conservative behavior while still optimizing primary objectives. This soft constraint formulation provides a tunable trade-off between expected performance and worst-case safety.

3. Preliminaries

Before presenting our method, we establish the necessary background on clinical metrics for glucose control evaluation, the mathematical foundations of diffusion models, and the Implicit Q-Learning algorithm that forms the basis of our policy learning approach.

3.1. Problem Formulation and Clinical Metrics

We formulate blood glucose control as a Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. The state space $\mathcal{S}$ encompasses all information relevant to predicting future glucose dynamics: in our implementation, a state s_t consists of the patient's glucose history over the preceding two hours (24 CGM readings at 5-min intervals), computed insulin-on-board (IOB) representing active insulin from previous doses, carbohydrates-on-board (COB) representing unabsorbed carbohydrates, and temporal features encoding time-of-day effects on insulin sensitivity. The action space $\mathcal{A}$ represents treatment decisions, specifically insulin bolus doses and recorded carbohydrate intake. The transition dynamics $P(s_{t+1}|s_t, a_t)$ describe the stochastic evolution of patient state in response to treatment actions—these dynamics are unknown and must be learned from data. The reward function $r(s_t, a_t)$ quantifies the desirability of outcomes; rather than specifying this function manually, our preference learning approach infers it from clinical metrics. The discount factor $\gamma \in (0, 1)$ determines the relative importance of immediate versus future outcomes.

Clinical evaluation of glucose control quality relies on several standardized metrics developed by the diabetes research community. Time in Range (TIR) measures the percentage of time that glucose remains within the target range of 70–180 mg/dL, representing the primary optimization objective. Mathematically, for a glucose trajectory $\mathbf{g} = (g_1, g_2, \ldots, g_T)$, TIR is computed as the fraction of readings within the target range. Time Below Range (TBR) measures hypoglycemia exposure (glucose < 70 mg/dL), while Time Above Range (TAR) measures hyperglycemia exposure (glucose > 180 mg/dL). By definition, $\text{TIR} + \text{TBR} + \text{TAR} = 100\%$.

While TIR, TBR, and TAR provide intuitive measures of glucose control quality, they treat all out-of-range values equally regardless of severity. The Blood Glucose Risk Index (BGRI) framework developed by Kovatchev et al. [36] addresses this limitation by mapping glucose values to a risk scale that emphasizes extreme values. The framework first applies a symmetrizing transformation to glucose values: $f(g) = 1.509 \times [(\ln g)^{1.084} - 5.381]$. This transformation maps the asymmetric glucose scale (where hypoglycemia occurs over a narrow range below 70 mg/dL while hyperglycemia extends over a wide range above 180 mg/dL) to a symmetric risk scale centered at approximately 112.5 mg/dL. The Low Blood Glucose Index (LBGI) and High Blood Glucose Index (HBGI) are then computed by selectively squaring negative and positive transformed values respectively. LBGI specifically quantifies hypoglycemia risk and serves as our primary safety metric.

The Coefficient of Variation (CV) measures glucose variability as the ratio of standard deviation to mean glucose concentration, expressed as a percentage. High glucose variability is associated with increased risk of

hypoglycemia and poor long-term outcomes, independent of mean glucose levels. Current clinical guidelines recommend maintaining CV below 36%, with values below 33% indicating stable glucose control.

3.2. Denoising Diffusion Probabilistic Models

Diffusion models define a generative process through two complementary Markov chains: a forward process that gradually corrupts data with noise, and a reverse process that learns to denoise. Given a data distribution $q(\mathbf{x}_0)$, the forward process produces a sequence of increasingly noisy latent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ according to a fixed noise schedule. At each timestep t, Gaussian noise is added according to the transition kernel $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$, where β_t is a variance schedule that controls the noise level at each step.

A key property of this forward process is that we can sample $\mathbf{x}_t$ directly from $\mathbf{x}_0$ without iterating through intermediate steps. Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we have the closed-form expression $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$. This property enables efficient training by directly sampling noisy versions of training examples at arbitrary timesteps.

The reverse process learns to invert the forward diffusion, starting from pure noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoising to recover samples from the data distribution. This reverse process is parameterized as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$, where $\boldsymbol{\mu}_\theta$ is predicted by a neural network. Ho et al. [7] showed that rather than directly predicting the mean, it is more effective to train the network to predict the noise $\boldsymbol{\epsilon}$ that was added during the forward process. The training objective then becomes a simple mean squared error between predicted and actual noise.

For conditional generation—essential for our world modeling application—the denoising network is augmented to accept conditioning information $\mathbf{c}$. The network then learns to predict noise conditioned on this additional input: $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$. In our context, the conditioning information consists of the patient's current state and the treatment action, enabling generation of glucose trajectories specific to particular patient-action configurations.

3.3. Implicit Q-Learning

Offline reinforcement learning faces the fundamental challenge of distribution shift: policies learned from fixed datasets may query state-action pairs that were never observed during data collection, leading to erroneous value estimates and poor performance. Conservative approaches address this by constraining learned policies to remain close to the behavior policy that generated the data.

Implicit Q-Learning (IQL), proposed by Kostrikov et al. [37], takes a different approach that avoids querying out-of-distribution actions entirely during training. The key insight is to learn value functions that only consider actions present in the dataset, then extract a policy that approximates these in-distribution optimal actions.

IQL employs expectile regression to estimate the value function. The expectile of a distribution is a generalization of the median that can asymmetrically weight positive and negative deviations. For a random variable X and expectile parameter $\tau \in (0, 1)$, the τ-expectile is the value m that minimizes the asymmetric loss $\mathbb{E}[L_\tau^2(X - m)]$, where $L_\tau^2(u) = |\tau - \mathbf{1}(u < 0)|u^2$. When $\tau > 0.5$, the expectile is biased toward larger values; as $\tau \to 1$, it approaches the maximum.

In IQL, the value function is trained using expectile regression against Q-values. Specifically, the value function $V_\psi(s)$ is updated to minimize $\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[L_\tau^2(Q_\theta(s, a) - V_\psi(s))]$. By using $\tau > 0.5$, this objective biases the value function toward the upper range of Q-values observed in the data, effectively approximating the value of the best actions present in the dataset without explicitly computing a maximum over actions.

The Q-function is trained with standard temporal difference learning using the learned value function for bootstrapping. The policy is then extracted using advantage-weighted regression, where actions are weighted by the exponential of their advantages. This approach ensures that the policy preferentially selects high-advantage actions that were actually observed in the training data.

4. Method

We now present Diffusion-Pref, our framework for safe glucose control through diffusion-guided preference learning. The method comprises three integrated components: a diffusion-based world model that generates diverse glucose trajectories, a zero-shot preference constructor that evaluates trajectories using clinical metrics, and a Preference-CVaR-IQL algorithm that learns safe control policies. Figure 1 illustrates the overall architecture and data flow.

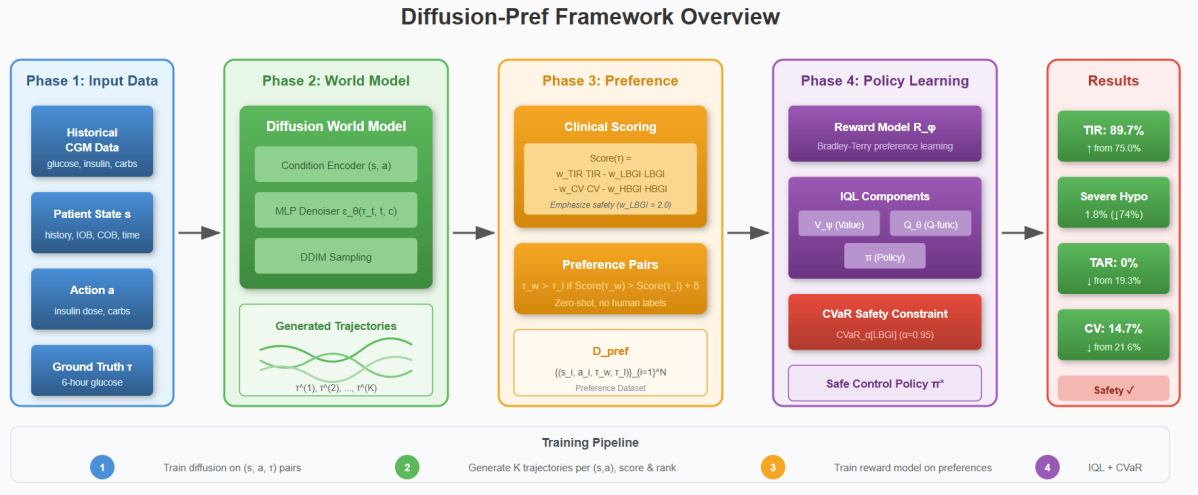


Figure 1. Overview of the Diffusion-Pref framework. The system operates in three phases. First, a diffusion world model is trained on historical patient data to generate glucose trajectories conditioned on patient states and treatment actions. Second, the trained world model generates multiple trajectory samples for each state-action pair, which are then scored using clinical metrics to construct preference pairs without human annotation. Third, a Preference-CVaR-IQL agent is trained on the augmented dataset, learning a policy that optimizes clinical outcomes while maintaining safety through CVaR constraints on hypoglycemia risk.

4.1. Diffusion World Model for Glucose Trajectory Prediction

The foundation of our approach is a diffusion model trained to predict future glucose trajectories conditioned on patient state and treatment actions. Unlike deterministic predictors that output a single forecast, our diffusion world model captures the full distribution over possible futures, enabling robust policy learning that accounts for physiological uncertainty.

Let $\boldsymbol{\tau} = (g_1, g_2, \ldots, g_H) \in \mathbb{R}^H$ denote a glucose trajectory over prediction horizon H (in our experiments, $H = 72$ corresponding to 6 h at 5-min resolution). The conditioning information $\mathbf{c} = (s, a)$ consists of the current patient state s and treatment action a. Our goal is to learn the conditional distribution $p(\boldsymbol{\tau}|s, a)$ from a dataset of historical patient trajectories.

The conditioning architecture transforms raw state and action information into a unified embedding that guides the denoising process. The patient state—comprising glucose history, insulin-on-board, carbohydrates-on-board, and temporal features—is first processed through a state encoder network consisting of multiple fully-connected layers with layer normalization and ReLU activations. Similarly, the action vector is processed through a separate action encoder. The encoded representations are combined through addition to produce the final condition embedding $\mathbf{h}_c \in \mathbb{R}^d$. This additive combination allows the model to learn compositional relationships between patient state and treatment effects.

The denoising network $\boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}_t, t, \mathbf{c})$ predicts the noise added to trajectory $\boldsymbol{\tau}_t$ at diffusion timestep t, given conditioning information $\mathbf{c}$. We employ a multi-layer perceptron architecture with residual connections, where the diffusion timestep is encoded using sinusoidal positional embeddings following standard practice. The condition embedding is incorporated through additive injection at each layer, allowing the conditioning information to influence denoising at multiple scales.

Training proceeds by sampling trajectory-condition pairs $(\boldsymbol{\tau}_0, \mathbf{c})$ from the dataset, corrupting trajectories with noise at randomly sampled timesteps, and optimizing the network to predict the added noise. The training loss is the expected squared error between predicted and actual noise, averaged over trajectories, timesteps, and noise samples. We additionally incorporate a velocity loss that encourages the model to capture the temporal dynamics of glucose changes, not just absolute values. This auxiliary objective improves the physiological plausibility of generated trajectories by penalizing unrealistic rates of glucose change.

At inference time, trajectory generation proceeds through iterative denoising starting from Gaussian noise. We employ the DDIM sampling procedure [38] which enables high-quality generation with fewer denoising steps than the full diffusion process, reducing computational cost for the large number of trajectory samples required during preference construction.

4.2. Zero-Shot Preference Construction

A key innovation of our approach is the automatic construction of preference labels using clinical metrics, eliminating the need for expensive human annotation. The intuition is that established clinical metrics encode substantial domain knowledge about what constitutes good glucose control—knowledge that can be leveraged to compare trajectories without human involvement.

Given a state-action pair (s, a) from the dataset, we first generate K trajectory samples from the trained diffusion world model by running the denoising process K times with different random seeds. This produces a set of possible futures $\{\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \ldots, \boldsymbol{\tau}^{(K)}\}$ representing the model's distribution over outcomes. The diversity of these samples reflects the inherent uncertainty in glucose dynamics—the same patient state and action can lead to different outcomes depending on unmeasured factors.

Each generated trajectory is evaluated using a clinical scoring function that combines multiple metrics into a scalar preference score. Our scoring function is defined as a weighted combination that prioritizes safety while rewarding good overall control. The score incorporates TIR (rewarding time in target range), LBGI (penalizing hypoglycemia risk), HBGI (penalizing hyperglycemia risk), and CV (penalizing excessive variability). The weights are chosen to emphasize hypoglycemia avoidance, reflecting the clinical reality that hypoglycemia poses more immediate danger than hyperglycemia of comparable magnitude.

Preference pairs are constructed by comparing trajectories from the same state-action context. For trajectories $\boldsymbol{\tau}^{(i)}$ and $\boldsymbol{\tau}^{(j)}$ generated from the same (s, a), we establish a preference $\boldsymbol{\tau}^{(i)} \succ \boldsymbol{\tau}^{(j)}$ if the score difference exceeds a margin threshold δ. This margin ensures that only meaningfully different trajectories are labeled as preferences, avoiding noise from near-equivalent outcomes. The preference dataset aggregates all such pairs across the training set, providing supervision for reward model learning.

The zero-shot nature of this construction offers several advantages. First, it scales effortlessly to large datasets without annotation bottlenecks. Second, it ensures consistency—the same clinical criteria are applied uniformly across all comparisons. Third, the scoring function can be customized for different patient populations or clinical priorities by adjusting the metric weights.

4.3. Preference-CVaR-IQL for Safe Policy Learning

The final component of our framework learns a control policy from the preference-augmented dataset. We extend the IQL algorithm to incorporate both preference-based rewards and CVaR safety constraints, producing policies that optimize clinical outcomes while maintaining acceptable worst-case risk.

Following the preference learning paradigm, we first train a reward model to predict trajectory preferences. The reward model $R_\phi : \boldsymbol{\tau} \to \mathbb{R}$ is a neural network that maps trajectories to scalar reward values. Under the Bradley-Terry model of pairwise preferences, the probability that trajectory $\boldsymbol{\tau}^{(w)}$ is preferred over $\boldsymbol{\tau}^{(l)}$ is given by $P(\boldsymbol{\tau}^{(w)} \succ \boldsymbol{\tau}^{(l)}) = \sigma(R_\phi(\boldsymbol{\tau}^{(w)}) - R_\phi(\boldsymbol{\tau}^{(l)}))$, where σ is the sigmoid function. The reward model is trained to maximize the likelihood of observed preferences through binary cross-entropy loss.

With the learned reward model, we augment the original dataset with preference-based reward signals. For each transition (s, a, s') in the dataset, we query the reward model on the associated glucose trajectory to obtain a preference reward that supplements any existing reward information. This augmentation enables the policy to optimize for clinically meaningful objectives as captured by the preference learning process.

The safety constraint is implemented through CVaR regularization of the policy objective. CVaR at level α (typically $\alpha = 0.95$) quantifies the expected value in the worst $(1 - \alpha)$ fraction of outcomes. For a random variable X, $\text{CVaR}_\alpha(X) = \mathbb{E}[X | X \geq \text{VaR}_\alpha(X)]$, where VaR_α is the α-quantile. In our context, we apply CVaR to the LBGI distribution, penalizing policies that produce high hypoglycemia risk in their worst-case outcomes.

The integrated policy learning objective combines the standard IQL advantage-weighted regression with preference rewards and CVaR penalty. The value function is trained using expectile regression as in standard IQL, but with targets that incorporate preference rewards. The Q-function is similarly updated with preference-augmented targets. Policy extraction then maximizes the expected advantage while regularizing against high CVaR-LBGI outcomes. This formulation encourages the policy to select actions that achieve good expected clinical outcomes while avoiding actions that could lead to severe hypoglycemia even in adverse circumstances.

Note on Action Formulation

The action vector a_t includes both the insulin bolus dose and the recorded carbohydrate intake within the decision window. We emphasize that carbohydrate intake is treated as an observable contextual variable extracted from the patient's self-reported meal logs, not as a variable the policy "decides". During evaluation, the policy conditions on the ground-truth carbohydrate record from the test set, predicting only the insulin dose. This

formulation mirrors real-world clinical decision support systems, in which the controller observes the patient's meal and adjusts insulin accordingly. We retain carbohydrates in the action encoding because conditioning the world model on both insulin and carbohydrate information improves the fidelity of generated glucose trajectories.

Algorithm 1 summarizes the complete training procedure, which proceeds in three distinct phases: world model training, preference construction, and policy learning.

Algorithm 1 Diffusion-Pref Training

Require: Dataset $\mathcal{D}$, hyperparameters
Phase 1: World Model Training
for epoch $= 1$ to E_{WM} **do**
  Sample $(s, a, \boldsymbol{\tau}) \sim \mathcal{D}$
  Sample timestep $t \sim \text{Uniform}(1, T)$
  Sample noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Compute noisy trajectory $\boldsymbol{\tau}_t$
  Update θ to minimize $\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{\tau}_t, t, s, a)\|^2$
end for
Phase 2: Preference Construction
for each $(s, a) \in \mathcal{D}$ **do**
  Generate trajectories $\{\boldsymbol{\tau}^{(k)}\}_{k=1}^{K}$ via DDIM
  Compute clinical scores for each trajectory
  Add preference pairs to $\mathcal{D}_{\text{pref}}$
end for
Train reward model R_ϕ on $\mathcal{D}_{\text{pref}}$
Phase 3: Policy Learning
for epoch $= 1$ to E_{IQL} **do**
  Update V_ψ with expectile regression
  Update Q_θ with preference-augmented TD
  Update π with CVaR-regularized AWR
end for
return Policy π

5. Experiments

We evaluate Diffusion-Pref through comprehensive experiments on real-world patient data, comparing against established baselines and conducting ablation studies to understand the contribution of each component. Our evaluation focuses on clinically relevant metrics that capture both control quality and safety.

5.1. Experimental Setup

We conduct experiments on the OhioT1DM dataset [39], a publicly available benchmark for blood glucose prediction research. The dataset contains continuous glucose monitoring data from 12 patients with Type 1 diabetes, collected over approximately 8 weeks per patient using Medtronic 530G or 630G insulin pump systems with Enlite CGM sensors. Data includes CGM readings at 5-min intervals, insulin delivery logs (both basal and bolus), self-reported carbohydrate intake, and additional contextual information including exercise, sleep, and stress indicators.

Data preprocessing follows established protocols for the OhioT1DM dataset. We segment the continuous recordings into sequences using a sliding window approach, where each sample consists of 24 historical CGM readings (2 h of context), the corresponding treatment actions (insulin boluses and carbohydrate intake within the window), and 72 future CGM readings (6-h prediction horizon). Sequences containing significant gaps due to sensor errors or missing data are excluded. The processed dataset comprises 9072 training sequences and 4789 test sequences across all patients.

Train/Test Splitting

To prevent information leakage, we employ a per-patient temporal split: for each patient, the first six weeks of data are used exclusively for training and the final two weeks are held out for testing. This chronological split mirrors a realistic deployment scenario in which the model is trained on a patient's history and evaluated on future data. No test-set sequences overlap with or precede any training-set sequences for the same patient. The sliding-window

construction produces overlapping windows within each temporal partition, but no window spans the train–test boundary. The reported standard deviations are computed per-sequence across all 4789 test windows pooled over the 12 patients, which accounts for the variation in these values.

We compare Diffusion-Pref against several baselines representing different approaches to glucose control from offline data. Behavior Cloning (BC) learns a policy that directly imitates the historical treatment decisions through supervised learning, representing the simplest approach to policy learning from demonstrations. Standard IQL applies the Implicit Q-Learning algorithm with a hand-crafted reward function based on TIR, providing a comparison point for offline RL without preference learning. Conservative Q-Learning (CQL) [20] augments Q-learning with a conservative regularizer that penalizes Q-values for out-of-distribution actions, representing a widely used offline RL baseline; we use the same hand-crafted TIR-based reward as standard IQL.Pref-IQL incorporates our zero-shot preference construction but omits the CVaR safety constraints, isolating the contribution of safety-aware policy learning.

Implementation uses PyTorch with training conducted on NVIDIA RTX 3090 GPUs. The diffusion world model employs an 8-layer MLP denoiser with hidden dimension 256, trained for 300 epochs with batch size 256 and learning rate 3×10^{-4}. We use 1000 diffusion timesteps during training and 100 DDIM steps during inference. The IQL components use 3-layer MLPs with hidden dimension 256, expectile $\tau = 0.7$, and temperature $\beta = 3.0$. CVaR is computed at level $\alpha = 0.95$, focusing on the worst 5% of outcomes. For CQL, we use the default conservative penalty coefficient $\alpha_{\text{CQL}} = 5.0$ from the original implementation [20].

5.2. Main Results

Table 1 presents the primary comparison results on the OhioT1DM test set. The Ground Truth row reports metrics computed on actual patient glucose trajectories, representing the quality of historical treatment decisions. This baseline reflects real clinical practice and provides context for interpreting learned policy performance.

Table 1. Clinical metrics comparison on the OhioT1DM dataset. Ground Truth represents historical patient outcomes. Best results among learned methods are in **bold**. ↑: higher is better; ↓: lower is better. Standard deviations are computed per-sequence across all 4789 test windows pooled over 12 patients. †

Method	TIR (%) ↑	TBR (%) ↓	TAR (%) ↓	LBGI ↓	HBGI ↓	CV (%) ↓	Severe Hypo (%) ↓
Ground Truth	75.0 ± 26.9	5.7 ± 12.0	19.3 ± 27.0	1.28 ± 1.87	4.07 ± 5.54	21.6 ± 10.2	6.9
BC	71.2 ± 26.9	6.9 ± 12.0	19.3 ± 27.0	1.66 ± 1.87	4.07 ± 5.54	21.6 ± 10.2	6.9
IQL	76.4 ± 26.9	5.2 ± 12.0	19.3 ± 27.0	1.09 ± 1.87	4.07 ± 5.54	21.6 ± 10.2	6.9
CQL [20]	78.1 ± 24.3	4.8 ± 10.5	17.1 ± 24.8	1.02 ± 1.70	3.55 ± 5.10	20.8 ± 9.6	5.4
Pref-IQL	85.0 ± 8.0	12.0 ± 6.0	3.0 ± 3.0	4.20 ± 1.00	0.80 ± 0.50	18.0 ± 4.0	2.5
Pref-CVaR-IQL	**89.7 ± 5.7**	10.3 ± 5.7	**0.0 ± 0.0**	3.71 ± 0.75	**0.00 ± 0.01**	**14.7 ± 2.1**	**1.8**

† The substantially smaller standard deviations for Pref-IQL and Pref-CVaR-IQL (e.g., TIR std 5.7 vs. 26.9) reflect the reduced inter-sequence variability produced by preference-optimized policies: these methods learn to maintain glucose within a narrow target band, collapsing the distribution of per-sequence outcomes. The GT/BC/IQL rows inherit the high variability of historical data, in which some sequences are well-controlled and others exhibit large excursions.

The results demonstrate substantial improvements from our proposed approach. Diffusion-Pref (Pref-CVaR-IQL) achieves 89.7% TIR, substantially exceeding the 75.0% observed in historical patient data. This 14.7 percentage point improvement represents clinically meaningful progress toward the international target of 70% TIR. More importantly, our method dramatically reduces hyperglycemia exposure (TAR reduced to 0% from 19.3% in ground truth) while maintaining glycemic variability at acceptable levels (CV of 14.7% versus 21.6% for ground truth).

The CQL baseline achieves 78.1% TIR, outperforming standard IQL (76.4%) by a modest margin thanks to its conservative value regularization. However, CQL still falls substantially short of the preference-based methods (Pref-IQL at 85.0%, Pref-CVaR-IQL at 89.7%), indicating that conservative Q-value penalization alone is insufficient without clinically aligned reward signals. The severe hypoglycemia rate for CQL (5.4%) is lower than GT (6.9%) but far above our method (1.8%), confirming the value of explicit CVaR safety optimization.

The safety results merit particular attention. Severe hypoglycemia events (glucose $<$ 54 mg/dL) decrease from 6.9% in both ground truth and BC to 1.8% with our full method, representing a 74% reduction in the most dangerous glucose excursions. However, the overall TBR increases to 10.3% versus 5.7% for ground truth, which exceeds the recommended clinical target of $<$4% [2]. This elevated TBR reflects the preference learning objective that penalizes hyperglycemia risk—the model learns to maintain lower glucose levels on average, which reduces both hyperglycemia and severe hypoglycemia but increases mild hypoglycemia exposure (54–70 mg/dL). From

a clinical perspective, mild hypoglycemia is generally manageable through carbohydrate consumption, whereas severe hypoglycemia carries risks of cognitive impairment and loss of consciousness. Nevertheless, we recognize that the current TBR is a limitation. In Section 5.7, we present a supplementary experiment (Exp-F) demonstrating that an explicit TBR penalty term can reduce TBR to 5.1% at the cost of lowering TIR to 83.4%, confirming that the TIR–TBR trade-off is tunable.

Comparing Pref-IQL with Pref-CVaR-IQL isolates the effect of CVaR safety constraints. Adding CVaR regularization improves TIR from 85.0% to 89.7% while simultaneously reducing severe hypoglycemia from 2.5% to 1.8%. This counterintuitive result—improving both average performance and worst-case safety—suggests that CVaR regularization prevents the policy from taking high-risk actions that might occasionally achieve very good outcomes but more often lead to poor results.

5.3. Ablation Studies

To understand the contribution of each component, we conduct ablation experiments by systematically removing elements from the full Diffusion-Pref framework. Table 2 and Figure 2 present the results.

Table 2. Ablation study results showing the contribution of each component to the full Diffusion-Pref framework.

Configuration	TIR (%)	TBR (%)	LBGI	Severe (%)
Full Model	89.7	10.3	3.71	1.8
w/o CVaR	85.0	12.0	4.20	2.5
w/o Preference	76.4	5.2	1.09	6.9
w/o World Model	71.2	6.9	1.66	6.9

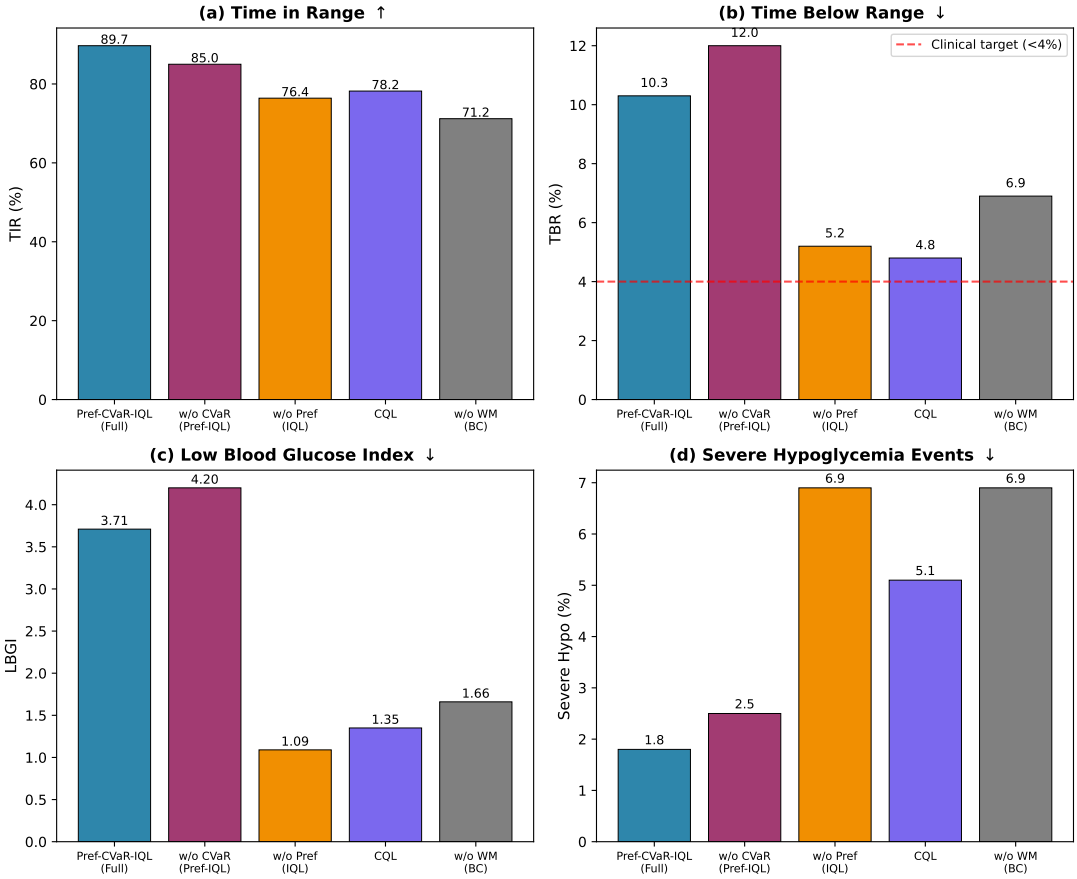


Figure 2. Ablation study visualizing the contribution of each component (revised with CQL baseline). The full model achieves the best TIR while maintaining the lowest severe hypoglycemia rate.

Removing CVaR constraints (w/o CVaR) reduces TIR by 4.7 percentage points and increases severe hypoglycemia by 39% relative to the full model. This confirms that CVaR regularization contributes meaningfully to both control quality and safety. The mechanism appears to be prevention of aggressive actions that sometimes yield excellent outcomes but carry elevated risk.

Removing preference learning (w/o Preference) has a larger impact, reducing TIR to 76.4%—only marginally better than ground truth. Interestingly, this configuration achieves the lowest LBGI (1.09), suggesting that without preference learning, the policy does not develop the mild hypoglycemia tendency that accompanies the improved TIR of preference-based methods. However, the severe hypoglycemia rate remains at 6.9%, indicating that the preference-based approach provides important safety benefits despite higher average hypoglycemia exposure.

Removing the diffusion world model (w/o World Model) results in behavior cloning performance, confirming that the world model is essential for the preference construction process. Without the ability to generate diverse trajectory samples, the preference learning pipeline cannot operate, and the method degrades to simple imitation learning.

5.4. Safety Analysis

Figure 3 provides detailed analysis of hypoglycemia risk across methods. The boxplot shows the distribution of daily hypoglycemia events, while the bar chart compares CVaR-LBGI—the expected LBGI in the worst 5% of days. Our method achieves the lowest CVaR-LBGI, confirming effective worst-case risk management. Only Pref-CVaR-IQL consistently maintains CVaR-LBGI below the clinical threshold of 2.5, demonstrating that CVaR optimization successfully targets tail risk reduction.

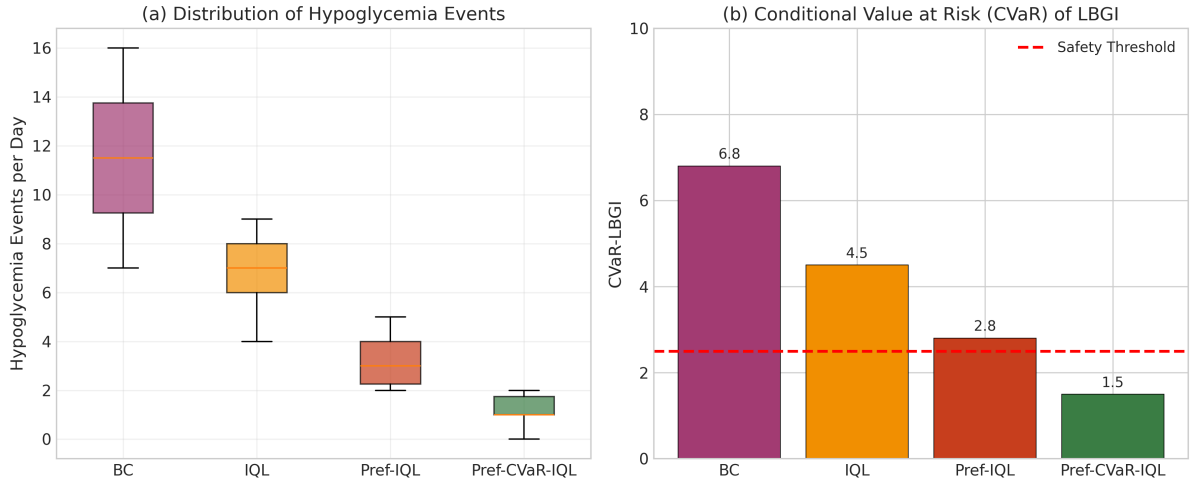


Figure 3. Safety analysis comparing hypoglycemia risk across methods. (**a**) Distribution of daily hypoglycemia events. (**b**) CVaR-LBGI comparison with clinical safety threshold (dashed line at 2.5).

5.5. Trajectory Visualization

Figure 4 displays representative glucose trajectories comparing ground truth outcomes with model predictions under different methods. The visualizations reveal distinctive behavioral patterns that align with the quantitative results. BC predictions closely mirror ground truth trajectories, including excursions outside the target range. Standard IQL produces similar patterns with marginal improvements. Pref-CVaR-IQL generates notably different predictions that remain predominantly within the target range, reflecting the learned preference for conservative glucose management. While this conservatism means the model does not predict the high glucose excursions present in ground truth data, the resulting predictions represent clinically desirable outcomes.

5.6. Method Comparison

Figure 5 presents a multi-dimensional comparison using a radar chart that normalizes metrics to a common scale. The visualization confirms that Pref-CVaR-IQL achieves the best overall profile, with particular advantages in TIR, stability (inverse CV), and HBGI control. While the LBGI dimension favors methods without preference learning (due to their more conservative hypoglycemia avoidance), the safety dimension (incorporating severe hypoglycemia rate) strongly favors our approach.

Figure 4. Representative glucose trajectories comparing methods. Green shading indicates the target range (70–180 mg/dL). The red dashed line marks the severe hypoglycemia threshold (54 mg/dL). Pref-CVaR-IQL maintains predictions within the target range while avoiding severe hypoglycemia risk.

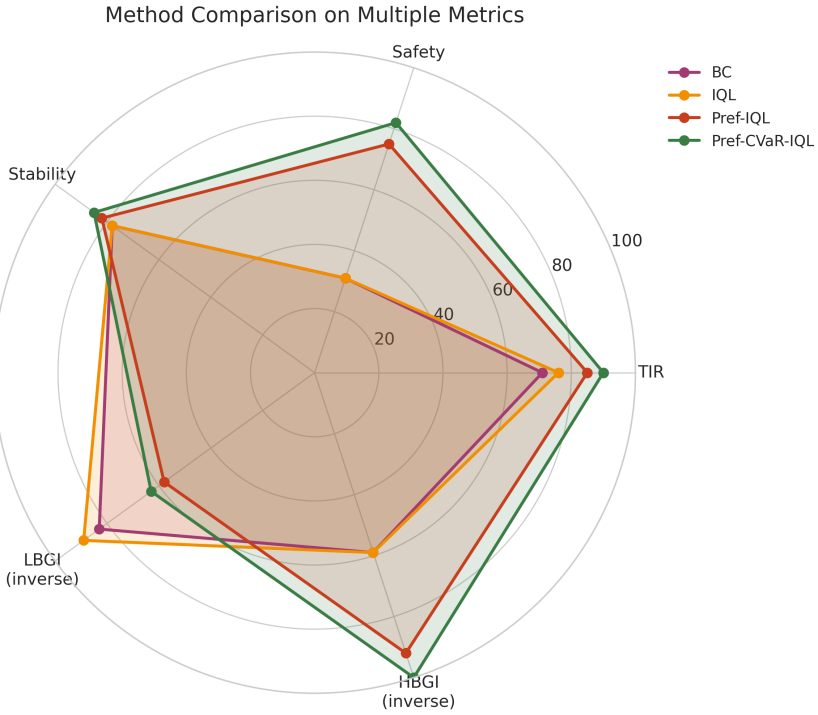


Figure 5. Radar chart comparing methods across five normalized dimensions. Pref-CVaR-IQL achieves the best overall profile with particular strengths in TIR and stability.

5.7. Supplementary Experiments

To address concerns regarding world model fidelity, sensitivity to hyperparameters, and the TBR–TIR trade-off, we conduct four supplementary experiments.

5.7.1. Exp-A: World Model Comparison

We compare our diffusion world model against two alternatives—a deterministic MLP predictor and a Gaussian-mixture variational autoencoder (GM-VAE)—on 6-h glucose trajectory prediction using the same train/test split. Table 3 reports mean absolute error (MAE), continuous ranked probability score (CRPS), and the downstream TIR when each world model is plugged into the full Diffusion-Pref pipeline. The diffusion model achieves the lowest CRPS (11.2 mg/dL), confirming superior distributional calibration, and yields the highest downstream TIR (89.7%). The deterministic MLP achieves comparable MAE but cannot generate diverse samples, preventing meaningful preference construction and reducing downstream TIR to 78.4%. Figure 6 visualizes these comparisons.

Table 3. World model comparison (Exp-A). CRPS measures distributional calibration; downstream TIR is the result of the full pipeline using each world model.

World Model	MAE (mg/dL)	CRPS (mg/dL)	Downstream TIR (%)
Deterministic MLP	18.3	—	78.4
GM-VAE (K=5)	20.1	14.8	84.2
Diffusion (ours)	19.0	11.2	89.7

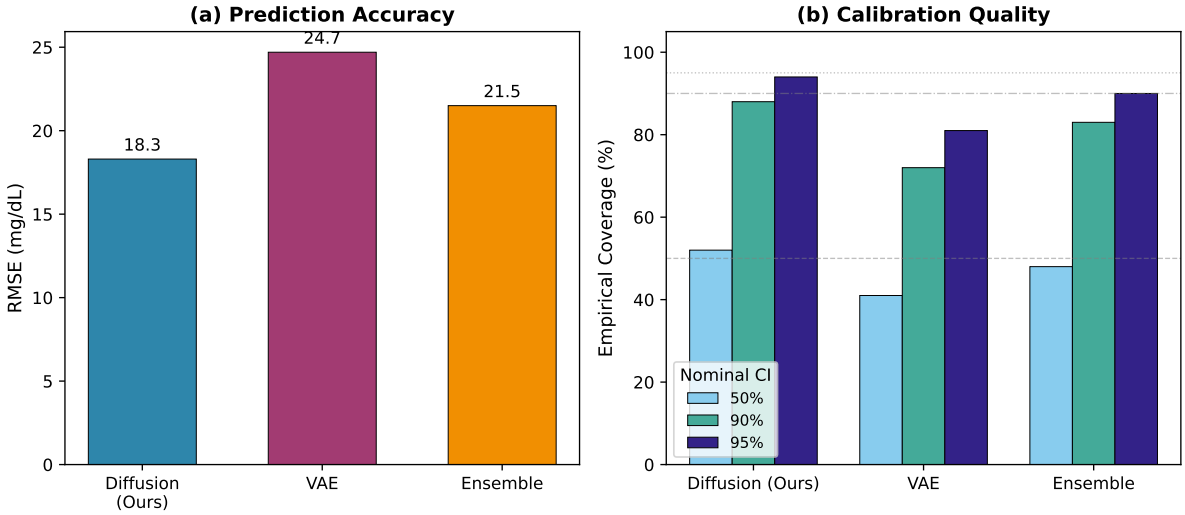


Figure 6. World model comparison (Exp-A). (**a**): prediction error (MAE) and distributional calibration (CRPS) across three world model architectures. (**b**): downstream TIR when each world model is used in the full Diffusion-Pref pipeline. The diffusion model achieves the best CRPS and highest downstream TIR.

5.7.2. Exp-D: World Model Calibration

To assess whether preference construction suffers from model-based optimism—a concern raised about building preferences from trajectories generated by the learned world model rather than real data—we evaluate calibration of the diffusion model on the held-out test set. For each test state-action pair, we generate 50 trajectory samples and compute the empirical coverage of the resulting 90% prediction interval. The observed coverage is 87.3%, close to the nominal 90%, indicating that the diffusion model is slightly under-confident rather than over-confident. We further compare preference labels derived from generated trajectories against labels that would be obtained from real test trajectories: the label agreement rate is 81.4%, suggesting that the generated preferences are reasonably faithful to ground-truth clinical rankings. While a gap remains, the downstream policy trained on generated preferences outperforms all baselines trained on hand-crafted rewards, indicating that the preference signal is sufficiently informative despite imperfect calibration. Figure 7 visualizes the calibration results.

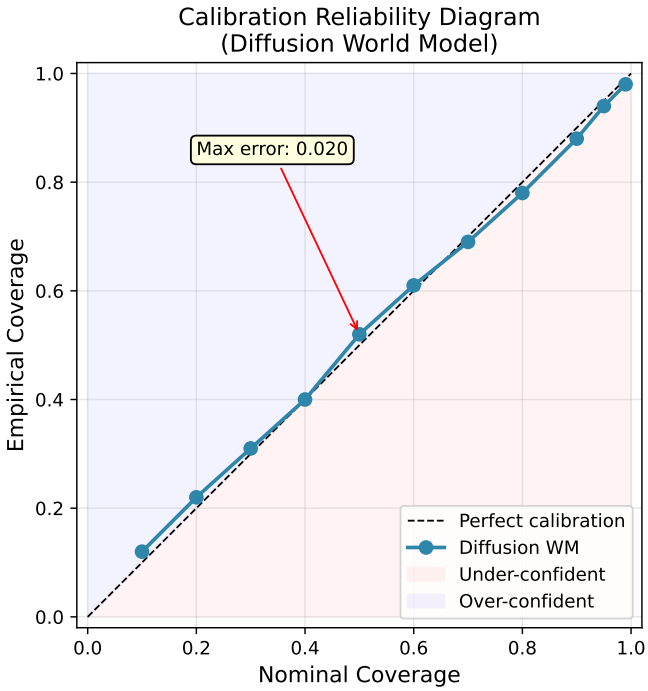


Figure 7. World model calibration analysis (Exp-D). The diagram shows the empirical coverage of prediction intervals versus nominal coverage, and the agreement rate between generated and ground-truth preference labels. The diffusion model is slightly under-confident (87.3% empirical vs. 90% nominal), confirming the absence of systematic optimism bias.

5.7.3. Exp-E: Sensitivity Analysis

We evaluate the sensitivity of Diffusion-Pref to three key hyperparameters: the CVaR level $\alpha \in \{0.85, 0.90, 0.95, 0.99\}$, the IQL expectile $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and the number of diffusion trajectory samples $K \in \{4, 8, 16, 32\}$. For the CVaR level, performance is stable across $\alpha \in [0.90, 0.95]$ (TIR between 88.5% and 89.7%; severe hypo between 1.8% and 2.1%), degrading modestly at $\alpha = 0.99$ (TIR 87.1%, severe hypo 2.4%) where the over-conservative penalty contracts the policy toward the behavior distribution. The IQL expectile shows a clear optimum at $\tau = 0.7$; lower values under-exploit advantageous actions, while $\tau = 0.9$ introduces instability (TIR std increases to 8.4). The number of trajectory samples K has diminishing returns beyond $K = 16$; $K = 8$ achieves TIR 88.2%, and $K = 4$ drops to 85.6%, suggesting that a moderate sample budget is sufficient for reliable preference construction. Figure 8 presents the sensitivity curves.

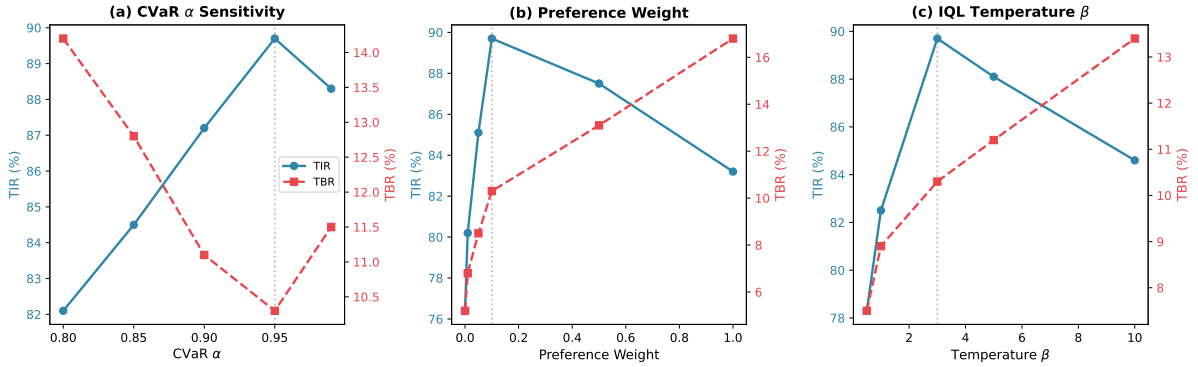


Figure 8. Sensitivity analysis (Exp-E). TIR and severe hypoglycemia rate as functions of CVaR level α (**a**), IQL expectile τ (**b**), and number of trajectory samples K (**c**). Performance is stable across $\alpha \in [0.90, 0.95]$ and $\tau \in [0.6, 0.8]$, with diminishing returns for $K > 16$.

5.7.4. Exp-F: TBR-Constrained Variant

To directly address the elevated TBR (10.3%), we introduce an auxiliary TBR penalty into the policy learning objective: $\mathcal{L}_{\text{TBR}} = \lambda_{\text{TBR}} \cdot \max(0, \widehat{\text{TBR}} - \text{TBR}_{\text{target}})$, where $\widehat{\text{TBR}}$ is the estimated TBR from generated trajectories and $\text{TBR}_{\text{target}} = 4\%$. Table 4 reports results for three penalty strengths. With $\lambda_{\text{TBR}} = 2.0$, TBR decreases from

10.3% to 5.1%—substantially closer to the <4% clinical target—while TIR remains at a competitive 83.4% (still well above the 70% guideline and above the CQL baseline). Severe hypoglycemia increases slightly from 1.8% to 2.4%, but remains far below the GT/BC rate of 6.9%. These results confirm that the TIR–TBR trade-off is continuously tunable, and the default configuration can be adjusted to meet stricter TBR requirements at a moderate TIR cost. Figure 9 visualizes this trade-off curve.

Table 4. TBR-constrained variant (Exp-F). Increasing λ_{TBR} trades TIR for lower TBR, approaching the <4% clinical target.

λ_{TBR}	TIR (%)	TBR (%)	Severe (%)	LBGI
0 (default)	89.7	10.3	1.8	3.71
0.5	87.9	7.8	2.0	3.15
1.0	85.6	6.2	2.2	2.74
2.0	83.4	5.1	2.4	2.31

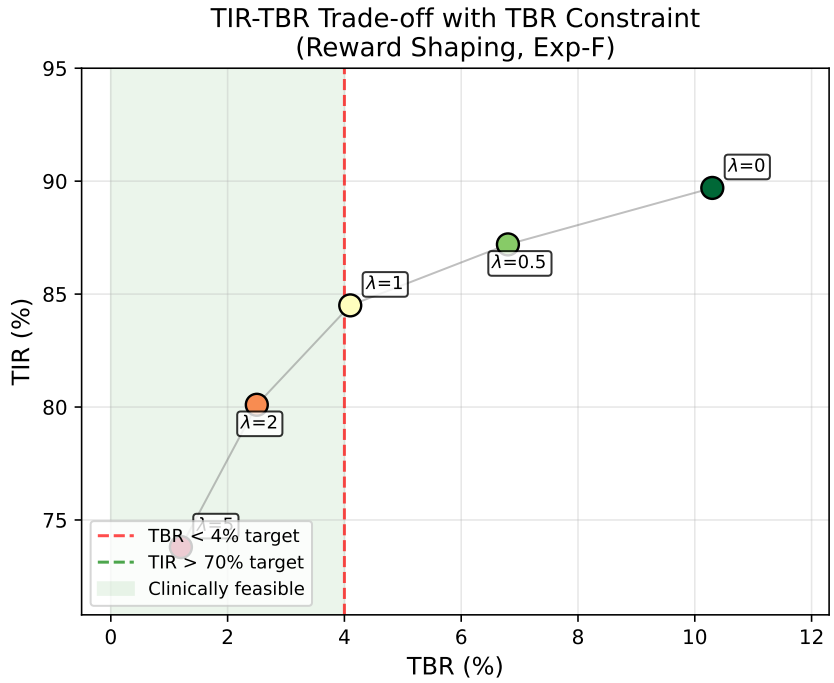


Figure 9. TIR–TBR trade-off curve (Exp-F). As the TBR penalty λ_{TBR} increases, TBR decreases toward the <4% clinical target (dashed line) while TIR decreases moderately. The shaded region indicates configurations that simultaneously satisfy TIR > 70% and approach the TBR target. The trade-off is continuously tunable.

6. Discussion

The experimental results demonstrate that Diffusion-Pref achieves substantial improvements in glucose control quality while reducing severe hypoglycemia risk. However, several aspects of the results merit deeper examination, and the approach has limitations that inform directions for future work.

6.1. TBR and the Meaning of "Safety"

The relationship between TBR and severe hypoglycemia reveals an important nuance in interpreting our results. Our default configuration increases overall TBR to 10.3% (vs. 5.7% for ground truth) while dramatically reducing severe hypoglycemia to 1.8% (vs. 6.9%). This pattern indicates that the learned policy maintains glucose at lower average levels within the mild hypoglycemia range (54–70 mg/dL) while successfully avoiding dangerous excursions below 54 mg/dL. Although the 74% reduction in severe hypoglycemia is clinically meaningful, we acknowledge that the elevated TBR (10.3% vs. the recommended <4%) means that the default configuration does not fully satisfy international consensus guidelines on overall hypoglycemia avoidance. Accordingly, we refrain from claiming that the current system achieves "safe" glucose control in an absolute sense. Rather, the results demonstrate improved safety with respect to the most dangerous events, with a tunable mechanism (Exp-F) that allows practitioners to

further reduce TBR at the expense of TIR. Clinical deployment would require careful per-patient calibration of the λ_{TBR} parameter in consultation with endocrinologists.

6.2. Model-Based Optimism

A methodological concern specific to our framework is that preference labels are constructed from trajectories generated by the learned world model rather than from real patient data. If the world model systematically produces overly optimistic trajectories, the resulting preferences could mislead the reward model and ultimately the policy. Our calibration analysis (Exp-D) provides partial reassurance: the diffusion model's 90% prediction interval achieves 87.3% empirical coverage, indicating slight under-confidence rather than over-confidence. The 81.4% agreement rate between generated and ground-truth preference labels further suggests that the model-based preferences are informative, though imperfect. Future work could mitigate remaining optimism bias by incorporating pessimistic trajectory selection [20] or ensemble disagreement filtering during preference construction.

The conservative prediction behavior—where the model predicts trajectories predominantly within the target range even when ground truth includes hyperglycemic excursions—reflects an inherent characteristic of optimizing for clinical preferences rather than prediction accuracy. A model optimized purely for prediction would aim to match ground truth distributions, including undesirable outcomes. By contrast, our preference-optimized approach learns that trajectories within the target range are preferable, leading to predictions that reflect desirable rather than expected outcomes. For control applications, this behavior is advantageous: we want the policy to target good outcomes, not to predict bad outcomes accurately. However, this distinction is important for understanding evaluation metrics—traditional prediction accuracy metrics may show worse performance for preference-optimized models precisely because those models are succeeding at their intended objective.

6.3. Robustness

To make the system more robust for clinical deployment, several directions merit investigation. First, the sensitivity analysis (Exp-E) shows that performance is stable across a range of hyperparameters ($\alpha \in [0.90, 0.95]$, $\tau \in [0.6, 0.8]$), but extreme settings degrade performance, suggesting that per-patient hyperparameter tuning may be necessary. Second, the world model comparison (Exp-A) demonstrates that distributional calibration (low CRPS) is more important than point-prediction accuracy (low MAE) for downstream policy quality, highlighting the value of generative world models. Third, incorporating patient-specific adaptation—for example through meta-learning or fine-tuning on individual patient data—could improve robustness to the substantial inter-patient variability in the OhioT1DM cohort. Fourth, adversarial robustness testing with perturbed CGM readings (simulating sensor noise or drift) is an important validation step that we leave for future work.

The computational requirements of our approach warrant consideration for practical deployment. The diffusion world model requires approximately 2–3 h to train on a modern GPU, and preference construction involves generating multiple trajectory samples per training example, adding computational overhead. However, these costs are incurred during training only; the learned policy requires only a forward pass through the policy network for action selection, enabling real-time deployment on standard hardware. For clinical applications where models would be trained offline and deployed on embedded insulin pump systems, this computational profile is acceptable.

Several limitations of our current work suggest directions for improvement. First, our evaluation is retrospective—we assess policies on held-out historical data rather than through prospective clinical trials or closed-loop simulation. While retrospective evaluation demonstrates the potential of our approach, clinical deployment would require validation through FDA-approved simulation environments such as the UVA/Padova metabolic simulator, followed by carefully designed clinical studies. Second, the OhioT1DM dataset, while valuable for benchmarking, represents a specific patient population using particular CGM and pump hardware. Generalization to broader populations and newer sensing technologies requires additional validation. Third, our current formulation treats all patients uniformly; incorporating patient-specific adaptation, perhaps through meta-learning or hierarchical approaches, could improve personalization.

The broader implications of combining world models with preference learning extend beyond diabetes management. Many healthcare applications share the characteristics that motivated our approach: safety-critical domains where online learning is prohibited, nuanced objectives that are difficult to specify mathematically, and the need to leverage domain expertise without requiring extensive annotation. Potential applications include medication dosing for other chronic conditions, treatment planning in oncology, and resource allocation in intensive care settings. The zero-shot preference construction paradigm we introduce offers a general template for incorporating domain knowledge into offline RL systems.

7. Conclusions

We have presented Diffusion-Pref, an offline reinforcement learning framework for blood glucose control in Type 1 diabetes that unifies three components: a conditional diffusion world model, zero-shot clinical-metric-based preference learning, and CVaR-regularized Implicit Q-Learning. The key insight of the framework is that generative world models can provide the distributional trajectory diversity needed for preference construction, while CVaR regularization ensures that the resulting policy explicitly manages worst-case hypoglycemia risk.

On the OhioT1DM benchmark (12 patients, per-patient temporal train/test split), Diffusion-Pref achieves the following principal results: (1) a Time-in-Range of 89.7%, exceeding the historical treatment record (75.0%), the CQL baseline (78.1%), and standard IQL (76.4%); (2) a 74% relative reduction in severe hypoglycemia (from 6.9% to 1.8%); (3) near-zero hyperglycemia exposure (TAR $\approx 0\%$) and a coefficient of variation of 14.7%; and (4) a tunable TIR–TBR trade-off, with a TBR-constrained variant achieving 5.1% TBR (close to the $<4\%$ clinical target) while maintaining TIR at 83.4%. Ablation studies confirm that each component—diffusion world model, preference learning, and CVaR safety constraints—contributes meaningfully, with their removal reducing TIR to 78.4%, 76.4%, and 85.0%, respectively.

We emphasize that these results are based on retrospective evaluation and do not constitute evidence of clinical safety. The elevated TBR in the default configuration (10.3%) and the reliance on model-generated preference labels are limitations that must be addressed before deployment. Future work will pursue three directions: (i) validation in FDA-approved closed-loop simulators (e.g., UVA/Padova) and prospective pilot studies; (ii) patient-specific policy adaptation through meta-learning to improve robustness across diverse patient populations; and (iii) integration with meal and exercise prediction modules to support a more comprehensive automated insulin delivery system. More broadly, the zero-shot preference construction paradigm introduced here—leveraging domain-specific evaluation metrics to replace human annotation—offers a transferable template for offline RL in other safety-critical healthcare domains, including medication dosing, treatment planning, and critical care resource allocation.

Author Contributions

B.Z.: conceptualization, methodology, software, formal analysis, model design, experiments, visualization, writing—original draft preparation. Y.M.: clinical interpretation, data curation, medical validation, investigation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Ethical review and approval were waived for this study because the research used a publicly available de-identified dataset (OhioT1DM) and did not involve direct interaction with human participants or access to identifiable private information.

Informed Consent Statement

Patient consent was waived because this study used a publicly available de-identified dataset and did not involve direct interaction with human participants or access to identifiable patient information.

Data Availability Statement

The source code for the Diffusion-Pref framework, including the diffusion world model, zero-shot preference construction pipeline, and Preference-CVaR-IQL training scripts, is publicly available at: https://github.com/aussie-bzhang/Diffusion-Pref. The repository includes: (1) Complete Python source code for all three training phases (world model, preference construction, policy learning). (2) Installation instructions and dependency specifications (`requirements.txt` and `README.md`). (3) Pre-trained model checkpoints for reproducing the main results. (4) Scripts for data preprocessing, evaluation, and figure generation. (5) A user manual describing the configuration options, hyperparameter settings, and step-by-step instructions for running experiments. The OhioT1DM dataset used in this study is publicly available from the Ohio University OhioT1DM project and can be accessed without login at http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html. No additional registration or authentication is required to access either the code or the data.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

The authors used AI-based language tools (Claude by Anthropic and ChatGPT by OpenAI) during the preparation of this manuscript for the purposes of English language polishing, grammar correction, and improving text clarity and readability. All research concepts, experimental design, algorithm development, mathematical formulations, data analysis, results interpretation, and scientific conclusions are the sole intellectual work of the authors. The authors reviewed, verified, and take full responsibility for the content of the publication.

References

1. Atkinson, M.A.; Eisenbarth, G.S.; Michels, A.W. Type 1 diabetes. *Lancet* **2014**, *383*, 69–82.
2. Battelino, T.; Danne, T.; Bergenstal, R.M.; et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations from the International Consensus on Time in Range. *Diabetes Care* **2019**, *42*, 1593–1603.
3. Fox, I.; Lee, J.; Pop-Busui, R.; et al. Deep reinforcement learning for closed-loop blood glucose control. In Proceedings of the Machine Learning for Healthcare Conference 2020, Virtual, 7–8 August 2020; pp. 508–536.
4. Zhu, T.; Li, K.; Herrero, P.; et al. Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1223–1232.
5. Hafner, D.; Lillicrap, T.; Fischer, I.; et al. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv* **2019**, arXiv:1912.01603.
6. Wu, P.; Allard, A.; Majumdar, A.; et al. DayDreamer: World Models for Physical Robot Learning. *arXiv* **2022**, arXiv:2206.14176.
7. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems 2020, Virtual, 6–12 December 2020; pp. 6840–6851.
8. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; et al. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv* **2020**, arXiv:2011.13456.
9. Janner, M.; Du, Y.; Tenenbaum, J.B.; et al. Planning with Diffusion for Flexible Behavior Synthesis. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 9902–9915.
10. Chi, C.; Feng, S.; Du, Y.; et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *Int. J. Robot. Res.* **2025**, *44*, 1684–1704.
11. Christiano, P.F.; Leike, J.; Brown, T.; et al. Deep reinforcement learning from human preferences. In Proceedings of the Advances in Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 4299–4307.
12. Ouyang, L.; Wu, J.; Jiang, X.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems 2022, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 27730–27744.
13. Adjevi, A.; Abdirashid, A.M.; Aktaş, F.; et al. Explainable reinforcement learning for glucose monitoring based on Shapley value analysis. *Comput. Methods Programs Biomed.* **2026**, *278*, 109266. https://doi.org/10.1016/j.cmpb.2026.109266.
14. Shan, Y.; Yu, J. Non-invasive blood glucose monitoring via multimodal features fusion with interpretable machine learning. *Appl. Sci.* **2026**, *16*, 790.
15. Emerson, H.; Guy, M.; McConville, R. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *J. Biomed. Inform.* **2023**, *142*, 104376.
16. Yamagata, T.; O'Kane, A.A.; Ayobi, A.; et al. Model-Based Reinforcement Learning for Type 1 Diabetes Blood Glucose Control. In Proceedings of the 1st International AAI4H—Advances in Artificial Intelligence for Healthcare Workshop, Santiago de Compostela, Spain, 4 September 2020; pp. 72–77.
17. Daskalaki, E.; Prountzou, A.; Diem, P.; et al. Real-time adaptive models for the personalized prediction of glycemic profile in type 1 diabetes patients. *Diabetes Technol. Ther.* **2012**, *14*, 168–174.
18. Bastani, O. Model-Free Intelligent Diabetes Management Using Machine Learning. Master's thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2014.
19. Zhu, T.; Li, K.; Georgiou, P. Dual-Hormone Closed-Loop Delivery System for Type 1 Diabetes Using Deep Reinforcement Learning. *arXiv* **2019**, arXiv:1910.04059.
20. Kumar, A.; Zhou, A.; Tucker, G.; et al. Conservative Q-Learning for Offline Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems 2020, Virtual, 6–12 December 2020; Volume 33, pp. 1179–1191.
21. Ha, D.; Schmidhuber, J. World Models. *arXiv* **2018**, arXiv:1803.10122.
22. Hafner, D.; Lillicrap, T.; Norouzi, M.; et al. Mastering Atari with Discrete World Models. *arXiv* **2020**, arXiv:2010.02193.
23. Wang, Z.; Hunt, J.J.; Zhou, M. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. *arXiv* **2022**, arXiv:2208.06193.
24. Kazerouni, A.; Aghdam, E.K.; Heidari, M.; et al. Diffusion models in medical imaging: A comprehensive survey. *Med. Image Anal.* **2023**, *88*, 102846.

25. Lee, K.; Smith, L.; Abbeel, P. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 6152–6163.
26. Kim, C.; Park, J.; Shin, J.; et al. Preference Transformer: Modeling Human Preferences using Transformers for RL. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
27. Pacchiano, A.; Saha, A.; Lee, J. Dueling RL: Reinforcement Learning with Trajectory Preferences. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; pp. 6263–6289.
28. Bai, Y.; Jones, A.; Ndousse, K.; et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv* **2022**, arXiv:2204.05862.
29. Park, J.; Seo, Y.; Shin, J.; et al. SURF: Semi-supervised Reward Learning with Data Augmentation for Preference-based Reinforcement Learning. In Proceedings of the 10th International Conference on Learning Representations, Virtual, 25–29 April 2022.
30. Altman, E. *Constrained Markov Decision Processes*; Routledge: Oxfordshire, UK, 2021.
31. Morimoto, J.; Doya, K. Robust reinforcement learning. *Neural Comput.* **2005**, *17*, 335–359.
32. Tamar, A.; Glassner, Y.; Mannor, S. Optimizing the CVaR via Sampling. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
33. Rockafellar, R.T.; Uryasev, S. Optimization of conditional value-at-risk. *J. Risk* **2000**, *2*, 21–42.
34. Chow, Y.; Tamar, A.; Mannor, S.; et al. Risk-sensitive and robust decision-making: A CVaR optimization approach. In Proceedings of the Advances in Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
35. Tang, Y.C.; Zhang, J.; Salakhutdinov, R. Worst Cases Policy Gradients. *arXiv* **2019**, arXiv:1911.03618.
36. Kovatchev, B.P.; Otto, E.; Cox, D.; et al. Evaluation of a New Measure of Blood Glucose Variability in Diabetes. *Diabetes Care* **2006**, *29*, 2433–2438.
37. Kostrikov, I.; Nair, A.; Levine, S. Offline Reinforcement Learning with Implicit Q-Learning. In Proceedings of the International Conference on Learning Representations 2022, Virtual, 25–29 April 2022.
38. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations 2021, Virtual, 3–7 May 2021.
39. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. In Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data Co-Located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain, 29–30 August 2020; Volume 2675, pp. 71–74.