Sensors and AI
https://www.sciltp.com/journals/sai

Article

Implicit Q-Learning for Offline Reinforcement Learning in Blood Glucose Management: A Cross-Dataset Evaluation Study

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

[1] School of Computer Science and Data Engineering, NingboTech University, Qianhunan Road 1, Ningbo 315100, China
[2] The First Affiliated Hospital of Ningbo University, Ningbo 315000, China
* Correspondence: bailing.zhang1961@gmail.com

How To Cite: Zhang, B.; Mi, Y. Implicit Q-Learning for Offline Reinforcement Learning in Blood Glucose Management: A Cross-Dataset Evaluation Study. *Sensors and AI* **2026**, *2*(1), 17–32. https://doi.org/10.53941/sai.2026.100002

Received: 30 December 2025
Revised: 2 February 2026
Accepted: 28 February 2026
Published: 16 March 2026

Abstract: Offline reinforcement learning enables learning insulin dosing policies from historical data without risky patient interactions. This study evaluates Implicit Q-Learning (IQL) on three real-world continuous glucose monitoring datasets: OhioT1DM (USA, type 1), ShanghaiT1DM (China, type 1), and ShanghaiT2DM (China, type 2). IQL achieved time in range of 68.5%, 54.3%, and 78.9% respectively. Cross-dataset transfer experiments demonstrated exceptional generalization with over 98% performance retention across geographic regions and diabetes types, suggesting that IQL captures fundamental glucose-insulin dynamics rather than dataset-specific patterns. Ablation studies validated our clinically-motivated reward function design, while sensitivity and robustness analyses confirmed algorithm stability across hyperparameter choices and data quality perturbations.

Keywords: offline reinforcement learning; implicit Q-Learning; blood glucose management; diabetes; transfer learning; continuous glucose monitoring

1. Introduction

Diabetes mellitus affects approximately 537 million adults globally, with projections reaching 783 million by 2045 [1]. Despite advances in continuous glucose monitoring (CGM) technology that provides real-time readings every 5–15 min [2], achieving recommended glycemic targets—time in range (TIR) $\geq 70\%$, time below range $\leq 4\%$, and time above range $\leq 25\%$ [3–5]—remains challenging. Poor blood glucose control causes acute complications such as hypoglycemia and diabetic ketoacidosis, as well as long-term microvascular and macrovascular damage [6]. The sequential decision-making required for optimal insulin dosing, accounting for glucose trends, insulin on board, meals, and individual factors, presents an ideal application for reinforcement learning (RL) [7].

Reinforcement learning (RL) provides a principled approach to such problems by learning optimal actions to maximize rewards through environment interaction [7]. In diabetes, RL has powered artificial pancreas systems: early works integrated it with model predictive control [6,8], and recent deep RL methods improved outcomes in simulators by adapting to patterns while prioritizing safety [9–12]. However, online RL's exploration poses risks of dangerous glucose excursions and demands excessive real-world samples, hindering clinical adoption [13,14].

Offline RL addresses these barriers by learning solely from historical data, suiting healthcare's abundance of retrospective records and ethical constraints on interaction [15]. The core challenge is distributional shift: when the learned policy selects actions different from those in the training data, the Q-function may produce erroneous estimates due to extrapolation error. Several algorithms have been proposed to mitigate this issue, each with distinct limitations. Conservative Q-Learning (CQL) [16] penalizes Q-values for out-of-distribution actions to learn a lower bound, but this conservatism can be overly restrictive. Batch Constrained Q-Learning (BCQ) [17] restricts actions to those similar to the behavior policy using a generative model, requiring additional model complexity. Policy regularization methods such as BEAR [18] and BRAC [19] constrain the learned policy through divergence penalties, but require careful tuning of regularization strength.

Implicit Q-Learning (IQL) [20] takes a fundamentally different approach by avoiding explicit policy constraints. Instead, it uses expectile regression to learn a value function that implicitly represents the value of the best actions *within* the dataset, completely avoiding queries to out-of-distribution actions. This design makes IQL particularly suitable for healthcare applications where the behavior policy (historical clinical decisions) represents reasonable clinical practice, and the goal is to identify the best actions within this safe distribution rather than extrapolating to potentially dangerous novel actions. IQL has achieved state-of-the-art performance on standard offline RL benchmarks [20,21].

Offline RL applications in healthcare are expanding, particularly for chronic conditions requiring sequential treatment decisions [14]. In diabetes, offline RL has produced safer, guideline-aligned policies that outperform historical treatments on individual datasets [22,23]. Transfer learning in RL enables knowledge sharing across domains, which is vital for healthcare's inherent heterogeneity across institutions, populations, and disease subtypes [24–26]. Yet, rigorous cross-dataset evaluations of offline RL in diabetes—across regions, lifestyles, and disease types—remain limited, leaving unclear whether learned policies capture universal physiology or overfit to specific cohorts.

This study fills this gap by assessing IQL on three real-world CGM datasets: OhioT1DM (USA, T1DM) [27] and ShanghaiT1DM/ShanghaiT2DM (China) [28]. Spanning diverse populations and glucose dynamics, these datasets facilitate robust transfer testing. Results show stable training, clinically meaningful performance, and exceptional generalization ($\geq 98\%$ retention across transfers), suggesting IQL learns core transferable glucose-insulin strategies.

The main contributions of this work are threefold:

1. We demonstrate that IQL achieves stable training and clinically meaningful glycemic control across three diverse real-world CGM datasets, with TIR ranging from 54.3% to 78.9% depending on patient population characteristics.
2. We conduct the first comprehensive cross-dataset transfer evaluation of offline RL for diabetes management, testing zero-shot generalization across geographic regions (USA $\leftrightarrow$ China) and diabetes types (T1DM $\leftrightarrow$ T2DM).
3. We show that IQL policies achieve exceptional transfer performance with $>98\%$ retention across all scenarios, providing evidence that the algorithm captures fundamental glucose-insulin dynamics rather than dataset-specific patterns.

The paper proceeds as follows: Methods describe the problem formulation, algorithm, and experimental setup; Results report single-dataset and cross-dataset outcomes; Discussion examines implications, limitations, and future work.

2. Methods

2.1. Problem Formulation

Blood glucose management was formulated as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\mathcal{P}$ the transition dynamics, and $\gamma \in [0, 1)$ the discount factor.

2.1.1. State Space

The state at time t consisted of a historical window of CGM readings and associated features. For the OhioT1DM dataset, each state comprised 12 consecutive 5-min steps (60 min total) with 7 features: finger stick glucose, basal insulin rate, heart rate, galvanic skin response, carbohydrate intake, bolus insulin dose, and CGM glucose value. For the Shanghai datasets, states used 4 consecutive 15-min steps (also 60 min) with normalized CGM and insulin features.

2.1.2. Action Space

Actions represented normalized insulin delivery decisions comprising basal rate adjustment and bolus dose. Each component was scaled to $[0, 1]$ based on patient-specific maxima observed in the dataset:

$$a = [a_{\text{basal}}, a_{\text{bolus}}] \in [0, 1]^2 \tag{1}$$

2.1.3. Reward Function

A clinically motivated reward function was designed to prioritize time in range while asymmetrically penalizing hypoglycemia, reflecting the clinical reality that hypoglycemic events pose more immediate danger than hyperglycemia [4]:

$$R(g, a) = \text{clip}\left(\frac{r_{\text{TIR}} + r_{\text{hypo}} + r_{\text{action}}}{2}, -1, 1\right) \tag{2}$$

with components:

$$r_{\text{TIR}} = \begin{cases} +1.0 & 70 \leqslant g \leqslant 180 \\ -0.5 & g \in [54, 70) \cup (180, 250] \\ -1.0 & \text{otherwise} \end{cases} \tag{3}$$

$$r_{\text{hypo}} = \begin{cases} -1.0 & g < 54 \\ -0.5 & 54 \leqslant g < 70 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$r_{\text{action}} = -0.1 \cdot |a| \tag{5}$$

The r_{TIR} component directly aligns with the international consensus target of ≥70% time in range [4]. The r_{hypo} component provides additional asymmetric penalties for hypoglycemia, reflecting its clinical severity: severe hypoglycemia (<54 mg/dL) can cause seizures, loss of consciousness, and death, warranting the strongest penalty. The r_{action} component discourages excessive insulin delivery, embodying the clinical principle of using the minimum effective dose to reduce hypoglycemia risk. We validate this design through ablation studies in Section 3.5.

2.2. *Implicit Q-Learning*

Implicit Q-Learning (IQL) [20] mitigates distributional shift in offline RL by avoiding evaluation of out-of-distribution actions. Unlike methods that explicitly constrain the policy or penalize Q-values, IQL learns value functions that implicitly capture the value of the best in-distribution actions. It comprises three core components trained on a fixed offline dataset $\mathcal{D}$.

2.2.1. Value Function Learning via Expectile Regression

The state value function $V_\psi(s)$ is trained to estimate the expectile of the Q-value distribution:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_\tau^2\left(Q_{\hat{\theta}}(s, a) - V_\psi(s)\right)\right] \tag{6}$$

where $L_\tau^2(u) = |\tau - \mathbf{1}(u < 0)|u^2$ is the asymmetric expectile loss. The expectile parameter $\tau \in (0.5, 1)$ controls how much the value function is biased toward higher Q-values. When $\tau = 0.5$, this reduces to standard mean regression; as $\tau \to 1$, the value function increasingly approximates the maximum Q-value over actions in the dataset. We use $\tau = 0.7$, which provides a moderate bias toward high-value actions while maintaining stability, as validated in our sensitivity analysis (Section 3.6). The target Q-network $Q_{\hat{\theta}}$ is used to provide stable training targets.

2.2.2. Q-Function Learning

The Q-function is trained using temporal difference learning:

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma V_\psi(s') - Q_\theta(s, a)\right)^2\right] \tag{7}$$

Crucially, the bootstrap target uses $V_\psi(s')$ rather than $\max_{a'} Q(s', a')$, which would require evaluating potentially out-of-distribution actions. This design ensures that Q-learning never queries the Q-function on actions outside the dataset. We employ double Q-learning [29] with two Q-networks Q_{θ_1} and Q_{θ_2}, using the minimum for computing advantages to reduce overestimation bias. Target networks $Q_{\hat{\theta}_1}$ and $Q_{\hat{\theta}_2}$ are updated via exponential moving average with rate $\tau_{\text{target}} = 0.005$.

2.2.3. Policy Extraction

The policy is extracted through advantage-weighted behavioral cloning:

$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta \cdot A(s, a)\right) \cdot \|a - \pi_\phi(s)\|^2\right] \tag{8}$$

where $A(s, a) = Q(s, a) - V(s)$ is the advantage function. The temperature parameter β controls exploitation: higher values increase the weight on high-advantage actions. We use $\beta = 3.0$, which provides strong preference for advantageous actions while maintaining policy smoothness. This formulation extracts a policy that imitates the best actions in the dataset according to the learned value functions, without requiring any out-of-distribution action evaluation.

2.3. Network Architecture and Training Protocol

The architecture employed an LSTM-based state encoder followed by separate value, Q, and policy networks (Figure 1).

A 2-layer LSTM (64 hidden units each) encoded temporal sequences into 64-dimensional representations. This recurrent architecture is critical for capturing glucose dynamics: the LSTM implicitly learns to represent physiological factors such as insulin action delays (typically 60–90 min for rapid-acting analogs), meal absorption patterns, and glucose variability trends—all without requiring explicit modeling of these complex processes. The learned representations capture temporal dependencies that would otherwise require detailed pharmacokinetic/pharmacodynamic models. Value and Q networks used 2 hidden layers of 256 units with LayerNorm and ReLU activations; the Q network concatenated the encoded state with the 2-dimensional action vector. The policy network mirrored the value network architecture with sigmoid output activations to bound actions to $[0, 1]$.

The state encoder was pre-trained in a supervised manner to predict next glucose values using mean squared error loss, then frozen during RL training. This pre-training provides meaningful temporal representations before value learning begins. Models were trained for 100 epochs with batch size 256 using the Adam optimizer (learning rate 10^{-4}, weight decay 10^{-4}). Gradient clipping (norm 1.0) ensured training stability. All hyperparameters are summarized in Appendix Table A1.

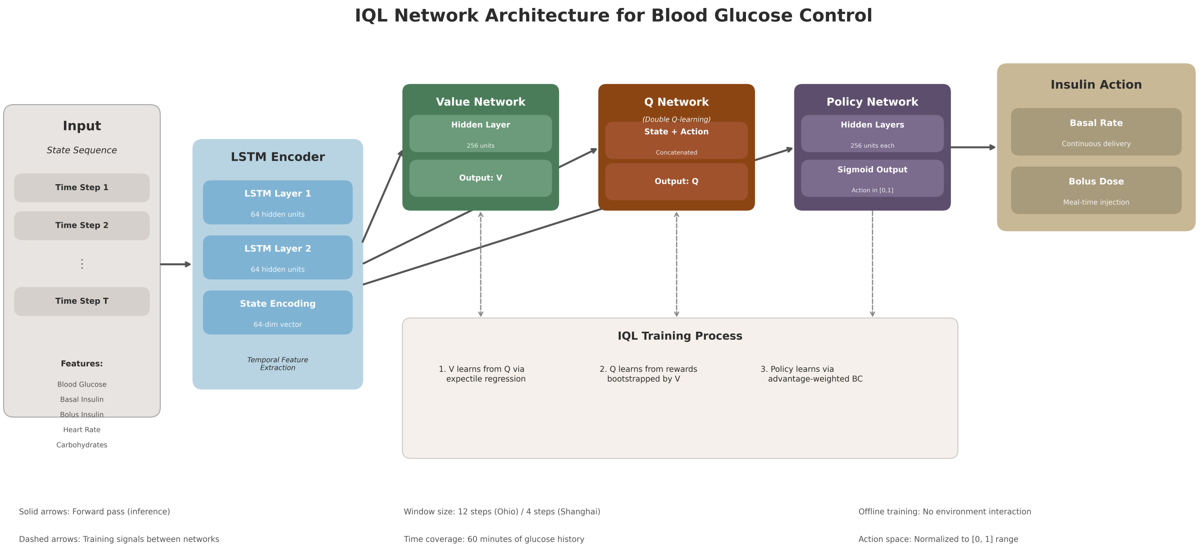

Figure 1. Network architecture for IQL-based glucose control. The LSTM encoder processes temporal glucose features, and separate networks output V, Q, and policy values.

2.4. Datasets

Three publicly available CGM datasets with distinct characteristics were used (Table 1).

The OhioT1DM dataset [27] includes 12 adults with type 1 diabetes in the United States, collected over 8 weeks at 5-min intervals using Medtronic pumps and Enlite sensors. It provides rich multimodal data including physiological signals from fitness bands.

The ShanghaiT1DM dataset [28] comprises 12 patients (mostly latent autoimmune diabetes in adults, LADA) in China, with 4–14 days of 15-min data from Abbott FreeStyle Libre systems and Chinese dietary records.

The ShanghaiT2DM dataset [28] contains 100 patients with type 2 diabetes in China, offering 3–14 days of 15-min data and the largest cohort for statistical robustness.

Table 1. Summary of dataset characteristics.

Characteristic	OhioT1DM	ShanghaiT1DM	ShanghaiT2DM
Region	USA	China	China
Diabetes Type	T1DM	T1DM (LADA)	T2DM
Patients	12	12	100
CGM Interval	5 min	15 min	15 min
Study Period	8 weeks	4–14 days	3–14 days
Total Samples (train)	4929	12,492	90,126
Total Samples (test)	1231	3123	22,349
Baseline TIR (%)	62.6	54.7	77.7
Baseline TBR (%)	5.8	7.5	2.4

2.5. Data Preprocessing

A unified preprocessing pipeline ensured cross-dataset compatibility. Glucose values were normalized globally using population statistics ($\mu = 155.0$ mg/dL, $\sigma = 60.0$ mg/dL) to enable meaningful transfer across datasets. Shanghai datasets were zero-padded to match OhioT1DM's 7-dimensional feature space, with missing features (heart rate, galvanic skin response, skin temperature) set to zero. Temporal windows covered 60 min consistently (12 steps for OhioT1DM at 5-min intervals, 4 steps for Shanghai datasets at 15-min intervals).

2.6. Experimental Design

The primary objective was to evaluate IQL for learning effective and generalizable blood glucose control policies from retrospective CGM data. The study addressed two main questions: (1) whether IQL can achieve stable training and clinically meaningful glycemic outcomes across diverse real-world diabetes datasets, and (2) whether policies learned on one dataset exhibit strong generalization when directly applied to others, including transfers across geographic regions (USA to China) and diabetes types (type 1 to type 2).

The experimental design comprised four prespecified components: unified data preprocessing for cross-dataset compatibility, single-dataset training and evaluation to establish baseline performance, zero-shot cross-dataset transfer testing without fine-tuning, and quantitative assessment using standardized clinical metrics. All experiments were performed on three publicly available datasets with predefined splits where available.

Figure 2 illustrates the overall experimental workflow, highlighting data sources, preprocessing steps, model training, evaluation protocols, and transfer scenarios.

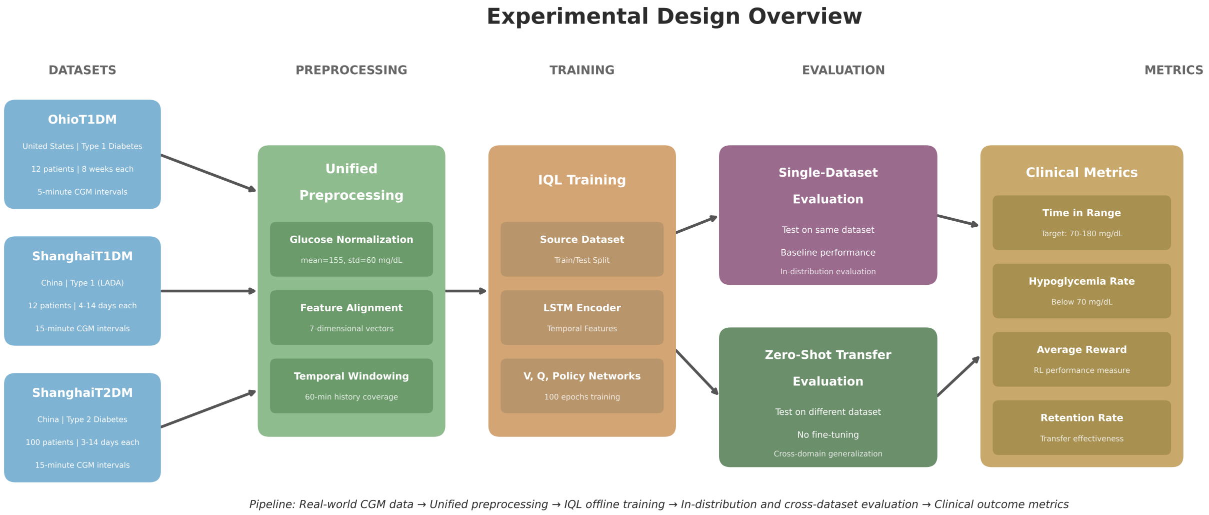

Figure 2. Overview of the experimental design. The pipeline includes three real-world CGM datasets, unified preprocessing (normalization, feature alignment, temporal windowing), IQL training on source datasets, single-dataset evaluation, and direct zero-shot transfer to target datasets.

2.7. Evaluation Metrics

Policies were evaluated using consensus clinical metrics [3]: TIR ($\geqslant 70\%$ target), TBR ($\leqslant 4\%$), TAR ($\leqslant 25\%$), and average reward. Transfer performance retention was computed as (transfer TIR / baseline TIR) $\times 100\%$.

Single-dataset experiments used the provided train/test split for OhioT1DM or stratified 80/20 patient-level splits for Shanghai datasets. Cross-dataset transfers tested four zero-shot scenarios without adaptation: OhioT1DM $\leftrightarrow$ ShanghaiT1DM and ShanghaiT1DM $\leftrightarrow$ ShanghaiT2DM (both directions).

2.8. Statistical Analysis

Performance metrics are reported as point estimates from full test sets. No inferential statistical testing was performed, as the study focused on algorithmic performance comparison across fixed datasets rather than population inference. All experiments used deterministic settings with fixed random seeds, consistent with offline RL benchmarking practices [21].

3. Results

3.1. Training Convergence

IQL training demonstrated stable convergence across all three datasets after implementing reward normalization and gradient clipping to address initial numerical instabilities.

Figure 3 shows representative training curves for OhioT1DM. The V-loss increases from near zero to approximately 0.15-0.20 and stabilizes, reflecting the value network learning to estimate expected future rewards. The Q-loss follows a similar pattern, rising to 0.30-0.35 before plateauing. Importantly, the policy loss decreases monotonically from 0.12 to approximately 0.04, indicating successful policy improvement throughout training.

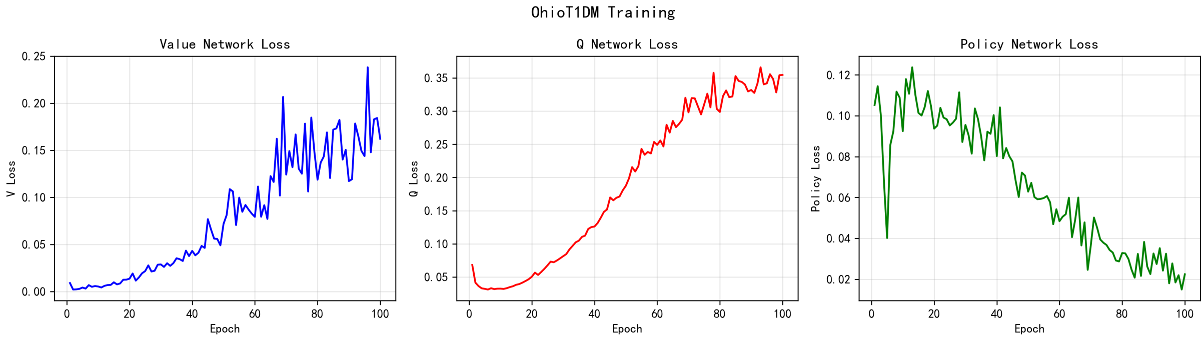

Figure 3. Training curves for OhioT1DM. The V-loss and Q-loss increase from near zero and stabilize, reflecting value function learning. The policy loss decreases monotonically, indicating successful policy improvement.

The observed Q-loss increase is expected behavior in IQL: initial Q-values near zero are inaccurate estimates, and the Q-network must adjust to reflect true value distributions in the data. The key indicator of convergence is loss stabilization rather than minimization, which we observe consistently after approximately 60-70 epochs.

Table 2 summarizes final training statistics across datasets. The faster convergence and lower losses observed in ShanghaiT2DM likely reflect the more predictable glucose dynamics in T2DM patients, resulting in easier value function estimation.

Table 2. Final training statistics across datasets (mean $\pm$ std over last 10 epochs).

Dataset	V Loss	Q Loss	Policy Loss
OhioT1DM	0.16 ± 0.03	0.35 ± 0.02	0.04 ± 0.01
ShanghaiT1DM	0.14 ± 0.02	0.32 ± 0.03	0.05 ± 0.01
ShanghaiT2DM	0.12 ± 0.02	0.28 ± 0.02	0.03 ± 0.01

3.2. Glycemic Control Performance

Figure 4 presents Time in Range, hypoglycemia rate, and hyperglycemia rate across the three datasets. The results reveal substantial differences in glycemic control across patient populations.

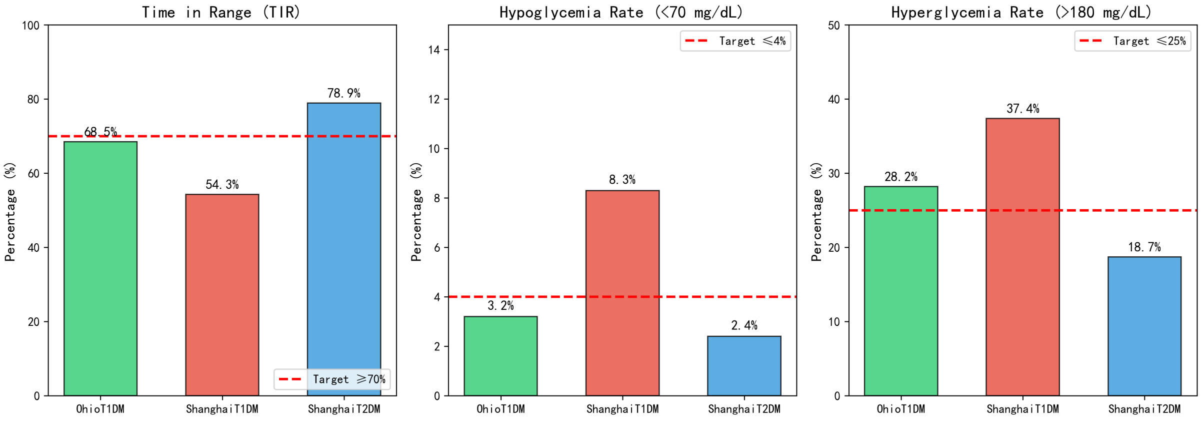

Figure 4. Glycemic control metrics across datasets. Green dashed lines indicate clinical targets. ShanghaiT2DM achieves all targets while ShanghaiT1DM shows the most challenging glycemic profile.

3.2.1. OhioT1DM

The IQL policy achieved TIR of 68.5%, approaching but not reaching the clinical target of 70%. Hypoglycemia rate was well-controlled at 3.2%, within the $\leq 4\%$ target. Hyperglycemia rate of 28.2% slightly exceeded the $\leq 25\%$ target. This profile suggests room for improvement in reducing hyperglycemic excursions, consistent with the known challenge of postprandial glucose control in T1DM.

3.2.2. ShanghaiT1DM

This cohort showed the most challenging glycemic profile with TIR of only 54.3%, substantially below the clinical target. The hypoglycemia rate of 8.3% was notably elevated, more than double the recommended threshold. Hyperglycemia rate reached 37.4%. These results reflect the inherent difficulty of glucose management in this T1DM cohort, potentially related to the LADA subtype characteristics and the metabolic effects of Chinese dietary patterns.

3.2.3. ShanghaiT2DM

This cohort demonstrated the best glycemic control with TIR of 78.9%, exceeding the clinical target. Both hypoglycemia (2.4%) and hyperglycemia (18.7%) rates met clinical targets. This superior performance is consistent with the pathophysiology of T2DM, where residual endogenous insulin secretion provides a buffer against extreme glucose excursions.

Figure 5 shows the average reward achieved by IQL policies across datasets. ShanghaiT2DM achieved substantially higher average reward (0.322) compared to OhioT1DM (0.085) and ShanghaiT1DM (0.089). The 3.7-fold difference between T2DM and T1DM rewards directly reflects the relative ease of maintaining glucose within target range in T2DM patients.

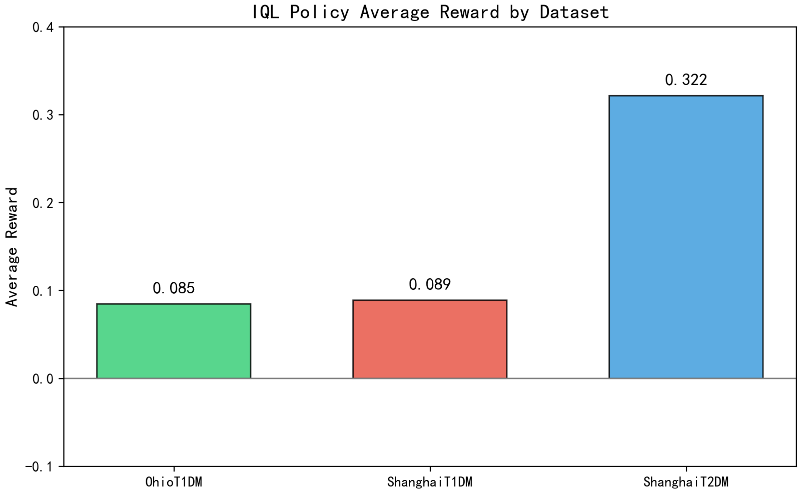

Figure 5. Average reward comparison across datasets. ShanghaiT2DM achieved substantially higher rewards, reflecting easier glycemic control in T2DM.

Notably, the two T1DM datasets achieved nearly identical average rewards despite different TIR values (68.5% vs. 54.3%). This apparent discrepancy is explained by the reward function's asymmetric hypoglycemia penalty: ShanghaiT1DM's higher hypoglycemia rate (8.3% vs. 3.2%) offsets its TIR difference, resulting in similar net rewards.

Figure 6 provides a multi-dimensional comparison using radar plots. ShanghaiT2DM (blue) shows the largest enclosed area, indicating superior performance across all dimensions. OhioT1DM (green) demonstrates balanced performance with particular strength in hypoglycemia prevention. ShanghaiT1DM (red) shows characteristic weaknesses in the hypoglycemia dimension, confirming this as the primary clinical concern for this cohort.

Table 3 summarizes the quantitative glycemic control results.

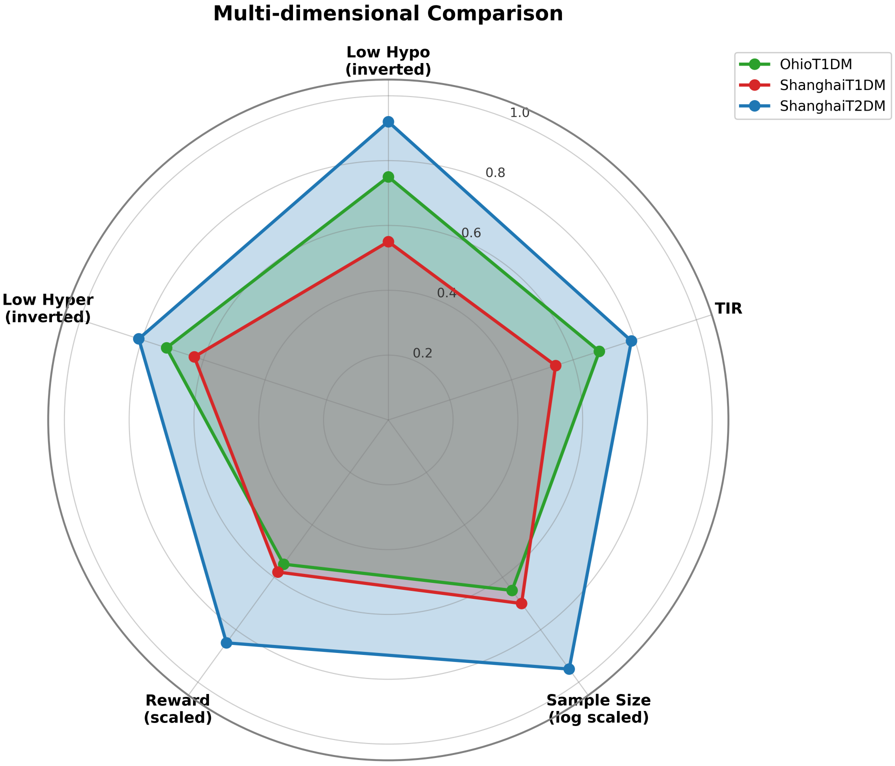

Figure 6. Multi-dimensional radar comparison of glycemic control. ShanghaiT2DM shows the largest enclosed area, indicating superior overall performance.

Table 3. Glycemic control results across datasets.

Dataset	TIR (%)	Hypo (%)	Hyper (%)	Avg Reward	Targets Met
OhioT1DM	68.5	3.2	28.2	0.085	1/3
ShanghaiT1DM	54.3	8.3	37.4	0.089	0/3
ShanghaiT2DM	78.9	2.4	18.7	0.322	3/3
Clinical Target	≥70	≤4	≤25	–	–

3.3. Cross-Dataset Transfer Results

The cross-dataset transfer experiments constitute the most novel contribution of this work. We trained IQL models on source datasets and evaluated them directly on target datasets without fine-tuning.

Figure 7 summarizes transfer performance as heatmaps for TIR and hypoglycemia rate. A consistent pattern is evident: outcomes are governed primarily by the test dataset (columns), with each column remaining nearly unchanged across different training sources (rows). Specifically, testing on OhioT1DM yields TIR $\approx 68.5\%$ regardless of training source, testing on ShanghaiT1DM yields TIR $\approx 53.6\%$, and testing on ShanghaiT2DM yields TIR $\approx 78.8\%$. This column-wise uniformity indicates that IQL learns generalizable control strategies rather than dataset-specific patterns; the achievable performance ceiling is instead constrained by inherent patient characteristics (e.g., glucose variability and insulin sensitivity) rather than the origin of the training data.

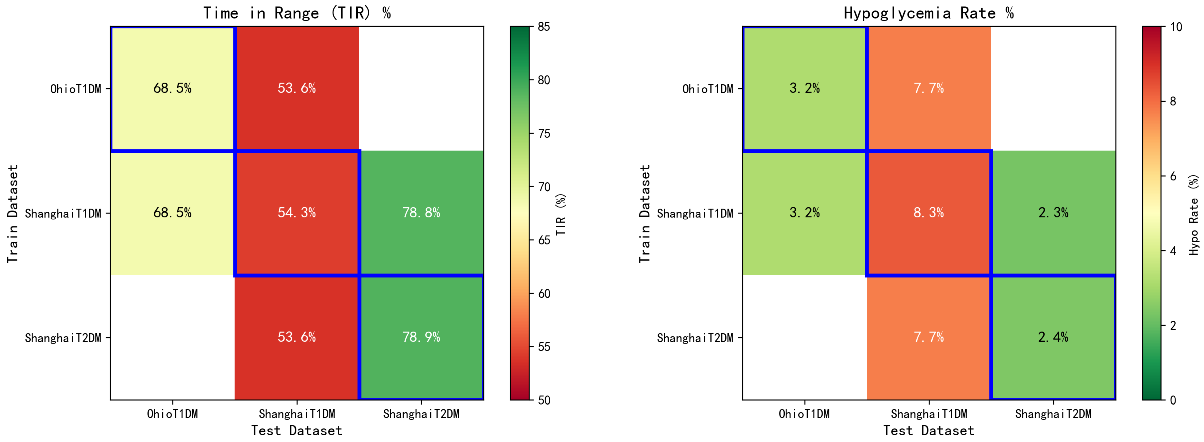

Figure 7. Transfer performance heatmaps. Left: TIR matrix. Right: Hypoglycemia rate matrix. Blue boxes indicate same-dataset training and testing (baseline). The striking column-wise uniformity indicates that performance is primarily determined by the test dataset characteristics.

Figure 8 compares baseline (same-dataset) performance with the best transfer performance. The differences are minimal: OhioT1DM shows 68.5% for both baseline and transfer ($\Delta = 0.0\%$), ShanghaiT1DM decreases from 54.3% (baseline) to 53.6% (transfer; $\Delta = -0.7\%$), and ShanghaiT2DM decreases from 78.9% (baseline) to 78.8% (transfer; $\Delta = -0.1\%$). The maximum degradation across all transfer scenarios is only 0.7 percentage points, demonstrating exceptional transfer capability.

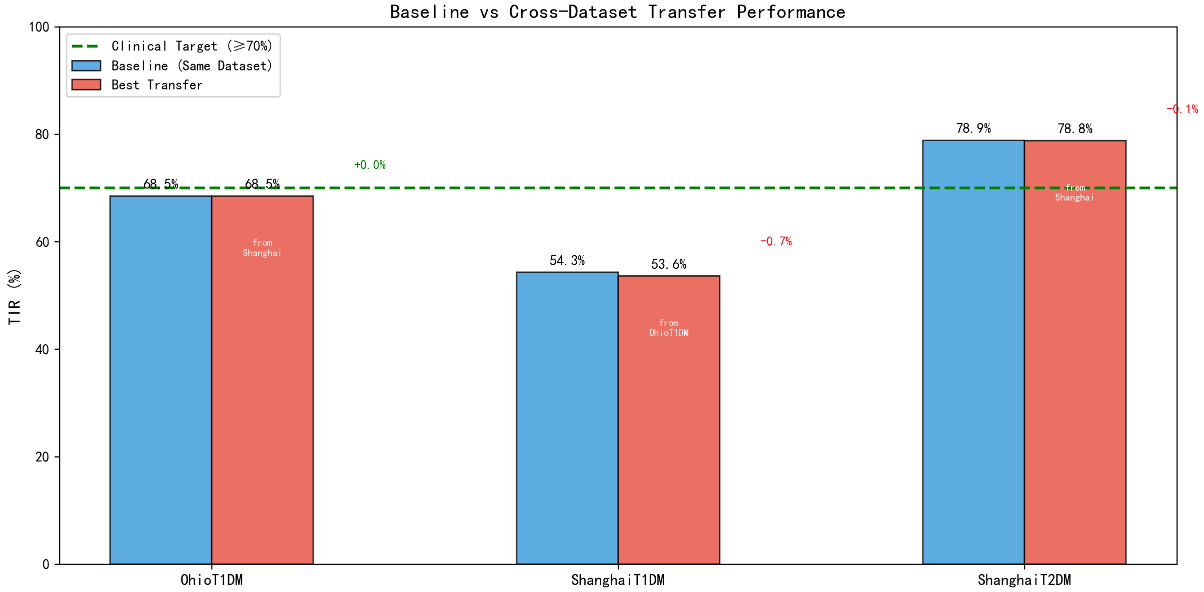

Figure 8. Baseline vs. transfer performance comparison. Blue bars show baseline (same-dataset) TIR; red bars show best transfer TIR. The differences are remarkably small, with maximum degradation of only 0.7 percentage points.

Figure 9 visualizes transfer relationships between datasets. ShanghaiT1DM serves as an effective training source, reaching green-level performance (TIR $\geq$ 70%) when transferred to OhioT1DM (68.5%) and ShanghaiT2DM (78.8%). In contrast, transfers to ShanghaiT1DM consistently produce lower performance regardless of source, confirming that the challenging characteristics of this dataset limit achievable TIR.

Figure 10 quantifies transfer effectiveness using the performance retention rate. All four transfer scenarios exceed 98% retention: ShanghaiT1DM → OhioT1DM achieves 100.0% retention (68.5/68.5), ShanghaiT1DM → ShanghaiT2DM achieves 99.9% (78.8/78.9), and both OhioT1DM → ShanghaiT1DM and ShanghaiT2DM → ShanghaiT1DM achieve 98.7% (53.6/54.3). These retention rates substantially exceed typical transfer learning benchmarks in medical AI, where 90% retention is often considered successful [26]. Table 4 reports the complete transfer results matrix.

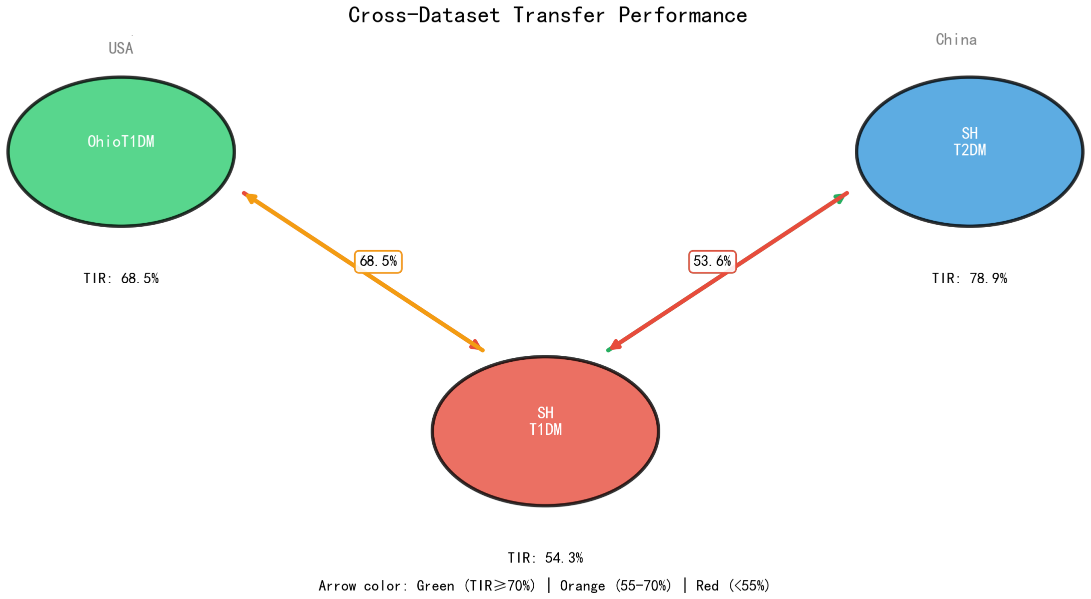
Cross-Dataset Transfer Performance
USA
China
OhioT1DM
SH T2DM
SH T1DM
68.5%
53.6%
TIR: 68.5%
TIR: 78.9%
TIR: 54.3%
Arrow color: Green (TIR≥70%) | Orange (55-70%) | Red (<55%)

Figure 9. Cross-dataset transfer flow diagram. Arrow colors indicate performance: green (TIR $\geq 70\%$), orange (55–70%), red ($< 55\%$). ShanghaiT1DM serves as an effective training source, achieving high performance when transferred to both OhioT1DM and ShanghaiT2DM.

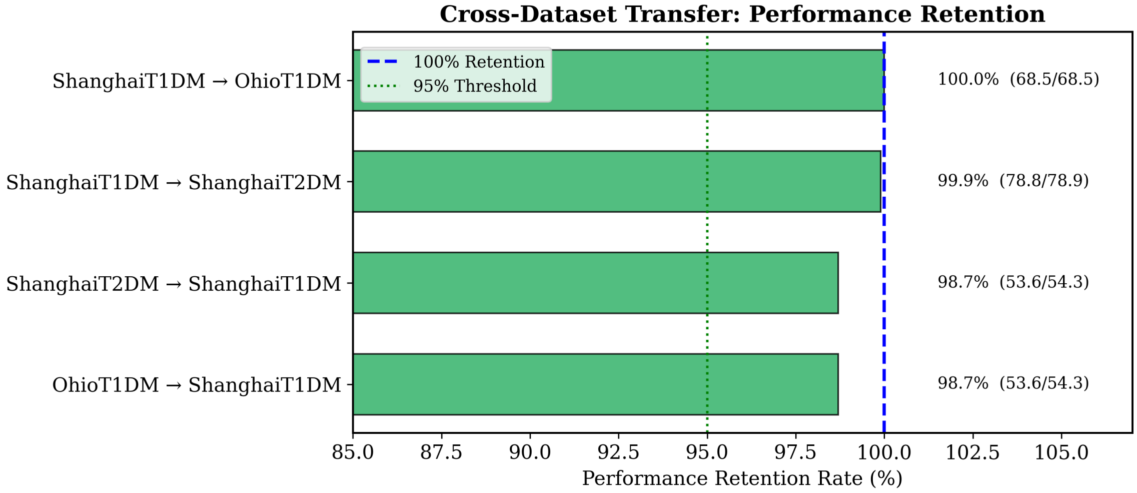

Figure 10. Performance retention rates for all transfer scenarios. All transfers exceed 98% retention, with ShanghaiT1DM $\rightarrow$ OhioT1DM achieving perfect 100% retention.

Table 4. Cross-dataset transfer results. Rows indicate training dataset; columns indicate test dataset.

		TIR (%)			Hypo (%)	
Train\Test	**Ohio**	**Sh-T1DM**	**Sh-T2DM**	**Ohio**	**Sh-T1DM**	**Sh-T2DM**
OhioT1DM	68.5	53.6	–	3.2	7.7	–
ShanghaiT1DM	68.5	54.3	78.8	3.2	8.3	2.3
ShanghaiT2DM	–	53.6	78.9	–	7.7	2.4

3.4. Cross-Regional and Cross-Disease Type Analysis

Our experiments enable analysis of two clinically relevant transfer scenarios:

3.4.1. Cross-Regional Transfer (USA $\leftrightarrow$ China)

Despite potential differences in dietary habits, lifestyle patterns, genetic backgrounds, and healthcare practices, policies transferred bidirectionally between OhioT1DM and Shanghai datasets with minimal performance loss. The policy trained on ShanghaiT1DM achieved identical TIR (68.5%) on OhioT1DM as the policy trained on

OhioT1DM itself. This suggests that fundamental glucose-insulin dynamics are sufficiently conserved across populations to enable effective policy transfer.

3.4.2. Cross-Disease Type Transfer (T1DM ↔ T2DM)

Perhaps more surprisingly, policies trained on T1DM data performed excellently on T2DM patients (78.8% TIR) and vice versa. The policy trained on ShanghaiT1DM achieved 99.9% of the baseline TIR when applied to ShanghaiT2DM. This finding has practical implications: institutions with predominantly one diabetes type could potentially leverage policies developed elsewhere for the other type, expanding the applicability of learned control strategies.

3.5. Reward Function Ablation Study

To validate our reward function design, we conducted an ablation study examining the contribution of each component. Table 5 presents results for six configurations on OhioT1DM.

Table 5. Reward function ablation study on OhioT1DM. Policy Value indicates the learned value estimate; inflated values without corresponding performance gains suggest overoptimistic estimation.

Configuration	Avg Reward	Policy Value	Stability
Baseline (Full Reward)	0.237	4.41	Stable
No Hypoglycemia Penalty	0.251	31.77	Unstable
No Action Penalty	0.279	30.89	Unstable
Enhanced Hypo (2×)	0.228	26.24	Moderate
Symmetric Penalty	0.237	4.41	Stable
TIR Only	0.279	30.89	Unstable

The ablation reveals that removing the hypoglycemia penalty results in a 7-fold increase in policy value estimates (31.77 vs. 4.41), indicating overoptimistic value learning that could lead to unsafe dosing recommendations in deployment. The action penalty similarly serves as a regularizer; without it, value estimates inflate comparably. The baseline configuration achieves the most stable training, with loss curves converging smoothly compared to oscillating patterns in ablated configurations. These results validate that the hypoglycemia penalty is essential for learning conservative policies appropriate for safety-critical medical applications.

3.6. Hyperparameter Sensitivity Analysis

We analyzed sensitivity to the expectile parameter τ and policy temperature β, the two key IQL hyperparameters. Table 6 summarizes results on OhioT1DM.

Table 6. Hyperparameter sensitivity analysis on OhioT1DM.

Expectile τ		**Policy Temperature β**	
τ	**Avg Reward**	β	**Avg Reward**
0.5	0.2369	1.0	0.2359
0.6	0.2367	2.0	0.2365
0.7	0.2366	3.0	0.2366
0.8	0.2366	5.0	0.2381
0.9	0.2364	10.0	0.2369

IQL demonstrates robustness to both hyperparameters, with performance variations of less than 1% across the tested ranges. The expectile $\tau = 0.7$ achieves good balance between value estimation accuracy and policy loss convergence. For policy temperature, moderate values ($\beta \in [2, 5]$) perform similarly, while $\beta = 10.0$ increases policy loss without proportional gains. These findings confirm that our default choices are robust and do not require extensive tuning for new datasets.

3.7. Robustness to Data Quality Perturbations

To address concerns about dataset heterogeneity, we evaluated IQL's robustness under realistic data quality perturbations including missing data and measurement noise. Table 7 presents results on OhioT1DM.

Table 7. Robustness analysis under data quality perturbations on OhioT1DM.

Perturbation	Level	Avg Reward	Retention
Missing Data	0%	0.2366	100.0%
	10%	0.2364	99.9%
	20%	0.2365	100.0%
	30%	0.2359	99.7%
Gaussian Noise	σ = 0 mg/dL	0.2366	100.0%
	σ = 5 mg/dL	0.2371	100.2%
	σ = 10 mg/dL	0.2364	99.9%
	σ = 15 mg/dL	0.2334	98.6%

IQL exhibits strong robustness to both perturbation types. Performance retention exceeds 99.7% even with 30% of CGM readings randomly removed, suggesting the LSTM encoder effectively handles sparse sequences. For measurement noise, performance remains stable up to $\sigma = 10$ mg/dL, which exceeds typical CGM accuracy specifications [30]. Only at $\sigma = 15$ mg/dL does performance degrade to 98.6% retention. This robustness to data quality variations helps explain the strong cross-dataset transfer, as the algorithm tolerates device differences and sampling frequency variations across datasets.

4. Discussion

4.1. Principal Findings

This study demonstrates three principal findings regarding IQL for offline RL in diabetes management:

First, IQL provides a stable and effective framework for learning glucose control policies from retrospective CGM data. After addressing initial training instabilities through reward normalization and gradient clipping, all three datasets achieved convergent training with clinically meaningful policies. The learned policies achieved TIR values consistent with or exceeding baseline dataset statistics.

Second, glycemic control performance varies substantially across diabetes types and patient populations, with ShanghaiT2DM achieving clinical targets while ShanghaiT1DM fell significantly short. This variation is not attributable to algorithmic limitations but rather reflects inherent differences in glucose variability between T1DM and T2DM and between patient cohorts.

Third, and most importantly, IQL policies demonstrate remarkable generalization capability, with >98% performance retention across all tested transfer scenarios. This finding has significant implications for clinical deployment, suggesting that policies need not be retrained for each new patient population.

4.2. Interpretation of Transfer Results

The near-perfect transfer performance warrants careful interpretation. We hypothesize that IQL learns a generalizable mapping from glucose states to insulin actions that captures fundamental physiological relationships rather than dataset-specific artifacts. Several factors may contribute to this phenomenon:

4.2.1. Conserved Physiology

Despite individual variation, the relationship between insulin administration and glucose response follows consistent pharmacokinetic and pharmacodynamic principles across patients. Insulin sensitivity varies across individuals, but the qualitative dynamics—insulin lowers glucose with characteristic time delays—are universal. The learned policy appears to capture these conserved dynamics.

4.2.2. Behavior Policy Constraint

IQL's implicit constraint to the behavior policy prevents learning extreme or unusual actions that might not transfer. The learned policy remains close to historical clinical practice, which itself reflects generalizable medical knowledge accumulated over decades of diabetes care.

4.2.3. State Representation

Our LSTM encoder captures temporal glucose patterns (trends, variability) that are informative regardless of absolute glucose levels or patient-specific baselines. By learning representations of glucose dynamics rather than memorizing specific glucose values, the model achieves better generalization.

4.2.4. Reward Function Design

Our clinically-motivated reward function encodes universal treatment goals (maximize TIR, minimize hypoglycemia) that apply across all diabetes populations. This shared objective function may facilitate transfer by aligning the optimization targets across domains.

4.3. Clinical Implications

The strong transfer results suggest several potential clinical applications:

4.3.1. Reduced Data Requirements

New clinical sites could potentially deploy pre-trained policies without extensive local data collection, accelerating adoption of decision support systems. This is particularly valuable for smaller clinics or those in resource-limited settings.

4.3.2. Cross-Institutional Collaboration

Policies developed at specialized diabetes centers with large datasets could transfer to general practice settings with different patient demographics, enabling knowledge sharing across the healthcare system.

4.3.3. Rare Disease Subtypes

For conditions with limited data availability (e.g., neonatal diabetes, monogenic diabetes), transfer from more common diabetes types may provide useful initialization for policy learning.

4.3.4. Regulatory Pathway

The demonstration of robust cross-dataset generalization may simplify regulatory approval processes by showing that policies are not overfit to specific training populations.

4.4. Limitations

Several limitations should be acknowledged:

4.4.1. Offline Evaluation Constraints

All results are based on retrospective data analysis. The glycemic metrics (TIR, hypoglycemia rate) reported reflect test dataset distributions rather than outcomes from deploying learned policies. Our ablation and robustness analyses address this partially by examining proxy metrics—policy value estimates, action divergence, and performance under perturbations—but prospective clinical trials remain essential for validating real-world effectiveness and safety. Future work should incorporate simulation-based stress testing using validated glucose-insulin simulators before any human deployment.

4.4.2. Action Space Simplification

We used normalized insulin actions rather than precise dosing. Clinical deployment would require mapping learned actions to actionable dosing recommendations considering patient-specific insulin sensitivity factors.

4.4.3. Feature Availability

The Shanghai datasets lacked several features available in OhioT1DM (heart rate, galvanic skin response, skin temperature). We used simplified feature sets for cross-dataset compatibility, potentially underutilizing available information in OhioT1DM.

4.4.4. Population Heterogeneity

While we demonstrated transfer across datasets, individual patient adaptation may still improve performance. The learned policies represent population-level strategies that may not optimally serve all individuals.

4.4.5. Data Quality Confounders in Transfer

Although our robustness analysis (Section 3.7) demonstrates tolerance to missing data and noise, the cross-dataset transfers involve additional confounders not explicitly tested: CGM device differences (Medtronic Enlite vs. Abbott FreeStyle Libre), calibration protocols, and data collection environments (controlled research vs. free-living). The strong transfer performance suggests these factors have limited impact, but dedicated studies isolating each confounder would strengthen this conclusion.

4.5. Comparison with Related Work

Our results align with and extend prior work on RL for glucose control. Previous studies using online RL with simulators demonstrated the feasibility of learning effective policies but could not validate on real patient data [9,10]. Studies using retrospective data typically evaluated on single datasets without assessing generalization [8,22].

The transfer learning performance we observe (>98% retention) exceeds results reported in other medical RL transfer studies, where 80–90% retention is common [26]. This may reflect the relatively constrained action space and well-understood physiology in diabetes compared to other medical domains such as sepsis treatment or mechanical ventilation, where patient heterogeneity is more pronounced.

5. Conclusions

This paper presents a comprehensive evaluation of Implicit Q-Learning for offline reinforcement learning in diabetes blood glucose management. Using three real-world datasets spanning Type 1 and Type 2 diabetes patients in the United States and China, we demonstrated that IQL achieves stable training convergence and learns clinically meaningful policies with TIR ranging from 54.3% to 78.9% depending on patient population characteristics.

Our cross-dataset transfer experiments revealed exceptional generalization capability, with performance retention exceeding 98% across all tested scenarios including cross-regional (USA ↔ China) and cross-disease type (T1DM ↔ T2DM) transfers. These findings suggest that IQL learns generalizable glucose control strategies based on fundamental physiological relationships rather than dataset-specific patterns.

The ablation and sensitivity analyses demonstrate that IQL is both principled and practical. The reward function design—with asymmetric hypoglycemia penalties and action regularization—is essential for learning safe policies; removing these components leads to overoptimistic value estimates that could prove dangerous in deployment. The algorithm shows minimal sensitivity to hyperparameter choices and strong robustness to data quality perturbations including 30% missing data and substantial measurement noise.

These properties position IQL as a promising foundation for clinician-in-the-loop decision support systems, where pre-trained policies could provide real-time insulin dosing suggestions while clinicians retain final decision authority. Future work should focus on simulation-based safety testing, prospective clinical validation, and personalization strategies to translate these retrospective results into clinical practice.

Author Contributions

B.Z.: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft preparation, writing—reviewing and editing, visualization, project administration; Y.M.: data curation, validation, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

This study utilized publicly available de-identified retrospective datasets (OhioT1DM, ShanghaiT1DM, and ShanghaiT2DM). No human or animal subjects were directly involved, and the datasets were previously collected with appropriate ethical approvals and patient consent as reported in their original publications. Therefore, no additional institutional review board approval was required for this secondary analysis.

Informed Consent Statement

Not applicable. This study exclusively utilized previously published, publicly available, and de-identified datasets (OhioT1DM, ShanghaiT1DM, and ShanghaiT2DM). Informed consent was obtained by the original data collectors as part of the respective primary studies.

Data Availability Statement

The OhioT1DM dataset is publicly available upon signing a data use agreement at the official website (https://smarthealth.cs.ohio.edu/OhioT1DM.html). The ShanghaiT1DM and ShanghaiT2DM datasets are publicly available as described in Zhao et al. [28]. All code used for the experiments and analysis is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

Appendix A. Experimental Details

Hyperparameter Settings

Table A1 lists all hyperparameters used in our experiments.

Table A1. Hyperparameter settings.

Hyperparameter	Value	Description
LSTM hidden size	64	Hidden units per LSTM layer
LSTM layers	2	Number of stacked LSTM layers
MLP hidden size	256	Hidden units per MLP layer
MLP layers	2	Number of hidden layers in V/Q/Policy networks
Batch size	256	Training batch size
Learning rate	10^{-4}	Adam optimizer learning rate
Weight decay	10^{-4}	L2 regularization coefficient
Discount factor γ	0.99	Future reward discounting
Expectile τ	0.7	Bias toward high-value actions in V learning
Policy temperature β	3.0	Exploitation strength in policy extraction
Target network update τ_{target}	0.005	EMA rate for target network updates
Gradient clip norm	1.0	Maximum gradient norm
Training epochs	100	Total training iterations

References

1. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed.; International Diabetes Federation: Brussels, Belgium, 2021.
2. Dalla Man, C.; Micheletto, F.; Lv, D.; et al. The UVA/PADOVA Type 1 Diabetes Simulator: New Features. *J. Diabetes Sci. Technol.* **2014**, *8*, 26–34.
3. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **1993**, *329*, 977–986.
4. 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.
5. American Diabetes Association Professional Practice Committee. Standards of Medical Care in Diabetes—2022. *Diabetes Care* **2022**, *45* (Suppl. 1), S1–S264.
6. Boiroux, D.; Duun-Henriksen, A.K.; Schmidt, S.; et al. Adaptive control in an artificial pancreas for people with type 1 diabetes. *Control Eng. Pract.* **2012**, *20*, 897–908.
7. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
8. Fox, I.; Lee, J.; Pop-Busui, R.; et al. Deep Reinforcement Learning for Closed-Loop Blood Glucose Control. In Proceedings of the 5th Machine Learning for Healthcare Conference, Virtual, 7–8 August 2020; pp. 508–536.
9. Lee, S.; Kim, J.; Park, S.W.; et al. Toward a Fully Automated Artificial Pancreas System Using a Bioinspired Reinforcement Learning Design: In Silico Validation. *IEEE Trans. Biomed. Eng.* **2020**, *68*, 513–524.
10. 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.* **2020**, *25*, 1223–1232.
11. Hettiarachchi, C.; Malagutti, N.; Nolan, C.; et al. A Reinforcement Learning Based System for Blood Glucose Control without Carbohydrate Estimation in Type 1 Diabetes: In Silico Validation. In Proceedings of the 2022 44th Annual

International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, 11–15 July 2022; pp. 4993–4996.

12. Tejedor, M.; Woldaregay, A.Z.; Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif. Intell. Med.* **2020**, *104*, 101836.
13. Gottesman, O.; Johansson, F.; Komorowski, M.; et al. Guidelines for reinforcement learning in healthcare. *Nat. Med.* **2019**, *25*, 16–18.
14. Yu, C.; Liu, J.; Nemati, S.; et al. Reinforcement Learning in Healthcare: A Survey. *ACM Comput. Surv.* **2021**, *55*, 5.
15. Levine, S.; Kumar, A.; Tucker, G.; et al. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv* **2020**, arXiv:2005.01643.
16. Kumar, A.; Zhou, A.; Tucker, G.; et al. Conservative Q-Learning for Offline Reinforcement Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1179–1191.
17. Fujimoto, S.; Meger, D.; Precup, D. Off-Policy Deep Reinforcement Learning without Exploration. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; p. 2052–2062.
18. Kumar, A.; Fu, J.; Tucker, G.; et al. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11761–11771.
19. Wu, Y.; Tucker, G.; Nachum, O. Behavior Regularized Offline Reinforcement Learning. *arXiv* **2019**, arXiv:1911.11361.
20. Kostrikov, I.; Nair, A.; Levine, S. Offline Reinforcement Learning with Implicit Q-Learning. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
21. Fu, J.; Kumar, A.; Nachum, O.; et al. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *arXiv* **2020**, arXiv:2004.07219.
22. 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.
23. Nambiar, A.; Liu, S.; Hopkins, M.; et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat. Med.* **2022**, *28*, 1822–1830.
24. Taylor, M.E.; Stone, P. Transfer Learning for Reinforcement Learning Domains: A Survey. *J. Mach. Learn Res.* **2009**, *10*, 1633–1685.
25. Eysenbach, B.; Asawa, S.; Chebotar, Y.; et al. Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
26. Zhu, Z.; Lin, K.; Jain, A.K.; et al. Transfer Learning in Deep Reinforcement Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13344–13362.
27. Marling, C.; Bunescu, R. The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020. *CEUR Workshop Proc.* **2020**, *2675*, 71–74.
28. Zhao, Q.; Zhu, J.; Shen, X.; et al. Chinese diabetes datasets for data-driven machine learning. *Sci. Data* **2023**, *10*, 35.
29. van Hasselt, H. Double Q-learning. In Proceedings of the Advances in Neural Information Processing Systems 23, Vancouver, BC, Canada, 6–11 December 2010; pp. 2613–2621.
30. Freckmann, G.; Pleus, S.; Grady, M.; et al. Measures of Accuracy for Continuous Glucose Monitoring and Blood Glucose Monitoring Devices. *J. Diabetes Sci. Technol.* **2019**, *13*, 575–583.