AIRC: ACTION-INNOVATION RESPONSE CERTIFICATES FOR RISK-AWARE OFFLINE POLICY IMPROVEMENT IN BLOOD GLUCOSE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline insulin-control learning must infer which actions are supported by retrospective treatment data, which direction improves control, and when an increase becomes unsafe. We introduce Action-Innovation Response Certificates (AIRC), an evidence architecture that converts state-relative action innovations into signed response certificates. Agreement across complementary data folds and clinically relevant prediction horizons determines the direction and confidence of policy updates, while a separate short-horizon pathway estimates the risk attached to positive insulin shifts. In SimGlucose, AIRC achieves the highest mean cumulative reward among six offline reinforcement learning baselines and improves time in range by 4.86% relative to the strongest reward baseline, IQL, while reducing the glycemia risk index by 10.77%. At the 240-minute analysis horizon, 54.67% of eligible AIRC events meet the postprandial target-range criterion, 10.2 percentage points more than with IQL. Component ablations separate the roles of response direction, asymmetric risk weighting, and innovation scaling. On matched held-out states from the public Loop dataset, AIRC improves trend–action correlation by 0.139 and directional consistency by 4.51 percentage points over IQL. Together, these results establish AIRC as a response-certified framework for risk-aware offline policy improvement in basal-insulin control. The source code and datasets are submitted as Supplementary Materials for Reproducibility.

1 INTRODUCTION

Improving insulin delivery from recorded treatment is directional control under asymmetric physiological risk. Delivery should rise when glucose pressure indicates insufficient insulin and recede when an increase would amplify hypoglycemia. Automated insulin delivery has improved glucose regulation in recent clinical trials and everyday use (Kovatchev et al., 2025; Wilmot et al., 2026; Lum et al., 2021). Glycemic trajectories after treatment initiation vary across patients (Hooven-Davis et al., 2026), making the resulting histories a source for patient-specific learning without exploratory dosing.

Offline reinforcement learning limits distribution shift through action support, policy regularization, and control of value extrapolation (Fujimoto et al., 2019; Fujimoto & Gu, 2021; Kumar et al., 2020; Kostrikov et al., 2022). These mechanisms govern policy deviation or induce preferences through learned return, yet do not independently establish the direction of improvement from observed response. AIRC treats stable signed response, rather than action proximity alone, as evidence for offline policy improvement.

Action-Innovation Response Certificates (AIRC) estimate signed response to future glucose burden from state-relative treatment variation. The method certifies its orientation across complementary folds and prediction horizons, while a separate low-glucose pathway regulates insulin increases. Its contributions are:

- **State-relative response construction.** AIRC constructs a state-relative innovation coordinate and estimates its signed relation to future hyperglycemic burden beyond the state prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Fold–horizon response certification.** Fold disagreement attenuates certificate confidence; cross-horizon sign agreement decides whether directional guidance is active.
- **Risk-decoupled policy optimization.** Certified direction and confidence shape the policy update, while short-horizon low-glucose susceptibility independently regulates positive shifts. The training-time evidence pathways are removed at deployment, leaving one encoder–actor pass.

Evaluation combines SimGlucose closed-loop control, meal-aligned trajectories, component interventions, and matched held-out Loop states. AIRC leads cumulative reward overall and within every SimGlucose age group, while showing greater trend–action monotonicity and directional consistency than IQL on observational states.

2 Related Work

2.1 Learning insulin control from treatment data

Validated simulators support repeatable closed-loop comparison (Dalla Man et al., 2014); clinical guidance defines treatment standards (American Diabetes Association Professional Practice Committee for Diabetes, 2026b), and natural-use data characterize delivery under everyday conditions (Lum et al., 2021). Learning-based insulin control spans reinforcement-learning dose personalization, Bayesian titration, and digital-twin-assisted automated delivery (Jafar et al., 2024; Kobayati et al., 2025; Kovatchev et al., 2025); offline policies have also been learned from simulated treatment records (Beolet et al., 2024). AIRC asks how retrospective treatment variation can provide explicit evidence for the direction of policy improvement.

2.2 Policy improvement under distribution shift

Behavior-constrained methods restrict unsupported actions or policy deviation (Fujimoto et al., 2019; Fujimoto & Gu, 2021); value-based methods induce preferences through conservative estimation or value-weighted policy extraction (Kumar et al., 2020; Kostrikov et al., 2022; Xu et al., 2025; Kim & Oh, 2026). Recent methods refine supported actions through adaptive neighborhoods or peer-state candidates (Mao et al., 2025; Zhang et al., 2026), adjust policy constraints (Tan et al., 2026), or selectively regulate out-of-distribution actions (Wang et al., 2026a). SCORE targets spurious correlations (Deng et al., 2024), whereas diffusion and flow policies expand policy expressiveness (Gao et al., 2025; Park et al., 2025; Nguimatsia Tiofack et al., 2026). Value-guided trajectory synthesis offers a complementary route through data augmentation (Wang et al., 2026b).

Risk-aware offline learning coordinates reward improvement with safety costs (Yan et al., 2025; Liu et al., 2026; Guo et al., 2025). These lines of work address action support, value extrapolation, conservative improvement, robustness, or explicit risk, but leave signed response orientation and its stability across folds and horizons implicit. AIRC separates three questions that are usually coupled: which direction is supported by response evidence, how reliable that evidence is, and how risky that direction is to execute.

2.3 Predictive knowledge for control

General value functions represent signal-specific predictive knowledge over multiple time scales (Sutton et al., 2011). Recent glucose models jointly forecast future levels and detect hypoglycemia (Hwang et al., 2025). AIRC uses long-horizon hyperglycemic predictions to define the residual outcome for response estimation, while short-horizon low-glucose predictions form a separate susceptibility pathway. Predictive knowledge thereby supports directional response estimation and asymmetric risk modulation before policy optimization.

3 Problem Formulation

For patient p, the offline dataset contains n_p transitions:

$$\mathcal{D}_p = \{(\boldsymbol{s}_t, a_t, r_t, \boldsymbol{s}_{t+1}, b_t)\}_{t=1}^{n_p}.$$

Here b_t marks an episode boundary. The policy observes $\boldsymbol{s}_t \in \mathbb{R}^{11}$, comprising nine glucose-history features, meal-on-board, and insulin-on-board. These variables summarize the available treatment history; physiological variables internal to the simulator do not enter the policy. Episode identifiers and timestamps support offline transition assembly but are excluded from the policy input. The action a_t controls continuous basal delivery, and a patient-specific map converts normalized action coordinates to physical dose. Meal boluses remain outside the policy output.

For patient dynamics P_p, the control objective is

$$J_p(\pi) = \mathbb{E}_{P_p,\pi}\left[\sum_{t=0}^{T-1} \gamma_{\mathrm{RL}}^t r_t\right], \qquad a_t \in \mathcal{A}_p, \tag{1}$$

where $\mathcal{A}_p$ is the permitted basal-action interval. The experiments set $r_t = -R_{\mathrm{M}}(g_{t+1})$, the negative Magni risk defined in Equation 10. Larger return therefore corresponds to lower accumulated glycemic risk. Policy learning is fully offline; no trajectories are collected under the learned policy during training. The reported control endpoint is the undiscounted sum of this reward over the evaluation rollout, accompanied by glucose-range and risk metrics; the discounted objective in Equation 1 defines policy learning.

All normalization statistics and data-derived action bounds are fitted on policy-training records. AIRC learns a deterministic policy; Section 4 specifies its predictive supervision and update rule.

4 AIRC: Action-Innovation Response Certificates

AIRC admits response evidence into policy optimization only when its orientation is stable across data partitions and prediction horizons. A separate short-horizon pathway represents direction-dependent low-glucose risk. Figure 1 connects these training pathways to the actor and deployed policy.

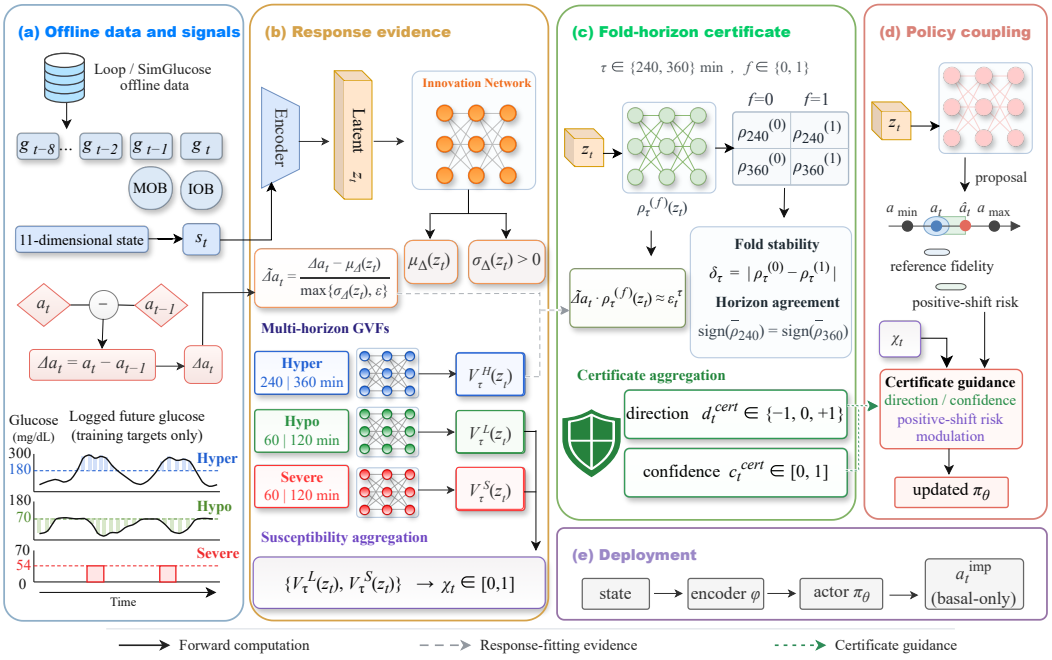

Figure 1: AIRC architecture and evidence flow: **(a)** offline data and signals, **(b)** response evidence, **(c)** fold–horizon certification, **(d)** policy coupling, and **(e)** deployment. Solid, gray dashed, and green dotted arrows denote forward computation, detached response-fitting evidence, and certificate guidance, respectively. Deployment retains the encoder, actor, and action projection.

4.1 OFFLINE DATA AND SIGNALS

For each offline transition, AIRC receives the control state $\boldsymbol{s}_t$, logged basal action a_t, successor state, and episode-boundary indicator b_t. The encoder produces $\boldsymbol{z}_t = \phi_\omega(\boldsymbol{s}_t)$. Consecutive logged actions define $\Delta a_t = a_t - a_{t-1}$ within contiguous episodes; starts and discontinuities are excluded from innovation-dependent losses. Recorded future glucose supervises the evidence pathways but never enters the deployed policy.

4.2 RESPONSE EVIDENCE

The response branch begins with a state-relative action coordinate. An innovation network estimates the conditional location $\mu_\Delta(\boldsymbol{z}_t)$ and positive deviation scale $\sigma_\Delta(\boldsymbol{z}_t)$ by robust regression; together they define the standardized action innovation

$$\widetilde{\Delta a}_t = \frac{\Delta a_t - \mu_\Delta(\boldsymbol{z}_t)}{\max\{\sigma_\Delta(\boldsymbol{z}_t), \epsilon\}}. \tag{2}$$

The implementation uses $\epsilon = 10^{-4}$, while the learned scale is also bounded away from zero. The sign of $\widetilde{\Delta a}_t$ records whether the logged basal change is larger or smaller than the adjustment expected in the current state. Conditional scaling makes these deviations comparable across states without discarding their local orientation.

The predictive branches distinguish persistent high glucose from near-term low glucose. For the next recorded glucose value g_{t+1}, in $\mathrm{mg/dL}$, the immediate signals are

$$\begin{aligned} c_t^{\mathrm{H}} &= \min\left\{10,\ \frac{\mathrm{softplus}^2((g_{t+1} - 180)/30)}{\kappa_{\mathrm{H}}}\right\}, \\ c_t^{\mathrm{L}} &= \mathrm{sigmoid}((90 - g_{t+1})/10), \\ c_t^{\mathrm{S}} &= \mathbf{1}[g_{t+1} < 54]. \end{aligned} \tag{3}$$

Here κ_{H} normalizes hyperglycemic burden using policy-training data; the smooth low-glucose channel anticipates the clinical $70\,\mathrm{mg/dL}$ threshold. Appendix A.2 gives the normalization rule.

For channel k and time constant τ, a general value function V_τ^k predicts the corresponding discounted signal. At nominal sampling interval Δ, its target is

$$y_{t,\tau}^k = c_t^k + \gamma_\tau(1 - b_t)\overline{V}_\tau^k(\overline{\phi}(\boldsymbol{s}_{t+1})), \qquad \gamma_\tau = \exp(-\Delta/\tau). \tag{4}$$

Bars denote slowly updated target networks. Long time constants $\mathcal{T}_{\mathrm{H}} = \{240, 360\}$ minutes describe sustained hyperglycemic response, whereas $\mathcal{T}_{\mathrm{L}} = \{60, 120\}$ minutes resolve early low-glucose susceptibility. These horizons distinguish near-term insulin action from more persistent postprandial effects, consistent with rapid-acting insulin pharmacodynamics and observed meal responses (American Diabetes Association Professional Practice Committee for Diabetes, 2026a; Jafar et al., 2024). Equation 4 converts each physical duration to a dataset-specific per-step discount.

4.3 FOLD–HORIZON RESPONSE CERTIFICATE

After the representation update, let $\bar{\boldsymbol{z}}_t = \mathrm{sg}[\phi_\omega(\boldsymbol{s}_t)]$. For each long horizon, the detached temporal-difference residual $\varepsilon_{t,\tau} = \mathrm{sg}[y_{t,\tau}^{\mathrm{H}} - V_\tau^{\mathrm{H}}(\bar{\boldsymbol{z}}_t)]$ measures future burden not explained by the state representation. Two response heads, fitted on complementary data folds $f \in \{0, 1\}$, relate this residual to the detached action innovation:

$$\mathcal{L}_{\mathrm{resp}} = \frac{1}{4}\sum_{\tau\in\mathcal{T}_{\mathrm{H}}}\sum_{f=0}^{1}\mathbb{E}_{t\in\mathcal{D}^{(f)}}\left[\ell_{\mathrm{SL1}}\left(\widetilde{\Delta a}_t\rho_\tau^{(f)}(\bar{\boldsymbol{z}}_t), \varepsilon_{t,\tau}\right) + \lambda_\rho\left(\rho_\tau^{(f)}(\bar{\boldsymbol{z}}_t)\right)^2\right]. \tag{5}$$

ℓ_{SL1} is the smooth-L_1 loss, and the factor $1/4$ averages the two horizons and folds. The response heads share the encoder and innovation model but are optimized on separate folds. Algorithm 1 specifies the complete update order and stop-gradient operations.

The response coefficient is defined on two state-relative quantities rather than raw action and glucose. Equation 2 places logged basal changes relative to the state-predicted adjustment, while $\varepsilon_{t,\tau}$

measures future burden beyond the state prediction. Their association yields $\rho_\tau^{(f)}(\bar{\boldsymbol{z}}_t)$, a state-conditioned signed measure of response orientation. Because future hyperglycemic burden is a cost, its sign is reversed when constructing the preferred direction.

Let $\bar{\rho}_\tau = (\rho_\tau^{(0)} + \rho_\tau^{(1)})/2$ and $\delta_\tau = |\rho_\tau^{(0)} - \rho_\tau^{(1)}|$. For long horizons τ_1, τ_2, the response certificate is

$$\begin{aligned} q_\tau &= \frac{|\bar{\rho}_\tau|}{|\bar{\rho}_\tau| + \delta_\tau + 0.25 s_\rho}, \\ A_t &= \mathbf{1}[\bar{\rho}_{\tau_1}\bar{\rho}_{\tau_2} > 0], \\ d_t^{\mathrm{cert}} &= -\operatorname{sign}(\bar{\rho}_{\tau_1} + \bar{\rho}_{\tau_2}) A_t, \\ c_t^{\mathrm{cert}} &= \min(q_{\tau_1}, q_{\tau_2}) A_t. \end{aligned} \tag{6}$$

The positive moving scale s_ρ stabilizes confidence near zero. Conditioning on $|\bar{\rho}_\tau|$ and s_ρ, confidence increases monotonically with fold agreement because

$$\frac{\partial q_\tau}{\partial \delta_\tau} = -\frac{|\bar{\rho}_\tau|}{\left(|\bar{\rho}_\tau| + \delta_\tau + 0.25 s_\rho\right)^2} \leq 0.$$

Fold disagreement continuously attenuates c_t^{cert}; a cross-horizon sign conflict sets $A_t = 0$ and withholds guidance. The estimated response is thus distinct from the direction and strength admitted to policy optimization.

4.4 Policy coupling

The actor receives detached evidence from the training-time pathways. Certificate direction and confidence control signed policy pressure, while short-horizon predictions independently form the low-glucose susceptibility

$$\chi_t = \min\left\{1, \max_{\substack{k\in\{\mathrm{L,S}\} \\ \tau\in\mathcal{T}_{\mathrm{L}}}} (1-\gamma_\tau) V_\tau^k(\boldsymbol{z}_t)\right\}. \tag{7}$$

The factor $1-\gamma_\tau$ places different discount scales on a comparable level. Write $\widehat{a}_t = \pi_\theta(\boldsymbol{z}_t)$ and $u_t = (\widehat{a}_t - a_t)/\widehat{\sigma}_t$, where $\widehat{\sigma}_t$ is the learned innovation scale restricted to the configured actor interval. With the evidence modules detached, the actor minimizes

$$\begin{aligned} \mathcal{L}_{\mathrm{actor}} = \mathbb{E}\big[&\ell_{\mathrm{SL1}}(\widehat{a}_t, a_t) + \lambda_{\mathrm{dir}} c_t^{\mathrm{cert}} \operatorname{softplus}(-d_t^{\mathrm{cert}} u_t) \\ &+ \lambda_{\mathrm{safe}} \chi_t [u_t]_+^2 + \lambda_{\mathrm{bd}} \ell_{\mathrm{bd}}(\widehat{a}_t)\big], \end{aligned} \tag{8}$$

where $[u]_+ = \max(u, 0)$ and ℓ_{bd} is the squared action-bound violation penalty.

Equation 8 keeps certified direction distinct from the low-glucose risk of positive shifts, alongside reference fidelity and action bounds. At a logged action within the action bounds, the reference, susceptibility, and bound terms have zero derivative with respect to u_t; the directional derivative is $-\lambda_{\mathrm{dir}} c_t^{\mathrm{cert}} d_t^{\mathrm{cert}}/2$. An active certificate therefore orients the initial descent step. For $u_t > 0$, susceptibility adds $2\lambda_{\mathrm{safe}}\chi_t u_t$, increasingly opposing insulin increases near the lower glucose tail.

4.5 Deployment

Training uses the innovation, predictive, response, and certificate pathways; deployment does not. The implemented action is

$$a_t^{\mathrm{imp}} = \Pi_{\mathcal{A}_p}[\pi_\theta(\phi_\omega(\boldsymbol{s}_t))],$$

which requires one encoder–actor forward pass followed by projection onto the patient-specific basal interval. Meal boluses remain outside the policy output. Appendix A.3 separates this deployment cost from the training-time evidence pathways.

Algorithm 1 Response-certified AIRC optimization

Require: $\mathcal{D} = \mathcal{D}^{(0)} \cup \mathcal{D}^{(1)}$; horizons $\mathcal{T}_{\mathrm{H}}, \mathcal{T}_{\mathrm{L}}$; updates N; actor period K

Initialize ϕ_ω, $(\mu_\Delta, \sigma_\Delta)$, $\{V_\tau^k\}$, $\{\rho_\tau^{(f)}\}$, π_θ, and target networks
for $n = 1, \ldots, N$ **do**
  Draw $\mathcal{B} \subset \mathcal{D}$; set $\boldsymbol{z}_t \leftarrow \phi_\omega(\boldsymbol{s}_t)$ and $\Delta a_t \leftarrow a_t - a_{t-1}$ on valid pairs
  Evaluate $\mu_\Delta(\boldsymbol{z}_t)$, $\sigma_\Delta(\boldsymbol{z}_t)$, and $\{V_\tau^k(\boldsymbol{z}_t)\}_{k,\tau}$
  Compute $y_{t,\tau}^k$ by Eq. 4
  Fit ϕ_ω, $(\mu_\Delta, \sigma_\Delta)$, and $\{V_\tau^k\}$ by robust regression
  $\bar{\boldsymbol{z}}_t \leftarrow \mathrm{sg}[\phi_\omega(\boldsymbol{s}_t)]$; detach $\widetilde{\Delta a}_t$ and $\varepsilon_{t,\tau}$ on valid transitions
  for $\tau \in \mathcal{T}_{\mathrm{H}}$, $f \in \{0, 1\}$ **do**
    Fit $\rho_\tau^{(f)}$ on $\mathcal{B} \cap \mathcal{D}^{(f)}$ by Eq. 5
  end for
  Update the moving scale s_ρ
  if $n \bmod K = 0$ **then**
    Evaluate and detach $(d_t^{\mathrm{cert}}, c_t^{\mathrm{cert}}, \chi_t)$ by Eqs. 6–7
    Update θ by Eq. 8
  end if
  Soft-update $\bar{\phi}$ and $\{\bar{V}_\tau^k\}$
end for
return $\Pi_{\mathcal{A}_p}[\pi_\theta(\phi_\omega(\boldsymbol{s}))]$

5 Experiments

The evaluation tests the response-certification principle through closed-loop control (RQ1), post-prandial recovery (RQ2), component attribution (RQ3), and directional response coherence on held-out Loop states (RQ4).

5.1 Datasets and evaluation design

SimGlucose evaluation. We evaluated closed-loop control with the UVA/Padova physiology models implemented in SimGlucose (Dalla Man et al., 2014; Xie, 2018). The cohort spanned adult, adolescent, and child physiology profiles. Each profile-specific policy was trained on offline trajectories from the reference Loop controller, and all methods received identical transitions and control states. After training and model selection, policies were evaluated for up to 4,800 three-minute steps, corresponding to ten days.

Loop evaluation. We used de-identified records from the public Loop observational dataset to examine policy behavior on real treatment states (Lum et al., 2021; Jaeb Center for Health Research, 2023). The cohort spanned adult, adolescent, and child participants. IQL served as the focused real-data comparator because it led the six SimGlucose baselines in mean cumulative reward, the prespecified primary control outcome. AIRC and IQL policies were trained on the first 80% of each participant's chronological record; the remaining 20% was excluded from training and reserved for analysis.

Hyperparameter selection. Each SimGlucose offline dataset was partitioned by whole episode, assigning 80% to policy fitting and 20% to validation. For every method, one method-specific hyperparameter was tuned over the prespecified three-value search space in Table 4. Minimax Q-function learning (MQL) (Uehara et al., 2020) was fitted on the policy-fitting split and used to estimate each candidate's discounted return from validation-episode initial states. Candidate configurations were ranked by the profile-balanced validation estimate, and the top-ranked configuration was selected before closed-loop evaluation.

Comparisons and outcomes. The SimGlucose baselines were BCQ (Fujimoto et al., 2019), TD3+BC (Fujimoto & Gu, 2021), CQL (Kumar et al., 2020), IQL (Kostrikov et al., 2022), SCORE (Deng et al., 2024), and FQL (Park et al., 2025). Every method received a budget of 100,000 updates

and used the same offline transitions, state and action interfaces, and evaluation procedure; Appendix A reports the algorithm roles, search spaces, selection protocol, and implementation details. The primary outcome was cumulative reward. We also report time in range (TIR; 70–$180\,\mathrm{mg/dL}$), time below range (TBR; < 70), severe TBR (STBR; < 54), time above range (TAR; > 180), glycemia risk index (GRI), and clinical termination (Battelino et al., 2019; Klonoff et al., 2023). Profile-level means weight physiological profiles equally and retain clinically terminated evaluations. Failure is the percentage of evaluation rollouts that triggered clinical termination; numerical outcomes are displayed to two decimals.

5.2 Closed-loop control advantage (RQ1)

Table 1: Closed-loop SimGlucose comparison. Glucose-range metrics are percentages. Failure is the percentage of evaluation rollouts that triggered clinical termination. Bold marks the best displayed mean.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC	**-32333.15**	**69.48**	**0.00**	**0.00**	**30.52**	**32.41**	0.00
IQL	-35981.05	66.26	1.00	0.18	32.74	36.32	0.00
BCQ	-37003.08	65.14	0.70	0.09	34.16	36.86	0.00
TD3+BC	-39748.79	63.20	2.09	1.05	34.71	42.41	3.33
CQL	-38496.23	64.53	1.19	0.37	34.28	38.76	0.00
SCORE	-38091.04	58.02	0.07	0.01	41.91	43.97	0.00
FQL	-39082.88	55.77	0.12	0.02	44.11	46.22	0.00

AIRC achieved the highest mean cumulative reward, -32333.15, compared with -35981.05 for IQL; the profile-paired advantage was 3647.91 (95% interval 1012.93–6478.60). Relative to IQL, TIR was 4.86% higher and GRI 10.77% lower, with 0.00% TBR and STBR, the lowest TAR, and no clinical terminations. The reward advantage held across age groups (Appendix B); AIRC also led TIR and GRI in adults and children and retained the adolescent reward lead with 0.00% TBR and STBR. Clinical termination occurred only for TD3+BC, and complete-case analysis preserved every reward-comparison direction (Table 6).

5.3 Postprandial recovery dynamics (RQ2)

Meal-aligned analysis measured cumulative attainment of sustained target-range control under a prespecified common window shared by all methods. Events were averaged within profile before profile-balanced aggregation, and resampling preserved method pairing.

At 240 minutes, 54.67% of eligible AIRC events met the target-range criterion. The profile-paired advantage over IQL was 10.2 percentage points (95% CI 2.7–17.3); differences against the other baselines ranged from 13.6 to 33.6 percentage points, with every interval above zero. Figure 2 shows that this advantage emerged through earlier accumulation of sustained target-range control.

5.4 Component attribution (RQ3)

Four matched-protocol ablations removed directional guidance, the one-sided risk penalty, learned innovation scaling, or both actor-guidance terms. Exact interventions and age-stratified results appear in Appendix A.5.

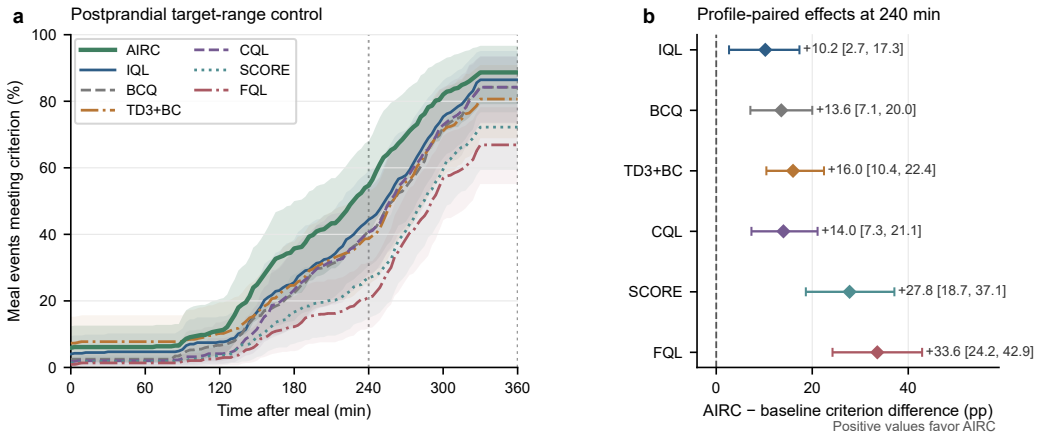

Figure 2: Meal-aligned postprandial target-range control in SimGlucose. **a**, Profile-balanced cumulative proportion of meal events meeting the target-range criterion by each horizon. An event met the criterion if glucose never exceeded $180\,\mathrm{mg/dL}$ or, following an excursion, returned to 70–$180\,\mathrm{mg/dL}$ and remained there for 30 consecutive minutes. Shading denotes 95% profile-bootstrap intervals. **b**, Mean profile-paired difference in criterion attainment at 240 minutes between AIRC and each baseline; whiskers denote 95% profile-bootstrap intervals. The same prespecified complete windows are used for all methods.

Table 2: Component ablations under the matched SimGlucose protocol. Glucose-range metrics are percentages. Failure is the percentage of evaluation rollouts that triggered clinical termination. Bold marks the best displayed mean.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC (full)	**-32333.15**	**69.48**	**0.00**	**0.00**	30.52	**32.41**	0.00
w/o direction	-44325.61	47.87	**0.00**	**0.00**	52.13	54.16	0.00
w/o risk penalty	-52731.57	68.72	9.75	5.41	**21.53**	42.50	13.33
w/o innovation scale	-41371.27	68.15	2.31	0.77	29.53	37.21	3.33
w/o both guidance terms	-38635.38	57.19	0.04	**0.00**	42.77	44.75	0.00

The last row removes both directional guidance and the risk penalty, retaining reference fidelity and action bounds.

Removing direction lowered TIR to 47.87% and raised TAR to 52.13%, while TBR remained 0.00%. Removing risk weighting produced the opposing pattern: TAR fell to 21.53%, whereas TBR reached 9.75%, severe TBR 5.41%, and failure 13.33%. Removing innovation scaling reduced reward to -41371.27, raised TBR to 2.31%, and yielded 3.33% failure. The interventions assign sustained-hyperglycemia control to direction, lower-tail protection to susceptibility, and response estimation to state-conditioned scaling; the full model retained the highest reward overall and across age groups.

5.5 Response coherence on held-out loop states (RQ4)

For each eligible held-out state, we imposed five recent glucose trends from -1 to $1\,\mathrm{mg/dL/min}$ by varying glucose at $t-30$ minutes while leaving every other input unchanged. AIRC and IQL received the same perturbed states. We assessed policy-response coherence through four properties: trend–action monotonicity, dose–response linearity across the five trend levels, directional consistency, and endpoint action separation scaled by the training-action IQR. Statistics were aggregated within participant before paired interval estimation; Appendix A.6 gives eligibility criteria and denominators.

Table 3: Out-of-sample policy response to controlled glucose-trend variation on matched held-out Loop states. Entries are participant-balanced point estimates; positive paired differences favor AIRC.

Method	Trend Spearman ρ	Dose–response R^2	Directional consistency (%)	Endpoint contrast in action IQR
AIRC	0.943	0.910	98.23	1.029
IQL	0.804	0.808	93.72	0.943
Paired Δ	$+0.139$	$+0.103$	$+4.51$	$+0.086$

AIRC exceeded IQL by 0.139 in trend–action Spearman correlation (95% CI 0.111–0.169) and 0.103 in R^2 (95% CI 0.083–0.124). Directional consistency was 4.51 percentage points higher (95% CI 3.40–5.74), and endpoint contrast increased by 0.086 training-action IQR (95% CI 0.005–0.170). The first three advantages remained positive across age groups (Appendix A.7). SimGlucose supplies matched closed-loop outcomes, while these held-out Loop states test whether the learned response remains graded under observational inputs.

6 Discussion and Future Work

The ablations expose distinct physiological effects of the policy components. Without directional guidance, sustained hyperglycemic exposure rises while time below range remains at zero; without risk weighting, lower-tail exposure and clinical terminations rise as time above range falls. Removing learned innovation scaling produces a third pattern, reducing reward while increasing both time below range and termination frequency. These signatures map to the architecture's separate decisions about adjustment direction, state-conditioned scaling, and the risk of executing an increase. Their combination yields the full model's reward lead without exchanging one glycemic tail for the other.

The meal-aligned and Loop analyses refine the interpretation of that control advantage. Earlier postprandial attainment shows that the gain extends beyond long-run averages to the timing of sustained target-range recovery. On held-out Loop states, changing recent glucose trend while holding all other inputs constant exposes each policy's dose–response curve. The four reported properties measure different aspects of that curve: trend ordering, linearity, directional consistency, and action separation. AIRC improves all four against IQL, the strongest SimGlucose reward baseline. Closed-loop outcomes and matched observational responses thus connect realized control with the structure of the learned action rule.

The same decomposition suggests a route to other ordered treatment decisions with state-dependent benefit and asymmetric harm. Predictive evidence can determine the direction of an adjustment, while a distinct risk pathway regulates whether that adjustment should be executed. Coupled basal–bolus control is a natural next setting: interacting actions call for multivariate response estimation, and delayed low-glucose effects require action-specific risk assessment. Testing whether fold–horizon agreement preserves directional ordering under those interactions would extend response certification to a richer control problem.

7 Conclusion

AIRC introduces response certification for risk-aware offline policy improvement. State-relative innovations and fold–horizon agreement separate improvement direction, confidence, and asymmetric risk before policy optimization. AIRC achieved the highest SimGlucose reward overall and within every age group, improved TIR and GRI over the strongest reward baseline, and produced stronger trend-responsive behavior than IQL on held-out Loop states. These results support response certification as a risk-aware policy-improvement principle for continuous-control problems with ordered actions and asymmetric directional consequences.

Ethics Statement

This study used SimGlucose for closed-loop policy evaluation and de-identified records from the public Loop observational dataset for offline training and held-out response analysis. No prospective human intervention or clinical deployment of the learned policies was conducted. The Loop data were handled under the applicable access and data-use conditions.

AI Use Statement

Generative AI tools assisted with manuscript language and presentation, including LaTeX organization, figure-label editing, and reference-metadata checks. The research questions, method and algorithm design, experimental design and execution, result verification, and scientific conclusions were developed and completed by the authors. All AI-assisted material was reviewed and finalized by the authors.

Reproducibility Statement

To support the reproducibility of our work, we have submitted the source code, datasets, implementation details, and necessary instructions as the Supplementary Materials. We believe that these resources provide sufficient information for reproducing the main experimental results presented in this paper.

References

American Diabetes Association Professional Practice Committee for Diabetes. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of care in diabetes—2026. *Diabetes Care*, 49 (Suppl. 1):S183–S215, 2026a. doi: 10.2337/dc26-S009.

American Diabetes Association Professional Practice Committee for Diabetes. 7. Diabetes Technology: Standards of care in diabetes—2026. *Diabetes Care*, 49(Suppl. 1):S150–S165, 2026b. doi: 10.2337/dc26-S007.

Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, Emanuele Bosi, Bruce A. Buckingham, William T. Cefalu, Kelly L. Close, et al. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 42(8):1593–1603, 2019. doi: 10.2337/dci19-0028.

Tristan Beolet, Alice Adenis, Erik Huneker, and Maxime Louis. End-to-end offline reinforcement learning for glycemia control. *Artificial Intelligence in Medicine*, 154:102920, 2024. doi: 10.1016/j.artmed.2024.102920.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1):26–34, 2014. doi: 10.1177/1932296813514502.

Zhihong Deng, Zuyue Fu, Lingxiao Wang, Zhuoran Yang, Chenjia Bai, Tianyi Zhou, Zhaoran Wang, and Jing Jiang. False correlation reduction for offline reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1199–1211, 2024. doi: 10.1109/TPAMI.2023.3328397.

Harry Emerson, Matthew Guy, and Ryan McConville. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, 2023. doi: 10.1016/j.jbi.2023.104376.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Chen-Xiao Gao, Chenyang Wu, Mingjun Cao, Chenjun Xiao, Yang Yu, and Zongzhang Zhang. Behavior-regularized diffusion policy optimization for offline reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 18630–18657, 2025. URL https://proceedings.mlr.press/v267/gao25q.html.

Junyu Guo, Zhi Zheng, Donghao Ying, Ming Jin, Shangding Gu, Costas J. Spanos, and Javad Lavaei. Don't trade off safety: Diffusion regularization for constrained offline RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5346.

Jayde Hooven-Davis, Christina M. Lalama, Sarah A. Syer, Scott D. Rothenberger, Ingrid M. Libman, and Christine A. March. What makes the difference? Predicting glycemic trajectories following initiation of automated insulin delivery systems in youth with type 1 diabetes. *Diabetes Care*, 49(9):1720–1728, 2026. doi: 10.2337/dc26-0634.

Minjoo Hwang, Vega Pradana Rachim, Junyoung Yoo, Yein Lee, and Sung-Min Park. Generalized multi task learning framework for glucose forecasting and hypoglycemia detection using simulation to reality. *npj Digital Medicine*, 8:612, 2025. doi: 10.1038/s41746-025-01994-4.

Jaeb Center for Health Research. An observational study of individuals with type 1 diabetes using the loop system for automated insulin delivery: Public dataset. Public Study Dataset 560, 2023. URL https://public.jaeb.org/dataset/560. Accessed 2026-08-25.

Adnan Jafar, Alessandra Kobayati, Michael A. Tsoukas, et al. Personalized insulin dosing using reinforcement learning for high-fat meals and aerobic exercises in type 1 diabetes: A proof-of-concept trial. *Nature Communications*, 15:6585, 2024. doi: 10.1038/s41467-024-50764-5.

Byeongchan Kim and Min-hwan Oh. Peng's Q(λ) for conservative value estimation in offline reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/14c1c1c1dabd8179f77895de25391ce9-Abstract-Conference.html.

David C. Klonoff, Jing Wang, David Rodbard, Michael A. Kohn, Chengdong Li, Dorian Liepmann, David Kerr, David Ahn, Anne L. Peters, et al. A glycemia risk index (GRI) of hypoglycemia and hyperglycemia for continuous glucose monitoring validated by clinician ratings. *Journal of Diabetes Science and Technology*, 17(5):1226–1242, 2023. doi: 10.1177/19322968221085273.

Alessandra Kobayati, Anas El Fathi, Natasha Garfield, et al. A Bayesian decision support system for automated insulin doses in adults with type 1 diabetes on multiple daily injections: A randomized controlled trial. *Nature Communications*, 16:8593, 2025. doi: 10.1038/s41467-025-63671-0.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Boris P. Kovatchev, Patricio Colmegna, Jacopo Pavan, Jenny L. Diaz Castañeda, Maria F. Villa-Tamayo, Chaitanya L. K. Koravi, Giulio Santini, Carlene Alix, Meaghan Stumpf, and Sue A. Brown. Human-machine co-adaptation to automated insulin delivery: A randomised clinical trial using digital twin technology. *npj Digital Medicine*, 8:253, 2025. doi: 10.1038/s41746-025-01679-y.

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

Zifan Liu, Xinran Li, Shibo Chen, and Jun Zhang. GAS: Enhancing Reward-Cost Balance of Generative Model-assisted Offline Safe RL. In *International Conference on Learning Representations*, 2026.

John W. Lum, Ryan J. Bailey, Victoria Barnes-Lomen, Diana Naranjo, Korey K. Hood, Rayhan A. Lal, Brandon Arbiter, Adam S. Brown, Daniel J. DeSalvo, Jeremy Pettus, et al. A real-world prospective study of the safety and effectiveness of the loop open source automated insulin delivery system. *Diabetes Technology & Therapeutics*, 23(5):367–375, 2021. doi: 10.1089/dia.2020.0535.

Yixiu Mao, Yun Qu, Qi Wang, and Xiangyang Ji. Adaptive neighborhood-constrained Q learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0877.

Franki Nguimatsia Tiofack, Théotime Le Hellard, Fabian Schramm, Nicolas Perrin-Gilbert, and Justin Carpentier. Guided flow policy: Learning from high-value actions in offline reinforcement learning. In *International Conference on Learning Representations*, 2026.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 48104–48127, 2025. URL https://proceedings.mlr.press/v267/park25f.html.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pp. 761–768, 2011. URL https://www.ifaamas.org/AAMAS/aamas2011/AcceptedFull_ShortPapers.html.

Jing Tan, Xiaorui Li, Chao Yao, Xiaojuan Ban, Yuetong Fang, Renjing Xu, and Zhaolin Yuan. Adaptive scaling of policy constraints for offline reinforcement learning. In *International Conference on Learning Representations*, 2026.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9659–9668, 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

Qingjun Wang, Hongtu Zhou, Hang Yu, Junqiao Zhao, Yanping Zhao, Chen Ye, Ziqiao Wang, and Guang Chen. Beyond penalization: Diffusion-based out-of-distribution detection and selective regularization in offline reinforcement learning. In *International Conference on Learning Representations*, 2026a.

Tieru Wang, Kunbao Wu, and Guoshun Nan. Trajectory generation with conservative value guidance for offline reinforcement learning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3072c1be85bb98d539bcb3a828ab41b1-Abstract-Conference.html.

Emma G. Wilmot, Jacques Beltrand, Bruno Guerci, et al. Tubeless automated insulin delivery versus multiple daily injections in children and adults with type 1 diabetes with elevated HbA1c (RADIANT): A multicentre, international, parallel-group, open-label, randomised, controlled trial. *The Lancet Diabetes & Endocrinology*, 14(4):305–316, 2026. doi: 10.1016/S2213-8587(25)00364-X.

Jinyu Xie. Simglucose v0.2.1: A type 1 diabetes simulator implemented in Python. GitHub repository, 2018. URL https://github.com/jxx123/simglucose.

Haoran Xu, Liyuan Mao, Hui Jin, Weinan Zhang, Xianyuan Zhan, and Amy Zhang. Uni-RL: Unifying online and offline RL via implicit value regularization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4436.

Runze Yan, Xun Shen, Akifumi Wachi, Sebastien Gros, Anni Zhao, and Xiao Hu. Offline guarded safe reinforcement learning for medical treatment optimization strategies. In *Advances in Neural Information Processing Systems*, volume 38, pp. 108662–108703, 2025. doi: 10.52202/085713-3275.

Yujia Zhang, Lin Li, Wei Wei, Jianguo Wu, Yi Ma, and Jiye Liang. Efficient offline reinforcement learning via peer-influenced constraint. In *International Conference on Learning Representations*, 2026.

A Experimental and Implementation Details

A.1 SimGlucose data processing and evaluation

Each profile-specific policy uses offline transitions generated by the reference Loop data-collection controller. The official preprocessing produces an 11-dimensional state comprising nine glucose-history features, meal-on-board, and insulin-on-board. Each transition record retains the normalized basal action, reward, successor state, episode boundary, timestep, and the patient-specific action transformation. Meal boluses are generated by the environment and are not policy outputs.

Whole episodes are assigned to the 80% policy-fitting split or the 20% validation split. State normalization, action normalization, and data-derived dose bounds are estimated on policy-fitting records. Complete policy-fitting episodes are alternated between the two response folds. When episode alternation yields severe fold imbalance, the block-balanced assignment described in Appendix A.2 is applied. Predecessor actions are recovered from verified contiguous rows. Episode starts, terminal transitions, sequence tails, and detected discontinuities are masked from losses that require temporal adjacency.

Closed-loop evaluation is disjoint from offline dataset construction. After training and model selection, policies act in newly generated SimGlucose trajectories for the same physiology profiles; evaluation transitions never enter the training data. The evaluation cohort is balanced across adult, adolescent, and child physiology groups.

Every reported model is trained for 100,000 updates and evaluated for up to 4,800 three-minute steps. The archives record the dataset hash, run manifest, requested and observed rollout length, and clinical termination flag. All clinically terminated rollouts are retained, with reward and glucose metrics calculated over their observed trajectories.

Table 4 gives the method-specific hyperparameter search spaces used for the SimGlucose comparison.

Table 4: SimGlucose hyperparameter search spaces.

Method	Tuned parameter	Search space
AIRC	Direction weight	$\{0.30, 0.50, 0.70\}$
IQL	Expectile	$\{0.50, 0.70, 0.90\}$
BCQ	Perturbation scale ϕ	$\{0.025, 0.050, 0.100\}$
TD3+BC	Behavior weight α	$\{1.25, 2.50, 5.00\}$
CQL	Minimum-Q weight	$\{1.00, 5.00, 10.00\}$
SCORE	β	$\{0.10, 0.20, 0.40\}$
FQL	α	$\{1.00, 3.00, 10.00\}$

MQL used a $12 \to 32 \to 32 \to 1$ Q-network, a discount of 0.99, minibatches of 500 transitions, and 100,000 optimization steps at a learning rate of 0.005. Its kernel bandwidth was the median block-balanced training distance divided by 7.5. Candidate policies were ranked by the profile-balanced mean final MQL discounted-return estimate across validation profiles. This MQL-based screening was confined to the validation partition; closed-loop evaluation did not enter the ranking. The top-ranked configuration for each method was selected before evaluation. The comparison otherwise shares offline transitions, state and action interfaces, update budget, and evaluator across methods.

Baseline implementations. The comparison covers complementary offline-policy mechanisms. BCQ constrains action selection through a learned behavior model and bounded perturbations (Fujimoto et al., 2019); TD3+BC regularizes a deterministic actor toward logged actions (Fujimoto & Gu, 2021); CQL applies a conservative action-value objective (Kumar et al., 2020); and IQL combines expectile value fitting with advantage-weighted policy extraction (Kostrikov et al., 2022). SCORE targets spurious offline correlations (Deng et al., 2024), while FQL parameterizes the policy through conditional flow matching (Park et al., 2025). All implementations use the same state and action interfaces, data partition, update budget, and evaluator. Run manifests record architecture, optimizer, and preprocessing metadata.

A.2 AIRC IMPLEMENTATION

Networks and optimization. AIRC is implemented as a shared-representation, multi-branch policy network. The encoder feeds the state-relative innovation pathway, multi-horizon clinical-value heads, fold-specific response heads, and actor. AIRC uses an $11 \to 128 \to 128$ SiLU encoder and a $128 \to 128 \to 1$ actor. Six positive-output clinical-value heads predict the three channels at two time constants each. Four response heads cover the two long horizons and two data folds. The innovation network predicts a location and a softplus-transformed deviation scale with an additive lower bound of 0.03. Minibatches contain 256 transitions. AdamW optimizers update the representation, response heads, and actor separately; the actor updates every second global step. Target networks use a soft-update rate of 0.005. The online networks contain 130,445 trainable parameters; target copies add 67,974 parameters updated through soft synchronization. Deployment requires the 34,689 encoder and actor parameters.

Table 5: AIRC optimization and loss settings for the reported models.

Setting	Value
Encoder / actor hidden width	128 / 128
Clinical and response-head hidden width	64
Minibatch size	256
Representation / response learning rate	3×10^{-4} / 3×10^{-4}
Actor learning rate	1×10^{-4}
AdamW weight decay / gradient-norm cap	1×10^{-5} / 10
Target soft-update rate / actor-update period	0.005 / 2 steps
One-sided risk weight	1.25
Action-bound penalty weight λ_{bd}	5.00
Innovation location / scale weight	0.50 / 0.50
Low- / severe-glucose prediction weight	0.50 / 0.25

Representation supervision. The location head regresses Δa_t with smooth-L_1 loss. The scale head regresses $|\Delta a_t - \mathrm{sg}(\mu_\Delta(\boldsymbol{z}_t))|$ with the same loss, yielding a learned deviation scale. Clinical-value losses supervise the high, low, and severe channels. These tasks update the shared encoder. Response fitting uses detached representations, innovation estimates, and temporal-difference residuals. Actor fitting also detaches the evidence pathways.

Discount scales. Both datasets use high-glucose time constants of 240 and 360 minutes and low/severe time constants of 60 and 120 minutes. Equation 4 maps each physical duration to the dataset's nominal sampling interval:

Dataset	Nominal interval	High-glucose time constants	Low/severe time constants
SimGlucose	3 min	240, 360 min	60, 120 min
Loop	5 min	240, 360 min	60, 120 min

For a time constant τ, the discount is $\exp(-\Delta/\tau)$. Thus the two sampling rates yield different per-step discounts while preserving the same time scales in minutes.

The hyperglycemic signal is normalized by the 95th percentile of its unnormalized training values, with a lower bound of 0.5. A degenerate percentile is resolved sequentially with the training standard deviation and a unit scale. Bootstrap targets are capped at 256. The actor's innovation scale is restricted to $[0.03, 0.50]$. In Equation 2, $\epsilon = 10^{-4}$ implements the response-denominator floor. Response-head regularization is 10^{-4}. The moving response scale is initialized to 1, updated at rate 0.01, and bounded below by 10^{-4}.

Data partition and certificate. Episodes are assigned alternately to two folds. If the second fold contains less than 10% or more than 90% of the policy-fitting data, the assignment procedure instead alternates contiguous 256-transition blocks. Innovation validity is checked separately at episode

starts and discontinuities. Both folds use the shared representation and innovation model. Fold-specific disagreement continuously attenuates confidence through its denominator. Cross-horizon sign disagreement sets both directional output and confidence to zero.

Reference fidelity and execution. For normalized action bounds $(a_{\min}, a_{\max})$, $\ell_{\mathrm{bd}}(\widehat{a}) = [a_{\min} - \widehat{a}]_+^2 + [\widehat{a} - a_{\max}]_+^2$. The learned action is projected into the permitted patient-specific interval before conversion to physical basal delivery.

A.3 Computational complexity

Let B be the minibatch size; C_{E}, C_{N}, C_{V}, C_{R}, and C_{A} denote the per-example costs of the encoder, innovation network, one predictive head, one response head, and actor. With six predictive heads, two long horizons, and two folds, a global training update costs

$$\mathcal{O}(B\,(C_{\mathrm{E}} + C_{\mathrm{N}} + 6C_{\mathrm{V}} + 4C_{\mathrm{R}} + \mathbf{1}_{n \bmod K=0} C_{\mathrm{A}}))\,, \tag{9}$$

up to constant recomputation and backward-pass factors. Certificate aggregation costs $\mathcal{O}(B|\mathcal{T}_{\mathrm{H}}|F)$, while susceptibility aggregation costs $\mathcal{O}(B|\mathcal{T}_{\mathrm{L}}|)$. Treating the horizon and fold counts as constants, both scale linearly with B. Training memory is $\mathcal{O}(P + Bd)$ for model and optimizer parameters P and stored activations of width d, excluding storage of the offline dataset shared by every method.

Deployment discards the innovation, predictive, response, and target networks. Its cost is one $11 \to 128 \to 128$ encoder followed by one $128 \to 128 \to 1$ actor and scalar clipping. Omitting biases and activations, this path uses 34,304 multiply–accumulate operations per state and stores 34,689 parameters. It performs no action-grid search, critic evaluation, or certificate construction at inference time.

A.4 Clinical metrics and aggregation

The offline-transition and evaluation rewards follow the Magni glucose-risk function used in prior offline blood-glucose control work (Emerson et al., 2023):

$$\begin{aligned} R_{\mathrm{M}}(g) &= 10\left[3.5506\left((\ln \bar{g})^{0.8353} - 3.7932\right)\right]^2, \\ \bar{g} &= \max\{1, g\}, \qquad r_t = -R_{\mathrm{M}}(g_{t+1}). \end{aligned} \tag{10}$$

The numerical lower bound keeps the logarithm defined and is inactive over the physiological glucose range. Cumulative reward is the sum of r_t over the observed evaluation trajectory.

Time-in-range percentages use the thresholds in Section 5.1. Let STAR denote the percentage above $250\,\mathrm{mg/dL}$. The glycemia risk index is computed for each profile as (Klonoff et al., 2023)

$$\mathrm{GRI} = \min\{100,\, 3\,\mathrm{STBR} + 2.4(\mathrm{TBR} - \mathrm{STBR}) + 1.6\,\mathrm{STAR} + 0.8(\mathrm{TAR} - \mathrm{STAR})\}. \tag{11}$$

We average profile-level GRI after applying the upper bound, preserving each profile's contribution to the cohort estimate.

The reporting unit is the profile. Numerical summaries are presented to two decimals, and relative comparisons use the tabulated means. Let $T_j = 1$ when evaluation rollout j triggers clinical termination and $T_j = 0$ otherwise. For N_{eval} rollouts, $\mathrm{Failure}(\%) = 100 N_{\mathrm{eval}}^{-1} \sum_{j=1}^{N_{\mathrm{eval}}} T_j$. Each rollout therefore contributes one binary outcome, irrespective of its observed length. The main comparison and ablations share the same full-model summary for each cohort. Source summaries and archived run records are retained with the analysis materials.

Table 6: Complete-case sensitivity to clinical termination. Positive cumulative-reward differences favor AIRC; intervals are profile-paired bootstrap 95% intervals.

Baseline	Mean Δ reward	95% interval
IQL	+3647.91	1012.93–6478.60
BCQ	+4669.93	1595.48–8357.60
TD3+BC	+6155.16	1585.13–13115.42
CQL	+6163.08	1755.29–12009.07
SCORE	+5757.89	3028.95–8654.07
FQL	+6749.74	3545.29–10306.89

A.5 Ablation interventions

The completed ablation study changes one specified component or a defined combination while retaining the remaining training and evaluation protocol.

A1: w/o direction. Remove the directional loss by setting $\lambda_{\mathrm{dir}} = 0$.

A2: w/o risk penalty. Remove the one-sided risk penalty by setting $\lambda_{\mathrm{safe}} = 0$.

A3: w/o innovation scale. Remove the scale-regression loss and the learned-scale denominators from both response residualization and the actor shift. Location residualization remains active.

A4: w/o both guidance terms. Set both λ_{dir} and λ_{safe} to zero. Reference fidelity, dose bounds, and representation training remain.

The age-stratified tables retain all variants and clinical endpoints, with cumulative reward as the primary comparison.

A.6 Loop trend-response protocol

Data partition and held-out policy evaluation. The analysis uses de-identified observational records processed into the same 11-dimensional state interface. IQL is the single baseline comparator because it had the highest mean cumulative reward among the six SimGlucose baselines. Each participant record is ordered chronologically; the first 80% is used for policy training and training-derived normalization, while the final 20% is held out. Participant-specific AIRC and IQL policies receive the same training partition and update budget. The reported models are selected without access to held-out records and evaluated on exactly matched held-out states; held-out records do not influence model selection.

Trend intervention. Each held-out state yields five recent glucose trends at -1, -0.5, 0, 0.5, and $1\,\mathrm{mg/dL/min}$. Current glucose and the remaining state variables stay at their observed values, while glucose at $t - 30$ minutes is set to $g_t - 30m$. A state enters the analysis when all perturbed versions fall inside the participant-specific 1st–99th percentile support estimated from training data and all glucose-history channels remain within 40–$400\,\mathrm{mg/dL}$. This retains 181,844 of 242,973 held-out states overall.

Table 7: Trend-response eligibility of held-out states from the Loop observational dataset.

Cohort	Held-out states	Eligible states	Retained (%)
Overall	242,973	181,844	74.84
Adult	87,051	62,007	71.23
Adolescent	76,806	59,308	77.22
Child	79,116	60,529	76.51

Response metrics. The four metrics describe distinct properties of the policy response to the same five-level trend intervention. Spearman's ρ measures within-state monotonicity between trend and action. Dose–response R^2 summarizes the linearity of actions across the five imposed trend levels. Directional consistency records whether the action under $m = 1$ exceeds that under $m = -1$. Endpoint contrast measures the participant-level median action difference between these endpoints, normalized by the interquartile range of that participant's training actions. The first three metrics are averaged within participant; all four then receive equal participant weight in cohort aggregation.

Uncertainty. The participant is the independent unit. State-level measurements are aggregated within participant before method differences are formed. Paired participant-level percentile intervals therefore quantify between-participant variation while preserving the AIRC–IQL match.

A.7 AGE-STRATIFIED LOOP COMPARISON

The paired AIRC–IQL differences in trend–action correlation, dose–response R^2, and directional consistency were positive in every age group (Table 8). Endpoint-contrast estimates were also positive across groups; their groupwise intervals overlapped zero, whereas the participant-balanced overall interval remained above zero.

Table 8: Age-stratified paired AIRC–IQL differences under the matched Loop protocol. Brackets give participant-level 95% confidence intervals; positive values favor AIRC.

Cohort	Trend Spearman ρ	Dose–response R^2	Directional consistency (pp)	Endpoint contrast in action IQR
Adult	$0.103\ [0.067, 0.142]$	$0.091\ [0.057, 0.132]$	$2.76\ [1.48, 3.99]$	$0.086\ [-0.111, 0.295]$
Adolescent	$0.141\ [0.103, 0.176]$	$0.108\ [0.081, 0.137]$	$4.58\ [3.23, 5.79]$	$0.084\ [-0.015, 0.174]$
Child	$0.173\ [0.115, 0.238]$	$0.108\ [0.073, 0.149]$	$6.21\ [3.84, 9.11]$	$0.087\ [-0.011, 0.186]$

B AGE-STRATIFIED SIMGLUCOSE RESULTS

The three age groups are reported separately, with reward retained as the primary outcome. All subgroup tables follow the metric definitions, aggregation, and formatting conventions of Tables 1 and 2.

Table 9: Adult subgroup: AIRC and baselines.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC	**-15817.77**	**80.45**	**0.00**	**0.00**	**19.55**	**16.45**	0.00
IQL	-16709.39	74.44	**0.00**	**0.00**	25.56	21.70	0.00
BCQ	-17289.32	73.42	0.09	**0.00**	26.49	22.74	0.00
TD3+BC	-31347.43	68.21	5.49	3.08	26.30	38.56	10.00
CQL	-18306.66	70.16	0.01	**0.00**	29.83	26.37	0.00
SCORE	-19559.15	64.92	**0.00**	**0.00**	35.08	30.44	0.00
FQL	-20398.82	62.12	**0.00**	**0.00**	37.88	32.92	0.00

Table 10: Adolescent subgroup: AIRC and baselines.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC	**-34564.37**	65.08	**0.00**	**0.00**	34.92	39.09	0.00
IQL	-41043.86	**65.50**	2.81	0.53	**31.69**	40.39	0.00
BCQ	-36435.74	65.36	1.43	0.28	33.21	**38.56**	0.00
TD3+BC	-37220.48	63.50	0.43	**0.00**	36.07	40.32	0.00
CQL	-41152.95	64.41	3.25	1.06	32.34	42.50	0.00
SCORE	-38613.01	57.90	**0.00**	**0.00**	42.10	46.16	0.00
FQL	-37939.01	58.04	**0.00**	**0.00**	41.96	45.70	0.00

Table 11: Child subgroup: AIRC and baselines.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC	**-46617.30**	**62.90**	**0.00**	**0.00**	**37.10**	**41.71**	0.00
IQL	-50189.91	58.85	0.20	0.01	40.95	46.87	0.00
BCQ	-57284.18	56.64	0.60	0.01	42.76	49.29	0.00
TD3+BC	-50678.45	57.88	0.37	0.06	41.75	48.34	0.00
CQL	-56029.07	59.03	0.30	0.05	40.67	47.40	0.00
SCORE	-56100.97	51.24	0.21	0.02	48.55	55.32	0.00
FQL	-58910.81	47.15	0.37	0.07	52.48	60.03	0.00

C AGE-STRATIFIED ABLATION RESULTS

These tables use the interventions defined in Appendix A.5. The full model achieves the highest reward in all three groups, while individual clinical metrics reveal the different effects of removing direction, risk weighting, or scale normalization.

Table 12: Adult subgroup: component ablations.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC (full)	**-15817.77**	80.45	**0.00**	**0.00**	19.55	**16.45**	0.00
w/o direction	-25254.32	52.37	**0.00**	**0.00**	47.63	42.09	0.00
w/o risk penalty	-28867.80	**81.53**	6.02	1.40	**12.45**	25.58	10.00
w/o innovation scale	-22075.62	80.65	0.67	0.03	18.68	17.25	0.00
w/o both guidance terms	-19563.06	65.07	0.01	**0.00**	34.93	30.48	0.00

Table 13: Adolescent subgroup: component ablations.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC (full)	**-34564.37**	65.08	**0.00**	**0.00**	34.92	39.09	0.00
w/o direction	-43221.44	52.42	**0.00**	**0.00**	47.58	52.86	0.00
w/o risk penalty	-41640.39	**68.86**	2.58	0.23	**28.55**	37.27	0.00
w/o innovation scale	-36117.65	66.49	0.11	**0.00**	33.40	**36.87**	0.00
w/o both guidance terms	-37243.74	58.95	**0.00**	**0.00**	41.05	44.68	0.00

Table 14: Child subgroup: component ablations.

Method	Reward↑	TIR↑	TBR↓	STBR↓	TAR↓	GRI↓	Failure (%)
AIRC (full)	**-46617.30**	**62.90**	**0.00**	**0.00**	37.10	**41.71**	0.00
w/o direction	-64501.07	38.83	**0.00**	**0.00**	61.17	67.55	0.00
w/o risk penalty	-87686.52	55.77	20.64	14.59	**23.59**	64.64	30.00
w/o innovation scale	-65920.54	57.31	6.16	2.27	36.53	57.51	10.00
w/o both guidance terms	-59099.35	47.56	0.11	**0.00**	52.33	59.08	0.00