

Research paper

A safety-supervised dual-agent reinforcement learning framework for pediatric automated insulin delivery

Giada Lops^{a,b,*}, Luca De Cicco^{a,b}, Saverio Mascolo^{a,b}

^a Department of Electrical and Information Engineering, Polytechnic University of Bari, Via Orabona 4, Bari, 70125, Italy

^b Quavlive S.r.l., c/o Department of Electrical and Information Engineering, Polytechnic University of Bari, Via Orabona 4, Bari, 70125, Italy

ARTICLE INFO

Keywords:

Reinforcement learning
Automated insulin delivery
Safe reinforcement learning
Dual-Agent control architecture
Biomedical control systems
Pediatric glucose regulation
Maskable PPO

ABSTRACT

This study presents a safety-supervised dual-agent reinforcement-learning framework for pediatric automated insulin delivery. The architecture combines a primary Maskable Proximal Policy Optimization controller with predictive correction, learned insulin-command modulation, and deterministic runtime constraints. The framework was evaluated *in silico* using ten virtual pediatric subjects, five independent training seeds, and 200 matched stochastic 24-h scenarios per subject and seed.

Relative to an information-matched single-agent Maskable PPO baseline, the dual-agent framework increased time in range (TIR) in all ten subjects, raising the cohort mean from 42.86% to 69.96% (paired difference: 27.10 percentage points; 95% CI: [20.80, 33.40]; $p < .001$). Cohort-average time below range decreased from 25.56% to 11.99%, while time above range decreased from 31.58% to 18.05%. The Blood Glucose Risk Index decreased from 22.31 to 8.82, corresponding to a 60.5% relative reduction ($p = .007$), and inter-subject TIR variability decreased from 33.90% to 13.11%. Component-wise ablation showed complementary contributions from learned modulation and deterministic runtime constraints, while subject-specific analyses demonstrated adaptive responses across heterogeneous glycemic profiles.

Overall, the results provide simulation-based evidence that hierarchical integration of learned and deterministic supervision can improve glycemic regulation and consistency across the evaluated pediatric profiles and stochastic meal scenarios. The proposed framework offers a modular basis for further investigation of safety-aware reinforcement learning in automated insulin delivery.

1. Introduction

Type 1 diabetes mellitus (T1D) is a chronic autoimmune disease characterized by the progressive destruction of pancreatic β -cells and the resulting deficiency of endogenous insulin. People with T1D therefore require lifelong exogenous insulin therapy. Glycemic regulation is commonly assessed through the proportion of time spent within the target glucose range of 70.00 mg/dL to 180.00 mg/dL, referred to as *Time in Range* (TIR). Current recommendations generally target TIR above 70.00 %, time below range (TBR; < 70.00 mg/dL) below 4.00 %, and time below 54.00 mg/dL below 1.00 % [1].

Achieving these targets remains particularly challenging in pediatric populations. Children with T1D exhibit pronounced physiological variability, age-dependent insulin sensitivity, heterogeneous carbohydrate absorption, and rapid pre- and postprandial changes in glucose–insulin dynamics [2]. From a control perspective, *Automated Insulin Delivery* (AID) must therefore regulate a nonlinear physiological system subject to delayed sensing and actuation, uncertain disturbances, and substan-

tial inter- and intra-subject variability. *Continuous Glucose Monitoring* (CGM) measurements are affected by sensing delay and noise, while subcutaneous insulin delivery is characterized by comparatively slow and variable pharmacokinetics [3,4]. Meals, physical activity, stress, and variations in insulin sensitivity further complicate closed-loop regulation [3].

Commercial hybrid closed-loop systems combine CGM, insulin pumps, and control algorithms to automatically adjust insulin delivery while generally retaining some degree of user interaction, particularly for meal-related dosing and physical activity [5,6]. Systems such as MiniMed 780G, Control-IQ, and Omnipod 5 have demonstrated improvements in TIR and reductions in hypoglycemic exposure compared with conventional insulin therapy [7–9], but full automation remains constrained by subcutaneous insulin delay, CGM imperfections, disturbance uncertainty, and dependence on user-provided information [3]. These challenges are especially relevant in pediatric AID, where changing physiology and large glycemic variability can make target-range regulation more difficult [10].

* Corresponding author.

E-mail addresses: g.lops@phd.poliba.it (G. Lops), luca.decicco@poliba.it (L. De Cicco), saverio.mascolo@poliba.it (S. Mascolo).

<https://doi.org/10.1016/j.rineng.2026.112962>

Received 4 June 2026; Received in revised form 6 September 2026; Accepted 11 September 2026

Available online 13 September 2026

2590-1230/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1.1. Automated insulin delivery: state of the art

Classical control approaches, particularly *proportional–integral–derivative* (PID) control and *model predictive control* (MPC), have been extensively investigated for AID and represent the two dominant algorithmic philosophies in commercial and research systems [3]. PID provides a comparatively simple feedback structure requiring low computational effort, whereas MPC incorporates model-based prediction and explicit optimization of future control actions at the cost of higher computational burden and dependence on an accurate patient model [3]. Their performance can nevertheless depend on controller tuning, model accuracy, disturbance information, and the ability of the underlying representation to capture time-varying physiological dynamics. A systematic review and meta-analysis of MPC-based artificial pancreas systems reported a significant increase in overnight TIR relative to conventional insulin therapy, while also noting substantial heterogeneity in reported outcomes across studies [11].

Limitations related to delayed insulin action, CGM imperfections, infusion-system faults, and dependence on meal or activity information remain relevant across both PID- and MPC-based systems, particularly in pediatric subjects characterized by greater glycemic variability and changing insulin requirements [8,12–14].

These limitations motivate continued investigation of adaptive control strategies capable of handling nonlinear physiological dynamics and stochastic disturbances while retaining explicit mechanisms for constraining insulin delivery.

1.2. Reinforcement learning: background and control applications

Reinforcement Learning (RL) is a data-driven control paradigm in which an agent interacts with an environment and learns a policy that maximizes expected cumulative reward [15]. The interaction is commonly formulated as a *Markov Decision Process* (MDP), characterized by a state space S , an action space A , transition dynamics, and a reward function. At each decision step, the agent observes the current state and selects an action according to a policy $\pi(a|s)$, which is iteratively optimized from the resulting state transitions and rewards.

In control applications, RL is attractive because it can approximate nonlinear decision policies, account for delayed effects, and optimize long-horizon objectives under stochastic dynamics, without requiring an explicit dynamic model of the controlled system. Unlike model predictive control, a trained model-free RL policy does not solve an online optimization problem at every control step. Its performance nevertheless depends strongly on the training environment, reward formulation, exploration strategy, and distribution of operating conditions encountered during policy optimization.

These characteristics are particularly relevant to glucose regulation, where insulin actions have delayed physiological effects and the controller must operate under uncertain disturbances and heterogeneous subject dynamics. At the same time, unrestricted exploratory learning is unsuitable for direct application in human subjects because inappropriate insulin actions may produce clinically unacceptable glucose excursions. Simulation-based offline training therefore provides an appropriate engineering framework for developing and evaluating RL-based AID strategies before any translational investigation can be considered.

1.3. Reinforcement learning for automated insulin delivery

RL has increasingly been investigated for closed-loop glucose regulation. Baldisseri et al. [16] developed a DDPG-based controller for a glucose–insulin model and compared it with PID and nonlinear MPC. Atanasious et al. [17] and Lops et al. [18] investigated deep-RL approaches using the UVa/Padova simulation environment, including cross-subject behavior without direct meal information. Yamagata et al. [19] proposed a model-based RL framework combining learned glucose prediction and predictive control, while Zhu et al. [20] and Vi-

roonluecha et al. [21] investigated deep value-based and recurrent PPO strategies for automated glucose regulation.

Safety-oriented RL for AID has been addressed through several complementary, individually-established mechanisms. Asymmetric and restricted-action-space reward formulations that penalize hypoglycemic excursions more severely than hyperglycemic ones, combined with a discretized and constrained action space, have been used to shape value-based and recurrent PPO policies [20]. Predictive information has been incorporated through model-based RL architectures that couple a learned glucose predictor with a predictive control layer [19]. More recently, supervisory correction through test-time action shielding, in which candidate RL actions are filtered by a learned dynamics model before delivery, has been shown to restore safety margins that training-time constraints alone fail to preserve under distribution shift [22]. Action masking, reward shaping, predictive dosing, and runtime supervision are therefore not individually novel; each has been investigated, at least in part, in the works cited above. A remaining methodological question is how these mechanisms can be integrated while keeping regulation-oriented learning, learned supervisory modulation, and deterministic command constraints explicitly distinguishable, a distinction that, to our knowledge, has not been systematically addressed in the pediatric AID literature.

This distinction is particularly important when interpreting performance improvements. If a learned controller is combined with deterministic suspension, dose limitation, or cumulative-dose constraints, improved closed-loop outcomes cannot automatically be attributed to the learned policy itself. A component-wise evaluation is therefore required to identify the contribution of each layer.

1.4. Research gap and contribution of this work

The present work builds on our previous investigation of Maskable PPO for automated insulin delivery [23]. That study already incorporated invalid action masking into a single-policy Maskable PPO controller and evaluated the approach on three virtual physiological profiles representing an adult, an adolescent, and a child under stochastic variations in meal timing and carbohydrate intake. Consequently, neither invalid action masking nor stochastic meal evaluation is claimed here as a new contribution.

The present study extends that framework in several substantive ways. First, action admissibility is conditioned jointly on the current glucose concentration and its rate of change, rather than predominantly on glucose level alone. Second, a short-horizon predictive supervisory signal is introduced to condition candidate insulin delivery. Third, a separately trained PPO-based policy, termed the *Safety Adjuster*, is added to learn bounded modulation of the command generated by the primary Maskable PPO controller. Finally, deterministic low-glucose suspension, predictive dose limitation, and short-term cumulative insulin constraints are applied after the learned modulation and remain functionally independent of the second RL policy.

This raises the question of whether separating regulation-oriented learning from a second, independently trained supervisory policy yields measurable benefits over a single-agent controller equipped with the same deterministic constraints, or whether the deterministic layer alone accounts for most of the observed improvement. Addressing this question requires evaluating each component individually, which motivates the component-wise ablation presented in Section 3.

The use of two separately trained policies is motivated by the separation between regulation-oriented optimization and supervisory command modulation. The primary Maskable PPO controller first learns the glucose-regulation task. Its parameters are then frozen, after which the Safety Adjuster learns how to modify the output of that fixed regulation policy using additional glycemic, predictive, and insulin-exposure information. This sequential organization avoids simultaneous co-adaptation of the two learned components and permits the contribution of the secondary policy to be analyzed independently. The additional agent also

increases training and computational complexity relative to a single-policy controller; its usefulness must therefore be demonstrated empirically rather than assumed from the architectural design.

Importantly, the Safety Adjuster is distinct from the deterministic runtime supervisory mechanisms. The Safety Adjuster is a learned command-modulation policy, whereas low-glucose suspension, predictive dose limitation, and cumulative-dose enforcement are fixed control rules applied after learned modulation. The predictive dose limitation is described as *barrier-inspired* only in the engineering sense that the admissible insulin command decreases as a predicted lower glucose boundary is approached. It is not formulated as a formally verified control barrier function, and no forward-invariance or formal safety guarantee is claimed.

The main contributions of this work are summarized as follows:

- A sequential two-policy RL architecture is developed for pediatric AID, comprising a primary Maskable PPO regulation policy and a separately trained PPO-based Safety Adjuster responsible for bounded command modulation.
- The glucose-dependent masking mechanism introduced in our previous work is extended by incorporating glucose rate-of-change information, allowing insulin-action admissibility to depend jointly on current glycemia and its short-term trend.
- A low-complexity short-horizon predictive supervisory mechanism is integrated together with deterministic low-glucose suspension, predictive dose limitation, and short-term cumulative insulin constraints.
- Learned supervisory modulation is explicitly separated from deterministic runtime constraints, enabling their respective contributions to be analyzed independently.
- A two-stage optimization strategy is adopted in which the primary regulation policy is frozen before training the Safety Adjuster, thereby separating regulation-oriented learning from supervisory-policy learning.
- A component-wise ablation analysis is performed to isolate the contributions of the primary policy, action masking, predictive supervision, deterministic runtime constraints, and the learned Safety Adjuster.
- The framework is evaluated across the complete ten-subject pediatric virtual cohort using multiple independent RL training seeds and held-out stochastic 24-h meal scenarios, extending the pediatric evaluation of our previous work.
- The proposed architecture is compared primarily with a single-agent Maskable PPO baseline. PID and MPC are additionally considered as contextual classical-control references, while differences in controller information availability are explicitly acknowledged in the interpretation.

The study is entirely *in silico*. Its objective is to investigate whether the proposed decomposition of regulation learning, learned supervisory modulation, and deterministic runtime constraints improves selected glycemic and risk-related outcomes relative to a single-agent (SA) Maskable PPO controller under controlled pediatric simulation conditions. The term *safety-supervised* refers to this layered engineering architecture and should not be interpreted as evidence of clinical safety, formal closed-loop safety, or translational readiness.

The remainder of the paper is organized as follows. Section 2 describes the proposed architecture, predictive and deterministic supervisory mechanisms, sequential training procedure, and experimental protocol. Section 3 presents the *in silico* evaluation, including glycemic performance, training-seed variability, predictive accuracy, component-wise ablation, sensitivity analysis, and comparison with reference controllers. Section 4 discusses the interpretation, robustness, scalability, limitations, and implications of the findings. Section 5 summarizes the main conclusions and their scope.

2. Methods

This work investigates a safety-supervised dual-agent reinforcement learning architecture for AID. The proposed framework comprises three functional components, of which two are independently trained reinforcement learning policies and one is a deterministic, rule-based layer; together they separate regulation-oriented learning, learned supervisory modulation, and deterministic runtime constraints into distinct functional layers. The overall control organization is illustrated in Fig. 1.

The controller comprises three main components:

- a *primary policy*, based on *Maskable Proximal Policy Optimization* (Maskable PPO) [24], which maps the current glucose observation to a candidate discrete insulin micro-bolus command u_{cmd} ;
- a second learned policy, referred to as the *Safety Adjuster*, which modulates the command generated by the primary policy according to current glycemia, glucose trend, short-horizon prediction, and recent insulin exposure;
- a deterministic runtime supervisory layer, which independently applies low-glucose suspension, predictive dose limitation, and cumulative insulin-delivery constraints before the command is delivered to the virtual subject.

The term *dual-agent* (DA) refers specifically to the two independently optimized reinforcement learning policies, the primary Maskable PPO policy and the Safety Adjuster, which have distinct state-action representations and control objectives. The primary policy is responsible for regulation-oriented insulin selection, whereas the Safety Adjuster learns context-dependent modulation of the command it receives from the primary policy. The deterministic runtime supervisory layer is not learned by, or part of, either policy: it applies fixed physiological safety rules and remains active independently of the Safety Adjuster's selected action. This distinction is maintained throughout the remainder of the paper: only the primary policy and the Safety Adjuster are referred to as *agents*, while low-glucose suspension, predictive dose limitation, and cumulative insulin-delivery constraints are referred to as *deterministic constraints* or the *runtime supervisory layer*.

The two policies are trained sequentially in two dedicated environments, denoted by $\mathcal{E}_1$ and $\mathcal{E}_2$. Environment $\mathcal{E}_1$ is used to optimize the primary Maskable PPO policy and contains the physiological simulator, the discrete insulin action space, state-dependent action masking, and the predictive hyperglycemia-correction mechanism described below. Once Phase 1 training is completed, the parameters of the primary policy are frozen. Environment $\mathcal{E}_2$ is subsequently used to optimize the Safety Adjuster while the primary regulation policy remains fixed.

This sequential design implements, at the technical level, the two-stage separation between regulation-oriented and supervisory learning motivated in Section 1. In practice, the Safety Adjuster does not relearn the complete insulin-delivery problem, but instead learns how to modify the output of the already-optimized primary policy. This separation reduces simultaneous co-adaptation and makes the contribution of learned modulation accessible through component-wise ablation. It nevertheless increases training, implementation, and validation complexity relative to a single-policy controller. The computational platform and observed training time are reported in Section 2.4, whereas the performance contribution of each component is evaluated in Section 3.5.

2.1. Maskable PPO and invalid action masking

The primary policy is trained using Maskable PPO [24], an extension of Proximal Policy Optimization (PPO) [25] for discrete-action problems in which action admissibility depends on the current state. In conventional reinforcement learning, a policy may initially explore actions that are undesirable before learning to avoid them through the reward signal. Such unconstrained exploration is particularly problematic in insulin-delivery control, where administration of insulin during low or rapidly decreasing glucose conditions can increase simulated hypoglycemic risk.

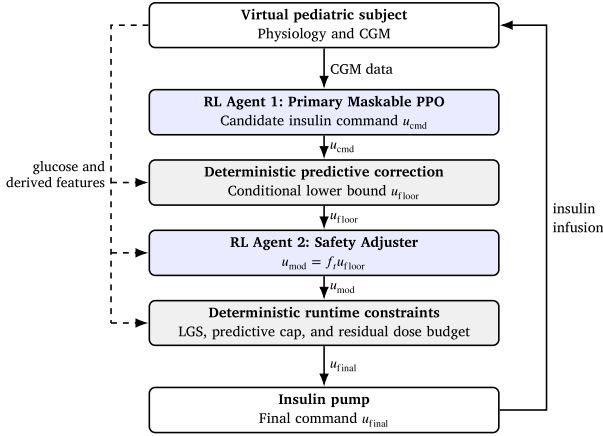


Fig. 1. High-level closed-loop organization of the proposed safety-supervised controller. CGM information from the virtual pediatric subject is processed by the primary Maskable PPO policy, a deterministic predictive-correction stage, the learned Safety Adjuster, and deterministic runtime constraints. These stages respectively produce u_{cmd} , u_{floor} , u_{mod} , and the final insulin command u_{final} . Dashed arrows indicate glucose and derived physiological information used by the supervisory components.

Invalid action masking incorporates state-dependent admissibility information directly into the action-selection process. At each decision step, actions classified as inadmissible are assigned zero probability before sampling. Consequently, the policy optimizes its probability distribution only over the subset of currently admissible insulin actions.

The application of invalid action masking to AID was previously investigated in our earlier work [23], where admissibility depended on G_t alone. The present formulation extends this mechanism to depend jointly on G_t and the estimated glucose rate of change $\dot{G}_t$, allowing the controller to distinguish the same glucose concentration observed during falling, approximately stable, or rising trajectories.

2.1.1. Discrete insulin action space

The primary policy operates on the discrete action-index set $\mathcal{A} = \{0, 1, \dots, n_T - 1\}$, where $n_T = 21$ denotes the number of available insulin-action levels.

The action space includes both zero insulin and a predefined maximum primary-policy command $u_{max}^{(prim)}$. The action resolution is therefore $\Delta a = \frac{u_{max}^{(prim)}}{n_T - 1}$.

Action index $a \in \mathcal{A}$ is mapped to the corresponding insulin command according to $u_{cmd}(a) = a \Delta a$, $a \in \{0, \dots, n_T - 1\}$.

Consequently, $a = 0$ corresponds to zero insulin and $a = n_T - 1 = 20$ corresponds exactly to the maximum command. In the present implementation, $u_{max}^{(prim)} = 0.40$ U, resulting in $\Delta a = 0.02$ U.

The same action-to-dose mapping is used during training, validation, ablation, sensitivity analysis, and final evaluation.

2.1.2. Glucose rate-of-change estimation

At decision step t , the glucose rate of change is estimated from consecutive glucose observations according to

$$\dot{G}_t = \frac{G_t - G_{t-1}}{\Delta t}, \quad (1)$$

where $\Delta t = 3.00$ min is the fixed simulation and controller sampling interval.

The present implementation assumes regularly sampled glucose measurements provided by the simulation environment. Missing CGM observations, timestamp loss, sensor dropout, and irregular sampling intervals are therefore not reconstructed within the controller.

If one of these conditions occurred in a real implementation, Eq. (1) could not be applied directly without an additional estimation or fallback mechanism. Possible implementations include calculation from the

most recent valid timestamped observations, filtering over a short historical glucose window, or transition to a predefined conservative control mode. These mechanisms are not evaluated in the present study and are therefore considered limitations rather than implemented safety features.

2.1.3. State-dependent action masking

At each decision step, the mask determines the subset of admissible actions $\mathcal{A}_v(G_t, \dot{G}_t)$:

$$\mathcal{A}_v(G_t, \dot{G}_t) = \begin{cases} \{0\}, & G_t < G_{hp} \\ & \vee (G_t < G_{sl} \wedge \dot{G}_t < \dot{G}_{th}), \\ \{a \in \mathcal{A} \mid a \geq a_{md}\}, & G_t > G_{ml} \\ & \wedge \dot{G}_t > \dot{G}_{rs}, \\ \{a \in \mathcal{A} \mid a \geq a_{hf}\}, & G_t > G_{sv}, \\ \mathcal{A}, & \text{otherwise.} \end{cases} \quad (2)$$

The corresponding binary mask is

$$M(a \mid G_t, \dot{G}_t) = \begin{cases} 1, & a \in \mathcal{A}_v(G_t, \dot{G}_t), \\ 0, & a \notin \mathcal{A}_v(G_t, \dot{G}_t). \end{cases} \quad (3)$$

Actions for which $M(a \mid G_t, \dot{G}_t) = 0$ are excluded from the policy distribution before sampling.

When glucose is below G_{hp} , or when glucose is below G_{sl} and decreasing faster than $\dot{G}_{th}$, the only admissible action is zero insulin. Conversely, when glucose is elevated and rising, the lower portion of the action space is excluded. When $G_t > G_{sv}$, only action indices equal to or greater than a_{hf} remain admissible. In all other conditions, the complete action set remains available.

These rules represent physiologically motivated engineering constraints on policy exploration. They do not establish formal closed-loop safety.

All masking thresholds and rate-of-change limits used in Eq. (2) are fixed prior to training and remain constant throughout each simulated episode and across the entire evaluation protocol; none of these quantities is updated online at runtime.

2.1.4. Controller parameters and validation strategy

The masking and supervisory parameters used throughout the study are summarized in Table 1. Parameters are divided conceptually into two groups. The first group consists of the standard glycemic target boundaries (70 and 180 mg/dL) recommended by international consensus guidance [1,26], which delimit the reward and evaluation metrics. The second group consists of engineering design parameters, including all masking thresholds in Table 1, the predictive activation thresholds S_1, S_2 , and the proportional correction gains, which have no established clinical reference value and were instead selected empirically through simulation on the validation set described below.

Training, parameter validation, and final testing are performed using mutually exclusive stochastic meal realizations generated from non-overlapping random-seed sets. The validation scenarios are not used for reinforcement-learning policy optimization, and the final test scenarios are not used for either training or parameter selection.

Supervisory thresholds, proportional gains, refractory settings, and other empirically selected controller parameters are determined using the validation set. Once the nominal configuration is selected, all such parameters are frozen before final evaluation.

The same parameter set is shared across the ten virtual pediatric subjects and is not retuned on an individual-subject basis. In contrast, the reinforcement-learning policies themselves are trained separately for each virtual subject.

Table 1

Threshold and supervisory-control parameters used in the proposed framework.

Symbol	Description	Value
<i>Glycemic reference and predictive supervision</i>		
G^*	Nominal glucose reference	112 mg/dL
S_1, S_2	Mild/severe predicted hyperglycemia thresholds	220, 250 mg/dL
K_p^{lo}, K_p^{hi}	Low/high proportional correction gains	0.32, 0.52
u_{min}	Minimum enforced correction dose	0.02 U
τ_{ref}	Refractory interval	20 min
D_{30}^{max}	Maximum cumulative 30-min insulin dose	0.8 U
G_{pred}	Predictive lower-glucose constraint threshold	98 mg/dL
<i>State-dependent action masking</i>		
G_{hp}	Low-glucose masking threshold	90 mg/dL
G_{sl}	Slope-dependent masking threshold	120 mg/dL
G_{ml}	Mild hyperglycemia onset	160 mg/dL
G_{sv}	Severe hyperglycemia onset	200 mg/dL
$\dot{G}_{th}$	Rapid descent-rate threshold	-0.5 mg/dL/min
$\dot{G}_{rs}$	Rising-slope threshold	0.2 mg/dL/min
a_{nd}	Minimum action index during rising hyperglycemia	10
a_{hf}	Minimum action index during severe hyperglycemia	15

Because the same ten physiological simulator subjects are used during methodological development and final evaluation, this protocol provides scenario-level independence but does not constitute validation on an independent physiological cohort. Potential cohort-specific overfitting therefore remains a limitation of the study.

2.1.5. $\mathcal{E}_1$ implementation

Environment $\mathcal{E}_1$ was implemented using the Python *Simglucose* simulator [27], which provides a research-oriented implementation of the UVa/Padova physiological simulator [28]. The simulator contains 30 virtual subjects divided into pediatric, adolescent, and adult cohorts of ten subjects each.

Only the pediatric cohort is considered in the present study. The ten virtual pediatric subjects exhibit heterogeneous physiological profiles, including differences in age, body weight, carbohydrate ratio, insulin sensitivity, total daily insulin requirements, glucose absorption dynamics, and glucose distribution volume. Their simulator parameters are summarized in Table 2.

2.1.6. Short-horizon predictive correction

Environment $\mathcal{E}_1$ additionally incorporates a deterministic predictive correction intended to reduce projected hyperglycemic excursions. This mechanism is distinct from both learned policies.

A 30-min look-ahead glucose estimate is computed using first-order extrapolation of the current glucose trend:

$$\hat{G}_{t+30} = G_t + 30 \dot{G}_t. \quad (4)$$

This approximation is deliberately simple and is not intended to represent a complete physiological glucose-prediction model. Its role is to provide a low-complexity short-horizon signal for supervisory decision making.

When the predicted glucose exceeds the predefined hyperglycemic activation threshold, the positive prediction error relative to the nominal glucose reference is

$$e_t = \max(0, \hat{G}_{t+30} - G^*). \quad (5)$$

The associated correction dose is

$$u_{corr,t} = \max(u_{min}, K_p \frac{e_t}{CF}), \quad (6)$$

where K_p depends on the predicted hyperglycemic

$$K_p = \begin{cases} K_p^{lo}, & S_1 < \hat{G}_{t+30} \leq S_2, \\ K_p^{hi}, & \hat{G}_{t+30} > S_2. \end{cases} \quad (7)$$

No predictive correction is enforced when

$$\hat{G}_{t+30} \leq S_1. \quad (8)$$

In this in-silico study, CF is obtained from the fixed physiological parameter set of each virtual subject. It is used only in the deterministic correction mechanism and is not included in the primary-policy observation or reward. Treating CF as known and time invariant is a simulation-specific simplification; in a clinical implementation it would require estimation and periodic adaptation.

A refractory interval τ_{ref} prevents repeated enforced corrections within a short period. This mechanism is intended to reduce rapid repeated dosing and insulin stacking in view of the delayed pharmacodynamics of subcutaneous insulin [4,29,30].

The candidate command generated by the primary policy is then conditioned according to

$$u_{floor,t} = \max(u_{cmd,t}, u_{corr,t}), \quad (9)$$

whenever the predictive correction is active. Otherwise,

$$u_{floor,t} = u_{cmd,t}. \quad (10)$$

The predictive correction therefore acts only as a lower bound under projected hyperglycemic conditions. It is not a hypoglycemia-protection mechanism.

The thresholds S_1, S_2 , the gains K_p^{lo}, K_p^{hi} , the minimum dose u_{min} , and the refractory interval τ_{ref} are, like the masking thresholds in Section 2.1.3, fixed prior to training and are not updated at runtime or across decision steps within an episode.

2.1.7. Validation of the 30-min glucose forecast

Because the supervisory architecture uses the linear prediction in Eq. (4), the approximation is evaluated quantitatively on held-out stochastic scenarios.

For each eligible decision time t , the forecast residual is

$$\varepsilon_t = \hat{G}_{t+30} - G_{t+30}, \quad (11)$$

where G_{t+30} denotes the glucose concentration generated by the physiological simulator 30 min after the prediction time.

Prediction accuracy is quantified using the mean absolute error $MAE = \frac{1}{N} \sum_{t=1}^N |\varepsilon_t|$, the root-mean-square error $RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N \varepsilon_t^2}$, and the signed prediction bias $Bias = \frac{1}{N} \sum_{t=1}^N \varepsilon_t$.

Forecast performance is evaluated for the complete pediatric cohort and separately for each virtual subject.

Prediction errors are additionally stratified according to the glycemic condition at prediction time:

$$G_t = \begin{cases} \text{below range,} & G_t < 70 \text{ mg/dL,} \\ \text{target range,} & 70 \leq G_t \leq 180 \text{ mg/dL,} \\ \text{above range,} & G_t > 180 \text{ mg/dL.} \end{cases} \quad (12)$$

Errors are also stratified according to the current glucose trend. Specifically, three trend classes are considered: falling, approximately stable, and rising,

$$\text{trend}(t) = \begin{cases} \text{falling,} & \dot{G}_t < -0.5, \\ \text{stable,} & |\dot{G}_t| \leq 0.5, \\ \text{rising,} & \dot{G}_t > 0.5, \end{cases} \quad (13)$$

where $\dot{G}_t$ is expressed in mg/dL/min. Within the rising class, a further sub-threshold of 1.5 mg/dL/min is additionally used in Section 3.4 to isolate forecast accuracy during rapidly rising trajectories.

The resulting MAE, RMSE, and bias values are reported in Section 3.4. Forecast-error statistics computed on held-out scenarios are used only to characterize the predictive approximation and do not influence controller tuning.

Table 2

Physiological parameters of the ten virtual pediatric subjects used in the study. All values are rounded to two decimals.

Name	Age (y)	Weight (kg)	CR (g/U)	CF (mg/dL/U)	TDI (U/day)	k_{abs} (min ⁻¹)	V_g (dL/kg)
child#001	9	34.56	25	42.72	17.47	0.13	1.92
child#002	9	28.53	23	36.92	18.18	0.55	1.84
child#003	8	41.23	22	31.05	16.02	0.10	0.74
child#004	12	35.52	25	40.72	19.81	0.23	2.07
child#005	7	37.79	21	42.80	17.80	0.03	1.15
child#006	8	41.00	19	39.98	20.22	0.05	1.71
child#007	9	45.53	18	24.99	36.21	0.03	1.73
child#008	10	23.73	15	30.88	21.49	0.13	2.03
child#009	7	35.53	25	35.31	17.39	0.16	1.90
child#010	12	35.21	20	28.18	29.22	0.22	1.81

2.1.8. Primary-policy reward

The primary reward prioritizes glucose regulation while assigning asymmetric penalties to hypo- and hyperglycemic regions:

$$r_t^{(prim)} = 1 - \left(\frac{|G_t - G^*|}{60} \right)^2 - \mathbf{1}_{G_t > 180} [0.5 + 0.003(G_t - 180)] - \mathbf{1}_{G_t < 70} [1.5 + 0.01(70 - G_t)] - 8 \mathbf{1}_{(G_t < 50) \vee (G_t > 360)}. \quad (14)$$

Here, $\mathbf{1}_{(\cdot)}$ denotes the indicator function. The quadratic component penalizes deviation from the nominal reference. The asymmetric terms impose a stronger penalty for low-glucose states than for hyperglycemic states, consistent with the higher acute clinical risk associated with hypoglycemia relative to hyperglycemia [1,31]. The final term strongly penalizes entry into extreme glycemic regions.

The coefficients in Eq. (14) are fixed prior to training and remain constant throughout each simulated episode and across the entire evaluation protocol; none is updated online at runtime. These coefficients are controller-design parameters rather than clinically validated safety quantities. Their influence is therefore evaluated quantitatively through the sensitivity analysis described in Section 2.8.

2.2. Safety adjuster

The second learned component is a PPO-based *Safety Adjuster*. Its role is to modulate the insulin command generated by the primary regulation layer according to current glycemia, glucose dynamics, predicted short-term glucose, and recent insulin exposure.

The Safety Adjuster is explicitly distinguished from the deterministic runtime constraints described below. The Safety Adjuster is a learned policy whose action is a bounded multiplicative factor. Low-glucose suspension, predictive upper bounding, and cumulative dose-budget enforcement are deterministic rules applied after learned modulation.

At decision step t , the Safety Adjuster observes

$$x_t = [G_t, \dot{G}_t, \hat{G}_{t+30}, u_{f,loor,t}, IOB_t, dose_{30,t}]. \quad (15)$$

Here, G_t is the current glucose concentration, $\dot{G}_t$ is the estimated glucose rate of change, $\hat{G}_{t+30}$ is the short-horizon glucose prediction, $u_{f,loor,t}$ is the candidate command after predictive hyperglycemia correction, IOB_t is an insulin-on-board proxy, and $dose_{30,t}$ denotes cumulative insulin delivered during the preceding 30 min.

The insulin-on-board proxy is computed from recent delivered insulin commands using a finite-memory weighted sum, following standard IOB decay-curve approximations [32],

$$IOB_t = \sum_{j=1}^{N_{IOB}} w_j u_{f,final,t-j}, \quad (16)$$

where the coefficients w_j decrease monotonically with elapsed time $j \Delta t$ over a fixed memory horizon of N_{IOB} decision steps, approximating

the persistence of recent subcutaneous insulin action. The estimator is used as a compact exposure indicator rather than as a complete pharmacokinetic model.

The Safety Adjuster does not directly output an insulin dose. Its discrete action set is

$$F = \{0.50 + 0.05k \mid k = 0, \dots, 14\}, \quad (17)$$

which contains 15 multiplicative factors ranging from 0.50 to 1.20. The learned modulated command is

$$u_{mod,t} = f_t u_{f,loor,t}, \quad f_t \in F. \quad (18)$$

Factors smaller than one attenuate the primary command, $f_t = 1$ preserves it, and factors above one permit limited amplification. The bounded action range prevents the secondary policy from replacing the primary regulation policy with unrestricted absolute dosing decisions.

2.2.1. Deterministic runtime constraints

After learned modulation, $u_{mod,t}$ is subjected to deterministic runtime constraints. These rules act independently of the PPO policies and remain active regardless of the modulation factor selected by the Safety Adjuster.

Low-glucose suspension. The low-glucose suspension condition is

$$C_{LGS,t} = (G_t < 90 \wedge \dot{G}_t \leq -0.1) \vee (G_t < 100 \wedge \dot{G}_t \leq -0.3). \quad (19)$$

When $C_{LGS,t}$ is true, $u_{final,t} = 0$.

This deterministic rule overrides both learned policies and effects established predictive low-glucose suspend strategies, which typically operate on comparably short forecast horizons [32,33].

Predictive dose cap. When low-glucose suspension is inactive, an upper insulin bound is obtained from the short-horizon glucose prediction:

$$u_{pred,t}^{\max} = \max \left(0, \frac{\hat{G}_{t+30} - G_{pred}}{CF} \right). \quad (20)$$

The value $G_{pred} = 98$ mg/dL is reported in Table 1.

This mechanism is described as *barrier-inspired* in the same engineering sense introduced in Section 1: the admissible insulin command decreases as the predicted glucose trajectory approaches G_{pred} , without constituting a formally verified control barrier function, and without any forward-invariance or formal safety guarantee.

Residual 30-min insulin budget. The remaining admissible insulin budget is defined explicitly as a non-negative quantity, consistent with constraints on cumulative insulin delivery and insulin-on-board used to mitigate insulin stacking [30,32]:

$$D_{rem,t} = \max(0, D_{max}^{30} - dose_{30,t}). \quad (21)$$

Consequently, if cumulative insulin delivery during the preceding 30 min has already reached or exceeded D_{max}^{30} , the residual budget is exactly zero rather than negative.

The final insulin command is therefore

$$u_{\text{final},t} = \begin{cases} 0, \\ \min \left[\max(0, u_{\text{mod},t}), u_{\text{pred},t}^{\max}, D_{\text{rem},t} \right], & C_{\text{LGS},t}, \\ \text{otherwise.} \end{cases} \quad (22)$$

Eq. (22) explicitly guarantees a non-negative delivered insulin command, since each of its three arguments is non-negative by construction. Learned modulation and deterministic admissibility are therefore separate operations.

2.2.2. Safety adjuster reward

The Safety Adjuster reward combines target-range regulation, asymmetric penalties for hyper- and hypoglycemia, anticipated low-glucose risk, insulin parsimony, and conservative modification of the primary-policy command:

$$\begin{aligned} r_t^{(\text{saf})} = & 0.5 \mathbf{1}_{70 \leq G_{t+1} \leq 180} + 1 - 0.002 |G_{t+1} - 112| \\ & - 0.005 \max(0, G_{t+1} - 180) \\ & - 0.02 \max(0, 70 - G_{t+1}) \\ & - \lambda_{\text{pred}} \frac{\max(0, G_{\text{pred}} - \hat{G}_{t+30})}{30} \\ & - \lambda_{\text{ins}} u_{\text{final},t} - \lambda_{\text{dev}} |f_t - 1| \\ & - 8 \mathbf{1}_{(G_{t+1} < 50) \vee (G_{t+1} > 360)}. \end{aligned} \quad (23)$$

The nominal weighting coefficients are

$$\lambda_{\text{pred}} = 1.0, \quad \lambda_{\text{ins}} = 0.12, \quad \lambda_{\text{dev}} = 0.02. \quad (24)$$

The first two components encourage target-range regulation and proximity to the nominal reference. Hypoglycemia receives a stronger instantaneous penalty than hyperglycemia, for the same clinical-risk asymmetry described in Section 2.1.8. The predictive term penalizes trajectories approaching the lower predictive threshold. The insulin-parsimony term discourages unnecessary dosing, whereas the modulation penalty favors f_t values close to unity unless alteration of the primary-policy command provides sufficient reward benefit.

Like the primary-policy reward coefficients, these weighting parameters are engineering design quantities rather than clinically validated safety parameters, and are included in the sensitivity analysis of Section 2.8.

The Safety Adjuster should consequently be interpreted as a learned command-modulation policy rather than as an independent safety controller. Final command admissibility is determined by the subsequent deterministic rules.

2.3. Two-stage training and policy freezing

A separate two-policy controller is trained for each of the ten virtual pediatric subjects.

To quantify uncertainty arising from reinforcement-learning optimization, five independent training replicates are performed for each subject using initialization seeds {11, 22, 33, 44, 55}.

Random-number generators governing policy initialization and reinforcement-learning optimization are separated from those used to generate meal disturbances.

Phase 1 – Primary-policy training. The primary regulation policy is trained using Maskable PPO with a temporal resolution of 3.00 min. Its observation consists of the current glucose concentration G_t , and its action is one of the 21 discrete insulin commands defined in Section 2.1.1.

For each virtual subject and each independent training seed, Phase 1 consists of 50 stochastic 24-h training episodes.

Table 3

Training hyperparameters of the primary policy and Safety Adjuster.

Hyperparameter	Primary Phase 1	Safety Adjuster Phase 2
Algorithm	Maskable PPO	PPO
Sampling interval	3.00 min	3.00 min
Training episodes	50	150
n_{steps}	480	480
n_{epochs}	30	20
Learning rate	5×10^{-4}	3×10^{-4}
Discount factor γ	0.99	0.995
GAE parameter λ	0.95	0.95
Batch size	120	120
PPO clip	0.2	0.2
Observation	G_t	x_t
RL action	action index a_t	factor f_t

Phase 2 – Safety-Adjuster training. After completion of Phase 1, the parameters of the corresponding primary policy are frozen and embedded as a fixed component within environment $\mathcal{E}_2$.

The Safety Adjuster is then trained using standard PPO on the extended state representation in Eq. (15). Its reinforcement-learning action is the modulation factor f_t , rather than the final insulin dose.

For each virtual subject and each independent training seed, Phase 2 consists of 150 stochastic 24-h episodes.

The primary-policy parameters are not updated during Phase 2. Each complete training replicate therefore consists of one fixed primary-policy/Safety-Adjuster pair associated with one reinforcement-learning initialization seed.

The hyperparameters used in the two phases are summarized in Table 3.

The sequential training procedure enables the primary policy to optimize regulation before the secondary policy learns supervisory modulation. Freezing the primary policy eliminates simultaneous updates of the two learned components and reduces optimization interference between the regulation and supervisory objectives.

No reinforcement-learning parameters are shared across virtual subjects. The present design therefore evaluates patient-specific control under stochastic intra-subject variation rather than population-level policy generalization.

2.4. Implementation and reproducibility

Both reinforcement-learning policies were implemented in Python 3.14.3 using PyTorch 2.11.0, Stable-Baselines3 2.7.1, and Stable-Baselines3 Contrib 2.7.1 [34]. The physiological simulation environment used Simglucose 0.2.11. Numerical analysis and baseline tuning used NumPy 2.4.4, SciPy 1.17.1, and Optuna 4.8.0. The software environment was frozen in March 2026.

The actor and critic components use multilayer perceptron architectures with two hidden layers of 64 units each and hyperbolic-tangent activation functions. Policy parameters are initialized using the orthogonal initialization procedure of the Stable-Baselines3 implementation and optimized using Adam.

Full PPO hyperparameters, including the primary-policy and Safety-Adjuster learning rates, are reported in Table 3.

Experiments were executed on a workstation equipped with an AMD Ryzen 9 5900X CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3080 GPU running Ubuntu 20.04 LTS.

On the specified workstation, SA training required 14.2 ± 1.8 min per virtual subject, whereas complete two-stage DA training required 45.6 ± 4.3 min per subject across the five initialization seeds. The DA measurement includes both primary-policy training and subsequent Safety-Adjuster training. The complete five-seed pediatric-cohort experiment required approximately 10 h of wall-clock computation when independent simulations were parallelized.

Table 4

Training, validation, and evaluation protocol per virtual pediatric subject. All scenario sets use mutually exclusive random seeds.

Dataset	Episodes	Purpose
Primary training	50/seed	Primary-policy training
Safety training	150/seed	Safety-Adjuster training
Validation	50	Parameter selection
Final test	200	Performance evaluation

Run-time inference latency was measured over 1000 control decisions after 50 warm-up calls, excluding physiological-simulator execution and file I/O. SA required 1.2 ± 0.3 ms per decision, whereas DA required 2.4 ± 0.5 ms. Although evaluation of the second policy approximately doubled inference latency, DA execution remained negligible relative to the 3-min control interval.

Training seeds, meal-scenario seeds, analysis scripts, policy configurations, and simulator settings are maintained separately to ensure that reinforcement-learning initialization randomness is distinguishable from disturbance-scenario randomness. Reference code and configuration settings are provided through the repository cited in the Data Availability Statement.

2.5. Training, validation, and test scenario separation

Training, validation, and final evaluation use mutually exclusive stochastic meal realizations.

The meal-scenario random seeds were partitioned deterministically as follows: Phase 1 primary-policy training used seeds 1000–1049, Phase 2 Safety-Adjuster training used seeds 2000–2149, validation used seeds 3000–3049, and final testing used seeds 4000–4199. These non-overlapping intervals prevented meal realizations from being reused across the two training phases, parameter validation, and final evaluation. Within validation and final testing, corresponding controller configurations were evaluated using the same scenario seeds, thereby preserving controller pairing under identical meal disturbances.

For each virtual pediatric subject and each reinforcement-learning training seed:

- 50 stochastic 24-h episodes are used to optimize the primary policy;
- 150 stochastic 24-h episodes are used to optimize the Safety Adjuster.

A separate common validation set of 50 stochastic 24-h scenarios per virtual subject is used for supervisory-parameter selection. These scenarios are generated using meal seeds not used in either reinforcement-learning training phase.

After parameter selection, all controller-design parameters are frozen.

Final evaluation is then performed using a separate set of 200 stochastic 24-h scenarios per virtual subject. The validation and final-test meal seeds are mutually exclusive and are also distinct from all training meal seeds.

The same held-out validation and test scenarios are used across corresponding controller configurations and independent training replicates. This matched design allows controller differences to be evaluated under identical disturbance realizations.

Table 4 summarizes the experimental organization.

2.6. Testing protocol and stochastic meal scenarios

Each simulation episode spans 24.00 h, corresponding to 480 control decisions at the 3.00 min sampling interval.

The nominal meal schedule is adapted from our previous Maskable PPO study [23] and consists of five daily carbohydrate disturbances:

breakfast (40.00 g), morning snack (15.00 g), lunch (60.00 g), afternoon snack (10.00 g), and dinner (55.00 g).

For each stochastic episode, individual meal sizes are independently perturbed using zero-mean Gaussian noise with a standard deviation of 10.00 g. Meal times are independently shifted within a 30.00 min interval around their nominal occurrence times.

The reinforcement-learning controllers receive glucose measurements but no direct meal-announcement signal. Their response to carbohydrate disturbances therefore depends on observed and predicted glycemic evolution rather than explicit knowledge of meal carbohydrate content.

Unlike the three-subject evaluation (one adult, one adolescent, one child) in our previous study [23], the present protocol covers the complete ten-subject pediatric cohort.

Throughout the remainder of this paper, the proposed dual-agent controller is denoted DA, and the single-agent Maskable PPO baseline (Configuration 2 of the ablation study in Section 2.7: Maskable PPO with state-dependent action masking, without the predictive floor, Safety Adjuster, or deterministic hard constraints) is denoted SA.

2.7. Component-wise ablation study

A component-wise ablation study is performed to determine whether performance changes can be attributed specifically to the learned Safety Adjuster or instead arise predominantly from action masking, predictive correction, or deterministic runtime constraints.

Six controller configurations are evaluated:

1. *Primary PPO only*: a primary PPO controller without state-dependent action masking, predictive correction, Safety Adjuster, or deterministic runtime constraints;
2. *Primary + action masking*: the Maskable PPO regulation policy without additional predictive or runtime supervision;
3. *Primary + action masking + predictive floor*: the primary Maskable PPO policy combined with the deterministic hyperglycemia-oriented predictive correction;
4. *Primary + deterministic hard constraints*: the regulation architecture from Configuration 3 (Maskable PPO with action masking and the predictive floor), additionally equipped with deterministic low-glucose suspension, predictive upper bounding, and cumulative dose-budget enforcement, but without the learned Safety Adjuster;
5. *Primary + Safety Adjuster without hard constraints*: the regulation architecture from Configuration 3, followed by learned Safety-Adjuster modulation, with the deterministic hard constraints of Configuration 4 disabled;
6. *Full framework*: the complete architecture including Maskable PPO, action masking, predictive correction, learned Safety Adjuster, and deterministic runtime constraints.

All ablation configurations are evaluated using identical held-out stochastic meal scenarios. Where a learned component is present, independent reinforcement-learning training seeds are retained so that training variability is represented consistently across configurations.

The ablation analysis therefore separates the incremental contribution of policy-level action masking, predictive hyperglycemia correction, learned supervisory modulation, and deterministic runtime constraints.

2.8. Quantitative sensitivity analysis

A one-factor-at-a-time sensitivity analysis is performed to quantify the dependence of controller performance on the principal reward and supervisory parameters. Only one parameter is perturbed in each experiment while all remaining settings are maintained at their nominal values.

Dimensionless reward coefficients and proportional gains are evaluated at

$$\theta_i \in \{0.8\theta_i^{(0)}, \theta_i^{(0)}, 1.2\theta_i^{(0)}\}, \quad (25)$$

where $\theta_i^{(0)}$ denotes the nominal parameter value.

The sensitivity analysis includes the primary hyperglycemia penalty, the primary hypoglycemia penalty, the extreme-glycemia penalty, the proportional correction gains K_p^{lo} and K_p^{hi} , and the Safety Adjuster reward coefficients λ_{pred} , λ_{ins} , and λ_{dev} .

For glucose thresholds, absolute variations are used instead of percentage perturbations in order to preserve physiologically interpretable values. Principal glucose thresholds are evaluated at their nominal values and at ± 10 mg/dL.

The cumulative 30-min insulin budget is evaluated at

$$D_{\text{max}}^{30} \in \{0.64, 0.80, 0.96\} \text{ U}, \quad (26)$$

corresponding to a $\pm 20\%$ perturbation around the nominal setting.

Each perturbed configuration is evaluated using the same held-out stochastic scenarios as its corresponding nominal controller. Sensitivity is quantified through changes in TIR, TBR, TAR, mean glucose, coefficient of variation, and total insulin delivery.

For a generic performance metric Y , the sensitivity-induced deviation is

$$\Delta Y_i = Y(\theta_i) - Y(\theta_i^{(0)}). \quad (27)$$

Within the tested perturbation ranges, this analysis quantifies the sensitivity of controller performance to changes in individual parameters and identifies the parameters producing the largest variations in the reported outcomes.

2.9. Baseline controllers

In addition to the reinforcement-learning comparisons, proportional–integral–derivative (PID) and model predictive control (MPC) controllers were implemented as conventional control-system reference methods. These baselines are included to provide an engineering context for the performance of the proposed reinforcement-learning architecture rather than to establish a strictly information-matched comparison.

In particular, the baseline controllers may exploit information that is not provided to the reinforcement-learning policies. The proposed SA and DA controllers receive glucose measurements but no direct meal-announcement signal. By contrast, the classical baselines can use announced carbohydrate information through their corresponding meal-compensation mechanisms. Furthermore, the MPC controller uses the same physiological model family as the Simglucose/UVa–Padova simulation environment for short-horizon state prediction. These differences are explicitly considered when interpreting the comparative results.

2.9.1. PID Controller

The PID baseline computes a feedback correction from the glucose-regulation error

$$e_t = G_t - G^*. \quad (28)$$

The feedback component is represented as

$$u_{\text{fb},t}^{(\text{PID})} = K_p e_t + K_i I_t + K_d \dot{e}_t, \quad (29)$$

where I_t is the accumulated regulation error and $\dot{e}_t$ is its finite-difference derivative.

The basal and feedback contributions are combined and saturated independently of the meal-related contribution,

$$u_{\text{PID},t}^{\text{raw}} = \text{sat}(u_{\text{basal}} + u_{\text{fb},t}^{(\text{PID})}, 0, u_{\text{max}}^{(\text{PID})}) + u_{\text{meal},t}, \quad (30)$$

where $\text{sat}(\cdot, 0, u_{\text{max}}^{(\text{PID})})$ denotes saturation to the range $[0, u_{\text{max}}^{(\text{PID})}]$. The meal-related contribution $u_{\text{meal},t}$ is delivered as a separate bolus event

and is not subject to this saturation limit, consistent with standard hybrid closed-loop practice in which meal boluses are computed independently of the continuous basal-modulation loop. The complete command $u_{\text{PID},t}^{\text{raw}}$ is subsequently saturated to the admissible pump range before delivery.

The meal-bolus implementation $u_{\text{meal},t}$ delivers the standard computed dose CHO/CR exactly at the announced meal time. If the pre-meal glucose concentration exceeds 150 mg/dL, an active correction bolus of $(G_t - 150)/\text{CF}$ is added to the meal dose.

The PID gains were tuned using the Optuna Bayesian optimization framework [35] over 100 trials, with each trial evaluated on the validation set using the cohort-averaged Risk Index as the objective. The resulting optimal gains applied during evaluation are $K_p = 0.012$, $K_i = 0.0005$, and $K_d = 0.025$. The maximum permissible basal-plus-feedback command per 3-min control step is bounded by $u_{\text{max}}^{(\text{PID})} = 1.0 \text{ U}$.

The baseline basal delivery u_{basal} is implemented by scaling the subject's nominal hourly basal rate to the 3-min simulation interval. To mitigate integral windup during prolonged hyperglycemia or postprandial peaks, an anti-windup clamping rule is applied: the integration of I_t is paused when the basal-plus-feedback output saturates or when $e_t < 0$, and the maximum integral contribution $K_i I_t$ is strictly limited to $0.5 u_{\text{max}}^{(\text{PID})}$.

The PID gains and actuator limits are selected using the validation scenarios only and are frozen before final testing. The final 200 test scenarios are not used for PID parameter tuning.

2.9.2. Model predictive controller

The MPC baseline determines insulin delivery by repeatedly solving a finite-horizon constrained optimization problem. Future glucose evolution is predicted using the same physiological model family employed by the Simglucose/UVa–Padova simulation environment. The internal model uses the physiological parameters associated with the virtual subject being controlled, resulting in a nominal subject-matched modeling condition. This represents a favorable assumption for the MPC baseline and is explicitly considered when interpreting its comparison with the reinforcement-learning controllers.

At each control step, the MPC predicts glucose over a horizon of N_p samples and optimizes a sequence of N_c future insulin commands. The objective function is

$$J_t = \sum_{k=1}^{N_p} Q_G (\hat{G}_{t+k|t} - G^*)^2 + \sum_{k=0}^{N_c-1} R_u (u_{t+k|t} - u_{\text{basal}})^2 + \sum_{k=1}^{N_c-1} R_{\Delta u} (u_{t+k|t} - u_{t+k-1|t})^2. \quad (31)$$

The three terms penalize, respectively, predicted glucose deviation from the nominal target, excessive deviation of insulin delivery from the subject-specific basal requirement, and rapid changes in the control signal. Insulin delivery is constrained by $0 \leq u_{t+k|t} \leq u_{\text{max}}^{(\text{MPC})}$ throughout the control horizon.

For the nominal configuration, the prediction horizon is $N_p = 20$, corresponding to 60.00 min at the 3.00 min sampling interval, while the control horizon is $N_c = 5$, corresponding to 15.00 min. The nominal cost-function weights are $Q_G = 1.0$, $R_u = 0.08$, and $R_{\Delta u} = 0.15$, with a maximum insulin command of $u_{\text{max}}^{(\text{MPC})} = 0.50 \text{ U}$.

The constrained optimization problem is solved at every 3.00 min control step using a sequential quadratic programming implementation from the SciPy optimization library [36], with a maximum of 100 iterations and a numerical convergence tolerance of 10^{-6} .

Following the receding-horizon principle, only the first element of the optimized sequence, $u_{\text{MPC},t} = u_{t|t}^*$, is applied to the physiological simulator. The prediction and optimization procedure is then repeated at the following sampling instant using the updated system state.

Announced carbohydrate disturbances are included as known exogenous inputs to the MPC prediction model. For each meal, the controller

receives the announced carbohydrate amount and meal timing associated with the corresponding stochastic scenario, and the physiological model propagates the resulting disturbance over the prediction horizon. As noted in Section 2.9, this meal information is not available to the reinforcement-learning policies. In addition, the proposed DA controller does not use the physiological simulator model for prediction, but only the first-order 30-min glucose extrapolation defined in Eq. (4). The MPC baseline therefore benefits from both explicit subject-specific physiological-model knowledge and announced meal information.

The MPC prediction horizon, control horizon, objective-function weights, insulin bounds, and solver settings are selected using only the validation scenarios described in Section 2.5. All parameters are subsequently frozen before final evaluation and are not modified using final-test outcomes. For this reason, MPC results are interpreted as a contextual classical-control reference rather than as a strictly information-matched comparison with the proposed reinforcement-learning framework.

2.9.3. Baseline tuning and comparative interpretation

PID and MPC parameter selection was performed independently of the final test set. The same 50 validation scenarios used for controller-design validation were used to select the baseline parameters, whereas the final 200 stochastic scenarios remained held out until completion of all controller tuning.

Candidate PID and MPC configurations were evaluated primarily according to TIR and TBR, with mean glucose, glycemic variability, and total insulin delivery used as secondary criteria. Configurations producing lower TBR were preferred when candidate settings achieved similar TIR.

No baseline parameter was retuned for individual final-test meal realizations.

Whenever possible, PID, MPC, SA, and DA controllers were evaluated using the same stochastic meal realizations, allowing differences in glycemic outcomes to be compared under matched disturbance scenarios.

However, matched disturbance realization does not imply matched controller information: the PID and MPC baselines receive announced meal information and, for MPC, subject-specific physiological-model knowledge, whereas the reinforcement-learning policies rely only on measured glucose, its estimated rate of change, recent insulin-exposure variables, and the simple first-order forecast in Eq. (4).

For this reason, PID and MPC are treated as contextual classical-control references rather than strictly equivalent-information benchmarks. Their results are used to place the reinforcement-learning outcomes in an engineering control context, while the primary attribution of the proposed architectural contribution is based on the matched SA-DA comparison and the component-wise ablation analysis.

2.10. Outcome measures and statistical analysis

Controller performance is quantified using standard glycemic and control metrics, including Time in Range (TIR; 70–180 mg/dL), Time Below Range (TBR; < 70 mg/dL), Time Above Range (TAR; > 180 mg/dL), coefficient of variation (CV), mean glucose, glucose management indicator (GMI), mean absolute deviation (MAD), the Risk Index (combining the Low and High Blood Glucose Indices) [31], and total insulin delivery.

Severe hypoglycemic exposure is additionally characterized using the percentage of time below 54 mg/dL and the number of severe hypoglycemic episodes. A severe hypoglycemic episode is counted as a distinct excursion below 54 mg/dL separated from any previous event by return above the threshold.

Controller comparisons are paired at the stochastic-scenario level because corresponding controllers are evaluated under identical meal disturbances.

For each virtual subject and reinforcement-learning seed, paired scenario-level differences are computed as

$$d_{i,s,j} = Y_{i,s,j}^{(DA)} - Y_{i,s,j}^{(SA)}, \quad (32)$$

where i denotes the virtual subject, s the independent RL training seed, and j the matched evaluation scenario.

For percentage metrics such as TIR, TAR, and TBR, differences are reported in percentage points rather than as relative percentage changes unless explicitly stated otherwise.

The virtual subject was treated as the independent physiological unit for the primary cohort-level inference. For each subject, the matched differences are first averaged across the 200 scenarios within each training seed and then across the five training seeds. The resulting ten subject-level paired differences were used to calculate the cohort mean difference, paired t -tests, Wilcoxon signed-rank tests where reported, and paired standardized effect size d_z . This aggregation prevents the 10,000 simulated runs per controller from being treated as independent physiological observations.

Mean paired differences are accompanied by 95% confidence intervals obtained through paired patient-level bootstrap resampling [37]: virtual subjects are sampled with replacement, and the cohort mean of their precomputed paired differences is recalculated for each bootstrap replicate. Controller matching is therefore preserved within every resampled subject. For all bootstrap analyses, 100,000 replicates were generated using a fixed random seed of 2026. The 95% confidence limits were defined as the 2.5th and 97.5th percentiles of the bootstrap distribution (percentile method).

Effect sizes and confidence intervals are reported together with p -values rather than relying solely on null-hypothesis significance testing.

Scenario-level variability and training-seed variability are reported separately. Standard deviations calculated across stochastic evaluation scenarios therefore characterize disturbance-related variability and are not interpreted as uncertainty associated with reinforcement-learning optimization.

For severe hypoglycemic event counts, event rates are reported together with 95% confidence intervals.

A two-sided significance level of $\alpha = 0.05$ is used for inferential analyses.

3. Results

The proposed controllers were evaluated across the complete cohort of ten virtual pediatric subjects, using the five independent training replicates and the held-out test protocol described in Section 2.5. Each of the five independently trained replicates per subject was evaluated on the same set of 200 held-out stochastic 24-h test scenarios. Corresponding controller configurations were evaluated on these same meal realizations, enabling paired scenario-level comparisons. Unless otherwise stated, differences in TIR, TBR, and TAR are expressed in percentage points rather than relative percentages.

Unless otherwise specified, cohort-level summary statistics are computed hierarchically: for each virtual subject and training replicate, a per-replicate mean is first obtained across the 200 held-out scenarios; cohort-level values are then obtained by averaging these per-replicate means across subjects and training replicates. This procedure avoids implicitly weighting subjects or training replicates unequally due to differences in outcome variance, and is consistent with the separate treatment of scenario-level and training-seed variability introduced in Section 3.3.

The evaluation includes overall and subject-specific glycemic performance, variability across independently trained policies and stochastic scenarios, short-horizon forecast accuracy, component-wise ablation, parameter sensitivity, insulin-delivery behavior, CVGA, comparison with classical control references, and severe hypoglycemia analysis.

Table 5

Cohort-level comparison between the single-agent Maskable PPO controller (SA) and the proposed dual-agent framework (DA). Differences in TIR, TBR, TAR, and inter-patient CV_{TIR} are expressed in percentage points. For TIR, TBR, and TAR, confidence intervals, paired effect sizes, and p -values were calculated from the ten subject-level paired outcomes. The Risk Index interval was obtained through patient-level bootstrap resampling, whereas its p -value was obtained from the paired t -test. The CV_{TIR} interval was obtained through paired subject-level bootstrap resampling and is reported without a standardized effect size or hypothesis-test p -value.

Metric	SA	DA	Diff.	95% CI	d_z	p
TIR (%)	42.86	69.96	27.10	[20.80, 33.40]	3.08	< 0.001
TBR (%)	25.56	11.99	-13.57	[-30.77, 3.63]	-0.56	0.108
TAR (%)	31.58	18.05	-13.53	[-33.81, 6.75]	-0.48	0.166
Risk Index	22.31	8.82	-13.49	[-21.49, -7.47]	-1.10	0.007 ^a
Inter-patient CV_{TIR} (%)	33.90	13.11	-20.79	[-30.22, -10.91]	–	–
Mean glucose (mg/dL)	172.66	136.29	-36.37	[-115.56, 42.82]	-0.33	0.326

^a Paired t -test ($t = -3.48$, corresponding to $d_z \approx -1.10$ for $n = 10$). The Wilcoxon signed-rank test yielded $p = 0.002$.

3.1. Overall glycemic performance

The primary comparison between the dual-agent framework (DA) and the single-agent Maskable PPO controller (SA) was performed using the paired outcomes of the ten virtual pediatric subjects. The corresponding inferential statistics are summarized in Table 5.

DA increased cohort-average TIR from 42.86% under SA to 69.96%, corresponding to a paired improvement of 27.10 percentage points (95% CI: [20.80, 33.40]; $d_z = 3.08$; paired t -test $p < 0.001$). TIR was higher under DA in all ten subjects, indicating a large and consistent increase in the proportion of time spent within the target glucose range.

Cohort-average TBR decreased from 25.56% to 11.99% (paired difference: -13.57 percentage points; 95% CI: [-30.77, 3.63]; $d_z = -0.56$; $p = 0.108$), while TAR decreased from 31.58% to 18.05% (paired difference: -13.53 percentage points; 95% CI: [-33.81, 6.75]; $d_z = -0.48$; $p = 0.166$). Although both cohort averages favored DA, neither difference reached statistical significance at the subject level because of substantial inter-subject heterogeneity.

The composite Blood Glucose Risk Index (BGRI) decreased from 22.31 under SA to 8.82 under DA, corresponding to an absolute reduction of 13.49 and a relative reduction of approximately 60.5%. The patient-bootstrap 95% CI excluded zero ([-21.49, -7.47]), and the reduction was supported by both the paired t -test ($t = -3.48$, $d_z \approx -1.10$, $p = 0.007$) and the Wilcoxon signed-rank test ($p = 0.002$).

Inter-patient TIR variability also decreased, with CV_{TIR} declining from 33.90% under SA to 13.11% under DA. The paired subject-level bootstrap difference was -20.79 percentage points (95% CI: [-30.22, -10.91]), indicating more homogeneous TIR performance across the cohort.

Overall, DA improved TIR, composite glycemic risk, and inter-patient consistency. Nevertheless, the residual cohort-average TBR and the heterogeneous subject-level responses indicate that guideline-level regulation was not achieved uniformly across the pediatric cohort.

3.2. Subject-specific glycemic outcomes

The effect of the supervisory architecture was heterogeneous across the ten virtual subjects. DA increased TIR relative to SA in all 10 subjects, increasing the cohort-average TIR from 42.86% to 69.96%. This corresponds to an absolute improvement of 27.10 percentage points. Cohort-average TBR decreased from 25.56% under SA to 11.99% under DA, although a subject-level reduction was observed in only 6 of 10 subjects. Cohort-average TAR also decreased, from 31.58% to 18.05%, with subject-level reductions observed in 4 of 10 subjects. Thus, the favorable cohort averages concealed substantial differences in the balance

between hypo- and hyperglycemic exposure across individual physiological profiles.

The subject-level results reveal two distinct limitations of the SA baseline. For child#001–child#006, SA was characterized primarily by substantial hypoglycemic exposure, with TBR ranging from 29.31% to 59.19%. In contrast, child#007–child#010 exhibited negligible TBR (0.01%) under SA but pronounced hyperglycemia, with TAR ranging from 55.78% to 86.11%. Consequently, the near-absence of hypoglycemia in these four subjects should not be interpreted as satisfactory regulation because it was accompanied by a marked loss of time in range.

DA shifted all ten subjects toward a larger proportion of glucose values within the target range. The highest DA TIR was observed for child#005, increasing from 62.38% under SA to 80.98% under DA. In this subject, TBR decreased from 29.31% to 2.41%, whereas TAR increased from 8.31% to 16.61%. This result illustrates that a substantial improvement in TIR may still be accompanied by a redistribution between low- and high-glucose exposure.

Marked improvements were also observed in subjects dominated by hyperglycemia under SA. For child#007, TIR increased from 28.61% to 66.55%, while TAR decreased from 71.38% to 31.06%. Similarly, child#010 improved from 36.11% to 73.52% TIR, with TAR decreasing from 63.88% to 19.09%.

The most challenging physiological profile remained child#008. Under DA, this subject achieved a TIR of 46.52%, compared with 13.88% under SA. The substantial reduction in TAR, from 86.11% to 22.69%, was accompanied by an increase in TBR from 0.01% to 30.79%. Therefore, although DA shifted a large fraction of the glucose distribution away from hyperglycemia and into the target range, it did not achieve an acceptable balance between hypo- and hyperglycemic exposure for this subject.

The metabolic indicators exhibited similarly subject-dependent patterns. Cohort-average GMI decreased from 7.44% under SA to 6.57% under DA, while MAD decreased from 63.73 to 48.01 mg/dL. The reduction in cohort-average MAD was driven primarily by the subjects characterized by marked hyperglycemia under SA. In particular, large MAD reductions were observed for child#007–child#010. Conversely, MAD increased under DA for child#001, child#002, child#003, and child#006, demonstrating that the cohort-average improvement was not uniform.

GMI should be interpreted together with TIR, TBR, and TAR rather than in isolation. Several subjects exhibited low GMI under SA despite substantial hypoglycemic exposure. A lower GMI therefore did not necessarily correspond to superior glycemic regulation in this cohort.

Mean insulin exposure increased from 5.71 to 6.61 U/day at the cohort level. However, the direction and magnitude of this change varied across subjects. Insulin exposure decreased for child#001, child#002, child#003, child#005, and child#006, remained unchanged for child#004, and increased for child#007–child#010. The increases in the latter group are consistent with the greater insulin requirements needed to counter the pronounced hyperglycemia observed under SA.

Overall, the subject-specific analysis demonstrates a consistent improvement in TIR under the dual-agent architecture and substantial reductions in cohort-average TBR and TAR. Nevertheless, these improvements were not uniform across the individual components of glycemic exposure. DA reduced TBR primarily in subjects dominated by hypoglycemia under SA, whereas it reduced TAR primarily in subjects dominated by hyperglycemia under SA. Accordingly, the results do not support the conclusion that DA consistently achieved guideline-level glycemic regulation across the complete cohort. In particular, child#008 remained a notable failure case requiring further controller refinement.

3.3. Training-seed and scenario variability

Performance variability was analyzed separately with respect to stochastic meal realizations and independent RL optimization. For each

Table 6

Subject-specific glycemic and control outcomes for the single-agent Maskable PPO controller (SA) and the proposed dual-agent framework (DA). Central estimates were obtained across five independently trained controller replicates, each evaluated on 200 matched held-out stochastic 24-h scenarios.

(a) Time-in-range metrics. For TIR and TAR, the more favorable value for each subject is shown in bold. TBR values are reported without subject-level ranking because low TBR in isolation may coexist with severe hyperglycemia.

Subject	TIR (%) ↑		TBR (%) ↓		TAR (%) ↓	
	SA	DA	SA	DA	SA	DA
child#001	50.88	68.05	43.61	17.91	5.51	14.04
child#002	44.99	75.48	54.39	14.41	0.62	10.11
child#003	36.52	72.48	59.19	15.41	4.29	12.11
child#004	55.31	73.98	36.78	13.41	7.91	12.61
child#005	62.38	80.98	29.31	2.41	8.31	16.61
child#006	55.71	72.32	32.28	8.89	12.01	18.79
child#007	28.61	66.55	0.01	2.39	71.38	31.06
child#008	13.88	46.52	0.01	30.79	86.11	22.69
child#009	44.21	69.72	0.01	6.89	55.78	23.39
child#010	36.11	73.52	0.01	7.39	63.88	19.09
Cohort mean	42.86	69.96	25.56	11.99	31.58	18.05

(b) Metabolic and insulin-exposure metrics. MAD is reported in mg/dL and insulin exposure in U/day. GMI values are reported without ranking because their clinical interpretation depends on the accompanying hypo- and hyperglycemic exposure.

Subject	GMI (%)		MAD ↓		Insulin (U/day)	
	SA	DA	SA	DA	SA	DA
child#001	5.51	6.31	39.51	47.71	6.51	6.01
child#002	5.09	6.19	30.09	40.29	8.39	6.59
child#003	5.11	6.21	37.81	45.81	9.41	6.71
child#004	5.69	6.39	42.09	41.99	5.79	5.79
child#005	5.81	6.71	40.01	37.21	9.21	7.81
child#006	5.89	6.69	46.89	49.59	7.59	6.49
child#007	10.11	7.21	99.91	41.21	2.91	6.81
child#008	13.79	6.89	152.89	83.09	0.49	7.79
child#009	8.31	6.61	63.21	44.51	3.41	5.31
child#010	9.09	6.49	84.89	48.69	3.39	6.79
Cohort mean	7.44	6.57	63.73	48.01	5.71	6.61

Table 7

Between-training-seed variability of cohort-level glycemic performance across five independent RL training replicates.

Metric	SA	DA
TIR (%)	42.86(18)	69.96(12)
TBR (%)	25.56(24)	11.99(8)
TAR (%)	31.58(21)	18.05(14)
Risk Index	22.31(23)	8.82(9)

fixed trained controller, variability across the 200 held-out test scenarios quantifies sensitivity to stochastic meal timing and carbohydrate content. In contrast, variation among the five independently trained controller replicates quantifies dependence on RL initialization and optimization. As shown in Table 7, for the DA controller, cohort-level TIR across the five independent training replicates was 69.96%, with a between-seed standard deviation of 0.12 percentage points. The corresponding between-seed standard deviation for TBR was 0.08 percentage points. For SA, cohort-level TIR was 42.86%, with a between-seed standard deviation of 0.18 percentage points. These results indicate that the principal DA-SA difference (~27 percentage points in TIR) was stable across independent policy optimizations, with between-seed variability roughly two orders of magnitude smaller than the DA-SA gap. Importantly, the standard deviations reported across test scenarios should not be interpreted as estimates of RL training uncertainty.

Table 8

Thirty-minute glucose-forecast error stratified by glycemic condition and glucose trend. The bias indicates directional systematic errors (e.g., overestimation during rising trends and underestimation during falling trends).

Condition	MAE (mg/dL)	RMSE (mg/dL)	Bias (mg/dL)
Overall	21.4 ± 3.2	28.7 ± 4.1	2.1
$G_i < 70$ mg/dL	24.1 ± 3.8	31.5 ± 4.6	-2.5
$70 \leq G_i \leq 180$ mg/dL	18.5 ± 2.9	24.8 ± 3.5	1.4
$G_i > 180$ mg/dL	23.8 ± 4.0	30.2 ± 4.9	4.8
Falling trend	22.5 ± 3.5	29.8 ± 4.2	-5.2
Stable trend	17.2 ± 2.6	22.4 ± 3.1	0.6
Rising trend	28.3 ± 4.5	36.7 ± 5.8	7.4
(> 1.5 mg/dL/min)	34.5 ± 5.2	43.1 ± 6.5	12.5

3.4. Short-horizon glucose-forecast accuracy

The first-order 30-min glucose predictor was evaluated independently on held-out scenarios according to the procedure described in Section 2.1.7.

Across the pediatric cohort, the predictor achieved an overall MAE of 21.4 ± 3.2 mg/dL and an RMSE of 28.7 ± 4.1 mg/dL. The signed prediction bias was 2.1 mg/dL, indicating limited overall systematic displacement of the forecast at the cohort level. Across individual virtual subjects, overall MAE ranged from 18.2 ± 2.4 mg/dL to 25.6 ± 4.5 mg/dL, reflecting the heterogeneity of glucose dynamics within the pediatric population. Prediction accuracy was nevertheless strongly dependent on glucose dynamics.

Errors increased during rapidly changing glucose trajectories, with MAE exceeding 34.5 mg/dL when the rising glucose rate exceeded approximately 1.5 mg/dL/min.

The stratified analysis further showed different forecast accuracy and directional bias across below-range, target-range, and above-range conditions, as well as across falling, stable, and rising trajectories (Table 8).

These results confirm that the predictor provides a useful low-complexity supervisory signal but also identify a relevant limitation during rapidly rising glucose trajectories. The predictor should therefore not be interpreted as a high-fidelity physiological forecasting model.

3.5. Component-wise ablation analysis

A component-wise ablation study was conducted to characterize the incremental contributions of state-dependent action masking, predictive correction, learned Safety-Adjuster modulation, and deterministic runtime constraints, as defined in Section 2.7. Table 9 compares the six resulting configurations.

Relative to the Primary-PPO-only configuration, state-dependent action masking increased TIR from 35.65% to 42.86%, corresponding to an absolute improvement of 7.21 percentage points. Concurrently, TBR decreased by 7.27 percentage points, from 32.83% to 25.56%, while the Risk Index decreased from 35.08 to 22.31. TAR remained essentially unchanged, increasing by only 0.06 percentage points.

Adding the predictive floor to the SA configuration increased TIR by 11.39 percentage points and reduced TBR and TAR by 6.00 and 5.39 percentage points, respectively. However, the Risk Index increased from 22.31 to 25.26. This result indicates that improvements in aggregate time-in-range metrics were not accompanied by a corresponding reduction in the nonlinear glycemic-risk measure.

Configurations 4 and 5 isolate the contributions of the deterministic hard constraints and learned Safety Adjuster relative to the common action-masking-plus-predictive-floor configuration. Adding the deterministic constraints increased TIR by 5.32 percentage points and reduced TBR by 14.32 percentage points. This reduction in low-glucose exposure was accompanied by a 9.00 percentage-point increase in TAR, although the Risk Index decreased by 10.73.

Table 9

Component-wise ablation results across the held-out pediatric test set. To represent training variability consistently across configurations, the reported values are averages over the five independent reinforcement-learning initialization seeds. The “Primary + action masking” configuration corresponds to the single-agent (SA) baseline, whereas the “Full framework” corresponds to the dual-agent (DA) controller.

Configuration	TIR (%)	TBR (%)	TAR (%)	Risk Index	Insulin (U/day)
Primary PPO only	35.65	32.83	31.52	35.08	25.02
Primary + action masking (SA)	42.86	25.56	31.58	22.31	22.83
Primary + action masking + predictive floor	54.25	19.56	26.19	25.26	23.57
Primary + action masking + predictive floor + hard constraints	59.57	5.24	35.19	14.53	21.07
Primary + action masking + predictive floor + Safety Adjuster (no hard constraints)	65.13	12.27	22.70	18.09	22.51
Full framework (DA)	69.96	11.99	18.05	8.82	21.39

In contrast, adding the Safety Adjuster without the deterministic constraints increased TIR by 10.88 percentage points, reduced TBR by 7.29 percentage points, and reduced TAR by 3.49 percentage points relative to the same baseline. The corresponding Risk Index decreased by 7.17. Thus, the Safety Adjuster produced the larger isolated improvement in TIR and a more balanced reduction in both TBR and TAR, whereas the deterministic constraints produced the larger isolated reduction in TBR.

The complete DA framework achieved the highest TIR (69.96%), the lowest TAR (18.05%), and the lowest Risk Index (8.82) among the evaluated configurations. Relative to the Primary-PPO-only configuration, these values correspond to a 34.31 percentage-point increase in TIR, a 20.84 percentage-point reduction in TBR, a 13.47 percentage-point reduction in TAR, and a 26.26 absolute reduction in the Risk Index. The latter is equivalent to an approximately 74.9% relative reduction.

The interaction between the learned and deterministic layers was not purely additive. Compared with the Safety-Adjuster configuration without hard constraints, the full framework further increased TIR by 4.83 percentage points, reduced TAR by 4.65 percentage points, and reduced the Risk Index by 9.27, while TBR changed only marginally from 12.27% to 11.99%. Conversely, the hard-constraints-only configuration retained the lowest TBR (5.24%) but produced substantially higher TAR (35.19%) than the full framework. These results indicate that the complete architecture favored a more balanced regulation of hypo- and hyperglycemic exposure rather than minimizing TBR in isolation.

Overall, the ablation results attribute the larger isolated TIR improvement to learned Safety-Adjuster modulation and the larger isolated TBR reduction to deterministic runtime constraints. Their joint integration yielded the best overall TIR, TAR, and Risk Index, although residual TBR remained and the full framework did not attain the minimum TBR observed in the hard-constraints-only configuration.

3.6. Parameter sensitivity

The one-factor-at-a-time sensitivity analysis evaluated the effect of perturbing the principal reward and supervisory parameters around their nominal values. As specified in Section 2.8, sensitivity is quantified across all six principal glycemic and control metrics, summarized in Table 10.

Across the tested parameter variations, TIR changed by at most 2.3 percentage points, while TBR and TAR varied by at most 1.6 and 3.9 percentage points, respectively. Changes in total daily insulin delivery remained bounded within ± 2.2 U/day across all configurations.

Among the evaluated parameters, the controller was most sensitive to the 30-min cumulative insulin limit ($D_{\max}^{30}$). Its $\pm 20\%$ perturbation produced TIR changes ranging from -2.3 to $+2.0$ percentage points, TBR changes ranging from -1.6 to $+1.5$ percentage points, and the largest observed change in total insulin delivery, ranging from -1.8 to $+2.2$ U/day.

Variation of the 30-min cumulative insulin limit from 0.64 to 0.96 U produced the largest changes in metabolic control and slightly increased hypoglycemic exposure when the limit was relaxed. In compar-

ison, glucose-threshold perturbations of ± 10 mg/dL produced smaller changes in TIR (≤ 1.2 percentage points) and TBR (≤ 0.5 percentage points). The main performance trends were therefore preserved across the individual parameter variations evaluated in this analysis. Glucose CV remained comparatively stable, changing by at most 1.3 percentage points.

3.7. Insulin-delivery profile and supervisory-layer activity

The insulin-delivery trajectories were analyzed to characterize how the successive control layers modified the command generated by the primary policy, following the sequence $u_{\text{cmd}} \rightarrow u_{\text{floor}} \rightarrow u_{\text{mod}} \rightarrow u_{\text{final}}$ shown in Fig. 2.

The primary Maskable PPO policy first generates the candidate insulin command u_{cmd} . When the predictive-correction activation conditions are satisfied, the correction acts as a conditional lower bound, so that u_{floor} may exceed u_{cmd} ; otherwise, $u_{\text{floor}} = u_{\text{cmd}}$. The learned Safety Adjuster then selects a bounded multiplicative factor f_i and produces $u_{\text{mod}} = f_i u_{\text{floor}}$. Finally, deterministic low-glucose suspension, predictive upper bounding, and the residual 30-min insulin budget are independently enforced to obtain u_{final} .

Fig. 3 shows the population-level insulin delivery profiles and the temporal activity of the learned and deterministic supervisory mechanisms.

The Safety Adjuster selected a non-unity modulation factor ($f_i \neq 1$) during 54.8% of controller steps on average, making it the most frequently active supervisory component over the 24-h horizon. Its activity was sustained outside the immediate postprandial periods. The predictive correction increased the primary-policy command during 12.4% of steps, with activation concentrated around the three main meal disturbances.

The deterministic runtime constraints were binding less frequently. The predictive upper bound constrained the modulated command during 7.8% of steps, low-glucose suspension was triggered during 4.1% of steps, and the residual 30-min insulin budget limited delivery during 2.5% of steps.

This activation pattern is consistent with the component-wise ablation results in Section 3.5. The Safety Adjuster intervened frequently through bounded, discretized multiplicative modulation, whereas the deterministic runtime constraints were binding less frequently but produced a larger reduction in hypoglycemic exposure when present, consistent with their isolated contribution to TBR in the ablation analysis.

These activation rates describe how frequently each component modified or constrained the candidate insulin command. They should not, in isolation, be interpreted as evidence of clinical safety.

3.8. Subject-specific glucose and insulin dynamics

Fig. 4 shows the subject-specific glucose and final insulin-delivery profiles under the dual-agent framework (DA) and the single-agent Maskable PPO controller (SA). For each subject and controller, solid curves represent the median over 1000 held-out test trajectories, ob-

Table 10

Summary of the one-factor-at-a-time sensitivity analysis. Values represent changes relative to the nominal full-framework configuration. The row “Glucose thresholds (all)” summarizes the range of effects observed across six independent one-factor-at-a-time perturbations of G_{hp} , G_{sl} , G_{ml} , G_{sv} , S_1 , and S_2 (Table 1), each perturbed individually.

Parameter	Perturbation	ΔTIR (pp)	ΔTBR (pp)	ΔTAR (pp)	$\Delta \text{Mean gluc. (mg/dL)}$	ΔCV (pp)	$\Delta \text{Insulin (U/day)}$
Primary hyperglycemia penalty	-20% / +20%	-0.4 / +0.6	-0.1 / +0.2	+0.5 / -0.8	+1.2 / -1.6	+0.2 / -0.3	-0.5 / +0.8
Primary hypoglycemia penalty	-20% / +20%	+0.8 / -0.9	+0.6 / -0.8	-1.4 / +1.7	-2.0 / +2.2	-0.5 / +0.6	+0.9 / -1.1
Extreme-glycemia penalty (-8)	-20% / +20%	+0.3 / -0.4	-0.1 / +0.1	-0.2 / +0.3	-0.8 / +0.9	-0.1 / +0.2	+0.2 / -0.3
K_p^{lo}	-20% / +20%	-1.2 / +1.4	-0.3 / +0.5	+1.5 / -1.9	+2.7 / -3.2	+0.6 / -0.7	-1.0 / +1.3
K_p^{hi}	-20% / +20%	-1.4 / +1.6	-0.5 / +0.6	+1.9 / -2.2	+3.4 / -3.7	+0.8 / -0.9	-1.2 / +1.5
λ_{pred}	-20% / +20%	-0.9 / +1.1	+0.6 / -0.8	+0.3 / -0.3	-2.0 / +2.5	+0.4 / -0.5	+0.8 / -1.0
λ_{ins}	-20% / +20%	+1.3 / -1.5	-0.7 / +0.9	-0.6 / +0.6	-2.4 / +2.8	-0.6 / +0.7	+1.4 / -1.6
λ_{dev}	-20% / +20%	+0.4 / -0.5	-0.2 / +0.3	-0.2 / +0.2	-1.0 / +1.2	-0.2 / +0.3	+0.4 / -0.5
D_{max}^{30}	0.64 / 0.96 U	-2.3 / +2.0	-1.6 / +1.5	+3.9 / -3.5	+4.9 / -5.5	+1.1 / -1.3	-1.8 / +2.2
Glucose thresholds (all)	± 10 mg/dL	-1.1 / +1.2	-0.4 / +0.5	+1.5 / -1.7	+2.2 / -2.4	+0.5 / -0.6	-0.9 / +1.1

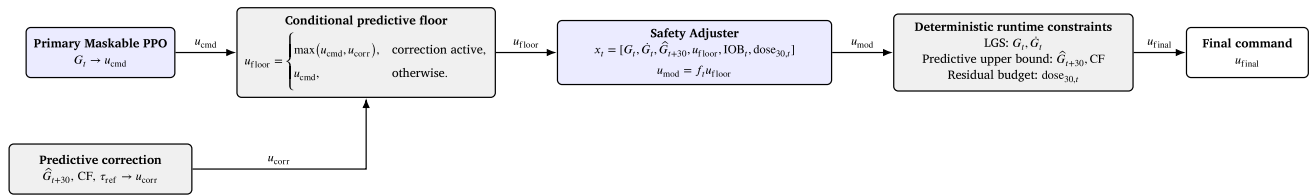


Fig. 2. Detailed runtime signal flow of the proposed controller. The primary Maskable PPO policy generates u_{cmd} . When the predictive-correction activation conditions are satisfied, the predictive correction acts as a lower bound and $u_{floor} = \max(u_{cmd}, u_{corr})$; otherwise, $u_{floor} = u_{cmd}$. The Safety Adjuster uses current glycemia, glucose trend, predicted glucose, recent insulin exposure, and u_{floor} to select the bounded modulation factor f_t . Low-glucose suspension, the predictive upper bound, and the non-negative residual insulin budget are then independently enforced to obtain u_{final} .

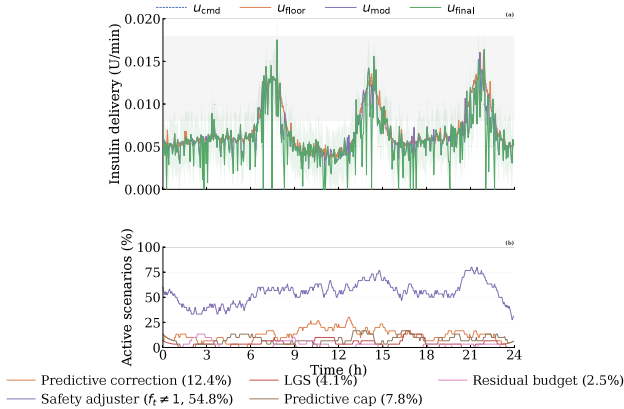


Fig. 3. Population-level insulin-delivery profiles and supervisory-layer activity across the virtual pediatric cohort. (a) Curves represent the cohort-level mean primary-policy command u_{cmd} , the command after conditional predictive correction u_{floor} , the learned Safety-Adjuster output u_{mod} , and the final delivered command u_{final} after deterministic runtime constraints. Commands are expressed as insulin-delivery rates after normalization by the 3-min control interval; shaded regions indicate the corresponding interquartile ranges across the evaluated outcomes. (b) At each decision time, curves indicate the percentage of evaluated outcomes in which the corresponding mechanism modified or constrained the candidate command. Values in parentheses report activation rates averaged over the complete 24-h evaluation horizon. Safety-Adjuster activation is defined by $f_t \neq 1$.

tained from 200 stochastic 24-h scenarios evaluated for each of five independent training seeds. Shaded regions denote the corresponding interquartile ranges.

The temporal profiles highlight two distinct patterns of baseline dysregulation. In child#001–child#006, the SA trajectories were characterized primarily by recurrent low-glucose excursions. DA generally shifted these profiles upward and reduced the depth and duration of the hypoglycemic periods, although the shift was accompanied by higher

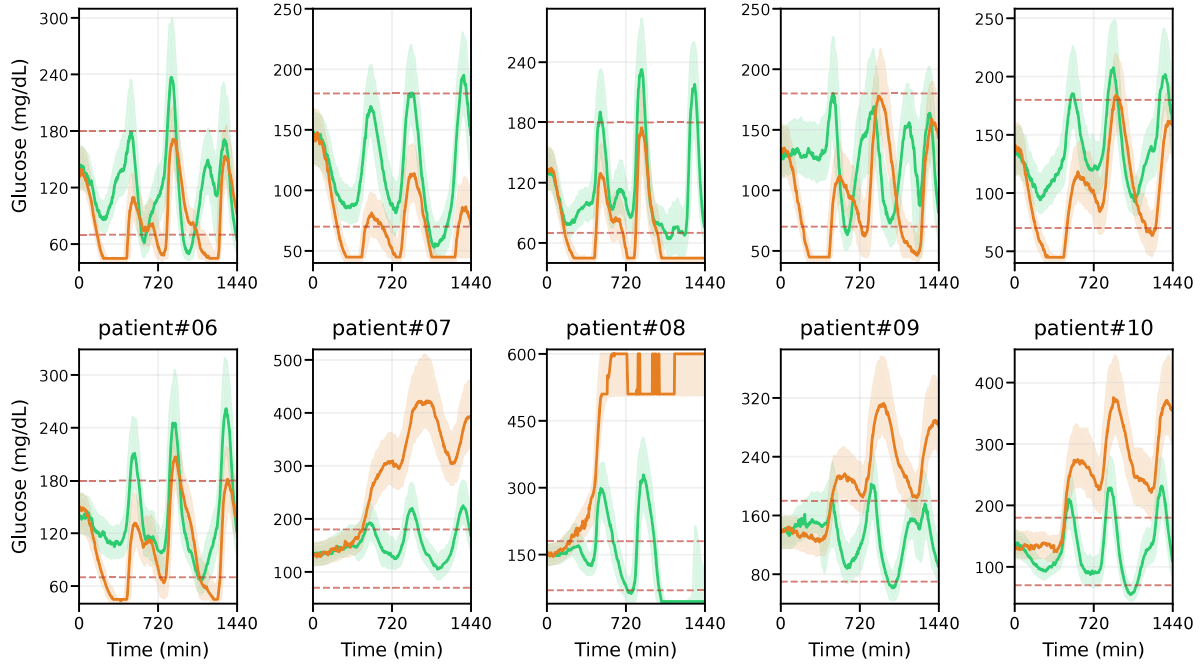
postprandial peaks in some subjects. In contrast, child#007–child#010 showed sustained or recurrent hyperglycemia under SA. For these subjects, DA substantially attenuated the hyperglycemic excursions and shifted a larger portion of the glucose trajectories toward the target range.

The temporal behavior of child#008 illustrates the main residual limitation of the proposed framework. DA markedly reduced the sustained and extreme hyperglycemia observed under SA, but the resulting profile retained broad interquartile variability and intermittent low-glucose excursions. Thus, the improved glucose distribution did not translate into uniformly stable regulation for this subject. This observation is consistent with the subject-level TIR, TBR, and TAR outcomes reported in Table 6.

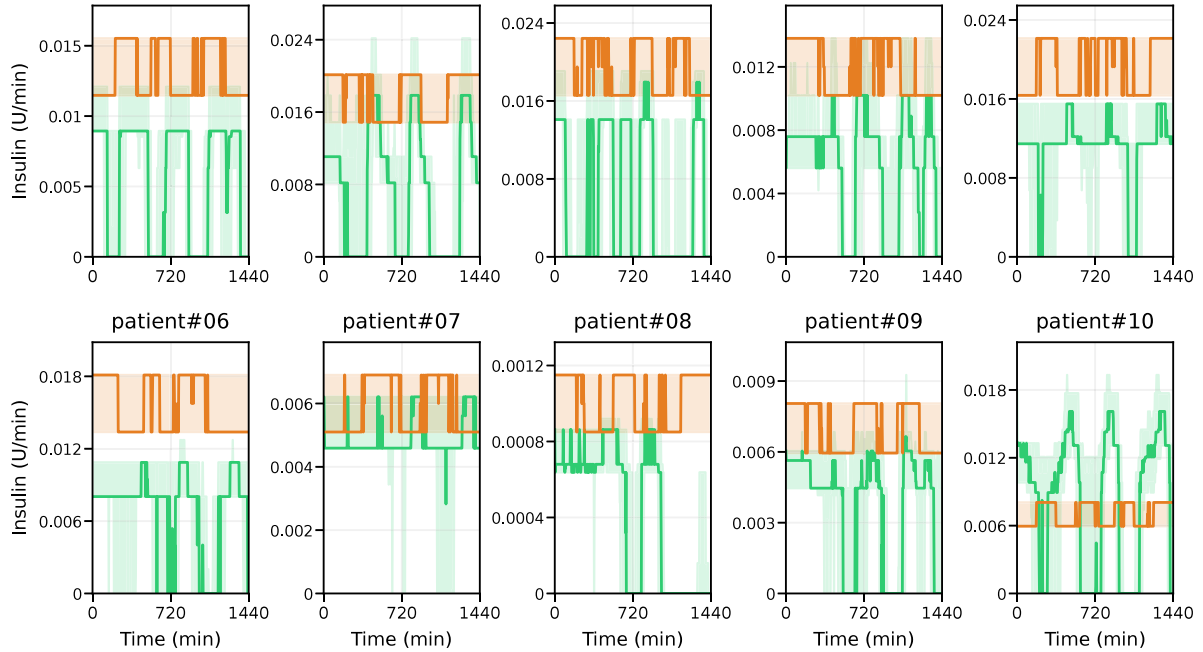
The final insulin-delivery profiles were also strongly subject dependent. For child#001–child#009, the DA median u_{final} was generally lower or more intermittent than the corresponding SA profile. This pattern was particularly evident in subjects for which SA produced substantial low-glucose exposure. Conversely, child#010 exhibited a higher and more structured DA insulin profile, consistent with the increased insulin delivery required to counter the pronounced hyperglycemic trajectory observed under SA.

Differences in u_{final} represent the combined effect of the primary policy, predictive correction, learned Safety-Adjuster modulation, and deterministic runtime constraints. They therefore cannot be attributed to any single supervisory component from these profiles alone. Moreover, because the insulin axes are autoscaled independently for each subject, the figure is intended to compare temporal patterns within subjects rather than absolute delivery magnitudes across subjects.

Overall, the paired glucose and insulin profiles show that DA modified insulin delivery differently according to the dominant baseline failure mode. The framework reduced low-glucose exposure in subjects with hypoglycemia-dominated SA profiles and attenuated hyperglycemia in subjects with hyperglycemia-dominated profiles. Nevertheless, the residual variability observed in child#008 confirms that neither improved cohort averages nor adaptive insulin modulation ensured uniformly satisfactory control across all physiological profiles.



(a) Subject-specific glucose profiles under DA (green) and SA (orange). Solid curves represent the median over 1000 held-out trajectories per subject and controller, and shaded regions denote the corresponding interquartile ranges. Dashed horizontal lines indicate the 70 and 180 mg/dL glycemic thresholds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(b) Corresponding subject-specific final insulin-delivery profiles u_{final} under DA (green) and SA (orange). Solid curves and shaded regions represent the median and interquartile range, respectively.

Fig. 4. Subject-specific glucose and final insulin-delivery dynamics under the dual-agent framework and the single-agent baseline across the 24-h evaluation horizon.

3.9. Glycemic risk distribution

Glycemic risk was further characterized using the Low Blood Glucose Index (LBGI), High Blood Glucose Index (HBGI), and their composite BGRI. These indices complement conventional time-in-range metrics by applying nonlinear penalties to low- and high-glucose deviations.

Fig. 5 shows the subject- and scenario-specific BGRI after averaging corresponding outcomes across the five training seeds. The risk distribution

was generally lower under DA, but the magnitude of the change remained strongly subject dependent. Under SA, child#007 exhibited high and persistent scenario-level risk, which was substantially attenuated under DA. Child#008 remained the main source of residual risk under DA, consistent with its unfavorable balance between TBR and TAR.

The two components of BGRI decreased numerically under DA. Cohort-average LBGI changed from 7.99 under SA to 3.79 under DA

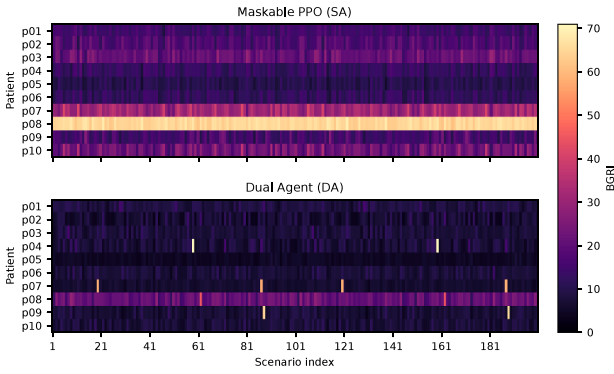


Fig. 5. BGRI distribution across ten virtual pediatric subjects and 200 test scenarios under the single-agent Maskable PPO baseline (SA) and the complete dual-agent framework (DA). Each cell reports the mean BGRI across five training seeds. A common color scale is used for both controllers.

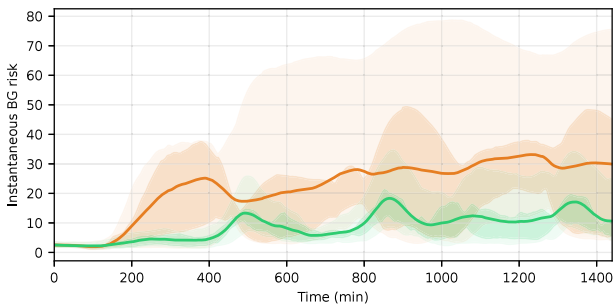


Fig. 6. Temporal evolution of instantaneous BG risk across the pediatric cohort under SA (orange) and DA (green). Solid curves denote the mean of the subject-specific mean-risk profiles. Dark and light shaded regions represent the interquartile and 5th–95th percentile ranges across subjects, respectively.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(paired difference: -4.20 ; 95% patient-bootstrap CI: $[-8.88, 0.77]$; paired t -test $p = 0.144$; Wilcoxon $p = 0.193$). HBGI changed from 14.32 to 5.04 (paired difference: -9.29 ; 95% patient-bootstrap CI: $[-21.55, 0.27]$; paired t -test $p = 0.153$; Wilcoxon $p = 0.557$). Neither component reached statistical significance when analyzed separately, despite their combined contribution to the significant BGRI reduction reported in Section 3.1.

Fig. 6 shows the temporal evolution of instantaneous glycemic risk. The controller profiles were initially similar, after which SA exhibited a sustained increase in mean risk. DA maintained a lower and more damped profile over most of the 24-h horizon, with residual peaks near 450, 850, and 1300 min. The wider SA uncertainty bands, particularly during the second half of the simulation, reflect the subject-specific heterogeneity also visible in the heatmap.

These results characterize the behavior of the complete DA architecture and do not isolate the contributions of learned modulation and deterministic runtime constraints. Component-specific attribution is addressed by the ablation analysis in Section 3.5.

3.10. Control variability grid analysis

Control Variability Grid Analysis (CVGA) [38] was used to characterize the joint occurrence of low- and high-glucose excursions. Each point is defined by the 2.5th and 97.5th percentiles of a 24-h glucose trajectory. For the pooled analysis, each marker represents one subject–scenario combination after averaging the CVGA coordinates across the five training seeds. Neither controller produced outcomes in zone A. Under SA, 58% of outcomes were assigned to zone C and the remaining 42% to zone D. Under DA, 18% of outcomes shifted to zone B and no

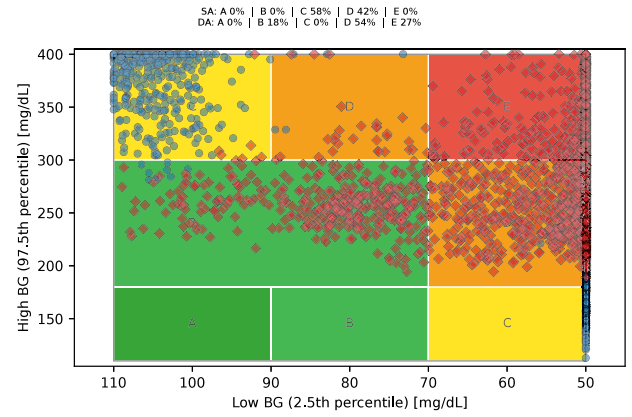


Fig. 7. Control Variability Grid Analysis comparing the single-agent Maskable PPO controller (SA; blue circles) and the complete dual-agent framework (DA; red diamonds). Each marker represents one subject–scenario outcome, defined by the 2.5th and 97.5th percentiles of the corresponding 24-h glucose trajectory and averaged across five training seeds. Background colors denote the CVGA zones: A–B in green, C in yellow, D in orange, and E in red.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

outcomes remained in zone C. However, this favorable shift was accompanied by a larger concentration in the higher-risk zones: 54% of DA outcomes were assigned to zone D and 27% to zone E.

Thus, the DA distribution did not exhibit a uniform shift toward safer CVGA regions. Although DA introduced a nonzero proportion of zone-B outcomes and eliminated zone-C outcomes, the combined proportion in zones D and E increased from 42% under SA to approximately 81% under DA. The appearance of zone-E outcomes under DA further indicates that the complete architecture did not prevent severe joint glucose excursions in a substantial subset of the evaluated conditions. Percentages may not sum to 100% because of rounding (Fig. 7).

3.11. Comparison with classical control references

DA and SA were additionally compared with PID and MPC references under the stochastic protocol described in Section 2.6. These comparisons provide engineering context rather than strictly information-matched benchmarks. PID and MPC received meal-announcement information unavailable to SA and DA, while MPC additionally relied on the subject-specific physiological model employed by the simulation environment.

As summarized in Table 11, DA achieved the highest mean TIR (69.96%), narrowly exceeding MPC (68.79%). DA also achieved the lowest TAR (18.05%) and the lowest Risk Index (8.82). MPC achieved the lowest TBR (2.85%) and the lowest inter-patient TIR variability (11.06%). SA exhibited the lowest cohort-average incidence of runs containing severe hypoglycemia (60.06%), narrowly below MPC (61.12%).

The classical controllers therefore provide useful contextual references, but their results should not be interpreted as equivalent-information comparisons. The DA-versus-SA analysis remains the principal assessment of the proposed supervisory architecture because both controllers operated under the same observation and meal-information structure.

3.12. Severe hypoglycemia analysis

Severe hypoglycemia was analyzed separately because TBR does not distinguish the number of severe episodes, their duration, or the proportion of 24-h evaluations affected by at least one excursion. A distinct event was defined as a contiguous sequence of one or more CGM measurements below 54 mg/dL.

Table 11

Comprehensive performance comparison of control strategies. Clinical metrics are reported as cohort mean $\pm$ inter-subject SD across 10 virtual pediatric subjects. For SA and DA, each subject was evaluated over 200 stochastic 24-h scenarios for each of five training seeds. Severe hypoglycemia incidence is the percentage of evaluation runs containing at least one glucose measurement below 54 mg/dL. Inter-patient CV_{TIR} is computed as $(SD_{TIR}/mean_{TIR}) \times 100$. The best value for each quantitative metric is shown in bold.

Metric / aspect	PID-based HCL	MPC	Single-agent PPO	Dual-agent RL (proposed)
<i>Clinical performance (cohort mean $\pm$ SD)</i>				
TIR (%) [70–180 mg/dL] $\uparrow$	62.13 $\pm$ 10.42	68.79 $\pm$ 7.61	42.86 $\pm$ 14.53	69.96 $\pm$ 9.17
TBR (%) [$<$ 70 mg/dL] $\downarrow$	4.71 $\pm$ 2.20	2.85 $\pm$ 2.10	25.56 $\pm$ 23.77	11.99 $\pm$ 8.49
TAR (%) [$>$ 180 mg/dL] $\downarrow$	33.16 $\pm$ 8.60	28.36 $\pm$ 6.50	31.58 $\pm$ 33.43	18.05 $\pm$ 6.39
Mean Risk Index (LBGI + HBGI) $\downarrow$	15.23 $\pm$ 4.61	12.08 $\pm$ 3.14	22.31 $\pm$ 16.17	8.82 $\pm$ 4.53
Severe hypoglycemia incidence ^a (%) $\downarrow$	78.02 $\pm$ 27.16	61.12 $\pm$ 49.47	60.06 $\pm$ 51.56	71.89 $\pm$ 34.55
<i>Robustness</i>				
Inter-patient CV_{TIR} (%) $\downarrow$	16.77	11.06	33.90	13.11
<i>Modeling and implementation</i>				
Explicit patient model required	No	Yes	No	No
Meal-announcement information	Yes	Yes	No	No
Prediction horizon	None	30 min	Implicit	Implicit + 30-min supervision
Offline training required	No	No	Yes	Yes (two-stage)
Run-time computational load	Very low	Moderate (QP)	Low (1 NN pass)	Low (2 NN passes)
Explicit safety constraints	LGS heuristic	Hard constraints	Reward-based only	Reward + runtime gating

^a Percentage of 24-h evaluation runs containing at least one glucose measurement below 54 mg/dL. The reported value is the cohort mean $\pm$ inter-subject SD of the patient-level incidence percentages.

Note: Meal-schedule randomization was identical across controllers. PID and MPC received carbohydrate announcements, whereas SA and DA did not. MPC additionally used the subject-specific physiological model employed by the simulation environment. The Risk Index was calculated from CGM measurements sampled at 3-min intervals using the Clarke–Kovatchev transformation [31].

Across the complete evaluation set, SA produced 24,705 distinct severe hypoglycemic events, compared with 15,357 under DA. DA therefore produced 9,348 fewer events, corresponding to a descriptive reduction of approximately 37.8%. Mean time below 54 mg/dL also decreased, from $18.75 \pm 18.38\%$ under SA to $7.47 \pm 7.77\%$ under DA.

In contrast, at least one severe low-glucose sample occurred in a larger number of DA evaluations. Severe hypoglycemia affected 6006 of 10,000 SA runs (60.06%; pooled Wilson 95% CI: [59.10, 61.02]%) and 7189 of 10,000 DA runs (71.89%; pooled Wilson 95% CI: [71.00, 72.76]%). Because runs were clustered within subjects, inferential interpretation was based primarily on patient-level estimates rather than on pooled binomial intervals.

The cohort-average patient-level incidence was $60.06 \pm 51.56\%$ under SA and $71.89 \pm 34.55\%$ under DA. The paired difference was 11.83 percentage points, with a patient-bootstrap 95% CI of [−22.97, 48.48] percentage points. Because this interval included zero, the observed increase in the proportion of affected runs was not statistically significant at the patient level (Table 12).

Taken together, DA produced fewer distinct events and shorter severe low-glucose exposure, but severe samples were distributed across a larger proportion of evaluation runs. The lower event count and exposure duration should therefore not be interpreted as evidence that DA reduced the probability of encountering at least one severe low-glucose sample.

4. Discussion

The present study evaluated a dual-agent reinforcement-learning architecture for automated insulin delivery under stochastic meal disturbances. Relative to the information-matched single-agent Maskable PPO baseline (SA), the complete dual-agent framework (DA) increased TIR in all ten virtual pediatric subjects, significantly reduced composite glycemic risk, and produced a patient-bootstrap-supported reduction in inter-patient TIR variability. Together, these findings indicate that combining learned command modulation with deterministic runtime supervision can improve both cohort-level regulation and consistency across heterogeneous pediatric profiles.

Table 12

Severe hypoglycemia outcomes on the final held-out test set. A distinct event was defined as a contiguous sequence of one or more CGM measurements below 54 mg/dL. Pooled incidence was calculated over 10,000 runs per controller. Patient-level values are reported as cohort mean $\pm$ inter-subject SD across 10 virtual pediatric subjects.

Outcome	SA	DA
Distinct events $<$ 54 mg/dL	24,705	15,357
Runs with ≥ 1 event	6,006/10,000	7,189/10,000
Pooled incidence (%)	60.06	71.89
Pooled Wilson 95% CI (%)	59.10–61.02	71.00–72.76
Patient-level incidence (%)	60.06 $\pm$ 51.56	71.89 $\pm$ 34.55
Patient-bootstrap 95% CI (%)	30.06–90.00	49.12–90.09
Time $<$ 54 mg/dL (%)	18.75 $\pm$ 18.38	7.47 $\pm$ 7.77

4.1. Glycemic performance and subject-specific adaptation

The most consistent improvement was observed for TIR, which increased from 42.86% under SA to 69.96% under DA, corresponding to a large paired effect. Cohort-average TBR and TAR also decreased, indicating that the TIR gain did not arise exclusively from a shift toward either lower or higher glucose values.

The subject-specific results further suggest that the supervisory architecture responded to different baseline failure modes. Under SA, child#001–child#006 were characterized primarily by hypoglycemic exposure, whereas child#007–child#010 exhibited negligible TBR but pronounced hyperglycemia. DA predominantly reduced TBR in the first group and TAR in the second. Thus, the same layered architecture altered insulin delivery differently according to the glycemic characteristics of each virtual subject, rather than producing a uniform population-level shift.

The BGRI results provide complementary evidence of improved regulation. Unlike threshold-based time-in-range metrics, BGRI applies nonlinear penalties to the magnitude of both low- and high-glucose deviations. Its significant reduction under DA therefore indicates attenuation of the overall severity of glycemic excursions, in addition to the observed increase in TIR. The numerical reductions in both LBGI and HBGI are

consistent with this interpretation, although the composite BGRI provided the clearest cohort-level statistical evidence.

GMI and MAD offered additional information on the redistribution of glucose exposure. Their cohort averages favored DA, with the largest improvements observed in subjects characterized by pronounced hyperglycemia under SA. Taken together, TIR, BGRI, GMI, and MAD indicate that the main benefit of DA was not limited to a single summary metric but was reflected across several complementary descriptions of glycemic regulation.

4.2. Contribution of the supervisory components

The ablation analysis demonstrates that the learned and deterministic components made distinct contributions to the complete-controller performance. State-dependent action masking improved TIR, reduced TBR, and lowered BGRI relative to the unconstrained Primary-PPO configuration. Adding the predictive floor produced a further increase in TIR and reductions in both TBR and TAR, showing that short-horizon predictive information provided a useful supervisory signal beyond the current glucose observation.

Relative to the common action-masking-plus-predictive-floor configuration, the learned Safety Adjuster produced the larger isolated improvement in TIR and reduced both TBR and TAR. The deterministic runtime constraints produced the larger isolated reduction in TBR, while shifting part of the glucose distribution toward higher values. These complementary effects reflect the different roles of the two supervisory layers: learned modulation performs frequent context-dependent reshaping of the primary command, whereas the deterministic mechanisms impose targeted limits when predefined conditions are reached.

Combining the learned and deterministic layers yielded the highest TIR and the lowest TAR and BGRI among the evaluated ablation configurations. Their interaction was not purely additive: the full framework did not retain the minimum TBR of the hard-constraints-only configuration, but substantially reduced the associated hyperglycemic exposure. The complete architecture therefore achieved the most favorable overall balance among the evaluated configurations rather than optimizing one glycemic endpoint in isolation.

The component-wise ablation also provides a more reliable basis for attribution than supervisory activation frequency alone. The final insulin command reflects the sequential action of the primary policy, predictive correction, learned modulation, and deterministic constraints. Accordingly, differences between intermediate and final commands characterize the combined control process, whereas the ablation results identify the incremental performance associated with each layer.

The 30-min predictor supplied this architecture with a computationally inexpensive short-horizon signal. Its overall MAE, RMSE, and limited cohort-level bias support its use as a lightweight supervisory mechanism rather than as a substitute for a physiological prediction model. Its condition-dependent accuracy also identifies a clear direction for future improvement through uncertainty-aware or adaptive forecasting.

4.3. Severe hypoglycemia and multidimensional risk

The severe-hypoglycemia analysis shows that DA changed both the persistence and distribution of low-glucose exposure. DA produced fewer distinct events and substantially reduced the mean time spent below 54 mg/dL. These findings indicate shorter or less recurrent severe excursions across the complete evaluation set. The run-level incidence metric showed a different pattern, with a higher DA point estimate for runs containing at least one low-glucose sample; however, the corresponding patient-level confidence interval included zero.

These outcomes illustrate the complementary meaning of the reported safety endpoints. Binary run-level incidence assigns the same outcome to a trajectory containing one low measurement and to one containing prolonged severe hypoglycemia, whereas event count and time below threshold capture persistence and recurrence. Reporting all

three measures therefore provides a more informative characterization than any single endpoint and shows that the principal DA benefit concerned the overall burden of severe low-glucose exposure.

The CVGA results provide an additional view of trajectory extremes. DA introduced zone-B outcomes and eliminated zone-C outcomes, while BGRI and inter-patient TIR consistency improved at the cohort level. At the same time, the distribution across zones D and E confirms that CVGA captures tail combinations that are not summarized by average risk measures. BGRI and CVGA should therefore be interpreted as complementary rather than competing assessments of controller behavior.

4.4. Robustness and comparison with classical references

The between-training-seed variability of the cohort-level outcomes was small relative to the DA-SA performance difference. In particular, the TIR gap of approximately 27 percentage points was substantially larger than the variation observed across the five independent policy initializations. This supports the reproducibility of the principal cohort-level finding across independent RL training runs.

The sensitivity analysis likewise showed comparatively limited changes over the tested one-factor-at-a-time perturbations. TIR, TBR, TAR, mean glucose, CV, and insulin delivery remained close to their nominal values across the evaluated reward coefficients, gains, and thresholds. The 30-min cumulative insulin limit produced the largest changes, consistent with its direct effect on short-horizon insulin availability. Overall, the main performance trends were preserved when each parameter was varied individually within the tested range.

The classical controllers provide an informative engineering context. DA achieved the highest mean TIR and the lowest TAR and BGRI, whereas MPC achieved the lowest TBR and inter-patient TIR variability. This endpoint-dependent ranking highlights the different trade-offs produced by learned and model-based control strategies. Because DA and SA used the same observation and meal-information structure, their matched comparison remains the primary assessment of the proposed architecture, while PID and MPC show how its performance relates to established control approaches under the reported experimental conditions.

4.5. Limitations

Several limitations define the scope of these findings. First, the evaluation was conducted entirely *in silico* using ten virtual pediatric subjects. The 200 stochastic scenarios and five training seeds improved within-subject precision but did not increase the number of independent physiological units. Consequently, the nonsignificant subject-level TBR and TAR comparisons should not be interpreted as evidence of equivalence.

Second, guideline-level regulation was not achieved uniformly. Although cohort-average TIR approached 70%, mean TBR remained above the recommended 4% threshold. Child#008 was the clearest residual failure case: DA substantially reduced the extreme hyperglycemia observed under SA and increased TIR, but the resulting redistribution included marked TBR. The higher point estimate for severe-hypoglycemia run incidence and the presence of unfavorable CVGA zones similarly show that improvements in aggregate metrics did not eliminate all adverse excursion patterns.

Third, each virtual subject was associated with dedicated trained controllers. The study therefore evaluated robustness to stochastic within-subject disturbances and training initialization rather than generalization to unseen subjects. Moreover, although parameter selection and final testing used non-overlapping meal scenarios, supervisory thresholds and reward parameters were developed using the same ten physiological profiles later used for evaluation. Scenario-level separation therefore limits test-scenario leakage but does not exclude cohort-specific overfitting.

Fourth, the simulator does not represent the complete range of physiological, behavioral, sensor, and actuator disturbances encountered

in clinical use. Exercise, illness, stress, treatment nonadherence, CGM dropout, calibration drift, compression artifacts, infusion-set occlusion, partial delivery, and pump malfunction were not explicitly evaluated. The glucose-trend estimate also assumes regular sampling and does not include an implemented fallback for missing measurements or time-stamps.

Fifth, the 30-min predictor is a first-order extrapolation whose error increased during rapidly changing glucose trajectories. In particular, its positive bias during steep rises could favor excessive predictive correction, although the present analysis does not establish a causal association between forecast error and individual adverse events. Similarly, the small between-subject variability of cohort averages may conceal subject-specific optimization sensitivity, and the one-factor-at-a-time analysis cannot characterize interactions among parameters or establish global robustness.

Finally, the PID and MPC comparisons were not information matched: both classical references received meal announcements unavailable to SA and DA, and MPC additionally used an explicit subject-specific physiological model. The layered mechanisms are engineering constraints rather than formally verified safety guarantees, and the study did not include real CGM data, embedded-hardware testing, or prospective clinical validation. The results should therefore be interpreted as comparative in-silico evidence rather than as evidence of clinical effectiveness or safety.

4.6. Future work

Future work should build on the complementary behavior identified by the ablation analysis. Adaptive insulin budgets, subject-specific supervisory thresholds, and uncertainty-aware predictive correction may further improve the balance between hypo- and hyperglycemic exposure, with child#008 providing a useful stress-test profile for controller refinement.

Evaluation should be extended to combined physiological and device-level disturbances, including exercise, illness, meal-estimation errors, sensor noise and dropout, insulin-absorption variability, infusion failures, and pump faults. Simultaneous multi-parameter perturbations would complement the present local sensitivity analysis, while explicit activation flags would allow each deterministic mechanism to be quantified independently.

Population-level pre-training, transfer learning, and domain adaptation should also be investigated to determine whether policies can be generalized and subsequently personalized to unseen subjects. Progression through larger digital cohorts, retrospective clinical data, hardware-in-the-loop testing, formal safety analysis, and staged prospective evaluation would provide the appropriate pathway for assessing the translational potential of the proposed architecture.

5. Conclusion

This study presented a safety-supervised dual-agent reinforcement-learning framework for pediatric automated insulin delivery, integrating a primary Maskable PPO policy with predictive correction, learned command modulation, and deterministic runtime constraints.

Across the evaluated in-silico pediatric cohort, the complete framework increased TIR relative to the information-matched single-agent baseline in all ten virtual subjects, significantly reduced composite glycemic risk, and improved inter-patient consistency. The ablation analysis demonstrated complementary contributions from the supervisory components: learned modulation provided the larger isolated improvement in TIR, whereas the deterministic constraints produced the larger isolated reduction in TBR. Their integration achieved the highest TIR and the lowest TAR and BGRI among the evaluated configurations.

Overall, these findings support the proposed separation of regulation-oriented learning, learned supervisory modulation, and deterministic runtime enforcement as a promising approach to pediatric

AID. Although the results remain simulation based and guideline-level outcomes were not achieved uniformly, the framework provides a modular foundation for broader robustness testing, generalization to unseen subjects, and progressive preclinical validation.

CRediT authorship contribution statement

Giada Lops: Writing – original draft, Software, Methodology, Conceptualization; **Luca De Cicco:** Supervision, Funding acquisition; **Save-rio Mascolo:** Supervision, Funding acquisition.

Ethical approval

This research did not involve human participants or animal experiments. All results were obtained using the open-source *Simglucose* simulator, which models glucose–insulin dynamics in silico.

Funding

This work has been co-funded by the European Union and the Swiss State Secretariat for Education, Research and Innovation (SERI), in the frame of the SHIELD (grant N° 101156751) project. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or SERI. Neither the European Union nor the granting authorities can be held responsible for them.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could influence this work.

Image and data integrity

All figures were generated directly from simulation data using Python's *matplotlib* and *seaborn* libraries. No features were enhanced, obscured, or removed from the original simulation outputs. Brightness and contrast adjustments were applied uniformly for clarity. No generative AI or AI-assisted image editing tools were used to create or alter any figures, in accordance with Elsevier's GenAI policy.

Sex- and Gender-Based Considerations

This study used virtual pediatric patient models from the UVA/Padova simulator, which represent physiological variability but do not encode sex or gender characteristics. Therefore, sex- and gender-based analyses could not be performed. This limitation should be considered when extrapolating findings to real-world cohorts, where biological sex and gender behaviors may influence insulin sensitivity and glucose regulation.

Data Availability Statement

Reference code and configuration settings are available on [GitHub](#). Additional code and data underlying the reported results are available from the corresponding author upon reasonable request.

References

- [1] R.I.G. Holt, J.H. DeVries, A. Hess-Fischl, I.B. Hirsch, M.S. Kirkman, T. Klupa, B. Ludwig, K. Nørgaard, J. Pettus, E. Renard, J.S. Skyler, F.J. Snoek, R.S. Weinstock, A.L. Peters, The management of type 1 diabetes in adults: a consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD), *Diabetes Care* 44 (11) (2021) 2589–2625. ADA/EASD consensus report, <https://doi.org/10.2337/dci21-0043>
- [2] S. Passanisi, A. Chobot, C. Piona, Glycemic control in children and adolescents with type 1 diabetes: from challenges to innovation, *Children* 12 (6) (2025) 708. <https://doi.org/10.3390/children12060708>
- [3] A. Cinar, Automated insulin delivery algorithms, *Diabetes Spectr.* 32 (3) (2019) 209–214. <https://doi.org/10.2337/ds18-0100>

- [4] R. Hovorka, V. Canonico, L.J. Chassin, U. Haueter, M. Massi-Benedetti, M. Orsini Federici, T.R. Pieber, H.-C. Schaller, L. Schaupp, T. Vering, M.E. Wilinska, Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes, *Physiol. Meas.* 25 (4) (2004) 905–920. <https://doi.org/10.1088/0967-3334/25/4/010>
- [5] Y.C. Kudva, R.M. Bergenstal, B.W. Bode, B. Draznin, S. Garg, G. Grunberger, I.B. Hirsch, A.L. Peters, O. Schnell, W.V. Tamborlane, J. Walsh, S.A. Weinzimer, M. Phillip, Y. Handelsman, J.H. DeVries, F.J. Doyle, E. Renard, S.J. Russell, o.b. o. t.E. Society, Approaches to closed-loop automated insulin delivery: consensus report of the endocrine society, *J. Clin. Endocrinol. Metab.* 107 (5) (2022) 1253–1268. <https://doi.org/10.1210/clinem/dgab865>
- [6] C.K. Boughton, R. Hovorka, Is an artificial pancreas (closed-loop system) for type 1 diabetes effective?, *N. Engl. J. Med.* 383 (9) (2020) 836–845. <https://doi.org/10.1056/NEJMr1908013>
- [7] S.A. Brown, B.P. Kovatchev, D. Raghinaru, J.W. Lum, B.A. Buckingham, Y.C. Kudva, L.M. Laffel, C.J. Levy, J.E. Pinsky, R.P. Wadwa, et al., Six-month randomized trial of a closed-loop system in type 1 diabetes, *N. Engl. J. Med.* 381 (18) (2019) 1707–1717. <https://doi.org/10.1056/NEJMoa1907863>
- [8] J.L. Sherr, L. Heinemann, G.A. Fleming, R.M. Bergenstal, D. Bruttomesso, H. Hanaire, R.W. Holl, J.R. Petrie, A.L. Peters, M. Evans, Automated insulin delivery: benefits, challenges, and recommendations. A consensus report of the joint diabetes technology working group of the European association for the study of diabetes and the American diabetes association, *Diabetes Care* 45 (12) (2022) 3058–3074. <https://doi.org/10.2337/dci22-0018>
- [9] J.L. Sherr, B.A. Buckingham, R.P. Wadwa, D.J. DeSalvo, E. Cengiz, S.A. Weinzimer, M.D. Breton, B.P. Kovatchev, G.P. Forlenza, Two-year outcomes with the omnipod 5 automated insulin delivery system in children, adolescents, and adults with type 1 diabetes, *Diabetes Technol. Ther.* 25 (11) (2023) 769–777. Long-term real-world follow-up study, <https://doi.org/10.1089/dia.2023.0213>
- [10] S. Di Molfetta, L. Di Gioia, I. Caruso, A. Cignarelli, S.C. Green, P. Natale, G.F.M. Strippoli, G.P. Sorice, S. Perrini, A. Natalicchio, L. Laviola, F. Giorgino, Efficacy and safety of different hybrid closed loop systems for automated insulin delivery in people with type 1 diabetes: a systematic review and network meta-analysis, *Diabetes/Metab. Res. Rev.* 40 (2024) e3842. <https://doi.org/10.1002/dmrr.3842>
- [11] S.L. Kang, Y.N. Hwang, J.Y. Kwon, S.M. Kim, Effectiveness and safety of a model predictive control (MPC) algorithm for an artificial pancreas system in outpatients with type 1 diabetes (T1D): systematic review and meta-analysis, *Diabetol. Metab. Syndr.* (2022). <https://doi.org/10.1186/s13098-022-00962-2>
- [12] B. Bombaci, S. Passanisi, F. Lombardo, G. Salzano, Clinical relevance of short-term glycaemic variability in children and adolescents with type 1 diabetes: a narrative review, *Transl. Pediatr.* 13 (7) (2024) 1231–1241. <https://tp.amegroups.org/article/view/126512/html>. <https://doi.org/10.21037/tp-24-114>
- [13] K. Dove, C. Boughton, M. Tauschmann, H. Thabit, L. Bally, J.M. Allen, C.L. Acerini, S. Arnolds, C. de Beaufort, R.M. Bergenstal, F. Campbell, A. Criego, D.B. Dunger, D. Elleri, M.L. Evans, E. Fröhlich-Reiterer, S. Hofer, T. Kapellen, L. Leelarathna, T.R. Pieber, B. Rami-Merhar, V.N. Shah, J. Sibayan, M.E. Wilinska, R. Hovorka, APCam11 Consortium, AP@Home Consortium, KidsAP Consortium, Young children have higher variability of insulin requirements: observations during hybrid closed-loop insulin delivery, *Diabetes Care* 42 (7) (2019) 1344–1347. <https://doi.org/10.2337/dc18-2625>
- [14] A.D. Monzon, L.B. Smith, S.W. Powers, L.M. Dolan, S.R. Patton, The association between glycaemic variability and macronutrient intake in youth with type 1 diabetes, *Front. Endocrinol.* 11 (2020) 543. <https://doi.org/10.3389/fendo.2020.00543>
- [15] R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2nd edition, 2018. <http://incompleteideas.net/book/the-book-2nd.html>
- [16] F. Baldisseri, D. Menegatti, A. Wrona, Deep deterministic policy gradient control of type 1 diabetes, in: *Proceedings of the European Control Conference (ECC), IEEE*, 2024, pp. 868–873. <https://doi.org/10.23919/ECC64448.2024.10591007>
- [17] M.M.H. Atanasiou, V. Becchetti, F. Baldisseri, D. Menegatti, A. Wrona, Deep reinforcement learning control of type-1 diabetes with cross-patient generalization, in: *Proceedings of the 32nd Mediterranean Conference on Control and Automation (MED)*, IEEE, 2024, pp. 1–6. <https://doi.org/10.1109/MED61351.2024.10566186>
- [18] G. Lops, V.A. Racanelli, G. Manfredi, L. De Cicco, S. Mascolo, A deep deterministic policy gradient control algorithm for automatic insulin delivery, *IFAC-Pap.* 59 (2) (2025) 13–18. 1st IFAC Workshop on Engineering Diabetes Technologies EDT 2025, <https://doi.org/10.1016/j.ifacol.2025.06.003>
- [19] T. Yamagata, A. O’Kane, A. Ayobi, D. Katz, K. Stawarz, P. Marshall, P. Flach, R. Santos-Rodríguez, Model-Based Reinforcement Learning for Type 1 Diabetes Blood Glucose Control, (2020). [arXiv:2010.06266](https://arxiv.org/abs/2010.06266) [cs.LG]
- [20] T. Zhu, K. Li, P. Herrero, P. Georgiou, Basal glucose control in type 1 diabetes using deep reinforcement learning: an in silico validation, *IEEE J. Biomed. Health Inform.* 25 (4) (2021) 1223–1232. <https://doi.org/10.1109/JBHI.2020.3014556>
- [21] P. Viroonluecha, E. Egea-Lopez, J. Santa, Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning, *PLoS ONE* 17 (9) (2022) e0274608. <https://doi.org/10.1371/journal.pone.0274608>
- [22] M. Kwon, J. Lamp, L. Feng, Safety Generalization Under Distribution Shift in Safe Reinforcement Learning: A Diabetes Testbed, 2026. Accepted at ICML 2026. [arXiv:2601.21094](https://arxiv.org/abs/2601.21094) [cs.LG]
- [23] G. Lops, G. Manfredi, V.A. Racanelli, L. De Cicco, S. Mascolo, A safety aware deep reinforcement learning technique for automated insulin delivery, in: 2025 33rd Mediterranean Conference on Control and Automation (MED), 2025, pp. 120–125. <https://doi.org/10.1109/MED64031.2025.11073297>
- [24] S. Huang, S. Ontañón, A Closer Look at Invalid Action Masking in Policy Gradient Algorithms, (2020). [arXiv:2006.14171](https://arxiv.org/abs/2006.14171) [cs.LG]
- [25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms, 2017. [arXiv:1707.06347](https://arxiv.org/abs/1707.06347)
- [26] T. Battelino, T. Danne, R.M. Bergenstal, et al., Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range, *Diabetes Care* 42 (8) (2019) 1593–1603. <https://doi.org/10.2337/dci19-0028>
- [27] J. Xie, SimGlucose v0.2.1: An Open-Source Simulation Environment for Reinforcement Learning in Type 1 Diabetes, 2018, (<https://github.com/jxx123/simglucose>). [Online; accessed 12 Nov. 2025]. <https://github.com/jxx123/simglucose>
- [28] C. Dalla Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, C. Cobelli, The UVA/PADOVA type 1 diabetes simulator: new features, *J. Diabetes Sci. Technol.* 8 (1) (2014) 26–34. <https://doi.org/10.1177/1932296813514502>
- [29] C. Cobelli, E. Renard, B.P. Kovatchev, Artificial pancreas: past, present, future, *Diabetes* 60 (11) (2011) 2672–2682. <https://doi.org/10.2337/db11-0654>
- [30] B.W. Bequette, Algorithms for a closed-loop artificial pancreas: the case for model predictive control, *J. Diabetes Sci. Technol.* 7 (6) (2013) 1632–1643. <https://doi.org/10.1177/193229681300700620>
- [31] W.L. Clarke, B.P. Kovatchev, Statistical tools to analyze continuous glucose monitor data, *Diabetes Technol. Therap.* 11 (Suppl 1) (2009) S45–S54. <https://doi.org/10.1089/dia.2008.0138>
- [32] A. Bertachi, C. Ramkissoon, J. Bondia, et al., The use of insulin-on-board constraints to improve safety of closed-loop systems, *J. Diabetes Sci. Technol.* 14 (3) (2020) 598–606. <https://doi.org/10.1177/1932296819898706>
- [33] B.A. Buckingham, D. Raghinaru, F. Cameron, et al., Effectiveness of predictive low-glucose suspend in preventing hypoglycemia in children, adolescents, and adults with type 1 diabetes, *Diabetes Care* 38 (6) (2015) 1106–1113. <https://doi.org/10.2337/dc14-2555>
- [34] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann, Stable-Baselines3: reliable reinforcement learning implementations, *J. Mach. Learn. Res.* 22 (268) (2021) 1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [35] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- [36] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, Í. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in python, *Nat. Methods* 17 (3) (2020) 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- [37] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.
- [38] L. Magni, D.M. Raimondo, L. Bossi, C. Dalla Man, G. De Nicolao, B.P. Kovatchev, C. Cobelli, Evaluating the efficacy of closed-loop glucose regulation via control variability grid analysis, *J. Diabetes Sci. Technol.* 2 (4) (2008) 630–635. <https://doi.org/10.1177/193229680800200413>