2026 European Control Conference (ECC)
July 7-10, 2026. Reykjavík, Iceland

Bridging Clinical Knowledge and Reinforcement Learning in Automated Insulin Delivery: An LLM-in-the-Loop Approach

Giada Lops[1], Taha Ramdan[2], Vito Andrea Racanelli[1], Luca De Cicco[1], and Saverio Mascolo[1]

***Abstract*— Automated insulin delivery (AID) requires controllers that are both adaptive and safe. This study proposes a hybrid control framework that combines a reinforcement learning (RL) agent with a language-guided advisory layer in the *SimGlucose* simulator of the FDA-approved UVA/Padova type 1 diabetes model. A Proximal Policy Optimization (PPO) agent learns insulin dosing via an asymmetric, safety-weighted reward, while a fine-tuned Falcon-RW-1B model provides guideline-consistent recommendations. A supervisory fusion rule merges both outputs according to policy uncertainty and medical constraints (suspension < 90 mg/dL, recovery cap > 70 mg/dL, rate limit 0.03 U/min). Across ten virtual adults and twenty stochastic meal scenarios, the hybrid RL+LLM controller achieved a time-in-range of $86\% \pm 7.3\%$ and reduced exposure to hypoglycemia compared with the considered baselines, while maintaining total insulin delivery within a limited deviation from reference levels. A quasi-counterfactual analysis indicated strong alignment between rule activations and action changes (fidelity $\approx$1.0, validity $\geq$90%). These results suggest that hybrid RL–LLM architectures are a promising direction for safe and adaptive closed-loop insulin control.**

I. Introduction

The regulation of blood glucose concentration in individuals with type 1 diabetes mellitus (T1DM) is a challenging control problem affected by disturbances such as meals, stress, and physical activity. The clinical objective is to maintain glucose within the recommended euglycemic range of $70\,\mathrm{mg/dL}$ to $180\,\mathrm{mg/dL}$ to prevent hypo- and hyperglycemia and reduce long-term complications [1].

Automated Insulin Delivery (AID) systems, also known as *artificial pancreas* technologies, address this problem through a feedback loop involving a Continuous Glucose Monitoring (CGM) sensor, an insulin pump, and a control algorithm that computes insulin dosing every few minutes [2]. As shown in Fig. 1, CGM measurements $G_m(t)$ are compared with a reference level G_{ref} to generate the control error $e(t)$, from which the insulin infusion rate $u(t)$ is determined. Early AID systems relied on proportional–integral–derivative (PID) controllers, while more recent approaches employ Model Predictive Control (MPC), now implemented in commercial hybrid closed-loop systems such as Medtronic 780G and Tandem Control-IQ [3].

This work has been partially funded from the European Union's Horizon Europe research and innovation programme under the project SHIELD, grant agreement No. 101156751.

[1]Department of Electrical and Information Engineering, Politecnico di Bari, Italy. g.lops@phd.poliba.it, {vitoandrea.racanelli, luca.decicco, saverio.mascolo}@poliba.it

[2]École Polytechnique Universitaire de Marseille (Polytech Marseille), Aix-Marseille University, France. taha.ramdan@etu.univ-amu.fr

Despite these advances, classical controllers require individualized models and careful tuning to handle physiological variability, insulin absorption delays, and meal disturbances. Model uncertainty and sensor noise can therefore lead to conservative or unsafe insulin delivery, while rule-based safety layers often improve safety at the expense of responsiveness [1].

Reinforcement Learning (RL) has emerged as a promising data-driven approach that learns insulin dosing policies directly from patient–environment interactions [4]. By optimizing long-term reward functions encoding clinical objectives, RL can achieve adaptive glucose regulation without explicit physiological models. However, RL policies typically rely on complex nonlinear function approximators, making their decisions difficult to interpret and clinically validate [5]. This limitation motivates hybrid control architectures that combine RL adaptability with structured supervisory mechanisms to improve transparency and safety.

The UVA/Padova model [6] and its open-source implementation SimGlucose [7] provide reproducible environments for evaluating closed-loop insulin controllers. Several studies have applied deep RL algorithms, including Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO), to these simulators [8], [9]. For example, Baldisseri *et al.* [10] proposed a DDPG-based controller validated on the Bergman model, while a subsequent study compared multiple DRL algorithms (DDPG, PPO, SAC, and TD3) for glucose regulation under meal uncertainty [11]. Although these approaches demonstrate promising *in silico* performance, their decision processes remain difficult to interpret and formally verify.

To address these limitations, safe and interpretable RL frameworks have been proposed, incorporating constraint-handling techniques such as Lyapunov-based methods, risk-sensitive optimization, and constrained policy updates [12]–[14]. Complementary work has explored hybrid symbolic–learning approaches where rule-based components improve safety and interpretability.

More recently, *Large Language Models* (LLMs) have been investigated as a mechanism for integrating structured reasoning into decision-making systems. LLMs can provide contextual recommendations and natural-language explanations [15], and have been incorporated into *language-in-the-loop* control frameworks where textual reasoning supports policy design or action interpretation [16]. Hybrid LLM–RL approaches have also been explored in domains such as autonomous driving and multi-task control [17], [18]. However, LLMs generally lack temporal state awareness and

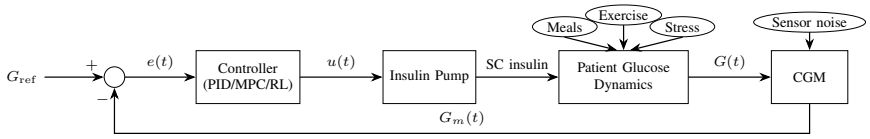

Fig. 1. Block diagram of an Automated Insulin Delivery (AID) system illustrating the feedback control loop among the CGM sensor, control algorithm (PID, MPC, or RL), insulin pump, and patient glucose dynamics. External disturbances such as meals, exercise, and stress motivate adaptive and safety-aware control strategies.

principled uncertainty estimation [19], limiting their use as standalone controllers in dynamical systems.

These observations motivate hybrid architectures in which learning-based control policies are complemented by structured advisory mechanisms capable of incorporating domain knowledge. Building on this idea, we propose a hybrid framework for safe and adaptive automated insulin delivery.

The main contributions of this work are:

- *Hybrid RL–language control:* a PPO-based insulin controller augmented by a language-based advisory module providing guideline-inspired recommendations.
- *Uncertainty-aware fusion:* a supervisory rule that merges RL and advisory actions according to policy dispersion and predefined safety limits.
- *Evaluation:* quantitative and qualitative analysis of RL-based and hybrid controllers across multiple virtual subjects and stochastic meal scenarios, including quasi-counterfactual safety validation.

The remainder of this paper is organized as follows. Section II describes the modelling framework and hybrid control design. Section III presents the experimental evaluation. Section IV concludes the paper and outlines future research directions.

II. Methodology

This section describes the experimental setup, control architecture, learning algorithm, and evaluation protocol used to assess the proposed insulin controllers.

A. Training Environment

All experiments were performed using the open-source *SimGlucose* simulator [7], derived from the UVA/Padova v.2008 T1DM model [6]. The simulator reproduces glucose–insulin dynamics and enables reproducible *in silico* testing. The virtual patient model includes subcutaneous glucose absorption, insulin kinetics, and endogenous glucose production with configurable meal events and sensor noise.

Continuous Glucose Monitoring (CGM) data are provided every 3 minutes, consistent with commercial sensor sampling rates. Each 24-hour simulation (480 control steps) included three main meals of 30, 60, and 40 g of carbohydrates at 06:00, 12:00, and 18:00, respectively, with a nominal basal rate of 1 U/h. To emulate realistic variability, a truncated Gaussian jitter was applied to meal timing (±30 min) and carbohydrate content (±7–10 g), using reproducible random seeds.

At each control step, the agent observes the state vector $s_t = [g_t,\ \Delta g_t,\ \mathrm{IOB}_t,\ \tau_t]$, where g_t is the current CGM glucose measurement (mg/dL), Δg_t denotes the short-term glucose trend, IOB_t represents insulin-on-board, and τ_t encodes contextual timing information such as time-of-day. Based on this observation, the controller outputs an insulin infusion command a_t (U/min) bounded in $[0, 2]$ to ensure physiological safety and compatibility with pump hardware limits [6].

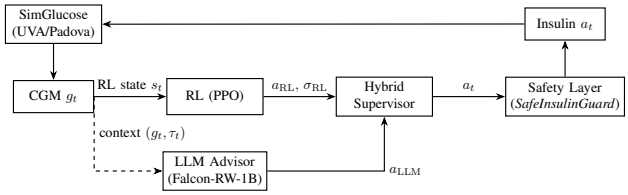

Fig. 2. Hybrid insulin control architecture combining the PPO-based RL controller, the language-based advisory module, and the *SafeInsulinGuard* safety layer. The PPO controller receives the RL state s_t, whereas the LLM advisor operates on a reduced contextual input. The hybrid supervisor fuses a_{RL} and a_{LLM} according to Eq. (7).

B. Reward Design and Safety Wrappers

The controller is trained to maintain glucose within the target range of 70–180 mg/dL while penalizing excessive or abrupt insulin delivery [20]. The instantaneous reward is defined as

$$r_t = -\,w(g_t)\left[\frac{(g_t - g^*)}{\sigma_g}\right]^2 - \lambda_a\,\bar{a}_t - \lambda_{\Delta a}\left|\bar{a}_t - \bar{a}_{t-1}\right|, \quad (1)$$

where $\bar{a}_t$ denotes the safety-filtered insulin rate actually applied. The weight $w(g_t)$ introduces asymmetric penalties to emphasize hypoglycemic risk:

$$w(g_t) = \begin{cases} 8, & g_t < g_{\text{hypo}}, \\ 3, & g_t > g_{\text{hyper}}, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

A supervisory module, termed *SafeInsulinGuard*, enforces state-dependent actuation limits [21]. For hypoglycemic states ($g_t < g_{\text{safe}}$), insulin delivery is suspended ($a_t = 0$). Within the near-euglycemic band ($g_{\text{safe}} \le g_t < g_{\text{high}}$), the command is capped at a conservative upper bound $\bar{a}_{\text{tight}}$, i.e., $a_t = \min(a_t, \bar{a}_{\text{tight}})$. For hyperglycemic conditions ($g_t \ge g_{\text{high}}$), the limit is relaxed to $\bar{a}_{\max}$.

A rate limiter enforces smooth actuation changes, $|\bar{a}_t - \bar{a}_{t-1}| \le \Delta\bar{a}_{\max}$, while, following any hypoglycemic episode ($g_t < g_{\text{hypo}}$), the maximum allowable rate is temporarily reduced to $\bar{a}_{\text{rec}}$ for N_{rec} steps. The coupling of the

TABLE I

NOMINAL PARAMETERS USED IN THE REWARD DESIGN, SAFETY LAYER, AND FUSION RULE.

Symbol	Description	Value / Unit
Reward		
g^*, σ_g	Target and normalization glucose	120, 60 mg/dL
$w(g_t)$	Glucose weighting factor	{1, 3, 8}
λ_a, $\lambda_{\Delta a}$	Action and rate penalties	30, 0.5
a_t, $\bar{a}_t$	Raw and filtered insulin rate	U/min
Safety		
g_{safe}, g_{high}	Lower and upper safety bounds	90, 110 mg/dL
g_{hypo}, g_{hyper}	Hypo- and hyperglycaemia thresholds	70, 180 mg/dL
$\bar{a}_{\text{tight}}$, $\bar{a}_{\max}$	Tight and maximum actuation rate	0.02, 2.00 U/min
$\bar{a}_{\text{rec}}$, $\Delta\bar{a}_{\max}$	Recovery cap and rate change limit	0.10, 0.03 U/min
N_{rec}	Recovery duration	10 steps
Fusion		
a_{RL}, a_{LLM}	RL and LLM insulin command	U/min
σ_{low}, σ_{high}	Policy dispersion thresholds	0.01, 0.05
δa	Action deviation	U/min
ε	Deviation tolerance	0.01 U/min
β	LLM weighting under high dispersion	0.6

asymmetric reward (1)–(2) with this safety logic promotes stable regulation while respecting physiological and actuator constraints [22].

C. Reinforcement Learning Controller

Mathematical Background: The insulin control task is formulated as a continuous-state Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where the state $s_t \in \mathcal{S}$ corresponds to the observation vector defined in Section II-A, the action $a_t \in \mathcal{A}$ denotes the insulin infusion rate (U/min), and $r_t = r(s_t, a_t)$ is the reward defined in (1). The objective is to maximize the expected discounted return

$$J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]. \tag{3}$$

Policy optimization follows the clipped surrogate objective of Proximal Policy Optimization (PPO) [23]:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big], \tag{4}$$

where $\rho_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and $\hat{A}_t$ is the advantage estimated via Generalized Advantage Estimation (GAE) [24]. Gradient ascent on (4) is performed with the Adam optimizer.

Implementation: The baseline controller employs PPO from *Stable-Baselines3* [25], chosen for its robustness and favourable bias–variance trade-off under nonlinear stochastic dynamics [26]. Both actor and critic networks consist of two fully connected layers (64 units, $\tanh$ activation). Training proceeds over sixteen simulated days (approximately 7680 control steps) using the environment described in Section II-A. At each step the agent outputs an insulin command a_t, which is filtered through the *SafeInsulinGuard* before actuation. Updates use $\gamma = 0.99$, $\lambda = 0.95$, learning rate 3×10^{-4}, clipping threshold $\epsilon = 0.2$, and ten epochs per update. Deterministic seeding ensures reproducibility.

D. Language-Based Advisor

Mathematical Background: Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ denote structured supervision pairs where $x^{(i)}$ encodes reduced physiological context (current glucose, announced carbohydrate intake, and time-of-day) and $y^{(i)}$ represents the insulin recommendation as text. The language model parameterized by θ defines a conditional distribution $p_\theta(y|x)$ and is trained by minimizing the negative log-likelihood

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,y)}\sum_{t=1}^{T} \log p_\theta(y_t|x, y_{<t}), \tag{5}$$

where $y_{<t}$ are preceding tokens [27]. Parameter-efficient adaptation is achieved through Low-Rank Adapters (LoRA) [28], which introduce a rank-r correction $\Delta W = BA$ to frozen attention weights W.

The advisor uses the *Falcon-RW-1B* transformer [29], fine-tuned with LoRA to encode guideline-inspired input–output mappings and produce clinically consistent recommendations. In this architecture the language model acts as a structured advisory module operating on a reduced contextual representation rather than as a full control policy.

Synthetic Data Generation: Supervision data were synthetically generated using XML files modelled after the *OhioT1DM Dataset* schema [30]. Each glucose reading g_t was paired with an insulin label

$$\hat{u}_t = 0.05(g_t - 100) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 0.5), \tag{6}$$

truncated to $\hat{u}_t \geq 0$. Each record was converted into a natural-language instruction pair such as "Glucose: 145 mg/dL, Carbs: 30 g, Time: 12:00" paired with "{insulin: 2.3}". This preserves the dataset schema while providing privacy-safe supervision.

Fine-Tuning Procedure: The *Falcon-RW-1B* checkpoint was fine-tuned for two epochs using the *transformers*, *trl*, and *datasets* libraries with a *PyTorch* backend. LoRA adapters were applied to attention projections with rank 8, scaling 16, and dropout 0.05. Optimization used AdamW (learning rate 10^{-4}, batch size 1, gradient accumulation 2, FP16 precision) and a 128-token context with early stopping on validation loss. The resulting model produces deterministic JSON-formatted insulin suggestions.

Inference and Control Integration: During closed-loop operation the advisor receives contextual input (g_t, c_t, τ_t), where g_t is the glucose measurement, c_t the announced carbohydrate amount when available, and τ_t the time-of-day. The advisor outputs $\{\text{insulin} : x\}$ where x (U) denotes the recommended dose. Assuming uniform delivery over 15 minutes, this corresponds to a constant rate of $x/15$ U/min saturated within $[0, 2]$ U/min. Additional safety logic suspends delivery for $g_t < 90$ mg/dL and limits correction frequency to one per hour.

E. Hybrid Supervisor and Fusion Logic

The hybrid supervisor combines the RL controller with the language-based advisor through an uncertainty-aware arbitration rule (Fig. 2). At each step the PPO agent outputs an insulin command a_{RL} together with a policy-dispersion measure σ_{RL}, while the advisor produces an auxiliary recommendation a_{LLM}. The dispersion σ_{RL} is computed as

the standard deviation of the stochastic action distribution parameterized by the PPO actor at inference time. It therefore represents a local measure of policy dispersion rather than a formal estimate of epistemic uncertainty.

The final command applied to the patient model is

$$a_t = \begin{cases} a_{\mathrm{RL}}, & \sigma_{\mathrm{RL}} \le \sigma_{\mathrm{low}} \text{ and } \delta a < \varepsilon \\ (1-\beta)a_{\mathrm{RL}} + \beta a_{\mathrm{LLM}}, & \sigma_{\mathrm{RL}} \ge \sigma_{\mathrm{high}} \\ \frac{1}{2}(a_{\mathrm{RL}} + a_{\mathrm{LLM}}), & \text{otherwise} \end{cases} \tag{7}$$

where $\delta a = |a_{\mathrm{RL}} - a_{\mathrm{LLM}}|$ and the parameters $\sigma_{\mathrm{low}}, \sigma_{\mathrm{high}}, \varepsilon, \beta$ are defined in Table I. The fused command a_t is subsequently constrained by the safety layer described in Section II-B.

F. Evaluation Protocol and Metrics

All controllers were tested on the adult virtual subjects of the UVA/Padova v.2008 cohort introduced in Section II-A. Three control configurations were evaluated: (i) RL-based control (PPO with safety layer), (ii) Hybrid RL+LLM (fusion rule in Eq. 7), (iii) PID control as a classical baseline.

A common evaluation set of 20 stochastic 24-hour scenarios was generated using the meal variability protocol described in Section II-A, together with CGM noise. This ensures that performance differences can be attributed to the control policy rather than to exogenous disturbances.

Controllers produced one action every 3 minutes. Performance was quantified using standard *in silico* metrics including time-in-range (TIR, 70–180 mg/dL), time-below-range (TBR), time-above-range (TAR), minimum glucose, mean glucose, and control action rate.

III. Results and Discussion

Results are reported using both quantitative and qualitative metrics to assess safety, adaptability, and overall control performance.

A. Quantitative Performance and Comparative Analysis

Table II summarizes the closed-loop outcomes obtained with the *RL-based*, *Hybrid RL+LLM*, and *PID* controllers across twenty randomized daily scenarios for ten virtual adults. The PID controller was tuned using the default *SimGlucose* release gains ($P = 0.001$, $I = 0.00001$, $D = 0.001$). It provided stable but comparatively sluggish responses, yielding an average TIR of $59.2 \pm 14.0\,\%$. Delayed recovery and overshoot were particularly evident in adults #004 and #006, where TIR dropped below $50\,\%$ despite moderate total insulin use ($\approx$ 17–33 U).

The RL-based controller achieved improved glycemic regulation, with a mean TIR of $71.8 \pm 12.5\,\%$ and reduced exposure to hypoglycemia (< 70 mg/dL fraction $5.3 \pm 5.4\,\%$). However, it exhibited noticeable inter-subject variability, particularly in adults with faster or less predictable glucose dynamics (e.g., #004, #006, #009), where transient oscillations and delayed recovery from hyperglycemia were observed.

The Hybrid RL+LLM controller achieved the most balanced overall performance, with a mean TIR of $86.0 \pm 7.3\,\%$.

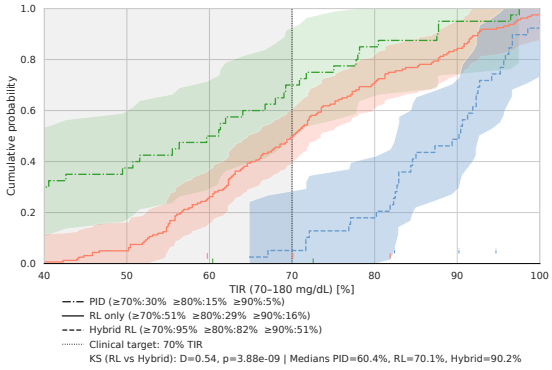

Fig. 3. Empirical cumulative distribution of the 24 h Time-in-Range (TIR, 70–180 mg/dL) across virtual adult scenarios under three controllers: PID, RL-based, and Hybrid RL+LLM. Shaded regions denote 95% Dvoretzky–Kiefer–Wolfowitz confidence bands of the empirical CDF. The vertical dotted line marks the clinical target of 70% TIR. The Hybrid RL+LLM controller shifts the distribution toward higher TIR values compared with the RL and PID baselines.

It reduced both hypo- ($2.4 \pm 3.0\,\%$) and hyperglycemic ($10.6 \pm 6.0\,\%$) exposure relative to the PID and RL baselines. A modest increase in total insulin delivery (56.9 ± 10.0 U) was also observed, reflecting a more proactive correction pattern in hyperglycemic conditions.

This point deserves careful interpretation. Part of the observed TIR improvement may be associated with the slightly higher total insulin delivery, rather than being attributable solely to the hybridization mechanism. At the same time, the increase remains limited and occurs under the same safety constraints and actuation limits applied to all controllers. The observed improvement should therefore be interpreted as the combined effect of the action-arbitration mechanism, improved correction timing, and moderately more proactive insulin delivery, rather than as evidence of improved performance independent of insulin usage.

Figure 3 further illustrates the distribution of 24 h TIR values across all runs. The Hybrid RL+LLM controller shifts the distribution toward higher TIR values compared with both RL-only and PID baselines. Specifically, 98% of Hybrid RL+LLM runs attained clinically desirable TIR levels ($\geq$70%), compared with 52% for RL and 30% for PID, with corresponding median TIRs of 90.2%, 70.1%, and 60.4%. The narrower confidence band and the significant KS difference ($D = 0.54$, $p = 3.9 \times 10^{-9}$) indicate improved inter-subject consistency and distributional superiority.

Representative glucose trajectories in Fig. 4 corroborate these findings, showing smoother transients and more stable responses under the hybrid policy. Paired Wilcoxon signed-rank tests [31], combined with Benjamini–Hochberg correction [32], showed significant performance gains of the hybrid approach relative to the baseline RL controller ($p_{\mathrm{BH}} = 0.041$), supporting the consistency of the observed trends.

B. Safety and Performance Trade-offs

Balancing responsiveness and safety remains a key challenge in automated insulin delivery. In the proposed frame-

TABLE II

COMPARISON OF CONTROL CONFIGURATIONS (*RL-based*, *Hybrid RL+LLM*, AND *PID*) OVER 20 SIMULATED SCENARIOS PER VIRTUAL ADULT. BEST VALUES FOR GLYCAEMIC REGULATION AND SAFETY ARE SHOWN IN BOLD. MEAN GLUCOSE AND TOTAL INSULIN ARE REPORTED AS DESCRIPTIVE INDICATORS TO SUPPORT INTERPRETATION OF CONTROLLER BEHAVIOUR.

Patient	TIR [%] (70–180 mg/dL) ↑			TBR [%] (<70 mg/dL) ↓			TAR [%] (>180 mg/dL) ↓			Mean Glucose [mg/dL]			Total Insulin [U]		
	RL	Hybrid	PID	RL	Hybrid	PID	RL	Hybrid	PID	RL	Hybrid	PID	RL	Hybrid	PID
adult001	**81.2 ± 11.0**	80.7 ± 8.4	67.4 ± 10.9	5.3 ± 4.9	**0.0 ± 0.0**	16.2 ± 12.6	**13.4 ± 10.6**	19.3 ± 8.4	16.4 ± 5.3	135.1 ± 8.9	154.6 ± 10.1	**130.6 ± 13.8**	51.1 ± 3.6	51.7 ± 9.2	15.9 ± 4.0
adult002	87.1 ± 7.4	**89.8 ± 5.8**	79.7 ± 8.4	0.5 ± 1.1	**0.0 ± 0.0**	6.4 ± 5.4	12.5 ± 7.4	**10.2 ± 5.8**	13.9 ± 6.6	135.8 ± 6.0	137.9 ± 8.3	**135.4 ± 4.7**	53.1 ± 2.4	65.1 ± 9.2	17.0 ± 1.5
adult003	78.0 ± 11.9	**89.2 ± 6.1**	44.0 ± 7.7	1.9 ± 2.4	**0.0 ± 0.0**	34.0 ± 10.0	20.1 ± 11.2	**10.9 ± 6.1**	22.1 ± 4.6	145.2 ± 8.9	134.5 ± 9.3	**127.7 ± 14.6**	53.8 ± 2.5	58.2 ± 9.3	22.7 ± 3.1
adult004	61.4 ± 9.1	**94.9 ± 3.6**	49.3 ± 9.3	20.7 ± 5.0	**0.0 ± 0.0**	28.4 ± 6.7	17.9 ± 6.7	**5.1 ± 3.6**	22.4 ± 4.1	127.5 ± 8.3	**122.5 ± 8.2**	132.6 ± 6.7	38.3 ± 1.8	71.4 ± 9.5	16.9 ± 2.8
adult005	72.7 ± 10.6	**74.2 ± 7.4**	70.5 ± 12.1	**0.3 ± 0.7**	11.3 ± 6.5	12.9 ± 7.1	27.0 ± 10.5	**14.6 ± 7.1**	16.6 ± 8.0	152.0 ± 8.8	**129.9 ± 9.2**	136.8 ± 6.9	56.9 ± 3.3	56.2 ± 9.0	20.8 ± 3.2
adult006	55.9 ± 10.7	**90.8 ± 5.5**	33.5 ± 7.0	1.7 ± 2.3	**0.0 ± 0.0**	39.6 ± 4.7	42.4 ± 10.1	**9.2 ± 5.5**	27.0 ± 4.4	172.8 ± 16.9	**133.6 ± 7.7**	134.3 ± 4.4	55.5 ± 2.9	70.8 ± 9.7	33.5 ± 3.4
adult007	65.3 ± 11.5	**84.5 ± 6.9**	26.4 ± 4.3	14.0 ± 5.0	**7.5 ± 6.3**	51.3 ± 5.7	20.8 ± 8.4	**8.0 ± 4.8**	22.3 ± 3.4	134.6 ± 10.9	122.4 ± 7.5	**121.5 ± 8.1**	42.7 ± 1.9	38.5 ± 9.0	25.7 ± 2.1
adult008	**94.4 ± 6.0**	77.0 ± 6.7	62.0 ± 9.9	**3.4 ± 4.2**	4.0 ± 5.8	19.7 ± 8.5	**2.3 ± 3.3**	19.1 ± 7.2	18.3 ± 6.9	**118.1 ± 4.2**	138.8 ± 10.6	131.7 ± 12.0	44.8 ± 2.2	42.8 ± 9.3	15.9 ± 2.8
adult009	57.8 ± 12.7	**94.9 ± 3.6**	80.8 ± 12.7	4.4 ± 4.6	**0.0 ± 0.0**	8.9 ± 9.9	37.8 ± 12.1	**5.1 ± 3.6**	10.2 ± 5.3	161.6 ± 12.9	**123.1 ± 7.8**	136.2 ± 9.6	55.2 ± 4.6	63.4 ± 9.6	20.3 ± 2.6
adult010	63.9 ± 8.9	**84.0 ± 6.9**	76.8 ± 11.2	**0.8 ± 1.4**	1.0 ± 2.5	9.5 ± 8.0	35.3 ± 8.6	15.0 ± 7.9	**13.8 ± 6.2**	159.9 ± 9.8	**136.0 ± 10.7**	137.2 ± 8.2	55.0 ± 3.2	51.1 ± 9.7	20.9 ± 2.7
Mean ± SD	71.8 ± 12.5	**86.0 ± 7.3**	59.2 ± 14.0	5.3 ± 5.4	**2.4 ± 3.0**	22.7 ± 10.4	25.5 ± 10.1	**10.6 ± 6.0**	18.3 ± 6.2	144.3 ± 13.6	133.3 ± 10.0	**132.1 ± 9.3**	49.6 ± 6.1	56.9 ± 10.0	20.7 ± 3.6

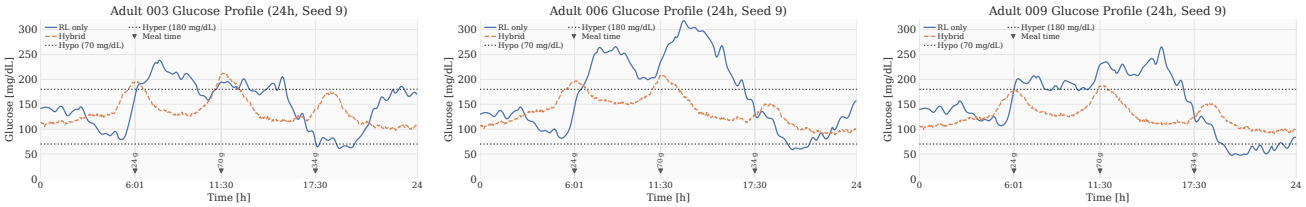

Fig. 4. Representative 24 h glucose trajectories under RL-only and Hybrid RL+LLM control for three adult virtual subjects: (a) 003, (b) 006, and (c) 009 (common meal schedule, seed 9). The Hybrid RL+LLM controller attenuates postprandial excursions and maintains glucose within the target range for longer durations in these representative examples.

work, the language-based module acts as a structured advisory layer that complements the RL controller through guideline-inspired logic and uncertainty-aware modulation.

The Hybrid RL+LLM architecture applies the fusion rule in (7), adjusting the influence of the advisory layer according to the RL policy dispersion. When policy confidence is high, actions are primarily determined by the RL controller; when dispersion increases, typically after unannounced meals or in the presence of sensor noise, the advisory layer provides corrective adjustments. This arbitration produces smoother glucose trajectories and fewer hypoglycemic events (Fig. 4), with a modest delay in hyperglycemia recovery. Such behaviour reflects a deliberately *risk-averse* control strategy consistent with safety-oriented formulations [33].

Quasi-counterfactual Safety Analysis: To evaluate the causal validity of the rule-based safety interventions, a quasi-counterfactual analysis inspired by [34] was conducted under matched stochastic conditions. At each decision step the variables $\{g_t, \tilde{a}_t, a_{\mathrm{LLM}}, a_t, \sigma_{\mathrm{RL}}, r_t^k\}$ were logged, where r_t^k denotes the activation of a safety rule (e.g., hypoglycemia guard, trend attenuation, recovery cap).

The consistency between rule activation and control adjustment was quantified through the *fidelity index*

$$\mathrm{Fid}_k = \Pr\left[\mathrm{sign}(a_t - \tilde{a}_t) = d_k \,\middle|\, r_t^k = 1\right], \tag{8}$$

where $d_k \in \{-1, 0, +1\}$ represents the expected control direction. For example, activation of the hypoglycemia guard implies $d_1 = -1$, corresponding to reduced insulin delivery relative to the baseline RL action. Values of Fid_k close to one therefore indicate consistent execution of the intended safety intervention.

In the present experiments, the hypoglycemia guard achieved $\mathrm{Fid}_k \approx 1.0$, confirming accurate rule execution.

TABLE III

QUASI-COUNTERFACTUAL SAFETY VALIDATION RESULTS.

Safety rule	Fid. index	CVR	ΔTBR [min]	$\Delta g_{\min}$ [mg/dL]
r_t^1 – Hypoglycemia guard	1.00	0.94±0.03	+18.2±5.6	−27.4±8.1
r_t^2 – Trend attenuation	0.97	0.89±0.05	+11.7±4.3	−15.9±6.2
r_t^3 – Recovery cap	0.95	0.91±0.04	+9.4±3.8	−12.6±5.1

The *Consistency–Validity Ratio (CVR)* was defined as

$$\mathrm{CVR} = \Pr\left[\text{valid intervention} \mid r_t^k = 1\right], \tag{9}$$

which measures the proportion of rule activations that produced beneficial safety outcomes relative to the baseline controller. Safety improvements were evaluated through changes in time-below-range (TBR) and minimum glucose $g_{\min}$, where positive $\Delta\mathrm{TBR}$ and negative $\Delta g_{\min}$ correspond to safer outcomes. Across 20 stochastic scenarios, all safety rules achieved high fidelity ($\mathrm{Fid}_k > 0.95$) and CVR above 0.9, indicating that the hybrid supervisor consistently mitigated unsafe tendencies of the baseline policy. These results support the causal coherence of the advisory layer within the considered simulation setting.

C. Limitations and Future Work

Although the proposed hybrid framework improves closed-loop stability and safety, several limitations remain. First, the study relies entirely on the *in silico* UVA/Padova adult cohort, which cannot fully reproduce behavioral variability, stress responses, or sensor artifacts [6]. Second, the LLM component was fine-tuned on synthetic supervision and has not yet been validated with clinical data. Third, the

advisor operates on a reduced contextual input and does not explicitly process the full temporal glucose–insulin trajectory available to the RL module.

Future work will extend the framework to adolescent and pediatric populations, incorporate online personalization of the advisory policy, and investigate how textual explanations affect clinician trust and intervention quality [35]. Formal verification of the safety guard through reachability or control-barrier-function analysis will also be considered to complement the empirical evidence with theoretical guarantees.

IV. Conclusion

A hybrid control framework combining reinforcement learning with a language-based advisory layer was developed for automated insulin delivery in type 1 diabetes. By integrating quantitative optimization with structured safety-oriented guidance, the controller exhibited adaptive and risk-aware behavior in simulation. *In silico* evaluation on the UVA/Padova cohort showed that the Hybrid RL+LLM configuration improved time-in-range and reduced hypoglycemic exposure relative to the considered baselines in the evaluated simulation scenarios, while maintaining stable actuation.

References

[1] B. P. Kovatchev, "The artificial pancreas in 2019: Systems, algorithms, and clinical outcomes," *Diabetes Care*, vol. 42, no. 5, pp. 872–879, 2019.

[2] B. W. Bequette, "Artificial pancreas control algorithms: A review of technology in development," *Journal of Diabetes Science and Technology*, vol. 7, no. 5, pp. 1255–1263, 2013.

[3] 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. Pinsker, R. P. Wadwa *et al.*, "Automated insulin delivery in adults with type 1 diabetes," *New England Journal of Medicine*, vol. 381, no. 18, pp. 1707–1717, 2019.

[4] K. H. Lee, C. D. Jun, and M. Kim, "Personalized reinforcement learning for blood glucose control in diabetes therapy," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 6, pp. 1899–1910, 2022.

[5] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 476–486, 2022.

[6] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, "In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes," *Journal of Diabetes Science and Technology*, vol. 3, no. 1, pp. 44–55, 2009.

[7] J. Xie, "Simglucose: An open-source simulation platform for blood glucose control research," in *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 16–19.

[8] G. Lops, V. A. Racanelli, G. Manfredi, L. De Cicco, and S. Mascolo, "A deep deterministic policy gradient control algorithm for automatic insulin delivery," *IFAC-PapersOnLine*, vol. 59, no. 2, pp. 13–18, 2025, 1st IFAC Workshop on Engineering Diabetes Technologies EDT 2025.

[9] G. Lops, G. Manfredi, V. A. Racanelli, L. De Cicco, and 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.

[10] F. Baldisseri, D. Menegatti, and A. Wrona, "Deep deterministic policy gradient control of type 1 diabetes," in *European Control Conference (ECC)*, 2024, pp. 868–873.

[11] F. Baldisseri *et al.*, "A quantitative comparison of deep reinforcement learning algorithms for type 1 diabetes control," in *2025 11th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2025, pp. 1073–1078.

[12] F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 908–918.

[13] J. Garcia and F. Fernandez, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[14] Y. Zhao, H. Liu, K. Yu *et al.*, "Safe and interpretable reinforcement learning for medical treatment policies," *Artificial Intelligence in Medicine*, vol. 118, p. 102147, 2021.

[15] K. Singhal, S. Azizi, T. M. Tu, and et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[16] Z. Liang, Y. Zhang, P. Sun *et al.*, "Llm-control: Embedding language models for human-interpretable decision-making in robotics," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1245–1252, 2024.

[17] Y. Chen, M. Xu, C. Zhao, and Q. Zhao, "Hcrmp: A llm-hinted contextual reinforcement learning framework for autonomous driving," *arXiv preprint arXiv:2505.15793*, 2025.

[18] Y. Li, Z. Wang, B. Xu, W. Zhang, and M. Tan, "Omni-thinker: Scaling multi-task reinforcement learning in llms with hybrid reward and task scheduling," *arXiv preprint arXiv:2507.14783*, 2025.

[19] S. Finlayson *et al.*, "Clinician readiness for ai-driven decision support: interpretability, reliability, and clinical trust," *NPJ Digital Medicine*, vol. 6, no. 1, p. 33, 2023.

[20] American Diabetes Association, "Standards of medical care in diabetes—2023," *Diabetes Care*, vol. 46, pp. S1–S154, 2023.

[21] V. N. Shah *et al.*, "Artificial pancreas systems: current progress and future perspectives," *Diabetologia*, vol. 64, no. 5, pp. 967–977, 2021.

[22] M. Lewicki *et al.*, "Reinforcement learning approaches for automated insulin delivery: a systematic review," *Artificial Intelligence in Medicine*, vol. 129, p. 102314, 2022.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[24] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2016.

[25] A. Raffin, A. Hill, A. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2022.

[29] Technology Innovation Institute (TII-UAE), "Falcon-rw-1b and falcon-llm: Open large language models from tii-uae," 2023.

[30] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data (KDH 2020)*, 2018, accessible via Ohio University SmartHealth Data Portal.

[31] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[32] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[33] B. Thananjeyan *et al.*, "Recovery rl: Safe reinforcement learning with learned recovery zones," in *IEEE Robotics and Automation Letters*, 2021.

[34] D. Gross and H. Spieker, "Enhancing rl safety with counterfactual llm reasoning," *arXiv preprint arXiv:2403.XXXX*.

[35] A. Holzinger *et al.*, "Causability and explainable ai in healthcare: A method for validation and benchmarking," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2022.