PHYSIOLOGICALLY AWARE FLOW MATCHING: JACOBIAN CONSTRAINTS FOR DOSE-CONSISTENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate forecasts under observed treatment plans do not guarantee physiologically consistent responses to altered ones: state-dependent dosing couples treatment to patient state, so models trained on such records learn associations that contradict physiology. We introduce *Physiologically Aware Flow Matching* (PAFM), a conditional flow-matching framework for probabilistic forecasting under planned treatments. PAFM regularizes the Jacobian of the velocity field with respect to pharmacokinetically smoothed treatment drivers, with physiological priors specifying the sign of each sensitivity and a band for its magnitude. The penalty is evaluated at sampled flow states and operates without a mechanistic outcome model or rollouts under alternative plans. We develop autoregressive (AR-PAFM) and joint-horizon (JH-PAFM) formulations, which constrain sensitivities at individual forecast steps and across forecast positions respectively, and evaluate both on two simulated closed-loop benchmarks. On glucose, every unregularized model integrates a glucose-raising response to added insulin; the penalty restores near-perfect directional and dose-ordering consistency in both formulations, and JH-PAFM achieves lower mean RMSE than all regularized baselines. On mean arterial pressure, both formulations reach full directional and dose-ordering consistency under propofol and norepinephrine infusions, with JH-PAFM lower than AR-PAFM in mean RMSE and calibration error.

1 INTRODUCTION

What should we demand of a physiological forecaster intended to inform treatment planning? Accurate trajectories and calibrated uncertainty are necessary but not sufficient: the evaluation must also interrogate how predictions respond to the treatment inputs on which they are conditioned. This holds equally across physiological applications, such as glucose forecasting conditioned on planned insulin delivery and carbohydrate intake (Andreassen et al., 1994), or mean arterial pressure (MAP) forecasting conditioned on propofol and norepinephrine infusions (Aubouin et al., 2023). Recent work documents this gap. Shakeri (2025) characterizes *driver-blindness*: treatment inputs add little predictive value beyond glucose history. (Namazi & Shakeri, 2026) show that models with strong forecasting scores can misrepresent glucose dynamics under insulin perturbation. Accuracy and calibration must therefore be assessed alongside physiological consistency: whether forecasts shift in the expected direction, exhibit sensitivity of plausible magnitude, and preserve the anticipated ordering across dose levels.

Learning physiological input sensitivities from observational data is challenging because dosing decisions depend on the patient's evolving condition (Bica et al., 2020). Insulin delivery often accompanies meals, allowing a model to associate insulin with rising glucose (Zou et al., 2024). During anesthesia, surgical stimulation can raise MAP while prompting increased propofol delivery to maintain hypnotic depth (Hans et al., 2010). Higher propofol infusion rates may therefore coincide with elevated MAP, even though propofol itself tends to lower blood pressure (Saugel et al., 2022). A model may reproduce these dosing patterns accurately without learning physiologically appropriate sensitivity to changes in its inputs. These considerations motivate physiological priors that guide sensitivity directions and magnitudes within relevant response windows and operating conditions.

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

Previous work incorporates physiological structure through mechanistic models, hybrid neural dynamics, and intervention-ranking objectives (Zou et al., 2024; Zou & Tian, 2026). Mechanistic and hybrid models encode physiological dependencies through system-specific equations or compartmental structure. Ranking objectives encourage appropriate dose ordering by comparing forecasts under multiple treatment plans and therefore require alternative-treatment rollouts during training. *What is missing is a way to constrain treatment sensitivity directly, without a mechanistic model of the outcome or rollouts under alternative treatment plans.*

We introduce *Physiologically Aware Flow Matching* (PAFM), a conditional flow-matching framework (Lipman et al., 2022) for probabilistic forecasting conditioned on patient history and pharmacokinetically smoothed planned treatment inputs. Its central component regularizes the Jacobian of the velocity field with respect to these smoothed drivers. Sign penalties encourage physiologically appropriate sensitivity directions, while magnitude penalties discourage negligible or excessive local sensitivity. The regularizer is evaluated at sampled flow states and requires neither numerical flow integration nor rollouts under alternative treatment plans. It thus encodes physiological prior knowledge through local sensitivity constraints, without requiring an explicit mechanistic model of the outcome dynamics.

The temporal structure of the velocity field determines which driver sensitivities can be regularized directly. We develop two formulations sharing the history encoder, driver smoother, response transformation, and base distribution. **AR-PAFM** generates trajectories autoregressively, regularizing each step's velocity with respect to the current smoothed drivers while a recurrent state carries information across steps. **JH-PAFM** uses a joint velocity field over the horizon, making each coordinate's dependence on current and earlier drivers explicit and supporting pairwise sign penalties and magnitude penalties on the aggregate sensitivity at each position.

We evaluate point forecast accuracy (RMSE), uncertainty calibration (empirical calibration error) (Kuleshov et al., 2018; Chung et al., 2021), and physiological consistency through response direction, dose ordering, and violation magnitude. The primary benchmark uses closed-loop glucose trajectories generated with the UVA/Padova simulator (Man et al., 2014) and the Control-IQ controller (Brown et al., 2019). A complementary study uses closed-loop trajectories from the Python Anesthesia Simulator (Aubouin et al., 2023) to compare AR-PAFM and JH-PAFM on mean arterial pressure (MAP) forecasting under propofol and norepinephrine infusions. We make three contributions:

- A Jacobian regularizer that constrains the sign and magnitude of treatment sensitivities in the flow-matching velocity field, discouraging implausible responses learned from state-dependent dosing without a mechanistic outcome model or alternative-treatment rollouts.
- Two formulations, AR-PAFM and JH-PAFM, enforce constraints at individual forecast steps and across forecast positions, respectively. The temporal parameterization determines which sensitivities the penalty reaches.
- Empirical evaluation across simulated glucose and MAP benchmarks: both formulations achieve near-perfect directional and dose-ordering consistency, while JH-PAFM attains lower mean RMSE than all regularized baselines on glucose and outperforms AR-PAFM in RMSE and calibration error on both benchmarks.

2 Related Work

Mechanistic and hybrid glucose forecasting. Hybrid glucose models combine mechanistic dynamics with learned components to incorporate physiological structure into prediction. H2NCM supplements predictive training with supervision based on known rankings of treatment responses, evaluated through forecasts under modified input sequences (Zou et al., 2024). Subsequent work investigates automatic selection of mechanistic states and sparsification of interactions in hybrid neural ODEs (Zou & Tian, 2026). These approaches demonstrate the necessity of embedding physiological domain priors beyond observational trajectory fitting, establishing both structural constraints and response-based supervision as viable inductive biases.

Derivative regularization and domain priors. Prior knowledge about input–output relationships can also be encoded through derivatives. Monteiro et al. (2022) study gradient penalties for mono-

tonicity and show that the sampling strategy influences where this property is encouraged. Kancheti et al. (2021) regularize model-implied effects using domain priors, including monotonicity and zero-effect requirements. Jacobian and kinetic regularization have also been used to encourage simpler neural ODE dynamics and reduce numerical integration cost (Finlay et al., 2020). PAFM extends derivative regularization to the generative velocity field, directly constraining velocity sensitivity with respect to continuous pharmacokinetic drivers.

Flow matching for probabilistic forecasting. Flow matching learns generative velocity fields through regression on conditional transport targets (Lipman et al., 2022). TSFlow adds Gaussian process priors to represent temporal structure in probabilistic forecasting (Kollovieh et al., 2025), while autoregressive flow matching models future observations through a sequence of conditional flows (ElGazzar & van Gerven, 2025). PAFM shows that physiological sensitivity regularization works within both autoregressive and joint-horizon formulations, and that temporal parameterization determines the scope and consistency of the constraints that can be imposed.

3 METHOD

We model the conditional distribution of a future response trajectory $y_{1:H}$, with $y_t \in (y_{\min}, y_{\max})$, given an observed history $X_{1:L} \in \mathbb{R}^{L\times d_{\text{in}}}$ and planned treatment drivers $U_{1:H} \in \mathbb{R}^{H\times d_u}$. The history contains past responses and driver channels, and $\hat{y}_{1:H}$ denotes a generated forecast. We use $\mathcal{C}^+$ and $\mathcal{C}^-$ for driver channels expected to raise and lower the response, respectively.

AR-PAFM and JH-PAFM share the components introduced below but differ in their velocity parameterizations and the temporal scope of their Jacobian response constraints (Sections 3.2 and 3.3). We omit batch indices except in the fidelity loss (Eq. 15), where B denotes the batch size. The operator $\text{sg}(\cdot)$ denotes stop-gradient, and LN denotes LayerNorm (Ba et al., 2016) without learnable affine parameters.

3.1 SHARED COMPONENTS

History encoder. A residual Mamba encoder (Gu & Dao, 2023) maps the history $X_{1:L}$ to a summary state through RMSNorm-preconditioned blocks (Zhang & Sennrich, 2019) and returns the initial encoded state $h_0 \in \mathbb{R}^{d_m}$, which conditions both methods. Full details are given in Appendix A.1.

Pharmacokinetic smoothing of future drivers (PK smoother). Drivers arrive as sparse impulses, while their physiological effects are delayed and distributed over time. We therefore transform the planned driver sequence $U_{1:H}$ into an effective smoothed sequence $D_{1:H} = (d_1, \ldots, d_H)$, where $d_t \in \mathbb{R}^{d_u}$. The PK smoother maintains learned per-channel exponential decay states whose combination produces a delayed-peak response that captures the temporal dynamics of driver effects. The resulting physiologically informed features d_t, rather than the raw impulses U_t, are provided to the velocity network in both methods. Full details are given in Appendix A.2.

Bounded latent response transformation. Rather than modeling the response in physical units, we map it to an unbounded latent variable z and decode through a bounded transform. For a target $y \in (y_{\min}, y_{\max})$,

$$z = \text{logit}\Big(\tfrac{y-y_{\min}}{y_{\max}-y_{\min}}\Big), \qquad y = y_{\min} + (y_{\max} - y_{\min})\,\sigma(z), \tag{1}$$

where σ is the logistic function. All flow dynamics operate in z, while decoded predictions remain within $(y_{\min}, y_{\max})$. The local slope of the decoder, used later to express driver sensitivities in physical units, is

$$\Lambda(z) = \frac{\partial y}{\partial z} = (y_{\max} - y_{\min})\,\sigma(z)\big(1-\sigma(z)\big). \tag{2}$$

Conditional flow matching (CFM). The conditional flow-matching objective trains the velocity network v_θ to transport samples from a base distribution to the target latent response distribution, conditioned on patient history and planned treatment inputs. Following the linear-interpolant formulation of flow matching (Lipman et al., 2022), we encode the ground-truth future response to latent

targets $z^\star_{1:H}$ through Eq. 1. Both methods draw their base from independent standard Gaussians, $z_{t,0} \sim \mathcal{N}(0,1)$ per forecast position, so all conditioning enters through the velocity field rather than the starting point. Given a base sample and independent flow times $\tau_t \sim \mathcal{U}(0,1)$, the linear interpolant and its target velocity are

$$z_{t,\tau} = (1-\tau_t)\, z_{t,0} + \tau_t\, z^\star_t, \qquad v^\star_t = z^\star_t - z_{t,0}, \tag{3}$$

where $z_{1:H,\tau} = (z_{1,\tau}, \ldots, z_{H,\tau})$ denotes the positions at their respective flow times $\tau_{1:H}$.

Both methods parameterize a single velocity field over the horizon,

$$v_\theta : \mathbb{R}^{H\times d_m} \times \mathbb{R}^{H\times d_u} \times \mathbb{R}^H \times \mathbb{R}^H \to \mathbb{R}^H, \qquad \big[v_\theta\big(\mathcal{H},\, D_{1:H},\, z_{1:H,\tau},\, \tau_{1:H}\big)\big]_t = v_{t,\tau}, \tag{4}$$

where $\mathcal{H} \in \mathbb{R}^{H\times d_m}$ collects the history conditioning for each forecast position: its t-th row is the recurrent state h_{t-1} in AR-PAFM and the encoded history h_0, shared across positions, in JH-PAFM, and $v_{t,\tau}$ is the scalar readout of each method (Eqs. 13 and 18). The two methods differ in which arguments each coordinate depends on: in AR-PAFM, $[v_\theta]_t$ is a function of $(h_{t-1}, d_t, z_{t,\tau}, \tau_t)$ alone; in JH-PAFM it depends on the drivers d_s at all positions $s \le t$ and on the latent states at neighbouring positions.

The velocity network is trained by

$$\mathcal{L}_{\mathrm{CFM}} = \frac{1}{H}\sum_{t=1}^{H} \Big(\big[v_\theta\big(\mathcal{H},\, D_{1:H},\, z_{1:H,\tau},\, \tau_{1:H}\big)\big]_t - v^\star_t\Big)^2. \tag{5}$$

The architecture of v_θ is specified per method in Sections 3.2 and 3.3.

Sampling. At inference, both methods draw a base sample $z_{t,0} \sim \mathcal{N}(0,1)$ and integrate the learned velocity field with K_f forward-Euler steps. Unlike training, where the flow times τ_t are drawn independently per position t, inference advances all positions together on a shared grid $\tau^{(k)} = k/K_f$ with $\Delta\tau = 1/K_f$, where the superscript indexes the solver step rather than a horizon position:

$$z_{t,\tau^{(k+1)}} = z_{t,\tau^{(k)}} + \Delta\tau\, \big[v_\theta\big(\mathcal{H},\, D_{1:H},\, z_{1:H,\tau^{(k)}},\, \tau^{(k)}\mathbf{1}\big)\big]_t, \qquad k = 0, \ldots, K_f - 1, \tag{6}$$

where $\mathbf{1} \in \mathbb{R}^H$ is the all-ones vector. The terminal state $\hat{z}_t = z_{t,\tau^{(K_f)}}$ is then decoded to physical units through Eq. 1. For paired comparisons across candidate plans we reuse the same base sample, so that trajectory differences reflect the planned drivers rather than independent sampling noise.

Driver-Jacobian response constraint. The two methods share a single response principle: constrain how the velocity field of Eq. 4 reacts to the smoothed drivers $D_{1:H}$. We detach every argument of the velocity network,

$$\bar{d}_s = \mathrm{sg}(d_s), \qquad \bar{z}_{t,\tau} = \mathrm{sg}(z_{t,\tau}), \qquad \bar{\tau}_t = \mathrm{sg}(\tau_t), \qquad \bar{\mathcal{H}} = \mathrm{sg}(\mathcal{H}), \tag{7}$$

and differentiate with respect to the smoothed drivers alone. Detaching d_s stops the gradient through the PK smoother, so that $\bar{D}_{1:H}$ enters as a free variable rather than as a function of the planned drivers $U_{1:H}$; the smoother therefore receives no gradient from this term and is trained only through $\mathcal{L}_{\mathrm{CFM}}$ and, in AR-PAFM, the fidelity loss. Differentiating the velocity solely through this pathway defines the latent sensitivity of the velocity at step t to driver channel c at step s,

$$J_{t,s,c} = \frac{\partial\, \big[v_\theta\big(\bar{\mathcal{H}},\, \bar{D}_{1:H},\, \bar{z}_{1:H,\tau},\, \bar{\tau}_{1:H}\big)\big]_t}{\partial \bar{d}_{s,c}}, \qquad J \in \mathbb{R}^{H\times H\times d_u}. \tag{8}$$

In AR-PAFM, given the recurrent state, the velocity at t depends only on d_t, so $J_{t,s,c}$ is nonzero only at $s = t$; in JH-PAFM it is nonzero on $s \le t$. Because the velocity operates in latent space while the constraints are specified in physical units, we convert each sensitivity through the decoder slope of Eq. 2 into an effect tensor of the same shape,

$$E_{t,s,c} = \varsigma_c\, \Lambda\big(\bar{z}_{t,\tau}\big)\, J_{t,s,c}, \tag{9}$$

where $\varsigma_c > 0$ is a fixed per-channel scale that places different driver units on a common response scale. We impose signed bounds $[b^-_c, b^+_c]$ per channel, with $0 < b^-_c < b^+_c$ for $c \in \mathcal{C}^+$ and $b^-_c < b^+_c < 0$ for $c \in \mathcal{C}^-$. Since each admissible interval excludes zero, the bounds also penalize the degenerate driver-ignoring solution $E \approx 0$. The two methods constrain different subsets of E, as specified in Sections 3.2 and 3.3.

Training objective. Both methods share the conditional flow-matching loss in Eq. 5 and a method-specific Jacobian regularizer $\mathcal{L}_J$. AR-PAFM additionally uses a rollout fidelity term $\mathcal{L}_{\text{fid}}$ (Eq. 15) to supervise compounding errors across its autoregressive recurrence. In contrast, JH-PAFM trains completely solver-free; its joint temporal parametrization models the full horizon simultaneously, bypassing sequential rollout drift. The two objectives are

$$\mathcal{L}^{\text{JH}} = \mathcal{L}_{\text{CFM}} + \lambda_J \mathcal{L}_J^{\text{JH}}, \qquad \mathcal{L}^{\text{AR}} = \mathcal{L}_{\text{CFM}} + \lambda_J \mathcal{L}_J^{\text{AR}} + \lambda_{\text{fid}} \mathcal{L}_{\text{fid}}, \tag{10}$$

with method-specific regularizers $\mathcal{L}_J^{\text{AR}}$ and $\mathcal{L}_J^{\text{JH}}$ defined below.

3.2 AR-PAFM: RECURRENT PER-STEP VELOCITY

Rollout. Physical time advances from the encoded history h_0 through a recurrence on the latent response alone, written with $\tilde{z}_t = z_t^\star$ under teacher forcing and $\tilde{z}_t = \hat{z}_t$, the integrated latent forecast, when the rollout is free-running — at inference and in the fidelity loss of Eq. 15:

$$h_t = \text{GRUCell}(\tilde{z}_t,\, h_{t-1}). \tag{11}$$

Because the recurrence consumes only the latent response, planned drivers modulate forecasts exclusively through the generative velocity field, isolating state transitions from direct driver influence.

Conditioning and velocity network. At each step t, AR-PAFM concatenates the recurrent state h_{t-1} with the instantaneous smoothed driver to form the conditioning vector $h_{t-1}^{\text{cond}} = [h_{t-1}; d_t] \in \mathbb{R}^{d_{\text{cond}}}$, $d_{\text{cond}} = d_m + d_u$. The flow time is encoded with K Fourier frequencies (Tancik et al., 2020), $\psi(\tau_t) \in \mathbb{R}^{1+2K}$, and a scalar velocity is produced by AdaLN (Zhang et al., 2025) modulation:

$$r_t = \text{Linear}_{2+2K\to d_f}\left([z_{t,\tau}; \psi(\tau_t)]\right), \qquad [\gamma_t; \beta_t] = \text{Linear}_{d_{\text{cond}}\to 2d_f}\left(h_{t-1}^{\text{cond}}\right), \tag{12}$$

$$\tilde{r}_t = \text{LN}\left(\text{SiLU}(r_t)\right) \odot (1+\gamma_t) + \beta_t, \qquad v_{t,\tau} = \text{Linear}_{d_f\to 1}\left(\text{MLP}_{d_f\to d_f}(\tilde{r}_t)\right), \tag{13}$$

where $r_t, \tilde{r}_t, \gamma_t, \beta_t \in \mathbb{R}^{d_f}$ and $\text{MLP}_{d_f\to d_f}$ is a stack of linear layers with SiLU activations. Since LN has no affine parameters, the driver enters transport only through the AdaLN parameters (γ_t, β_t); changing d_t directly reshapes the local dynamics.

Inference. Eq. 6 decouples across coordinates for AR-PAFM, since $[v_\theta]_t$ depends on the history and drivers only through h_{t-1}^{cond}, so it can be restricted to coordinate t. It is nonetheless applied one physical step at a time, because $\mathcal{H}$ is built as the rollout proceeds: each step closes a loop in which h_{t-1}^{cond} conditions the velocity at t, integrating it yields $\hat{z}_t$, and $\hat{z}_t$ enters Eq. 11 as $\tilde{z}_t$ to produce h_t. Only the state recurses; the driver is supplied fresh at each position. The full rollout costs HK_f velocity evaluations.

Teacher-forced conditioning. During training we advance Eq. 11 with the ground-truth latent response, $h_t^{\text{TF}} = \text{GRUCell}(z_t^\star, h_{t-1}^{\text{TF}})$, which provides a stable conditioning trajectory and removes the need to integrate the velocity field to obtain $\hat{z}_t$ for the state update. Stacking these states gives the history conditioning of Eq. 4, $\mathcal{H} = (h_0^{\text{TF}}, \ldots, h_{H-1}^{\text{TF}}) \in \mathbb{R}^{H\times d_m}$, and pairing each row with its smoothed driver yields $h_{\text{all}}^{\text{TF,cond}} = [\mathcal{H};\, D_{1:H}] \in \mathbb{R}^{H\times d_{\text{cond}}}$, which conditions both the flow-matching loss in Eq. 5 and the Jacobian regularizer below.

Pointwise response constraint. Because the conditioning at step t isolates the instantaneous driver d_t, cross-temporal sensitivities vanish. The effect tensor E is strictly diagonal, reducing the constraint to its diagonal entries $E_t^c \equiv E_{t,t,c}$. We apply a two-sided hinge that keeps each local sensitivity inside its signed band:

$$\mathcal{L}_J^{\text{AR}} = \frac{1}{H}\sum_{t=1}^{H}\sum_{c}\Big[\text{ReLU}\big(b_c^- - E_t^c\big) + \text{ReLU}\big(E_t^c - b_c^+\big)\Big]. \tag{14}$$

A wrong-signed response is thus penalized exactly at the step where it occurs. Evaluating Eq. 14 requires a single reverse-mode derivative with respect to the smoothed drivers, executing with zero numerical ODE integration overhead.

Fidelity loss. The local CFM and Jacobian terms do not penalize error that accumulates across the discrete rollout in Eq. 11. We therefore add a lightweight fidelity term, evaluated on a random subset of the batch, $\mathcal{S} \subseteq \{1, \dots, B\}$ with $|\mathcal{S}| = \max(1, \lfloor \rho_{\text{fid}} B \rfloor)$ and $\rho_{\text{fid}} \ll 1$, comparing the free-running latent forecast $\hat{z}_{1:H}$, obtained by integrating the velocity field with K_f forward-Euler steps per position under the recurrence in Eq. 11, with the target $z^{\star}_{1:H}$:

$$\mathcal{L}_{\text{fid}} = \frac{1}{|\mathcal{S}|H} \sum_{i \in \mathcal{S}} \sum_{t=1}^{H} \left(\hat{z}_t^{(i)} - z_t^{\star(i)} \right)^2, \tag{15}$$

where i indexes samples in the batch. Subsampling keeps this the only solver-based term in training, so the optimization remains dominated by the local CFM and Jacobian objectives while still receiving end-to-end rollout supervision.

3.3 JH-PAFM: JOINT-HORIZON VELOCITY FIELD

Conditioning and velocity network. JH-PAFM parameterizes the entire horizon jointly. The local flow state at each position combines the latent value $z_{t,\tau}$, a flow-time embedding $\psi(\tau_t) \in \mathbb{R}^{1+2K}$, and a position embedding $p_t \in \mathbb{R}^{d_p}$ encoding the location within the horizon:

$$r_t = \text{Linear}_{2+2K+d_p \to d_f} \left([z_{t,\tau}; \psi(\tau_t); p_t] \right), \qquad r_t \in \mathbb{R}^{d_f}. \tag{16}$$

The conditioning vector gathers history, instantaneous driver, and position, $\phi_t = [\mathcal{H}_t;\, d_t;\, p_t] \in \mathbb{R}^{d_{\text{cond}}}$ with $d_{\text{cond}} = d_m + d_u + d_p$. Every row of $\mathcal{H}$ equals h_0 here, so the horizon is distinguished by p_t and by the drivers rather than by the history conditioning.

The conditioning may optionally be extended with a driver context $\zeta_t^D \in \mathbb{R}^{d_\zeta}$, giving $\phi_t = [\mathcal{H}_t;\, d_t;\, \zeta_t^D;\, p_t]$ and $d_{\text{cond}} = d_m + d_u + d_\zeta + d_p$. It summarizes the drivers up to position t by their cumulative sum through a small network (Appendix A.3). It gives the constraint of Eq. 8 a direct route from the drivers to the velocity, so that satisfying the constraint need not reshape the temporal mixing that fits the trajectory. Whether to include it is selected per dataset on validation (Appendix F.4). The conditioning vector produces position-specific AdaLN parameters that modulate the local flow representation:

$$[\gamma_t; \beta_t] = \text{Linear}_{d_{\text{cond}} \to 2d_f}(\phi_t), \qquad \tilde{r}_t = \text{LN}\left(\text{SiLU}(r_t) \right) \odot (1 + \gamma_t) + \beta_t. \tag{17}$$

The drivers therefore reach transport only through (γ_t, β_t). Stacking the modulated representations as $M^{(0)} = (\tilde{r}_1, \dots, \tilde{r}_H) \in \mathbb{R}^{H \times d_f}$, we mix across positions with L_v residual blocks and read out a scalar velocity:

$$\begin{aligned} M^{(\ell)} &= M^{(\ell-1)} + \text{LConv}_2^{(\ell)} \Big(\text{SiLU}\left(\text{LConv}_1^{(\ell)}(\text{LN}(M^{(\ell-1)})) \right) \Big), \\ v_{t,\tau} &= \text{Linear}_{d_f \to 1}\left(\text{SiLU}(M_t^{(L_v)}) \right), \end{aligned} \tag{18}$$

for $\ell = 1, \dots, L_v$, where $M^{(\ell)} \in \mathbb{R}^{H \times d_f}$, and LConv is a one-dimensional temporal convolution of kernel width k_v with $k_v - 1$ zeros padded on the left only, so the output at position t never depends on positions $s > t$. Each block applies two such convolutions, giving a reach of $2L_v(k_v - 1)$ positions; we choose (k_v, L_v) so that this covers the horizon. The velocity at t therefore depends on the drivers and latent states at positions $s \le t$, and directly on every preceding driver when the driver context is included.

Cross-horizon response constraint. Since JH-PAFM's velocity can depend on drivers across positions, we constrain the full effect tensor $E_{t,s,c}$ from Eqs. 8–9. A pointwise sign term penalizes any wrong-signed sensitivity on the region $s \le t$,

$$\mathcal{L}_{\text{sign}} = \frac{1}{N_\triangle} \sum_{t=1}^{H} \sum_{s=1}^{t} \Big[\sum_{c \in \mathcal{C}^-} \text{ReLU}(E_{t,s,c}) + \sum_{c \in \mathcal{C}^+} \text{ReLU}(-E_{t,s,c}) \Big], \qquad N_\triangle = \tfrac{H(H+1)}{2}. \tag{19}$$

Summing the effect of every preceding driver on position t for channel c then gives the accumulated response, which a two-sided hinge keeps inside the signed band:

$$E_{t,c}^{\text{agg}} = \sum_{s=1}^{t} E_{t,s,c}, \qquad \mathcal{L}_{\text{agg}} = \frac{1}{H} \sum_{t=1}^{H} \sum_{c} \Big[\text{ReLU}(b_c^- - E_{t,c}^{\text{agg}}) + \text{ReLU}(E_{t,c}^{\text{agg}} - b_c^+) \Big]. \tag{20}$$

The JH-PAFM regularizer is the sum $\mathcal{L}_J^{\mathrm{JH}} = \mathcal{L}_{\mathrm{sign}} + \mathcal{L}_{\mathrm{agg}}$. The band $[b_c^-, b_c^+]$ therefore constrains a different quantity in each formulation: in AR-PAFM the response at t to the dose given at t, and in JH-PAFM the accumulated response at t to every dose at $s \leq t$. The latter is closer to how physiological sensitivity ranges are stated, and the two formulations use correspondingly different values.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two simulated physiological forecasting datasets, each comprising 100 virtual adults with patient-level 80/10/10 train/validation/test splits. On the PAS anesthesia benchmark, we forecast mean arterial pressure (MAP) over a 5-minute horizon ($H = 30$) from a 100-minute history ($L = 600$), conditioned on planned propofol and norepinephrine infusions; a synthetic closed-loop policy introduces confounding on both treatment channels. On the UVA/Padova Control-IQ benchmark, we forecast glucose over a 2-hour horizon ($H = 24$) from a 24-hour history ($L = 288$), conditioned on planned insulin and carbohydrate inputs; closed-loop insulin delivery and meal-associated boluses induce state-dependent treatment associations. Full simulation, confounding, and window-generation details are provided in Appendix C.

Baselines and Training. We compare against neural and hybrid baselines adapted from H2NCM (Zou et al., 2024). Because its mechanistic baselines are specific to the UVA/Padova glucose–insulin setting, the full baseline comparison is performed on Control-IQ; PAS serves to test transferability to a second physiological setting, where we compare AR-PAFM and JH-PAFM. We use α for each family's regularization weight: the H2NCM ranking-loss weight in Eq. 29 and the additive λ_J in Eq. 10, so matched α is not a matched accuracy penalty. Full baseline definitions, hyperparameter selection, and training details are in Appendices B and E.

4.2 Physiological-Consistency Metrics

Each treatment is evaluated at several dose levels, one per forecast: every dose is added on its own at the first step of the horizon, on top of the observed plan. A time point is evaluable when it lies inside that treatment's response window and the trajectory still has room to move before the boundary of the admissible range; support coverage Cov is the fraction retained. *Directional consistency* (DC) is the fraction of evaluable time–dose pairs whose response is correctly signed, where the expected direction is upward for carbohydrate and norepinephrine and downward for insulin and propofol. *Ordering consistency* (OC) is the fraction of adjacent dose pairs, at those same points, for which the larger dose moves the response further in the expected direction. The response chain is the ordering that runs from the largest outcome-lowering dose, through the unperturbed forecast, to the largest outcome-raising dose. The *graded violations* VD, VO, and VC are the mean hinge shortfalls of the directional, ordering, and response-chain conditions respectively, in outcome units and per dose for VO. Each condition is tested against a minimum-effect threshold, so the scores assess direction and ordering rather than magnitude; response shape is summarized separately by the dose-normalized peak effect $\overline{\nu}^{\mathrm{peak}}$, its lead time $\overline{t}^{\mathrm{peak}}$, and the area under the effect curve $\overline{\mathrm{AUEC}}$. Metrics are macro-averaged over forecast origins and then patients. Full definitions are in Appendix D.

4.3 Results

We report Control-IQ at $\alpha = 0.1$, the setting used for hyperparameter selection across all model families, and PAS at $\alpha = 0.3$ (Section 4.3.2). Sweeps over α, calibration results, and response profiles are provided in Appendices F.1 and F.2; effect-bound and fidelity ablations in Appendix F.3. Values at these settings are reported in Tables 1 and 2; results at other coefficients are cited to the corresponding appendix table or figure.

4.3.1 Results on the Control-IQ benchmark

JH-PAFM combines strong predictive accuracy with physiologically consistent responses. JH-PAFM achieves the lowest mean RMSE at all three horizons in the primary regularized com-

Table 1: Control-IQ comparison of neural and hybrid baselines with PAFM. Entries are means ± SD over four training seeds. RMSE and VC are in mg/dL; DC and OC are percentages. **Bold** marks the best mean per metric within each model group at $\alpha = 0.1$, including ties. Unregularized TCN is shown for reference and excluded from highlighting.

Group	Model	α	RMSE ↓			DC (%) ↑		OC (%) ↑		VC ↓
			30 min	60 min	120 min	Insulin	Carb.	Insulin	Carb.	
Neural	BNODE	0.1	$8.957_{\pm 0.182}$	$13.664_{\pm 0.266}$	$16.437_{\pm 0.686}$	$96.210_{\pm 1.331}$	$98.676_{\pm 0.350}$	$79.802_{\pm 12.389}$	$97.437_{\pm 0.577}$	$1.424_{\pm 0.982}$
	LSTM	0.1	$8.585_{\pm 0.784}$	$13.771_{\pm 1.302}$	$17.067_{\pm 2.487}$	$87.305_{\pm 6.643}$	$99.458_{\pm 0.441}$	$60.769_{\pm 13.042}$	$98.917_{\pm 0.443}$	$5.834_{\pm 4.328}$
	S4D	0.1	$13.676_{\pm 0.439}$	$15.565_{\pm 0.420}$	$17.179_{\pm 0.473}$	$55.599_{\pm 1.530}$	$95.479_{\pm 2.028}$	$47.997_{\pm 3.023}$	$94.936_{\pm 0.344}$	$12.544_{\pm 2.774}$
	TCN	0.1	$9.479_{\pm 0.380}$	$11.789_{\pm 0.619}$	$15.467_{\pm 1.010}$	$94.020_{\pm 3.660}$	$98.836_{\pm 0.495}$	$80.937_{\pm 10.015}$	$97.461_{\pm 1.126}$	$1.526_{\pm 0.925}$
Hybrid	Hovorka	0.1	$9.374_{\pm 0.346}$	$15.341_{\pm 0.169}$	$23.271_{\pm 0.550}$	$66.091_{\pm 20.863}$	$95.385_{\pm 0.038}$	$33.143_{\pm 9.637}$	$95.343_{\pm 0.103}$	$0.278_{\pm 0.148}$
	LP-Reduced	0.1	$63.653_{\pm 104.717}$	$53.812_{\pm 71.937}$	$50.686_{\pm 47.861}$	$91.139_{\pm 17.721}$	$94.175_{\pm 0.809}$	$91.230_{\pm 17.540}$	$94.172_{\pm 0.814}$	$0.006_{\pm 0.012}$
PAFM	AR-PAFM	0.1	$10.043_{\pm 0.760}$	$14.538_{\pm 0.742}$	$22.329_{\pm 1.038}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{99.999}_{\pm 0.002}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{99.999}_{\pm 0.002}$	$\mathbf{0.000}_{\pm 0.000}$
	JH-PAFM	0.1	$\mathbf{6.124}_{\pm 0.640}$	$\mathbf{9.408}_{\pm 0.531}$	$\mathbf{15.069}_{\pm 0.768}$	$\mathbf{100.000}_{\pm 0.000}$	$99.997_{\pm 0.006}$	$\mathbf{100.000}_{\pm 0.000}$	$99.975_{\pm 0.019}$	$0.001_{\pm 0.001}$
Reference	TCN	0.0	$4.665_{\pm 0.223}$	$6.166_{\pm 0.220}$	$8.920_{\pm 0.186}$	$34.064_{\pm 1.073}$	$49.219_{\pm 8.250}$	$19.071_{\pm 2.621}$	$72.769_{\pm 5.422}$	$16.823_{\pm 3.632}$

Table 2: PAS results at $\alpha \in \{0, 0.3, 1\}$: without regularization, at the selected coefficient, and at the strongest tested coefficient. Coverage is 100% for both treatments. Entries are means ± SD over four training seeds. RMSE and VC are in mmHg; consistency scores are percentages. **Bold** marks the lower mean RMSE and ECE at $\alpha = 0.3$.

Model	α	RMSE (mmHg) ↓			ECE ↓	DC (%) ↑		OC (%) ↑		VC ↓
		1 min	2.5 min	5 min		propo	nore	propo	nore	
AR-PAFM	0	$2.375_{\pm 0.114}$	$3.166_{\pm 0.182}$	$3.906_{\pm 0.298}$	$0.094_{\pm 0.064}$	$100.000_{\pm 0.000}$	$83.667_{\pm 4.170}$	$100.000_{\pm 0.000}$	$83.657_{\pm 4.167}$	$0.020_{\pm 0.013}$
JH-PAFM	0	$2.185_{\pm 0.063}$	$2.837_{\pm 0.048}$	$3.434_{\pm 0.045}$	$0.116_{\pm 0.008}$	$18.051_{\pm 16.505}$	$98.717_{\pm 1.789}$	$16.455_{\pm 14.219}$	$99.097_{\pm 1.366}$	$0.047_{\pm 0.019}$
AR-PAFM	0.3	$4.112_{\pm 1.432}$	$4.232_{\pm 1.107}$	$4.719_{\pm 0.791}$	$0.132_{\pm 0.035}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
JH-PAFM	0.3	$\mathbf{2.420}_{\pm \mathbf{0.156}}$	$\mathbf{3.064}_{\pm \mathbf{0.114}}$	$\mathbf{3.723}_{\pm \mathbf{0.105}}$	$\mathbf{0.050}_{\pm \mathbf{0.026}}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
AR-PAFM	1	$3.929_{\pm 1.547}$	$4.152_{\pm 1.134}$	$4.656_{\pm 0.786}$	$0.120_{\pm 0.027}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
JH-PAFM	1	$3.327_{\pm 0.821}$	$3.623_{\pm 0.312}$	$4.202_{\pm 0.130}$	$0.062_{\pm 0.036}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$0.000_{\pm 0.000}$

parison at $\alpha = 0.1$, attaining 6.12/9.41/15.07 mg/dL at 30/60/120 minutes (Table 1). This reduces regularized TCN's 9.48/11.79/15.47 mg/dL by 35.4/20.2/2.6%, the Hovorka-structured hybrid's 9.37/15.34/23.27 mg/dL by 34.7/38.7/35.2%, and AR-PAFM's 10.04/14.54/22.33 mg/dL by 39.0/35.3/32.5%, with corresponding reductions of 55.2/39.6/12.3% (S4D), 28.7/31.7/11.7% (LSTM), and 31.6/31.1/8.3% (BNODE). The gains are largest at 30 and 60 minutes; at 120 minutes the advantage over TCN is 0.40 mg/dL against seed standard deviations of 0.77 and 1.01 mg/dL. JH-PAFM's mean RMSEs also fall below the lowest baseline mean at each horizon across the tested nonzero α values, 8.59/11.79/15.47 mg/dL (Table 7). Unregularized TCN is more accurate (4.67/6.17/8.92 mg/dL) but reaches only 34.1% and 19.1% insulin directional and ordering consistency, so JH-PAFM combines the lowest errors among regularized comparators with near-perfect response consistency.

The physiological prior supports correct response directions and dose ordering on the confounded channel. At $\alpha = 0.1$, both PAFM formulations report 100% insulin directional and ordering consistency, compared with 55.60–96.21% and 33.14–91.23%, respectively, among the regularized baselines (Table 1). The carbohydrate gap is smaller: both PAFM formulations exceed 99.97% on both metrics, compared with 99.458/98.917% for the best baseline, LSTM. This asymmetry is consistent with confounding from glucose-dependent insulin dosing and meal-associated boluses, which can associate greater insulin delivery with higher glucose despite insulin's glucose-lowering physiological effect. No baseline exceeds 99.9% on all four consistency metrics at any tested coefficient (Table 6). LP-Reduced reaches 100% on both insulin metrics at $\alpha = 0.3$, but only 93.20/93.22% on carbohydrate (Table 6), with a 120-minute RMSE of 25.42 mg/dL (Table 7). The sweep supports the contribution of the driver-Jacobian prior: insulin directional/ordering consistency rises from 0/0% (AR-PAFM) and 30.99/24.25% (JH-PAFM) without the penalty to 100/100% at $\alpha = 0.1$ (Table 6). Negative sign-oriented insulin AUECs without the penalty further indicate wrong-signed integrated responses on average (Table 9).

PAFM generates forecasts with negligible mean response violations. At $\alpha = 0.1$, the reported mean violations are zero across all five graded measures for AR-PAFM and across the directional

and ordering measures on both channels for JH-PAFM (Table 5). JH-PAFM's mean response-chain violation is 0.001 mg/dL (Table 1), compared with unregularized values of 18.82 and 2.87 mg/dL for AR-PAFM and JH-PAFM, respectively (Table 5). Both formulations achieve lower mean chain violations than every baseline at this setting: the hybrid baselines report 0.006 mg/dL (LP-Reduced) and 0.278 mg/dL (Hovorka), while neural baseline values range from 1.424 to 12.544 mg/dL. JH-PAFM's reported mean is thus more than two orders of magnitude smaller than every baseline except LP-Reduced, whose small violation accompanies a 120-minute RMSE of 50.69 ± 47.86 mg/dL. These results show that local velocity-field regularization can yield negligible mean response violations without requiring alternative-treatment rollouts during training.

4.3.2 Results on the PAS anesthesia benchmark

The response constraint is effective in a second physiological setting. At $\alpha = 0.3$, both formulations attain 100% directional and dose-ordering consistency for propofol and norepinephrine across all four seeds (Table 2), with zero mean violations on all five graded measures (Figure 2a); this persists at all stronger tested coefficients (Table 2; Figures 1b and 2a). The models are retrained for 5-minute MAP forecasting using the same core formulations, with dataset-specific smoothing, physiological settings, and hyperparameters. Without the penalty, directional consistency for propofol/norepinephrine is $18.1/98.7\%$ for JH-PAFM and $100/83.7\%$ for AR-PAFM. JH-PAFM's mean sign-oriented propofol AUEC is wrong-signed and roughly an order of magnitude smaller in absolute value without the penalty (-0.155 versus 1.808 at $\alpha = 0.3$; Figure 2b). These channel-specific deficits are consistent with the imposed confounding (Appendix F.2).

JH-PAFM delivers superior accuracy and calibration over AR-PAFM. JH-PAFM consistently outperforms AR-PAFM in predictive accuracy across all tested coefficients and forecast horizons, exhibiting substantially reduced variance across training seeds (Figure 1a). At $\alpha = 0.3$, its RMSE at $1/2.5/5$ minutes is $2.42/3.06/3.72$ mmHg against $4.11/4.23/4.72$ mmHg for AR-PAFM, whose seed standard deviations are 7.5–9.7 times as large (Table 2). JH-PAFM also has lower mean ECE at every tested coefficient $\alpha \geq 0.3$, including 0.050 versus 0.132 at $\alpha = 0.3$. AR-PAFM has lower mean ECE only at $\alpha \in \{0, 0.1\}$ (Figure 1a), and at $\alpha = 0.1$ one of its four seeds has zero directional and ordering consistency on both channels (Appendix F.2; Figure 1b). For JH-PAFM, the selected coefficient $\alpha = 0.3$ raises 5-minute RMSE by 8.4% over the unregularized model ($3.434 \rightarrow 3.723$ mmHg). Full reported consistency is already reached at $\alpha = 0.1$, where the increase is only 3.8% ($3.434 \rightarrow 3.564$ mmHg; Figure 1a); we select $\alpha = 0.3$ as the smallest tested coefficient at which both formulations are consistent across all four seeds.

5 Limitations

Both benchmarks evaluate closed-loop simulation environments with parameterized mechanisms; assessing generalization across observational clinical recordings with irregular sampling and unobserved interventions represents the immediate next step. In addition, consistency is evaluated over prespecified doses and response windows with zero thresholds; selecting clinically appropriate thresholds and calibrating patient-specific sensitivity bounds both require prospective clinical investigation.

6 Conclusion

Physiologically Aware Flow Matching holds continuous generative forecasters to dose–response consistency without a mechanistic outcome model or alternative-treatment rollouts. Regularizing the velocity-field Jacobian with respect to pharmacokinetically smoothed drivers suppresses the inverted sensitivities induced by state-dependent dosing, at individual forecast steps in AR-PAFM and across forecast positions in JH-PAFM. On Control-IQ, JH-PAFM attains the lowest mean RMSE at all three horizons among the regularized comparators, with both formulations near-perfect on directional and dose-ordering consistency; on PAS, both reach full measured consistency under propofol and norepinephrine. Without the penalty, the models exhibit channel-specific wrong-signed responses, confirming that predictive objectives alone do not secure appropriate treatment sensitivity. The same mechanism carries to any setting where the direction of an intervention is known but its dosing is confounded.

REFERENCES

S Andreassen, JJ Benn, R Hovorka, et al. A probabilistic approach to glucose adjustment: description of metabolic study prediction and insulin dose model and pilot evaluation. *Computer Methods and Programs in Biomedicine*, 41:153–165, 1994.

Karl J Aström and Tore Hägglund. *Advanced PID control*. John Wiley & Sons, 2005.

Bob Aubouin, Thao Dang, Mirko Fiacchini, et al. Pas: A python anesthesia simulator for drug control. *Journal of Open Source Software*, 8(88):5480, 2023.

Michael S Avidan, Lini Zhang, Beth A Burnside, Kevin J Finkel, Adam C Searleman, Jacqueline A Selvidge, Leif Saager, Michelle S Turner, Srikar Rao, Michael Bottros, et al. Anesthesia awareness and the bispectral index. *New England journal of medicine*, 358(11):1097–1108, 2008.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

H Beloeil, J-X Mazoit, D Benhamou, and J Duranteau. Norepinephrine kinetics and dynamics in septic shock and trauma patients. *British journal of anaesthesia*, 95(6):782–788, 2005.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

Sue A Brown, Boris P Kovatchev, Dan Raghinaru, John W Lum, Bruce A Buckingham, Yogish C Kudva, Lori M Laffel, Carol J Levy, Jordan E Pinsker, R Paul Wadwa, et al. Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *New England Journal of Medicine*, 381(18):1707–1717, 2019.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

Chiara Dalla Man, Marc D Breton, and Claudio Cobelli. Physical activity into the meal glucose—insulin model of type 1 diabetes: In silico studies, 2009.

Douglas J Eleveld, Pieter Colin, Anthony R Absalom, and Michael MRF Struys. Pharmacokinetic–pharmacodynamic model for propofol for broad application in anaesthesia and sedation. *British journal of anaesthesia*, 120(5):942–959, 2018.

Ahmed ElGazzar and Marcel van Gerven. Probabilistic forecasting via autoregressive flow matching. *arXiv preprint arXiv:2503.10375*, 2025.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International conference on machine learning*, pp. 3154–3164. PMLR, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in neural information processing systems*, 35:35971–35983, 2022.

GA Hans, SM Lauwick, Abdourahmane Kaba, Vincent Bonhomme, MMRF Struys, PC Hans, ML Lamy, and JL Joris. Intravenous lidocaine infusion reduces bispectral index-guided requirements of propofol only during surgical stimulation. *British journal of anaesthesia*, 105(4):471–479, 2010.

Roman Hovorka, Valentina Canonico, Ludovic J Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R Pieber, Helga C Schaller, Lukas Schaupp, Thomas Vering, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, 25(4):905–920, 2004.

Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, and Amit Sharma. Matching learned causal effects of neural networks with domain priors. *arXiv preprint arXiv:2111.12490*, 2021.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 95435–95457, 2025.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pp. 47–54. Springer, 2016.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014.

João Monteiro, Mohamed Osama Ahmed, Hoseein Hajimirsadeghi, and Greg Mori. Monotonicity regularization: Improved penalties and novel applications to disentangled representation learning and robust classification. In *Uncertainty in Artificial Intelligence*, pp. 1381–1391. PMLR, 2022.

Alireza Namazi and Heman Shakeri. From prediction to practice: A task-aware evaluation framework for blood glucose forecasting. *arXiv preprint arXiv:2605.00645*, 2026.

Bernd Saugel, Elisa-Johanna Bebert, Luisa Briesenick, Phillip Hoppe, Gillis Greiwe, Dongsheng Yang, Chao Ma, Edward J Mascha, Daniel I Sessler, and Dorothea E Rogge. Mechanisms contributing to hypotension after anesthetic induction with sufentanil, propofol, and rocuronium: a prospective observational study. *Journal of Clinical Monitoring and Computing*, 36(2):341–347, 2022.

Heman Shakeri. The driver-blindness phenomenon: Why deep sequence models default to autocorrelation in blood glucose forecasting. *arXiv preprint arXiv:2511.20601*, 2025.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Fan Zhang, Yi Wei, Naye Ji, Jingmei Wu, Zhaohan Wang, Fuxing Gao, Liuqing Zhang, Zhenqing Ye, Leyao Yan, Lanxin Dai, et al. Dim-gestor: Co-speech gesture generation with adaptive layer normalization mamba-2. In *2025 International Conference on Computer Vision, Image Processing and Computational Photography (CVIP)(*, pp. 01–13. IEEE, 2025.

Bob Zou and Lu Tian. Automatic and structure-aware sparsification of hybrid neural odes with application to glucose prediction. In *International Conference on Learning Representations*, volume 2026, pp. 83832–83870, 2026.

Bob Junyi Zou, Matthew E Levine, Dessi P Zaharieva, Ramesh Johari, and Emily B Fox. Hybrid^2 neural ode causal modeling and an application to glycemic response. *arXiv preprint arXiv:2402.17233*, 2024.

A Architecture Details

A.1 Encoder Architecture

The encoder receives $X_{1:L} \in \mathbb{R}^{L \times d_{\text{in}}}$ and projects each time step to the model dimension d_m:

$$R_{1:L}^{(0)} = \text{Linear}_{d_{\text{in}} \to d_m}(X_{1:L}), \qquad R_{1:L}^{(0)} \in \mathbb{R}^{L \times d_m}. \tag{21}$$

The projected sequence is processed by L_e residual Mamba (Gu & Dao, 2023) blocks with RMSNorm(Zhang & Sennrich, 2019) pre-normalization:

$$\tilde{R}_{1:L}^{(\ell-1)} = \text{RMSNorm}\left(R_{1:L}^{(\ell-1)}\right), \tag{22}$$

$$R_{1:L}^{(\ell)} = R_{1:L}^{(\ell-1)} + \text{Mamba}_\ell\left(\tilde{R}_{1:L}^{(\ell-1)}\right), \qquad \ell = 1, \ldots, L_e, \tag{23}$$

$$\tilde{R}_{1:L}^{L_e} = \text{RMSNorm}\left(R_{1:L}^{(L_e)}\right), \qquad \tilde{R}_{1:L}^{L_e} \in \mathbb{R}^{L \times d_m}. \tag{24}$$

The initial hidden state conditioning the velocity network is the final output: $h_0 = \tilde{R}_L^{L_e} \in \mathbb{R}^{d_m}$.

A.2 Physiological Smoothing of Future Drivers

The smoother transforms the sparse planned driver impulses $U_{1:H}$ into an effective sequence $D_{1:H}$ whose impulse response is zero at the dose, peaks after a learned delay, and decays. We use two instantiations. The glucose variant (Appendix A.2.1) uses two decay states combined with fixed weights, so the response shape follows from the two learned rates alone. The MAP variant (Appendix A.2.2) adds a third state and learns the weights, giving the peak time and the tail decay separate degrees of freedom. Both are learned end to end through $\mathcal{L}_{\text{CFM}}$ and, in AR-PAFM, the fidelity loss.

A.2.1 Glucose PK-Smoother

Using per-channel fast and slow decay states,

$$g_t^{\text{fast}} = \tilde{s}^{\text{fast}} \odot g_{t-1}^{\text{fast}} + U_t, \qquad g_t^{\text{slow}} = \tilde{s}^{\text{slow}} \odot g_{t-1}^{\text{slow}} + U_t, \qquad d_t = \eta \odot \text{ReLU}\left(g_t^{\text{slow}} - g_t^{\text{fast}}\right), \tag{25}$$

where $\tilde{s}^{\text{fast}}, \tilde{s}^{\text{slow}} \in (0,1)^{d_u}$ are learned per-channel decay vectors satisfying $\tilde{s}^{\text{fast}} < \tilde{s}^{\text{slow}}$, and $\eta \in \mathbb{R}_{>0}^{d_u}$ is a learned gain vector. Because the slow state decays more gradually than the fast state, their difference produces a smooth delayed-peak response, exposing the temporally distributed effect of each driver to the velocity network. The smoother states are initialized as $g_0^{\text{fast}} = g_0^{\text{slow}} = \mathbf{0}$ at the beginning of each forecast window, while the positive gain vector η is initialized to $\mathbf{1}$ and learned jointly with the decay parameters.

A.2.2 MAP PK-Smoother

For each driver channel the smoother maintains P exponential accumulators of the dose history, combined through learned signed weights that sum to zero at initialization, so a dose has no instantaneous effect and the response rises from zero to a single positive peak before decaying. For $j = 1, \ldots, P$,

$$g_{t,j} = \tilde{s}_j \odot g_{t-1,j} + U_t, \quad \tilde{s}_j = \exp(-\Delta_{\text{bin}}/T_j), \quad g_{0,j} = \mathbf{0}, \qquad d_t = \eta \odot \text{ReLU}\left(\textstyle\sum_j w_j \odot g_{t,j}\right), \tag{26}$$

where the decay time constants $T_{1:P}$ (in seconds), the signed weights $w_{1:P}$, both in $\mathbb{R}^{d_u}$, and the gain $\eta \in \mathbb{R}_{>0}^{d_u}$ are learned, with η initialized to $\mathbf{1}$. All quantities in Eq. 26 lie in $\mathbb{R}^{d_u}$ and all operations are elementwise over channels, with $\Delta_{\text{bin}} = 10$ s. We use $P = 3$ and initialize the time constants to $(60, 411, 1200)$ s for propofol and $(45, 173, 600)$ s for norepinephrine, the middle values taken from the reference BIS effect-site equilibration time of 411 s (Eleveld et al., 2018) and from a reported two-minute norepinephrine half-life (Beloeil et al., 2005), each bracketed by a faster and a slower constant.Weights initialized at $(-1, 0, 1)$ make the response the difference between the slowest and fastest accumulators, decaying on the slowest initial timescale.

A.3 Prefix Driver Context

The driver context summarizes the effective drivers up to each forecast position by their cumulative sum, $q_t^D = \sum_{s \le t} d_s \in \mathbb{R}^{d_u}$, which is mapped to a d_ζ-dimensional representation by a two-layer network,

$$\zeta_t^D = \mathrm{LN}\Big(\mathrm{SiLU}\big(\mathrm{Linear}_{d_\zeta \to d_\zeta}\big(\mathrm{SiLU}(\mathrm{Linear}_{d_u \to d_\zeta}(q_t^D))\big)\big)\Big), \qquad \zeta_t^D \in \mathbb{R}^{d_\zeta}. \tag{27}$$

B Comparison Baselines

B.1 H2NCM Baselines

We compare against the baseline families considered in H2NCM (Zou et al., 2024): BNODE, LP-Reduced, LSTM, S4D (Gu et al., 2022), and TCN (Lea et al., 2016). BNODE is the black-box neural ODE baseline, which parameterizes the state dynamics entirely with a neural network. LP-Reduced is the reduced latent-parameter hybrid model from H2NCM, retaining the reduced UVA/Padova mechanistic dynamics (Man et al., 2014) while allowing selected simulator parameters to vary over time through learned latent dynamics.

We do not directly instantiate the MNODE variant from H2NCM. In that work, MNODE is constructed from the reduced UVA/Padova mechanistic dependency graph and includes exercise-response components (Dalla Man et al., 2009) with exercise-related covariates such as heart rate and step count. Our setting does not include exercise channels; the available inputs are glucose, insulin, and carbohydrates. We therefore retain the hybrid neural ODE design principle, but instantiate the mechanistic component with the Hovorka glucose–insulin model (Hovorka et al., 2004). As detailed in Appendix B.2, the Hovorka model is driven by insulin and carbohydrate inputs and does not require exercise modules, making it compatible with our forecasting setting while preserving a physiologically grounded glucose–insulin structure.

Glucose, insulin, and carbohydrate channels are normalized using training-set statistics, and prediction accuracy is optimized with a normalized glucose MSE loss ($\mathcal{L}_{\mathrm{pred}}$).

For the baselines, we use the driver-based intervention-ranking loss of (Zou et al., 2024). For each training example, one intervention category is sampled uniformly, and the ranking loss is computed from three forecasts generated by perturbing the observed future drivers. The intervention categories follow the original protocol: (i) carbohydrate-only, adding $\{0, 50, 100\}$ grams of carbohydrate; (ii) insulin-only, adding $\{0, 2.5, 5.0\}$ units of insulin; and (iii) mixed, using {no change, $+50$ grams carbohydrate, $+10$ units insulin}. Each branch is scored by its mean predicted glucose over the forecast horizon.

For branch b, define $\mu_b = \frac{1}{H}\sum_{t=1}^{H} \hat{y}_t^{(b)}$. Let $b^\star$ denote the branch that should be ranked highest under the selected intervention category. With default temperature $\varkappa = 10^4$, the ranking loss is

$$\mathcal{L}_{\mathrm{rank}} = -\log \frac{\exp(\varkappa\, \mu_{b^\star})}{\sum_{b=1}^{3} \exp(\varkappa\, \mu_b)}\,. \tag{28}$$

The full baseline objective is

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\mathrm{pred}} + \alpha \mathcal{L}_{\mathrm{rank}}\,, \qquad \alpha \in [0,1]\,, \tag{29}$$

where α controls the trade-off between observed-driver prediction accuracy and intervention-ranking supervision.

B.2 Hovorka-Structured Neural Dynamics

Following the mechanistic neural ODE (MNODE) framework (Zou et al., 2024), we model glucose evolution with a neural state-space model that preserves a mechanistic dependency graph while learning the state transition functions from data. We instantiate this graph using the compartmental structure of the Hovorka glucose–insulin model (Hovorka et al., 2004). The latent state is

$$\xi_t = [Q_1, Q_2, S_1, S_2, I, x_1, x_2, x_3, G_1, G_2]_t\,, \tag{30}$$

where Q_1, Q_2 are accessible and non-accessible glucose masses, S_1, S_2 are subcutaneous insulin absorption compartments, I is plasma insulin, and x_1, x_2, x_3 are remote insulin actions on glucose transport, disposal, and endogenous glucose production. The states G_1, G_2 are learned gut absorption compartments representing the two-stage carbohydrate absorption pathway.

Let $\xi_{t,j}$ denote the j-th component of the latent state vector ξ_t. Each component is updated by a neural transition function that receives only its physiologically admissible parent states, together with the planned inputs:

$$\xi_{t+1,j} = \xi_{t,j} + \Delta t\, f_j^{\Psi}\left(\xi_{t,\mathrm{Pa}_\xi(j)}, U_t\right), \qquad \Delta t = 1\,. \tag{31}$$

Here $U_t = [U_t^{\mathrm{ins}}, U_t^{\mathrm{carb}}]$ denotes insulin and carbohydrate inputs, $\mathrm{Pa}_\xi(j)$ is the parent state set for component j, and f_j^{Ψ} is the component-specific neural transition function with learnable parameters Ψ. Only f_{S_1} and f_{G_1} receive a driver input, as the graph below shows:

$$\Delta Q_1 = f_{Q_1}(Q_1, Q_2, x_1, x_3, G_2)\,, \qquad \Delta Q_2 = f_{Q_2}(Q_1, Q_2, x_1, x_2)\,, \tag{32}$$

$$\Delta S_1 = f_{S_1}(S_1, U^{\mathrm{ins}})\,, \qquad \Delta S_2 = f_{S_2}(S_1, S_2)\,, \tag{33}$$

$$\Delta I = f_I(S_2, I)\,, \qquad \Delta x_i = f_{x_i}(I, x_i)\,, \quad i = 1, 2, 3\,, \tag{34}$$

$$\Delta G_1 = f_{G_1}(G_1, U^{\mathrm{carb}})\,, \qquad \Delta G_2 = f_{G_2}(G_1, G_2)\,. \tag{35}$$

An LSTM encoder maps the observed history of glucose, insulin, and carbohydrate inputs to the initial latent state. The graph-structured decoder then rolls the state forward under future insulin and carbohydrate drivers. Glucose is read out from the accessible glucose compartment Q_1, divided by a learned positive scalar.

C Dataset Generation and Preprocessing

C.1 PAS Anesthesia Dataset

We generated MAP trajectories using the Python Anesthesia Simulator (PAS) (Aubouin et al., 2023), initialized near steady state under anesthesia. Each patient was simulated for 56,000 s, with outputs averaged into 10 s bins, yielding 560,000 observations. The forecasting channels were MAP (mmHg) and the propofol (mg/s) and norepinephrine (μg/s) infusion rates, with future MAP predicted under planned infusions. Sliding windows with a 5-minute stride (30 steps) yielded 16,600 examples.

Latent confounding. A nonnegative stimulation process combines randomly timed Gaussian pulses with slowly varying correlated drift. It raises MAP and the bispectral index (BIS) (Avidan et al., 2008), prompting the BIS controller to increase propofol delivery, and is therefore a common cause of higher MAP and greater propofol administration. For patient $i^\dagger$, a latent susceptibility $F_{i^\dagger} \sim \mathrm{Beta}(2, 5)$ lowers expected baseline MAP and cardiac output while increasing norepinephrine dosing, linking lower baseline pressure to greater vasopressor administration. These mechanisms induce observational associations opposing propofol's MAP-lowering and norepinephrine's MAP-raising effects.

Controller implementation. For patient $i^\dagger$ and dosing decision $j^\dagger$, taken every 60 s, let $\mathsf{BIS}^{\mathrm{obs}}_{i^\dagger,j^\dagger}$ and $\mathsf{MAP}^{\mathrm{obs}}_{i^\dagger,j^\dagger}$ denote the latest noisy measurements, and write $\mathsf{rate}^{\mathrm{propo}}_{i^\dagger,j^\dagger}$ and $\mathsf{rate}^{\mathrm{nore}}_{i^\dagger,j^\dagger}$ for the propofol and norepinephrine infusion rates, respectively. Proportional–integral (PI) controllers $\mathrm{PI}_{\mathrm{propo}}$ and $\mathrm{PI}_{\mathrm{nore}}$ (Aström & Hägglund, 2005) correct each rate around a patient-specific maintenance value, $\mathsf{rate}^{\mathrm{propo,base}}_{i^\dagger}$ or $\mathsf{rate}^{\mathrm{nore,base}}_{i^\dagger}$, computed to hold BIS and MAP near their targets. Since norepinephrine can only raise MAP, a cohort-wide MAP target is unreachable for some patients, so each patient's MAP target $\mathsf{MAP}^{\mathrm{tar}}_{i^\dagger}$ is set just above the pressure they reach on propofol alone, within

55–95 mmHg. With $\tilde{v}_{i^\dagger,j^\dagger} \sim \mathcal{N}(0, 0.02^2)$, the policy is

$$\mathsf{rate}^{\mathrm{propo}}_{i^\dagger,j^\dagger} = \mathrm{clip}_{[0,0.6]}\left(\mathsf{rate}^{\mathrm{propo,base}}_{i^\dagger} + \mathrm{PI}_{\mathrm{propo}}(\mathsf{BIS}^{\mathrm{obs}}_{i^\dagger,j^\dagger} - 50)\right), \tag{36}$$

$$\mathsf{rate}^{\mathrm{nore}}_{i^\dagger,j^\dagger} = \mathrm{clip}_{[0,0.5]}\Bigg(\mathrm{clip}_{[0,0.5]}\Big(\mathsf{rate}^{\mathrm{nore,base}}_{i^\dagger} + \mathrm{PI}_{\mathrm{nore}}(\mathsf{MAP}^{\mathrm{tar}}_{i^\dagger} - \mathsf{MAP}^{\mathrm{obs}}_{i^\dagger,j^\dagger}) + 0.25\mathsf{rate}^{\mathrm{propo}}_{i^\dagger,j^\dagger} + 0.03\mathbf{1}\{\mathsf{BIS}^{\mathrm{obs}}_{i^\dagger,j^\dagger} < 40\} + 0.06F_{i^\dagger}\Big) + \tilde{v}_{i^\dagger,j^\dagger}\Bigg). \tag{37}$$

The inner clip bounds the policy mean and the outer one the realized rate after exploration noise. $\mathrm{PI}_{\mathrm{propo}}$ uses a proportional gain of 2×10^{-3} and correction bounds of $[-0.25, 0.35]$, $\mathrm{PI}_{\mathrm{nore}}$ a gain of 4×10^{-3} and bounds of $[-0.15, 0.35]$. Both use an integral gain of 2×10^{-5}, with integral updates suspended whenever the proposed correction requires clipping.

C.2 UVA/Padova Control-IQ Dataset

We use closed-loop trajectories generated with the UVA/Padova Type 1 Diabetes Simulator (Man et al., 2014) and the Control-IQ controller (Brown et al., 2019). Each subject was simulated once for 56 days under Control-IQ with inter-day variability enabled, using profile-generation seed 7. Three repeating daily meal schedules were used, placing breakfast between 06:00 and 11:00, lunch between 11:00 and 13:00, and dinner between 18:00 and 20:00. Meals were defined relative to body weight, with carbohydrate amounts between 0.84 and 1.44 g/kg, so the absolute load depended on the subject. Meal timing varied with a standard deviation of 20 minutes and meal size by 15% around its nominal value. Hypoglycemia treatment was enabled with a detection threshold of 70 mg/dL. Sliding windows with a 6-hour stride (72 steps) yielded $22{,}000$ examples.

Latent confounding. Control-IQ raises insulin delivery when glucose is predicted to be high, and meal boluses accompany carbohydrate intake, so in the record greater insulin delivery coincides with higher glucose despite its glucose-lowering effect.

D Physiological-consistency metrics: full definitions

We test whether forecasts under altered driver plans move in the expected direction and preserve dose ordering for each treatment $\widetilde{a}$. Let $\hat{y}^{(0)}_{n,t}$ denote the prediction under the observed drivers at forecast origin n and lead time t, and let $\hat{y}^{(\widetilde{a},i)}_{n,t}$ denote the prediction when treatment $\widetilde{a}$ is added at dose $\delta_{\widetilde{a},i}$, with $\delta_{\widetilde{a},1} < \cdots < \delta_{\widetilde{a},N_{\widetilde{a}}}$. We set $\chi_{\widetilde{a}} = +1$ for treatments expected to raise the outcome (carbohydrate and norepinephrine) and $\chi_{\widetilde{a}} = -1$ for treatments expected to lower it (insulin and propofol). The sign-oriented effect is $\varepsilon^{(\widetilde{a},i)}_{n,t} = \chi_{\widetilde{a}}\left(\hat{y}^{(\widetilde{a},i)}_{n,t} - \hat{y}^{(0)}_{n,t}\right)$, which is positive exactly when the predicted response is in the expected physiological direction.

Metrics are first computed at each forecast origin, then averaged over origins within each patient and equally across patients. We denote this aggregation by $\underset{n}{\mathrm{MacroAvg}}$. All thresholds are set to zero, so the consistency scores assess response direction and dose ordering without imposing a positive minimum effect size. Since a clinically meaningful effect size varies across patients and across times of day, a zero threshold tests the weakest condition any physiologically appropriate response must satisfy.

Support-validity mask. The support-validity mask $W^{(\widetilde{a})}_{n,t}$ excludes lead times at which even a physiologically correct response could not be expressed, ensuring that the consistency scores reflect model behavior rather than situational constraints. It combines two conditions. First, the timing indicator $A^{(\widetilde{a})}_{n,t}$ specifies whether t lies within the prespecified response window for treatment $\widetilde{a}$, since no directional response is expected before the treatment takes effect or after its effect has worn off. Second, the headroom condition excludes points near the relevant boundary of the admissible range $(y_{\min}, y_{\max})$: a treatment expected to lower the outcome cannot express that response when the trajectory is already within a buffer $\varrho_{\widetilde{a}}$ of $y_{\min}$, and a treatment expected to raise the outcome cannot do so within $\varrho_{\widetilde{a}}$ of $y_{\max}$. We therefore define

$$W_{n,t}^{(\widetilde{a})} = A_{n,t}^{(\widetilde{a})} \cdot \Upsilon_{n,t}^{(\widetilde{a})}, \qquad \Upsilon_{n,t}^{(\widetilde{a})} = \begin{cases} \mathbb{I}[y_{n,t} > y_{\min} + \varrho_{\widetilde{a}}], & \chi_{\widetilde{a}} = -1, \\ \mathbb{I}[y_{n,t} < y_{\max} - \varrho_{\widetilde{a}}], & \chi_{\widetilde{a}} = +1, \end{cases} \tag{38}$$

with $\Upsilon_{n,t}^{(\widetilde{a})}$ the headroom indicator. A treatment–time pair is thus evaluable only when both the timing and headroom conditions hold.

Directional consistency (DC). The weakest requirement is that each treatment move the outcome the right way. $\mathrm{DC}_{\widetilde{a}}$ is the fraction of valid time–dose pairs whose response exceeds a minimum effect size $\vartheta_{\widetilde{a}}^{\mathrm{dir}} \geq 0$,

$$\mathrm{DC}_{\widetilde{a}} = \underset{n}{\mathrm{MacroAvg}} \left[\frac{\sum_t W_{n,t}^{(\widetilde{a})} \sum_{i=1}^{N_{\widetilde{a}}} \mathbb{I}\big[\varepsilon_{n,t}^{(\widetilde{a},i)} > \vartheta_{\widetilde{a}}^{\mathrm{dir}}\big]}{N_{\widetilde{a}} \sum_t W_{n,t}^{(\widetilde{a})}} \right], \tag{39}$$

A model can score well here while predicting nearly identical effects for every dose, so direction alone does not establish a usable dose–response relationship.

Ordering consistency (OC). We therefore also ask whether larger doses produce larger responses, which is what a clinician relies on when choosing between candidate doses. For adjacent doses, the sign-oriented dose–response slope is

$$\varphi_{n,t}^{(\widetilde{a},i)} = \chi_{\widetilde{a}} \frac{\hat{y}_{n,t}^{(\widetilde{a},i+1)} - \hat{y}_{n,t}^{(\widetilde{a},i)}}{\delta_{\widetilde{a},i+1} - \delta_{\widetilde{a},i}}, \qquad i = 1, \ldots, N_{\widetilde{a}} - 1, \tag{40}$$

so a positive slope means a larger dose produces a larger response in the expected direction. $\mathrm{OC}_{\widetilde{a}}$ is the fraction of valid time–dose pairs whose slope exceeds a threshold $\vartheta_{\widetilde{a}}^{\mathrm{ord}} \geq 0$,

$$\mathrm{OC}_{\widetilde{a}} = \underset{n}{\mathrm{MacroAvg}} \left[\frac{\sum_t W_{n,t}^{(\widetilde{a})} \sum_{i=1}^{N_{\widetilde{a}}-1} \mathbb{I}\big[\varphi_{n,t}^{(\widetilde{a},i)} > \vartheta_{\widetilde{a}}^{\mathrm{ord}}\big]}{(N_{\widetilde{a}} - 1) \sum_t W_{n,t}^{(\widetilde{a})}} \right]. \tag{41}$$

Complete response chain. The treatment-specific direction and ordering scores do not ensure that all inequalities hold jointly. Let $\widetilde{a} = -$ and $\widetilde{a} = +$ denote treatments with $\chi_{\widetilde{a}} = -1$ and $\chi_{\widetilde{a}} = +1$, respectively, and let N_- and N_+ be their numbers of doses. At origin n and lead time t, we require

$$\hat{y}_{n,t}^{(-,N_-)} < \cdots < \hat{y}_{n,t}^{(-,1)} < \hat{y}_{n,t}^{(0)} < \hat{y}_{n,t}^{(+,1)} < \cdots < \hat{y}_{n,t}^{(+,N_+)}. \tag{42}$$

Reading outward from $\hat{y}_{n,t}^{(0)}$, each larger dose must move the prediction further in its treatment's expected direction, and the two treatments must lie on opposite sides of the prediction under the observed drivers.

Graded violations (VD, VO, and VC). Consistency rates record how often the predicted responses satisfy the expected direction or dose ordering, but not how far a wrong response misses. We therefore also report graded hinge violations, which vanish for a correct response and grow with the size of the error. Writing $[x]_+ = \max(x, 0)$, the origin-level directional and ordering violations are

$$\omega_{n,\widetilde{a}}^{\mathrm{dir}} = \frac{\sum_t W_{n,t}^{(\widetilde{a})} \sum_{i=1}^{N_{\widetilde{a}}} \big[\vartheta_{\widetilde{a}}^{\mathrm{dir}} - \varepsilon_{n,t}^{(\widetilde{a},i)}\big]_+}{N_{\widetilde{a}} \sum_t W_{n,t}^{(\widetilde{a})}}, \qquad \omega_{n,\widetilde{a}}^{\mathrm{ord}} = \frac{\sum_t W_{n,t}^{(\widetilde{a})} \sum_{i=1}^{N_{\widetilde{a}}-1} \big[\vartheta_{\widetilde{a}}^{\mathrm{ord}} - \varphi_{n,t}^{(\widetilde{a},i)}\big]_+}{(N_{\widetilde{a}} - 1) \sum_t W_{n,t}^{(\widetilde{a})}}. \tag{43}$$

We report their patient-macro averages, $\mathrm{VD}_{\widetilde{a}} = \underset{n}{\mathrm{MacroAvg}}\big[\omega_{n,\widetilde{a}}^{\mathrm{dir}}\big]$ and $\mathrm{VO}_{\widetilde{a}} = \underset{n}{\mathrm{MacroAvg}}\big[\omega_{n,\widetilde{a}}^{\mathrm{ord}}\big]$.

The same construction applies to the response chain, which holds only when every adjacent relationship in Eq. 42 is correctly ordered. Its margin is therefore the weakest separation,

$$m_{n,t} = \min \Big\{ \hat{y}_{n,t}^{(0)} - \hat{y}_{n,t}^{(-,1)}, \quad \hat{y}_{n,t}^{(+,1)} - \hat{y}_{n,t}^{(0)}, \\ \min_{1 \leq i < N_-} \big(\hat{y}_{n,t}^{(-,i)} - \hat{y}_{n,t}^{(-,i+1)}\big), \quad \min_{1 \leq i < N_+} \big(\hat{y}_{n,t}^{(+,i+1)} - \hat{y}_{n,t}^{(+,i)}\big) \Big\}, \tag{44}$$

and the chain is consistent when $m_{n,t} > \vartheta^{\text{chain}}$. The chain violation is

$$\text{VC} = \underset{n}{\text{MacroAvg}} \left[\frac{\sum_t W^{(*)}_{n,t} \left[\vartheta^{\text{chain}} - m_{n,t}\right]_+}{\sum_t W^{(*)}_{n,t}} \right], \qquad W^{(*)}_{n,t} = W^{(-)}_{n,t} W^{(+)}_{n,t}, \tag{45}$$

where $W^{(*)}_{n,t} = 1$ only when both interventions are evaluable at (n, t).

Support coverage (Cov). Coverage reports what fraction of the response-window points of treatment $\widetilde{a}$ remain evaluable once the headroom condition is applied:

$$\text{Cov}_{\widetilde{a}} = \underset{n}{\text{MacroAvg}} \left[\frac{\sum_t W^{(\widetilde{a})}_{n,t}}{\sum_t A^{(\widetilde{a})}_{n,t}} \right], \tag{46}$$

where origins with no active response-window points are excluded. It indicates how much evidence the metrics rest on and, since the mask depends only on the observed trajectory, is a property of the data rather than of the model.

Dose-normalized response summaries. The consistency metrics test the direction and ordering of the response but say nothing about its shape, which is what a clinician would compare against known pharmacodynamics. Since effects scale with dose, we first normalize, $\nu^{(\widetilde{a},i)}_{n,t} = \varepsilon^{(\widetilde{a},i)}_{n,t} / \delta_{\widetilde{a},i}$, so that responses to different doses are comparable. Writing $\mathcal{T}^{(\widetilde{a})}_n = \{t : W^{(\widetilde{a})}_{n,t} = 1\}$ for the valid lead times, each origin and dose gives how strong the response becomes, when it arrives, and how much total effect it delivers:

$$\nu^{\text{peak}}_{n,\widetilde{a},i} = \max_{t \in \mathcal{T}^{(\widetilde{a})}_n} \nu^{(\widetilde{a},i)}_{n,t}, \quad t^{\text{peak}}_{n,\widetilde{a},i} = \arg\max_{t \in \mathcal{T}^{(\widetilde{a})}_n} \nu^{(\widetilde{a},i)}_{n,t}, \quad \text{AUEC}_{n,\widetilde{a},i} = \int_{\mathcal{T}^{(\widetilde{a})}_n} \nu^{(\widetilde{a},i)}_{n,t}\, dt, \tag{47}$$

the last being the area under the effect curve (AUEC), evaluated by the trapezoidal rule. For $\Xi \in \{\nu^{\text{peak}}, t^{\text{peak}}, \text{AUEC}\}$,

$$\overline{\Xi}_{\widetilde{a}} = \underset{n}{\text{MacroAvg}} \left[\frac{1}{N_{\widetilde{a}}} \sum_{i=1}^{N_{\widetilde{a}}} \Xi_{n,\widetilde{a},i} \right]. \tag{48}$$

Each intervention adds a single dose at the first step of the forecast horizon, on top of the observed driver plan; the doses used for each treatment are listed in Table 12. All branches for a given forecast origin share the same base sample (Section 3), and the consistency metrics are computed from the resulting complete trajectories.

E Training and Evaluation Protocol

Tables 13 and 14 give the search space and selected configuration for each model family on Control-IQ and PAS respectively, Table 15 lists the settings held fixed across all runs, and Table 12 gives the doses, response windows, and buffers used by the consistency metrics.

Architecture settings. All experiments use $K = 8$ Fourier frequencies. For JH-PAFM we set $(L_v, k_v) = (3, 5)$ on Control-IQ and $(3, 7)$ on PAS, so that the causal convolutional receptive field covers the 24- and 30-step forecast horizons respectively.

Training protocol. For each family we tune over 60 epochs with seed 42 at $\alpha = 0.1$, holding $\lambda_{\text{fid}} = 1$ and $\rho_{\text{fid}} = 0.1$ fixed for AR-PAFM. The selected configuration is retrained with seeds $\{1, 10, 42, 50\}$ for up to 150 epochs. Validation MSE drives tuning, early stopping (patience 15), and checkpoint selection, and all results are reported as mean$_{\pm\text{SD}}$ over the four seeds. The sensitivity-bound and fidelity ablations fix $\alpha = 0.1$.

Inference and calibration. Point forecasts use N_{MC} deterministically seeded base samples, where N_{MC} denotes the number of Monte Carlo samples. These samples are propagated through the conditional flow and Monte Carlo-averaged by the corresponding inference procedure, with latent

values mapped to the observation space through Eq. 1. For uncertainty evaluation, N_{MC} separately decoded one-sample forecasts are retained, yielding N_{MC} predictive trajectories per test window. Their empirical quantiles define central prediction intervals at nominal coverages from 10% to 90% in increments of 10%. Following standard regression calibration (Kuleshov et al., 2018; Chung et al., 2021), we report empirical calibration error (ECE) as the mean absolute difference between nominal and empirical coverage across these levels.

F Additional Results and Ablations

This section reports the full regularization sweeps for Control-IQ and PAS, together with three ablations: how the sensitivity bounds affect treatment-response magnitudes, how fidelity supervision affects AR-PAFM rollout accuracy, and whether the driver context helps or hinders on each dataset. Qualitative samples are available in Figures 3 and 4. The results show that both PAFM models produce physiologically meaningful forecasts that respond to the type of intervention and the administered dose, consistent with the physiological metrics.

F.1 Additional Control-IQ results and regularization analysis

Regularization improves consistency at varying costs in accuracy. At $\alpha = 0$, AR-PAFM has mean insulin directional and ordering consistency scores of 0% for both metrics, while JH-PAFM scores 30.990% and 24.252%, respectively; both formulations reach 100% on both metrics at $\alpha = 0.1$ (Table 6). The response summaries indicate wrong-signed integrated insulin responses on average. Unregularized AR-PAFM has $\overline{\nu}^{\mathrm{peak}}_{\mathrm{ins}} = -3.037\,\mathrm{mg/dL/U}$ and $\overline{\mathrm{AUEC}}_{\mathrm{ins}} = -911.467\,\mathrm{mg\,min/(dL\,U)}$, indicating a glucose-raising integrated response to additional insulin. The unregularized baselines also have negative mean insulin AUECs: -2027.512 for LSTM, -1494.697 for S4D, and -810.739 for Hovorka. Unregularized JH-PAFM has a positive mean peak ($1.761\,\mathrm{mg/dL/U}$) but a negative mean AUEC ($-91.378\,\mathrm{mg\,min/(dL\,U)}$), compatible with its partial directional consistency (Table 9). Increasing α from 0 to 0.1 changes the 120-minute mean RMSE from 21.83 to 22.33 mg/dL for AR-PAFM ($\uparrow 2.3\%$) and from 14.14 to 15.07 for JH-PAFM ($\uparrow 6.5\%$), compared with $8.92 \rightarrow 15.47$ for TCN ($\uparrow 73.4\%$), $11.48 \rightarrow 16.44$ for BNODE ($\uparrow 43.1\%$), and $19.05 \rightarrow 23.27$ for Hovorka ($\uparrow 22.1\%$) (Table 7).

PAFM maintains consistency across the regularization sweep. Both PAFM formulations retain mean directional and ordering scores above 99.9% on both treatment channels at every tested $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1\}$. Across these settings, 120-minute mean RMSE ranges from 15.07 to 17.89 mg/dL for JH-PAFM and from 22.12 to 23.70 mg/dL for AR-PAFM, showing that forecasting accuracy remains sensitive to the coefficient even when consistency scores remain near-perfect. For BNODE, increasing α from 0.1 to 0.5 raises mean insulin ordering consistency from 79.802% to 94.985%, while 120-minute mean RMSE increases from 16.44 to 25.23 mg/dL (Tables 6 and 7).

Calibration varies non-monotonically with regularization. JH-PAFM achieves lower mean ECE than AR-PAFM at every tested nonzero coefficient. Increasing α from 0.1 to 0.3 reduces JH-PAFM's ECE from 0.150 ± 0.027 to 0.059 ± 0.020, a 60.7% reduction in the mean, while its 120-minute mean RMSE increases by only 0.3% ($15.07 \rightarrow 15.12$ mg/dL). Further regularization does not improve calibration monotonically: ECE rises to 0.081 ± 0.039 at $\alpha = 0.5$ before falling to its lowest mean in the sweep, 0.031 ± 0.018, at $\alpha = 1$. At this last setting, however, 120-minute mean RMSE reaches 17.89 mg/dL, illustrating the accuracy cost associated with the lowest observed calibration error (Tables 7 and 8).

Support coverage. Support coverage is identical across models, with patient-macro averages of $\mathrm{Cov}_{\mathrm{ins}} = 99.777\%$ and $\mathrm{Cov}_{\mathrm{carb}} = 97.895\%$. The headroom condition therefore excludes only a small fraction of points within the prespecified response windows on average.

F.2 Additional PAS results

Unregularized treatment responses under confounded dosing. The simulator introduces confounding through unobserved stimulation for propofol and through patient susceptibility and state-

dependent dosing for norepinephrine (Appendix C.1). At $\alpha = 0$, JH-PAFM's mean propofol violations are $\mathrm{VD}_{\mathrm{propo}} = 0.105$ mmHg and $\mathrm{VO}_{\mathrm{propo}} = 0.088$ mmHg/dose. AR-PAFM instead exhibits norepinephrine violations of $\mathrm{VD}_{\mathrm{nore}} = 0.041$ mmHg and $\mathrm{VO}_{\mathrm{nore}} = 0.032$ mmHg/dose (Figure 2a). The response summaries distinguish these patterns further: JH-PAFM's mean sign-oriented propofol AUEC is negative (-0.155 mmHg·min/dose), indicating a wrong-signed integrated response on average, whereas AR-PAFM's mean norepinephrine AUEC remains positive (0.970 mmHg·min/dose) despite its nonzero directional and ordering violations (Figure 2b).

Selection of α on PAS. At $\alpha = 0.1$, AR-PAFM reports 75.0% mean consistency with a sample standard deviation of 50 percentage points on all four directional and ordering scores, consistent with three fully consistent runs and one run with zero consistency (Figure 1b). Although this setting yields AR-PAFM's lowest mean ECE (0.052 ± 0.059; Figure 1a), it does not satisfy the consistency requirement across all seeds. We therefore report $\alpha = 0.3$, the smallest tested coefficient at which both formulations achieve full reported directional and ordering consistency across all four seeds.

F.3 Effect-bound and fidelity ablations on Control-IQ

Sensitivity bounds alter response magnitude and forecasting accuracy. With carbohydrate bounds fixed, moving the insulin sensitivity band through the near-zero, reference, and high settings increases mean insulin AUEC monotonically across the tested configurations: $\overline{\mathrm{AUEC}}_{\mathrm{ins}}$ follows $71.82 \rightarrow 242.87 \rightarrow 495.92$ for AR-PAFM and $88.00 \rightarrow 321.68 \rightarrow 519.42$ for JH-PAFM. In the separate carbohydrate-bound sweep, with insulin bounds fixed, $\overline{\mathrm{AUEC}}_{\mathrm{carb}}$ follows $60.33 \rightarrow 76.03 \rightarrow 85.87$ for AR-PAFM and $61.27 \rightarrow 113.61 \rightarrow 107.40$ for JH-PAFM. For JH-PAFM, mean carbohydrate AUEC is highest at the reference setting, with a modest decrease under the high band (Table 10). Among the tested bound configurations, AR-PAFM achieves its lowest mean RMSE at all three horizons under the insulin near-zero setting: $9.99/14.65/21.60$ mg/dL at $30/60/120$ minutes, compared with $10.68/15.18/22.61$ at its reference configuration. JH-PAFM achieves its lowest mean RMSE at all three horizons under the carbohydrate near-zero setting: $5.26/8.58/14.06$ mg/dL, compared with $5.49/8.91/14.58$ at its reference configuration.

Fidelity supervision improves autoregressive rollout accuracy. In the fidelity-weight sweep ($\rho_{\mathrm{fid}} = 0.1$), setting $\lambda_{\mathrm{fid}} = 0$ raises AR-PAFM's 120-minute mean RMSE to 52.09 ± 7.84 mg/dL, compared with 21.72 ± 1.00 at $\lambda_{\mathrm{fid}} = 1$. The insulin response also exhibits greater variation across seeds: without fidelity supervision, $\overline{\nu}^{\mathrm{peak}}_{\mathrm{ins}} = 23.56 \pm 26.51$ mg/dL/U and $\overline{\mathrm{AUEC}}_{\mathrm{ins}} = 1071 \pm 1142\,\mathrm{mg\,min/(dL\,U)}$, compared with 7.67 ± 2.57 and 412 ± 142, respectively, at the reference configuration. At fixed $\lambda_{\mathrm{fid}} = 1$, increasing ρ_{fid} from 0.1 to 1 further reduces 120-minute mean RMSE from 21.72 ± 1.00 to 18.93 ± 1.23 mg/dL, while extending fidelity rollouts from 10% of each batch to all batch elements (Table 11).

F.4 Driver Context

Driver context improves Control-IQ accuracy but weakens PAS dose ordering. For JH-PAFM on Control-IQ, including ζ^D lowers mean RMSE and its across-seed standard deviation at every tested coefficient and horizon. At $\alpha = 0.3$, removing it increases 120-minute RMSE from 15.12 ± 0.80 to 24.22 ± 9.53 mg/dL, although both variants retain mean directional and ordering scores above 99.9% at every nonzero coefficient (Table 3). On PAS, omitting the context lowers mean RMSE at every horizon for every tested $\alpha > 0$ and yields 100% directional and ordering consistency on both channels (Table 4). At $\alpha = 0.3$, adding it increases 5-minute RMSE from 3.72 ± 0.11 to 4.81 ± 1.05 mmHg and reduces propofol and norepinephrine ordering scores to 47.359% and 78.309%, respectively, despite near-perfect directional scores; mean propofol ordering remains between 40.4 and 47.4% at every tested $\alpha \geq 0.3$. The context is included for Control-IQ and omitted for PAS based on validation MSE; the test results are consistent with these choices.

The direct route from cumulative drivers to the velocity field may ease the competition between trajectory fitting and sensitivity regularization on Control-IQ. Over the shorter PAS horizon, the smoothed drivers and temporal mixing may already provide an adequate treatment representation, so the additional nonlinear conditioning adds flexibility that the penalty does not constrain.

Table 3: Effect of driver context ζ^D on JH-PAFM on Control-IQ across regularization weights α. RMSE, VD, and VC are in mg/dL; VO is in mg/dL per dose unit. DC and OC are percentages. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best means within each α and metric, respectively; ties share bold.

Driver context	α	RMSE ↓			DC (%) ↑		OC (%) ↑		VD ↓		VO ↓		VC ↓
		30 min	60 min	120 min	Ins.	Carb.	Ins.	Carb.	Ins.	Carb.	Ins.	Carb.	
ζ^D	0	$\mathbf{5.209}_{\pm 0.323}$	$\mathbf{8.667}_{\pm 0.880}$	$\mathbf{14.143}_{\pm 1.196}$	$\mathbf{30.990}_{\pm 3.885}$	$98.476_{\pm 0.518}$	$\mathbf{24.252}_{\pm 5.739}$	$99.359_{\pm 0.234}$	$\mathbf{2.589}_{\pm 0.360}$	$0.101_{\pm 0.054}$	$\mathbf{1.383}_{\pm 0.214}$	$\mathbf{0.001}_{\pm 0.001}$	$\mathbf{2.871}_{\pm 0.558}$
	0.1	$\mathbf{6.124}_{\pm 0.640}$	$\mathbf{9.408}_{\pm 0.531}$	$\mathbf{15.069}_{\pm 0.768}$	$\mathbf{100.000}_{\pm 0.000}$	$99.997_{\pm 0.006}$	$\mathbf{100.000}_{\pm 0.000}$	$99.975_{\pm 0.019}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$0.001_{\pm 0.001}$
	0.3	$\mathbf{5.896}_{\pm 0.554}$	$\mathbf{9.691}_{\pm 0.541}$	$\mathbf{15.120}_{\pm 0.803}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$99.932_{\pm 0.103}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.5	$\mathbf{6.969}_{\pm 0.610}$	$\mathbf{10.610}_{\pm 0.778}$	$\mathbf{16.202}_{\pm 1.338}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.001}$	$\mathbf{100.000}_{\pm 0.000}$	$99.998_{\pm 0.004}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.7	$\mathbf{7.013}_{\pm 0.465}$	$\mathbf{10.778}_{\pm 0.684}$	$\mathbf{16.817}_{\pm 1.135}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$99.978_{\pm 0.044}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	1	$\mathbf{7.751}_{\pm 0.981}$	$\mathbf{11.817}_{\pm 0.632}$	$\mathbf{17.891}_{\pm 0.492}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
No ζ^D	0	$5.295_{\pm 0.596}$	$8.818_{\pm 1.219}$	$14.484_{\pm 2.183}$	$2.612_{\pm 1.003}$	$\mathbf{99.412}_{\pm 0.421}$	$2.593_{\pm 1.011}$	$\mathbf{99.568}_{\pm 0.322}$	$3.455_{\pm 0.764}$	$\mathbf{0.022}_{\pm 0.021}$	$1.849_{\pm 0.411}$	$\mathbf{0.001}_{\pm 0.001}$	$3.857_{\pm 0.804}$
	0.1	$7.535_{\pm 2.621}$	$11.405_{\pm 2.781}$	$18.007_{\pm 4.221}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.3	$11.193_{\pm 6.953}$	$15.654_{\pm 6.864}$	$24.218_{\pm 9.528}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.5	$25.909_{\pm 25.900}$	$32.080_{\pm 29.973}$	$45.685_{\pm 38.940}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.7	$45.917_{\pm 35.721}$	$53.460_{\pm 38.560}$	$70.789_{\pm 43.954}$	$99.998_{\pm 0.004}$	$\mathbf{100.000}_{\pm 0.000}$	$99.966_{\pm 0.069}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	1	$23.551_{\pm 23.735}$	$27.179_{\pm 23.024}$	$39.133_{\pm 29.776}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$

Table 4: Effect of driver context ζ^D on JH-PAFM on PAS across regularization weights α. RMSE is evaluated at 1, 2.5, and 5 minutes. RMSE, VD, and VC are in mmHg; VO is in mmHg per dose unit. DC and OC are percentages. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best means within each α and metric, respectively; ties share bold.

Driver context	α	RMSE ↓			DC (%) ↑		OC (%) ↑		VD ↓		VO ↓		VC ↓
		1 min	2.5 min	5 min	Propo.	Nore.	Propo.	Nore.	Propo.	Nore.	Propo.	Nore.	
ζ^D	0	$\mathbf{2.178}_{\pm 0.028}$	$2.849_{\pm 0.063}$	$3.439_{\pm 0.096}$	$\mathbf{58.195}_{\pm 7.009}$	$\mathbf{99.939}_{\pm 0.103}$	$\mathbf{48.594}_{\pm 10.970}$	$\mathbf{99.990}_{\pm 0.020}$	$0.109_{\pm 0.031}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.088}_{\pm 0.043}$	$\mathbf{0.000}_{\pm 0.000}$	$0.064_{\pm 0.020}$
	0.1	$3.501_{\pm 0.990}$	$3.946_{\pm 0.406}$	$4.769_{\pm 0.413}$	$96.563_{\pm 6.734}$	$96.161_{\pm 7.678}$	$84.457_{\pm 17.982}$	$74.675_{\pm 45.003}$	$0.035_{\pm 0.070}$	$0.020_{\pm 0.040}$	$0.091_{\pm 0.175}$	$0.294_{\pm 0.576}$	$0.196_{\pm 0.369}$
	0.3	$3.890_{\pm 1.273}$	$4.150_{\pm 1.011}$	$4.809_{\pm 1.051}$	$99.440_{\pm 1.115}$	$99.983_{\pm 0.035}$	$47.359_{\pm 25.558}$	$78.309_{\pm 10.051}$	$0.001_{\pm 0.001}$	$\mathbf{0.000}_{\pm 0.000}$	$0.280_{\pm 0.202}$	$0.292_{\pm 0.350}$	$0.548_{\pm 0.516}$
	0.5	$3.414_{\pm 0.892}$	$4.046_{\pm 0.786}$	$4.941_{\pm 1.045}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$40.406_{\pm 11.277}$	$66.289_{\pm 19.336}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$0.356_{\pm 0.211}$	$0.546_{\pm 0.499}$	$0.801_{\pm 0.516}$
	0.7	$4.538_{\pm 2.228}$	$4.763_{\pm 1.981}$	$5.076_{\pm 1.258}$	$99.996_{\pm 0.008}$	$\mathbf{100.000}_{\pm 0.000}$	$45.224_{\pm 26.539}$	$81.346_{\pm 23.102}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$0.359_{\pm 0.463}$	$0.225_{\pm 0.333}$	$0.508_{\pm 0.390}$
	1	$3.950_{\pm 1.147}$	$4.154_{\pm 0.918}$	$4.761_{\pm 1.004}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$43.588_{\pm 16.064}$	$82.517_{\pm 14.626}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$0.300_{\pm 0.171}$	$0.178_{\pm 0.192}$	$0.440_{\pm 0.265}$
No ζ^D	0	$2.185_{\pm 0.063}$	$\mathbf{2.837}_{\pm 0.048}$	$\mathbf{3.434}_{\pm 0.045}$	$18.051_{\pm 16.505}$	$98.717_{\pm 1.789}$	$16.455_{\pm 14.219}$	$99.097_{\pm 1.366}$	$\mathbf{0.105}_{\pm 0.048}$	$0.001_{\pm 0.002}$	$\mathbf{0.088}_{\pm 0.038}$	$0.001_{\pm 0.001}$	$\mathbf{0.047}_{\pm 0.019}$
	0.1	$\mathbf{2.253}_{\pm 0.061}$	$\mathbf{2.925}_{\pm 0.016}$	$\mathbf{3.564}_{\pm 0.130}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.3	$\mathbf{2.420}_{\pm 0.156}$	$\mathbf{3.064}_{\pm 0.114}$	$\mathbf{3.723}_{\pm 0.105}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.5	$\mathbf{2.840}_{\pm 0.622}$	$\mathbf{3.604}_{\pm 0.503}$	$\mathbf{4.340}_{\pm 0.600}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	0.7	$\mathbf{3.154}_{\pm 0.922}$	$\mathbf{3.573}_{\pm 0.449}$	$\mathbf{4.117}_{\pm 0.244}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$
	1	$\mathbf{3.327}_{\pm 0.821}$	$\mathbf{3.623}_{\pm 0.312}$	$\mathbf{4.202}_{\pm 0.130}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{100.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$

Table 5: Control-IQ directional (VD), ordering (VO), and complete-chain (VC) violation magnitudes across regularization weights α. VD and VC are in mg/dL; VO is in mg/dL per dose unit. Lower is better. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best means within each model and metric, respectively.

Model	α	$\mathrm{VD}_{\mathrm{ins}}$	$\mathrm{VD}_{\mathrm{carb}}$	$\mathrm{VO}_{\mathrm{ins}}$	$\mathrm{VO}_{\mathrm{carb}}$	VC
BNODE	0.0	$31.548_{\pm 5.564}$	$0.253_{\pm 0.123}$	$20.905_{\pm 3.775}$	$0.009_{\pm 0.004}$	$55.566_{\pm 13.212}$
BNODE	0.1	$0.127_{\pm 0.079}$	$0.025_{\pm 0.007}$	$0.434_{\pm 0.324}$	$0.002_{\pm 0.001}$	$1.424_{\pm 0.982}$
BNODE	0.3	$\mathbf{0.023_{\pm 0.029}}$	$0.090_{\pm 0.064}$	$\mathbf{0.107_{\pm 0.141}}$	$0.004_{\pm 0.003}$	$0.637_{\pm 0.793}$
BNODE	0.5	$0.075_{\pm 0.142}$	$0.020_{\pm 0.014}$	$0.136_{\pm 0.264}$	$\mathbf{0.001_{\pm 0.001}}$	$0.583_{\pm 1.139}$
BNODE	0.7	$0.169_{\pm 0.262}$	$\mathbf{0.010_{\pm 0.009}}$	$0.177_{\pm 0.186}$	$\mathbf{0.001_{\pm 0.001}}$	$\mathbf{0.548_{\pm 0.519}}$
Hovorka	0.0	$21.214_{\pm 11.430}$	$0.089_{\pm 0.138}$	$14.211_{\pm 7.669}$	$0.010_{\pm 0.014}$	$36.887_{\pm 22.846}$
Hovorka	0.1	$0.100_{\pm 0.112}$	$\mathbf{0.000_{\pm 0.000}}$	$0.127_{\pm 0.067}$	$\mathbf{0.000_{\pm 0.000}}$	$0.278_{\pm 0.148}$
Hovorka	0.3	$\mathbf{0.028_{\pm 0.053}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.063_{\pm 0.075}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.150_{\pm 0.189}}$
Hovorka	0.5	$0.095_{\pm 0.068}$	$\mathbf{0.000_{\pm 0.000}}$	$0.139_{\pm 0.091}$	$\mathbf{0.000_{\pm 0.000}}$	$0.362_{\pm 0.246}$
Hovorka	0.7	$0.037_{\pm 0.066}$	$0.001_{\pm 0.002}$	$0.082_{\pm 0.119}$	$\mathbf{0.000_{\pm 0.000}}$	$0.232_{\pm 0.308}$
LP-Reduced	0.0	$12.911_{\pm 10.004}$	$0.002_{\pm 0.001}$	$6.563_{\pm 5.081}$	$\mathbf{0.000_{\pm 0.000}}$	$11.349_{\pm 9.368}$
LP-Reduced	0.1	$0.006_{\pm 0.013}$	$0.003_{\pm 0.004}$	$0.003_{\pm 0.007}$	$\mathbf{0.000_{\pm 0.000}}$	$0.006_{\pm 0.012}$
LP-Reduced	0.3	$\mathbf{0.000_{\pm 0.000}}$	$0.005_{\pm 0.004}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
LP-Reduced	0.5	$\mathbf{0.000_{\pm 0.000}}$	$0.003_{\pm 0.001}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
LP-Reduced	0.7	$0.003_{\pm 0.006}$	$\mathbf{0.001_{\pm 0.002}}$	$0.001_{\pm 0.003}$	$\mathbf{0.000_{\pm 0.000}}$	$0.002_{\pm 0.004}$
LSTM	0.0	$44.449_{\pm 6.929}$	$0.256_{\pm 0.092}$	$27.097_{\pm 4.438}$	$0.006_{\pm 0.002}$	$63.646_{\pm 11.540}$
LSTM	0.1	$0.892_{\pm 0.722}$	$0.013_{\pm 0.017}$	$1.689_{\pm 1.205}$	$0.001_{\pm 0.001}$	$5.834_{\pm 4.328}$
LSTM	0.3	$0.524_{\pm 0.374}$	$0.006_{\pm 0.004}$	$0.629_{\pm 0.449}$	$\mathbf{0.000_{\pm 0.000}}$	$2.849_{\pm 1.935}$
LSTM	0.5	$0.422_{\pm 0.807}$	$\mathbf{0.003_{\pm 0.005}}$	$0.425_{\pm 0.708}$	$\mathbf{0.000_{\pm 0.000}}$	$1.922_{\pm 3.036}$
LSTM	0.7	$\mathbf{0.044_{\pm 0.020}}$	$0.004_{\pm 0.003}$	$\mathbf{0.129_{\pm 0.060}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.727_{\pm 0.356}}$
S4D	0.0	$31.155_{\pm 1.725}$	$0.345_{\pm 0.119}$	$17.322_{\pm 0.993}$	$0.009_{\pm 0.002}$	$46.716_{\pm 2.559}$
S4D	0.1	$9.192_{\pm 1.369}$	$\mathbf{0.134_{\pm 0.048}}$	$6.371_{\pm 1.098}$	$\mathbf{0.004_{\pm 0.001}}$	$12.544_{\pm 2.774}$
S4D	0.3	$7.634_{\pm 0.670}$	$0.187_{\pm 0.106}$	$4.750_{\pm 0.535}$	$0.005_{\pm 0.002}$	$9.611_{\pm 1.068}$
S4D	0.5	$5.357_{\pm 1.555}$	$0.144_{\pm 0.068}$	$3.255_{\pm 0.982}$	$\mathbf{0.004_{\pm 0.001}}$	$7.013_{\pm 1.826}$
S4D	0.7	$\mathbf{3.045_{\pm 0.604}}$	$0.178_{\pm 0.033}$	$\mathbf{1.787_{\pm 0.344}}$	$\mathbf{0.004_{\pm 0.001}}$	$\mathbf{4.459_{\pm 1.275}}$
TCN	0.0	$11.818_{\pm 1.504}$	$1.732_{\pm 0.344}$	$8.087_{\pm 1.176}$	$0.029_{\pm 0.004}$	$16.823_{\pm 3.632}$
TCN	0.1	$\mathbf{0.437_{\pm 0.352}}$	$\mathbf{0.018_{\pm 0.006}}$	$0.760_{\pm 0.493}$	$\mathbf{0.001_{\pm 0.001}}$	$1.526_{\pm 0.925}$
TCN	0.3	$1.237_{\pm 0.544}$	$0.021_{\pm 0.022}$	$1.346_{\pm 0.301}$	$\mathbf{0.001_{\pm 0.001}}$	$2.885_{\pm 0.415}$
TCN	0.5	$0.444_{\pm 0.526}$	$0.035_{\pm 0.035}$	$\mathbf{0.489_{\pm 0.549}}$	$0.002_{\pm 0.002}$	$\mathbf{0.967_{\pm 1.028}}$
TCN	0.7	$0.445_{\pm 0.610}$	$0.127_{\pm 0.230}$	$0.636_{\pm 0.582}$	$0.005_{\pm 0.010}$	$1.915_{\pm 0.945}$
AR-PAFM	0.0	$16.961_{\pm 1.040}$	$\mathbf{0.000_{\pm 0.000}}$	$9.116_{\pm 0.550}$	$\mathbf{0.000_{\pm 0.000}}$	$18.821_{\pm 1.052}$
AR-PAFM	0.1	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
AR-PAFM	0.3	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
AR-PAFM	0.5	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
AR-PAFM	0.7	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
AR-PAFM	1.0	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
JH-PAFM	0.0	$2.589_{\pm 0.360}$	$0.101_{\pm 0.054}$	$1.383_{\pm 0.214}$	$0.001_{\pm 0.001}$	$2.871_{\pm 0.558}$
JH-PAFM	0.1	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$0.001_{\pm 0.001}$
JH-PAFM	0.3	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
JH-PAFM	0.5	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
JH-PAFM	0.7	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$
JH-PAFM	1.0	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$

Table 6: Control-IQ directional (DC) and dose-ordering (OC) consistency across regularization weights α. Scores are percentages; higher is better. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best means within each model and metric, respectively.

Model	α	$\mathrm{DC}_{\mathrm{ins}}$ (%)	$\mathrm{DC}_{\mathrm{carb}}$ (%)	$\mathrm{OC}_{\mathrm{ins}}$ (%)	$\mathrm{OC}_{\mathrm{carb}}$ (%)
BNODE	0.0	$25.280_{\pm 4.804}$	$85.611_{\pm 8.968}$	$12.425_{\pm 1.149}$	$89.298_{\pm 4.502}$
BNODE	0.1	$96.210_{\pm 1.331}$	$98.676_{\pm 0.350}$	$79.802_{\pm 12.389}$	$97.437_{\pm 0.577}$
BNODE	0.3	$\mathbf{98.836_{\pm 1.731}}$	$97.276_{\pm 2.171}$	$91.698_{\pm 10.230}$	$95.001_{\pm 1.244}$
BNODE	0.5	$98.206_{\pm 2.201}$	$98.735_{\pm 0.829}$	$\mathbf{94.985_{\pm 7.761}}$	$97.980_{\pm 0.793}$
BNODE	0.7	$87.051_{\pm 20.204}$	$\mathbf{99.532_{\pm 0.222}}$	$73.551_{\pm 23.808}$	$\mathbf{98.522_{\pm 0.968}}$
Hovorka	0.0	$20.649_{\pm 7.515}$	$91.113_{\pm 5.705}$	$11.617_{\pm 5.492}$	$89.815_{\pm 6.802}$
Hovorka	0.1	$66.091_{\pm 20.863}$	$\mathbf{95.385_{\pm 0.038}}$	$33.143_{\pm 9.637}$	$\mathbf{95.343_{\pm 0.103}}$
Hovorka	0.3	$\mathbf{81.711_{\pm 14.808}}$	$95.372_{\pm 0.070}$	$\mathbf{41.792_{\pm 35.254}}$	$95.315_{\pm 0.107}$
Hovorka	0.5	$61.909_{\pm 10.817}$	$95.374_{\pm 0.037}$	$21.414_{\pm 21.229}$	$95.286_{\pm 0.094}$
Hovorka	0.7	$74.157_{\pm 14.093}$	$94.130_{\pm 2.565}$	$32.259_{\pm 21.785}$	$94.655_{\pm 1.468}$
LP-Reduced	0.0	$27.797_{\pm 47.425}$	$\mathbf{94.303_{\pm 0.735}}$	$28.659_{\pm 46.813}$	$93.875_{\pm 1.180}$
LP-Reduced	0.1	$91.139_{\pm 17.721}$	$94.175_{\pm 0.809}$	$91.230_{\pm 17.540}$	$\mathbf{94.172_{\pm 0.814}}$
LP-Reduced	0.3	$\mathbf{100.000_{\pm 0.000}}$	$93.204_{\pm 0.678}$	$\mathbf{100.000_{\pm 0.000}}$	$93.223_{\pm 0.654}$
LP-Reduced	0.5	$99.773_{\pm 0.441}$	$92.903_{\pm 0.765}$	$99.773_{\pm 0.443}$	$92.905_{\pm 0.762}$
LP-Reduced	0.7	$96.312_{\pm 7.376}$	$93.913_{\pm 1.324}$	$97.335_{\pm 5.331}$	$93.564_{\pm 1.277}$
LSTM	0.0	$14.801_{\pm 1.430}$	$86.834_{\pm 4.548}$	$12.539_{\pm 0.995}$	$89.147_{\pm 3.347}$
LSTM	0.1	$87.305_{\pm 6.643}$	$\mathbf{99.458_{\pm 0.441}}$	$60.769_{\pm 13.042}$	$98.917_{\pm 0.443}$
LSTM	0.3	$85.905_{\pm 9.517}$	$99.128_{\pm 0.534}$	$63.266_{\pm 24.567}$	$98.724_{\pm 0.710}$
LSTM	0.5	$90.942_{\pm 14.139}$	$99.355_{\pm 0.692}$	$74.920_{\pm 22.642}$	$99.208_{\pm 0.988}$
LSTM	0.7	$\mathbf{95.343_{\pm 1.615}}$	$99.270_{\pm 0.517}$	$\mathbf{78.092_{\pm 5.804}}$	$\mathbf{99.268_{\pm 0.492}}$
S4D	0.0	$6.304_{\pm 1.671}$	$83.773_{\pm 4.733}$	$7.067_{\pm 2.115}$	$87.998_{\pm 2.451}$
S4D	0.1	$\mathbf{55.599_{\pm 1.530}}$	$95.479_{\pm 2.028}$	$47.997_{\pm 3.023}$	$\mathbf{94.936_{\pm 0.344}}$
S4D	0.3	$51.468_{\pm 2.908}$	$95.336_{\pm 3.117}$	$\mathbf{51.947_{\pm 2.006}}$	$94.310_{\pm 1.138}$
S4D	0.5	$47.893_{\pm 7.282}$	$\mathbf{95.885_{\pm 2.646}}$	$49.357_{\pm 6.323}$	$94.519_{\pm 0.785}$
S4D	0.7	$44.743_{\pm 4.342}$	$94.501_{\pm 1.766}$	$47.751_{\pm 4.155}$	$94.164_{\pm 0.341}$
TCN	0.0	$34.064_{\pm 1.073}$	$49.219_{\pm 8.250}$	$19.071_{\pm 2.621}$	$72.769_{\pm 5.422}$
TCN	0.1	$\mathbf{94.020_{\pm 3.660}}$	$98.836_{\pm 0.495}$	$80.937_{\pm 10.015}$	$\mathbf{97.461_{\pm 1.126}}$
TCN	0.3	$85.643_{\pm 3.236}$	$\mathbf{98.966_{\pm 0.936}}$	$69.861_{\pm 1.189}$	$97.354_{\pm 2.538}$
TCN	0.5	$93.436_{\pm 7.329}$	$98.595_{\pm 1.067}$	$\mathbf{86.285_{\pm 14.686}}$	$96.364_{\pm 2.501}$
TCN	0.7	$92.902_{\pm 7.955}$	$97.152_{\pm 3.479}$	$80.015_{\pm 14.790}$	$95.589_{\pm 7.187}$
AR-PAFM	0.0	$0.000_{\pm 0.000}$	$99.999_{\pm 0.003}$	$0.000_{\pm 0.000}$	$99.998_{\pm 0.004}$
AR-PAFM	0.1	$\mathbf{100.000_{\pm 0.000}}$	$99.999_{\pm 0.002}$	$\mathbf{100.000_{\pm 0.000}}$	$99.999_{\pm 0.002}$
AR-PAFM	0.3	$99.998_{\pm 0.003}$	$\mathbf{100.000_{\pm 0.000}}$	$99.998_{\pm 0.002}$	$\mathbf{100.000_{\pm 0.000}}$
AR-PAFM	0.5	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.001}}$
AR-PAFM	0.7	$99.987_{\pm 0.019}$	$\mathbf{100.000_{\pm 0.000}}$	$99.991_{\pm 0.014}$	$99.997_{\pm 0.003}$
AR-PAFM	1.0	$99.957_{\pm 0.087}$	$99.998_{\pm 0.003}$	$99.964_{\pm 0.072}$	$99.974_{\pm 0.052}$
JH-PAFM	0.0	$30.990_{\pm 3.885}$	$98.476_{\pm 0.518}$	$24.252_{\pm 5.739}$	$99.359_{\pm 0.234}$
JH-PAFM	0.1	$\mathbf{100.000_{\pm 0.000}}$	$99.997_{\pm 0.006}$	$\mathbf{100.000_{\pm 0.000}}$	$99.975_{\pm 0.019}$
JH-PAFM	0.3	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$99.932_{\pm 0.103}$
JH-PAFM	0.5	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.001}}$	$\mathbf{100.000_{\pm 0.000}}$	$99.998_{\pm 0.004}$
JH-PAFM	0.7	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$99.978_{\pm 0.044}$
JH-PAFM	1.0	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$	$\mathbf{100.000_{\pm 0.000}}$

Table 7: Control-IQ forecasting RMSE at 30, 60, and 120 minutes across regularization weights α. RMSE is in mg/dL; lower is better. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best means within each model and forecast horizon, respectively.

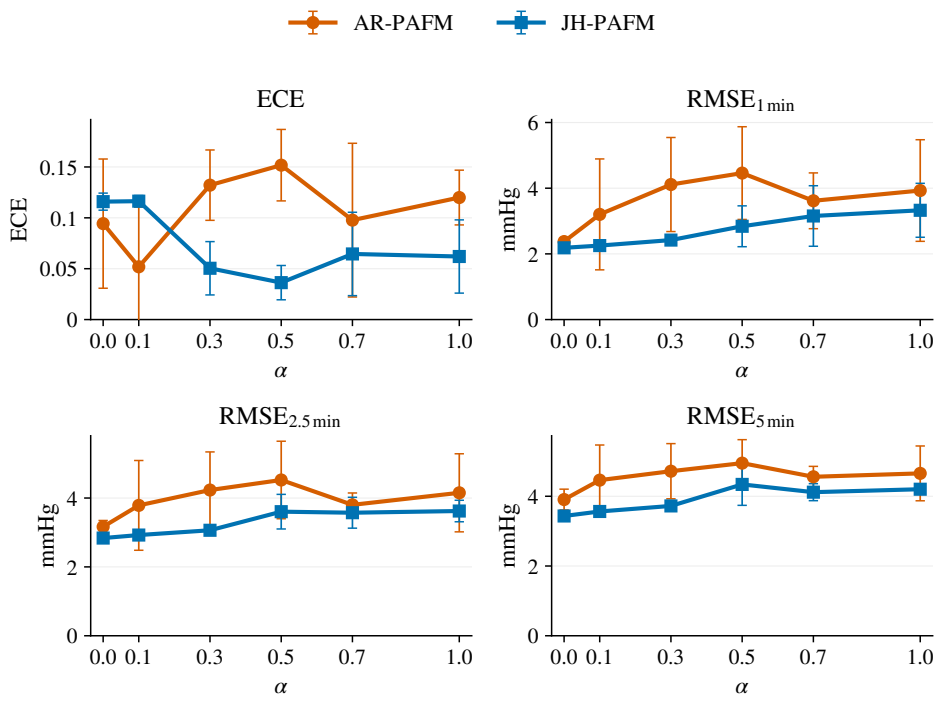
Model	α	$\mathrm{RMSE}_{30\mathrm{m}}$	$\mathrm{RMSE}_{60\mathrm{m}}$	$\mathrm{RMSE}_{120\mathrm{m}}$
BNODE	0.0	$\mathbf{6.409_{\pm 0.129}}$	$\mathbf{9.392_{\pm 0.106}}$	$\mathbf{11.484_{\pm 0.258}}$
BNODE	0.1	$8.957_{\pm 0.182}$	$13.664_{\pm 0.266}$	$16.437_{\pm 0.686}$
BNODE	0.3	$11.605_{\pm 1.429}$	$16.843_{\pm 1.466}$	$23.973_{\pm 2.312}$
BNODE	0.5	$12.059_{\pm 0.591}$	$17.409_{\pm 0.817}$	$25.232_{\pm 1.392}$
BNODE	0.7	$13.201_{\pm 1.796}$	$18.293_{\pm 1.611}$	$26.206_{\pm 1.810}$
Hovorka	0.0	$9.505_{\pm 0.207}$	$\mathbf{14.541_{\pm 0.348}}$	$\mathbf{19.053_{\pm 0.768}}$
Hovorka	0.1	$\mathbf{9.374_{\pm 0.346}}$	$15.341_{\pm 0.169}$	$23.271_{\pm 0.550}$
Hovorka	0.3	$9.930_{\pm 1.070}$	$15.935_{\pm 1.450}$	$24.165_{\pm 1.883}$
Hovorka	0.5	$10.131_{\pm 0.607}$	$16.056_{\pm 0.873}$	$24.187_{\pm 1.293}$
Hovorka	0.7	$10.467_{\pm 0.944}$	$16.789_{\pm 1.483}$	$25.359_{\pm 1.681}$
LP-Reduced	0.0	$11.336_{\pm 0.489}$	$17.713_{\pm 1.039}$	$25.352_{\pm 1.899}$
LP-Reduced	0.1	$63.653_{\pm 104.717}$	$53.812_{\pm 71.937}$	$50.686_{\pm 47.861}$
LP-Reduced	0.3	$10.655_{\pm 1.317}$	$16.903_{\pm 1.597}$	$25.423_{\pm 2.588}$
LP-Reduced	0.5	$\mathbf{10.407_{\pm 1.165}}$	$\mathbf{16.564_{\pm 1.631}}$	$\mathbf{24.634_{\pm 3.225}}$
LP-Reduced	0.7	$11.079_{\pm 0.813}$	$17.460_{\pm 0.928}$	$26.287_{\pm 1.569}$
LSTM	0.0	$\mathbf{6.795_{\pm 0.332}}$	$\mathbf{9.659_{\pm 0.187}}$	$\mathbf{11.215_{\pm 0.154}}$
LSTM	0.1	$8.585_{\pm 0.784}$	$13.771_{\pm 1.302}$	$17.067_{\pm 2.487}$
LSTM	0.3	$9.093_{\pm 0.198}$	$14.109_{\pm 0.640}$	$19.732_{\pm 1.564}$
LSTM	0.5	$9.579_{\pm 0.780}$	$14.790_{\pm 1.088}$	$20.935_{\pm 1.971}$
LSTM	0.7	$9.748_{\pm 0.445}$	$14.550_{\pm 0.560}$	$20.571_{\pm 1.026}$
S4D	0.0	$\mathbf{8.630_{\pm 0.170}}$	$\mathbf{10.034_{\pm 0.147}}$	$\mathbf{10.356_{\pm 0.089}}$
S4D	0.1	$13.676_{\pm 0.439}$	$15.565_{\pm 0.420}$	$17.179_{\pm 0.473}$
S4D	0.3	$17.355_{\pm 0.882}$	$19.140_{\pm 0.818}$	$21.303_{\pm 1.000}$
S4D	0.5	$19.403_{\pm 1.485}$	$21.209_{\pm 1.578}$	$24.098_{\pm 1.847}$
S4D	0.7	$23.392_{\pm 0.961}$	$24.634_{\pm 0.781}$	$27.833_{\pm 0.689}$
TCN	0.0	$\mathbf{4.665_{\pm 0.223}}$	$\mathbf{6.166_{\pm 0.220}}$	$\mathbf{8.920_{\pm 0.186}}$
TCN	0.1	$9.479_{\pm 0.380}$	$11.789_{\pm 0.619}$	$15.467_{\pm 1.010}$
TCN	0.3	$11.045_{\pm 0.388}$	$13.404_{\pm 0.336}$	$17.179_{\pm 0.273}$
TCN	0.5	$12.336_{\pm 0.522}$	$14.884_{\pm 0.743}$	$19.355_{\pm 1.112}$
TCN	0.7	$15.719_{\pm 1.735}$	$18.434_{\pm 2.282}$	$23.074_{\pm 2.743}$
AR-PAFM	0.0	$10.651_{\pm 1.257}$	$15.347_{\pm 0.959}$	$\mathbf{21.830_{\pm 0.893}}$
AR-PAFM	0.1	$10.043_{\pm 0.760}$	$14.538_{\pm 0.742}$	$22.329_{\pm 1.038}$
AR-PAFM	0.3	$\mathbf{9.864_{\pm 0.732}}$	$\mathbf{14.465_{\pm 1.044}}$	$22.120_{\pm 1.243}$
AR-PAFM	0.5	$10.789_{\pm 1.403}$	$15.112_{\pm 0.750}$	$22.614_{\pm 1.420}$
AR-PAFM	0.7	$11.036_{\pm 1.557}$	$15.401_{\pm 1.023}$	$23.702_{\pm 1.344}$
AR-PAFM	1.0	$10.029_{\pm 0.334}$	$14.596_{\pm 0.255}$	$22.575_{\pm 0.637}$
JH-PAFM	0.0	$\mathbf{5.209_{\pm 0.323}}$	$\mathbf{8.667_{\pm 0.880}}$	$\mathbf{14.143_{\pm 1.196}}$
JH-PAFM	0.1	$6.124_{\pm 0.640}$	$9.408_{\pm 0.531}$	$15.069_{\pm 0.768}$
JH-PAFM	0.3	$5.896_{\pm 0.554}$	$9.691_{\pm 0.541}$	$15.120_{\pm 0.803}$
JH-PAFM	0.5	$6.969_{\pm 0.610}$	$10.610_{\pm 0.778}$	$16.202_{\pm 1.338}$
JH-PAFM	0.7	$7.013_{\pm 0.465}$	$10.778_{\pm 0.684}$	$16.817_{\pm 1.135}$
JH-PAFM	1.0	$7.751_{\pm 0.981}$	$11.817_{\pm 0.632}$	$17.891_{\pm 0.492}$

Table 8: Control-IQ empirical calibration error (ECE) for AR-PAFM and JH-PAFM across regularization weights α. Lower is better. Entries are means $\pm$ SD over four training seeds. **Bold** and underlined values indicate the best and second-best mean ECE within each model, respectively.

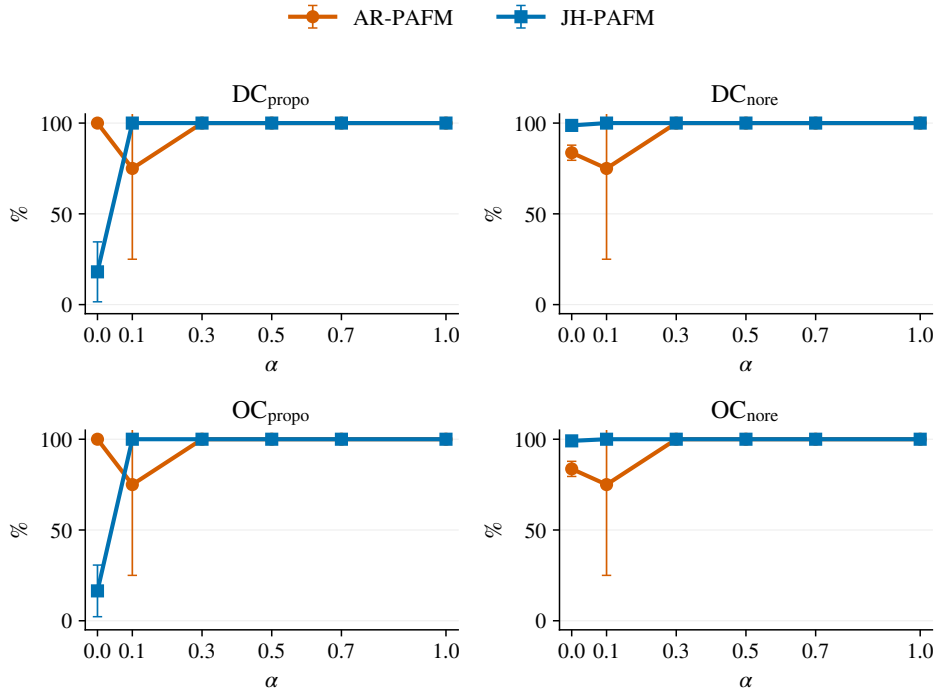
Model	α	ECE
AR-PAFM	0.0	$0.141_{\pm 0.025}$
AR-PAFM	0.1	$0.172_{\pm 0.049}$
AR-PAFM	0.3	$0.175_{\pm 0.059}$
AR-PAFM	0.5	$0.205_{\pm 0.072}$
AR-PAFM	0.7	$\underline{0.136}_{\pm 0.019}$
AR-PAFM	1.0	$\mathbf{0.130}_{\pm \mathbf{0.013}}$
JH-PAFM	0.0	$0.151_{\pm 0.065}$
JH-PAFM	0.1	$0.150_{\pm 0.027}$
JH-PAFM	0.3	$0.059_{\pm 0.020}$
JH-PAFM	0.5	$0.081_{\pm 0.039}$
JH-PAFM	0.7	$\underline{0.052}_{\pm 0.031}$
JH-PAFM	1.0	$\mathbf{0.031}_{\pm \mathbf{0.018}}$

Table 9: Control-IQ dose-normalized peak effects, peak times, and areas under the effect curve (AUECs) across regularization weights α. Insulin and carbohydrate peak effects are in $\mathrm{mg}/(\mathrm{dL}\cdot\mathrm{U})$ and $\mathrm{mg}/(\mathrm{dL}\cdot\mathrm{g})$, respectively. Peak times are in minutes; the corresponding AUECs are in $\mathrm{mg}\cdot\mathrm{min}/(\mathrm{dL}\cdot\mathrm{U})$ and $\mathrm{mg}\cdot\mathrm{min}/(\mathrm{dL}\cdot\mathrm{g})$. Entries are means $\pm$ SD over four training seeds.

		Insulin			Carbohydrate		
Model	α	$\overline{v}^{\mathrm{peak}}_{\mathrm{ins}}$	$\overline{t}^{\mathrm{peak}}_{\mathrm{ins}}$	$\overline{\mathrm{AUEC}}_{\mathrm{ins}}$	$\overline{v}^{\mathrm{peak}}_{\mathrm{carb}}$	$\overline{t}^{\mathrm{peak}}_{\mathrm{carb}}$	$\overline{\mathrm{AUEC}}_{\mathrm{carb}}$
BNODE	0.0	$1.898_{\pm 0.878}$	$103.988_{\pm 1.224}$	$-1170.829_{\pm 201.062}$	$0.319_{\pm 0.033}$	$50.937_{\pm 16.324}$	$15.807_{\pm 4.537}$
BNODE	0.1	$10.370_{\pm 2.702}$	$68.803_{\pm 9.146}$	$733.517_{\pm 171.724}$	$0.896_{\pm 0.152}$	$68.911_{\pm 4.145}$	$67.269_{\pm 12.503}$
BNODE	0.3	$4.257_{\pm 1.467}$	$79.760_{\pm 23.470}$	$342.114_{\pm 118.076}$	$1.138_{\pm 0.043}$	$100.333_{\pm 3.213}$	$88.005_{\pm 6.576}$
BNODE	0.5	$4.044_{\pm 1.187}$	$87.040_{\pm 15.255}$	$300.318_{\pm 104.315}$	$1.273_{\pm 0.122}$	$105.590_{\pm 3.209}$	$92.936_{\pm 13.456}$
BNODE	0.7	$2.285_{\pm 1.184}$	$66.036_{\pm 31.496}$	$156.080_{\pm 103.361}$	$1.171_{\pm 0.138}$	$105.559_{\pm 8.007}$	$85.258_{\pm 7.717}$
Hovorka	0.0	$1.588_{\pm 0.545}$	$47.198_{\pm 6.704}$	$-810.739_{\pm 459.741}$	$0.795_{\pm 0.217}$	$70.455_{\pm 4.247}$	$53.651_{\pm 15.128}$
Hovorka	0.1	$0.632_{\pm 0.226}$	$97.043_{\pm 16.266}$	$13.710_{\pm 13.754}$	$1.016_{\pm 0.239}$	$104.263_{\pm 3.176}$	$67.581_{\pm 15.970}$
Hovorka	0.3	$1.039_{\pm 0.511}$	$111.272_{\pm 15.216}$	$28.242_{\pm 17.962}$	$1.555_{\pm 0.292}$	$105.382_{\pm 3.501}$	$100.104_{\pm 8.849}$
Hovorka	0.5	$0.688_{\pm 0.379}$	$94.030_{\pm 18.899}$	$15.960_{\pm 11.615}$	$1.438_{\pm 0.474}$	$107.831_{\pm 4.558}$	$91.921_{\pm 18.613}$
Hovorka	0.7	$0.694_{\pm 0.325}$	$111.222_{\pm 12.896}$	$19.306_{\pm 13.697}$	$1.409_{\pm 0.573}$	$110.723_{\pm 4.177}$	$84.740_{\pm 24.822}$
LP-Reduced	0.0	$-0.194_{\pm 1.227}$	$48.000_{\pm 42.389}$	$-738.508_{\pm 622.366}$	$0.947_{\pm 0.213}$	$89.889_{\pm 23.436}$	$70.343_{\pm 12.812}$
LP-Reduced	0.1	$0.246_{\pm 0.121}$	$83.410_{\pm 41.529}$	$14.452_{\pm 9.845}$	$1.368_{\pm 0.385}$	$112.241_{\pm 4.586}$	$93.673_{\pm 21.849}$
LP-Reduced	0.3	$0.535_{\pm 0.531}$	$116.212_{\pm 3.382}$	$35.518_{\pm 37.458}$	$1.209_{\pm 0.073}$	$113.471_{\pm 5.172}$	$85.529_{\pm 3.783}$
LP-Reduced	0.5	$1.503_{\pm 0.782}$	$119.774_{\pm 0.451}$	$82.889_{\pm 44.580}$	$1.091_{\pm 0.204}$	$108.788_{\pm 12.553}$	$79.812_{\pm 12.543}$
LP-Reduced	0.7	$0.622_{\pm 0.673}$	$105.453_{\pm 19.071}$	$35.211_{\pm 36.573}$	$1.151_{\pm 0.120}$	$111.167_{\pm 10.653}$	$80.339_{\pm 8.259}$
LSTM	0.0	$-2.001_{\pm 2.737}$	$106.123_{\pm 1.805}$	$-2027.512_{\pm 320.911}$	$0.380_{\pm 0.026}$	$63.276_{\pm 3.128}$	$20.341_{\pm 2.114}$
LSTM	0.1	$9.985_{\pm 3.541}$	$84.218_{\pm 2.264}$	$678.013_{\pm 195.378}$	$0.995_{\pm 0.046}$	$66.561_{\pm 3.552}$	$74.066_{\pm 7.509}$
LSTM	0.3	$2.677_{\pm 0.961}$	$87.364_{\pm 6.856}$	$188.505_{\pm 82.531}$	$0.928_{\pm 0.093}$	$84.380_{\pm 8.545}$	$71.525_{\pm 6.651}$
LSTM	0.5	$1.798_{\pm 0.290}$	$72.750_{\pm 9.477}$	$132.580_{\pm 20.864}$	$0.972_{\pm 0.159}$	$89.213_{\pm 4.448}$	$73.796_{\pm 13.243}$
LSTM	0.7	$1.555_{\pm 0.636}$	$67.423_{\pm 4.863}$	$118.936_{\pm 49.048}$	$0.824_{\pm 0.112}$	$87.731_{\pm 3.768}$	$61.568_{\pm 10.236}$
S4D	0.0	$-0.293_{\pm 0.719}$	$107.863_{\pm 2.849}$	$-1494.697_{\pm 92.108}$	$0.231_{\pm 0.020}$	$52.061_{\pm 9.718}$	$9.783_{\pm 1.972}$
S4D	0.1	$14.449_{\pm 0.655}$	$100.407_{\pm 1.720}$	$129.243_{\pm 65.146}$	$0.624_{\pm 0.040}$	$78.374_{\pm 4.762}$	$40.584_{\pm 4.410}$
S4D	0.3	$9.121_{\pm 2.614}$	$107.461_{\pm 2.031}$	$-50.875_{\pm 69.558}$	$0.675_{\pm 0.065}$	$88.564_{\pm 9.444}$	$48.323_{\pm 5.093}$
S4D	0.5	$6.026_{\pm 3.704}$	$102.907_{\pm 1.737}$	$-64.656_{\pm 133.473}$	$0.697_{\pm 0.069}$	$88.390_{\pm 10.787}$	$49.887_{\pm 5.233}$
S4D	0.7	$2.802_{\pm 0.946}$	$97.644_{\pm 8.356}$	$-76.305_{\pm 23.725}$	$0.614_{\pm 0.066}$	$85.915_{\pm 8.882}$	$44.757_{\pm 4.839}$
TCN	0.0	$5.151_{\pm 0.698}$	$111.868_{\pm 1.436}$	$-352.701_{\pm 36.439}$	$0.055_{\pm 0.013}$	$98.586_{\pm 6.153}$	$-4.403_{\pm 2.322}$
TCN	0.1	$21.027_{\pm 1.607}$	$113.579_{\pm 1.623}$	$1372.397_{\pm 155.237}$	$0.513_{\pm 0.066}$	$91.878_{\pm 7.233}$	$43.686_{\pm 4.875}$
TCN	0.3	$14.243_{\pm 1.160}$	$113.706_{\pm 1.552}$	$763.567_{\pm 67.234}$	$0.501_{\pm 0.062}$	$86.620_{\pm 9.376}$	$42.794_{\pm 5.799}$
TCN	0.5	$14.024_{\pm 1.018}$	$110.983_{\pm 4.008}$	$877.862_{\pm 197.888}$	$0.558_{\pm 0.026}$	$94.049_{\pm 10.252}$	$48.032_{\pm 3.364}$
TCN	0.7	$10.434_{\pm 2.007}$	$98.279_{\pm 17.544}$	$705.066_{\pm 219.196}$	$0.643_{\pm 0.133}$	$88.798_{\pm 2.851}$	$49.906_{\pm 10.921}$
AR-PAFM	0.0	$-3.037_{\pm 0.469}$	$24.319_{\pm 2.713}$	$-911.467_{\pm 56.425}$	$0.726_{\pm 0.023}$	$81.310_{\pm 5.463}$	$56.996_{\pm 2.221}$
AR-PAFM	0.1	$7.182_{\pm 3.700}$	$117.992_{\pm 0.718}$	$382.188_{\pm 204.688}$	$1.170_{\pm 0.080}$	$113.020_{\pm 0.649}$	$79.872_{\pm 5.355}$
AR-PAFM	0.3	$5.906_{\pm 1.196}$	$118.630_{\pm 0.216}$	$316.609_{\pm 65.352}$	$1.045_{\pm 0.080}$	$110.865_{\pm 3.213}$	$75.421_{\pm 4.408}$
AR-PAFM	0.5	$7.129_{\pm 3.467}$	$119.188_{\pm 0.521}$	$381.530_{\pm 188.265}$	$1.096_{\pm 0.143}$	$112.778_{\pm 2.535}$	$77.924_{\pm 6.999}$
AR-PAFM	0.7	$10.497_{\pm 2.307}$	$118.752_{\pm 0.809}$	$569.513_{\pm 123.413}$	$1.216_{\pm 0.152}$	$114.990_{\pm 0.971}$	$83.751_{\pm 6.129}$
AR-PAFM	1.0	$7.661_{\pm 0.681}$	$118.620_{\pm 1.028}$	$414.435_{\pm 29.076}$	$1.099_{\pm 0.160}$	$111.683_{\pm 4.526}$	$77.541_{\pm 6.873}$
JH-PAFM	0.0	$1.761_{\pm 0.243}$	$95.856_{\pm 1.741}$	$-91.378_{\pm 18.056}$	$1.593_{\pm 0.043}$	$90.050_{\pm 2.518}$	$109.437_{\pm 1.658}$
JH-PAFM	0.1	$7.337_{\pm 0.638}$	$117.261_{\pm 0.436}$	$375.203_{\pm 41.746}$	$1.490_{\pm 0.058}$	$99.370_{\pm 1.766}$	$105.827_{\pm 3.850}$
JH-PAFM	0.3	$7.804_{\pm 1.633}$	$118.159_{\pm 0.614}$	$396.130_{\pm 90.600}$	$1.446_{\pm 0.072}$	$99.097_{\pm 1.769}$	$105.378_{\pm 5.116}$
JH-PAFM	0.5	$9.178_{\pm 2.744}$	$118.774_{\pm 0.388}$	$482.348_{\pm 180.137}$	$1.471_{\pm 0.094}$	$98.672_{\pm 0.910}$	$108.648_{\pm 6.572}$
JH-PAFM	0.7	$9.805_{\pm 3.261}$	$118.744_{\pm 0.509}$	$520.429_{\pm 182.466}$	$1.390_{\pm 0.111}$	$95.821_{\pm 1.798}$	$104.421_{\pm 9.520}$
JH-PAFM	1.0	$10.823_{\pm 2.525}$	$118.870_{\pm 0.252}$	$569.313_{\pm 140.593}$	$1.441_{\pm 0.110}$	$95.696_{\pm 2.027}$	$108.228_{\pm 8.103}$

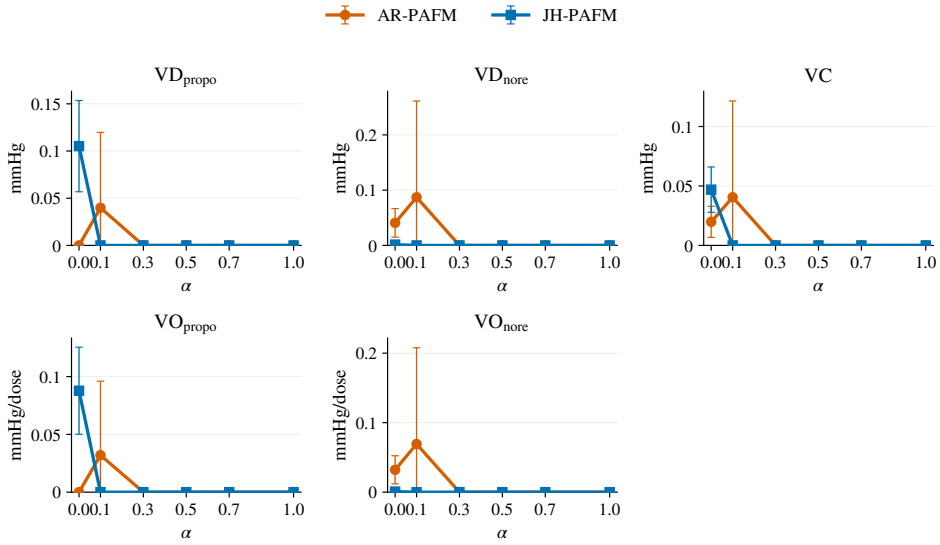


(a)

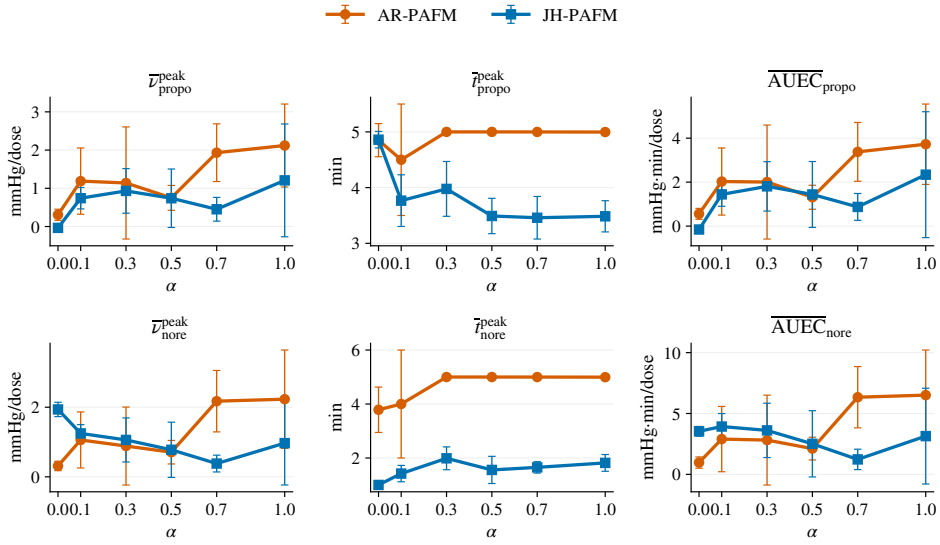


(b)

Figure 1: PAS forecasting performance and physiological consistency for AR-PAFM and JH-PAFM across regularization weights α: (a) empirical calibration error and RMSE at 1, 2.5, and 5 minutes; (b) directional and dose-ordering consistency for both treatments. RMSE is in mmHg; consistency scores are percentages. Points show means and error bars indicate ± 1 sample SD across four training seeds.



(a)



(b)

Figure 2: PAS treatment-response analysis for AR-PAFM and JH-PAFM across regularization weights α: (a) directional, ordering, and complete-chain violation magnitudes; (b) dose-normalized peak effects, peak times, and AUECs. Response windows are 3–5 minutes for propofol and 1–5 minutes for norepinephrine. Points show means and error bars indicate ± 1 sample SD across four training seeds.

Table 10: Effect of sensitivity bounds on PAFM response magnitudes and forecasting accuracy on Control-IQ. Bounds are varied separately for each channel from each model's reference configuration. Response units follow Table 9; RMSE is in mg/dL. Entries are means ± SD over training seeds. **Bold** and underlined values indicate the best and second-best mean RMSEs within each model and forecast horizon, respectively.

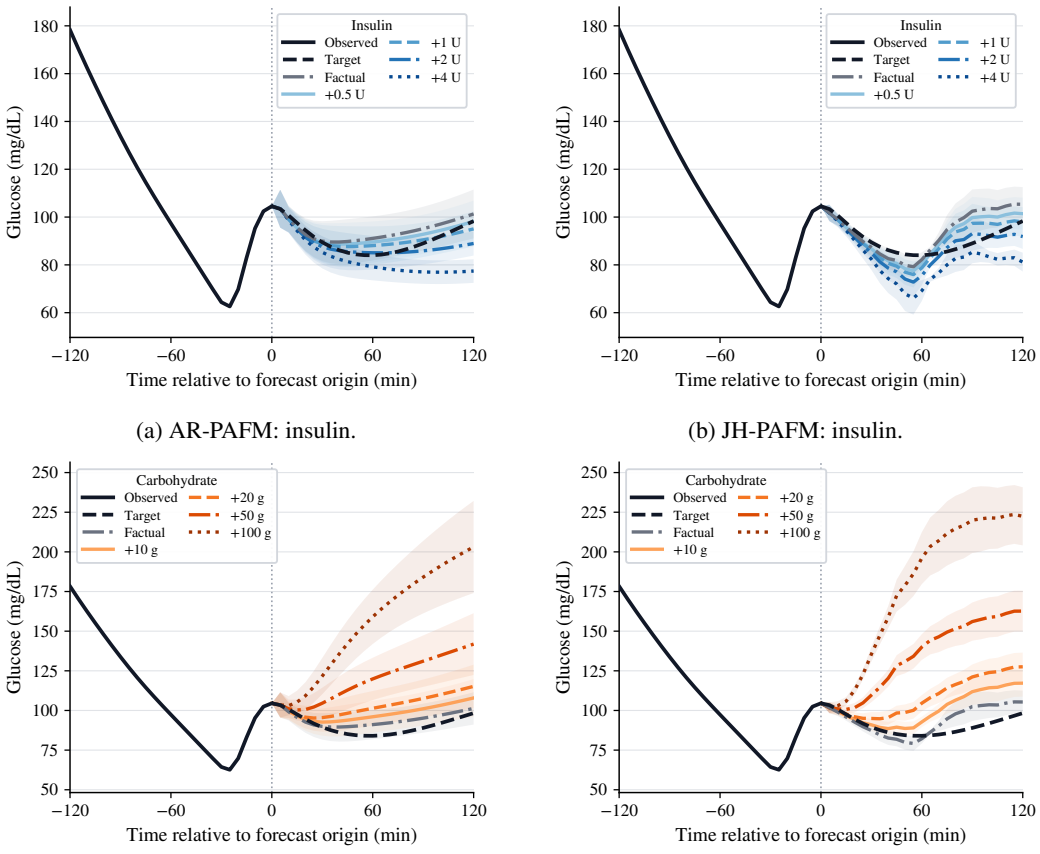
			Insulin response		Carbohydrate response		Forecasting RMSE		
Setting	Insulin bounds	Carbohydrate bounds	$\overline{v}^{\text{peak}}_{\text{ins}}$	$\overline{\text{AUEC}}_{\text{ins}}$	$\overline{v}^{\text{peak}}_{\text{carb}}$	$\overline{\text{AUEC}}_{\text{carb}}$	30 min	60 min	120 min
AR-PAFM									
Reference	$[-60, -30]$	$[5, 80]$	$4.613_{\pm 0.697}$	$242.869_{\pm 40.024}$	$1.094_{\pm 0.043}$	$76.033_{\pm 2.025}$	$10.679_{\pm 0.983}$	$15.184_{\pm 0.635}$	$22.610_{\pm 0.301}$
Insulin near-zero	$[-6, -3]$	$[5, 80]$	$1.304_{\pm 0.513}$	$71.819_{\pm 29.465}$	$1.177_{\pm 0.055}$	$81.008_{\pm 2.918}$	$\mathbf{9.990_{\pm 0.843}}$	$\underline{14.647}_{\pm 0.865}$	$\underline{21.602}_{\pm 0.715}$
Insulin high	$[-120, -60]$	$[5, 80]$	$9.574_{\pm 1.430}$	$495.924_{\pm 77.479}$	$1.178_{\pm 0.100}$	$81.180_{\pm 4.208}$	$10.659_{\pm 0.993}$	$15.331_{\pm 0.544}$	$23.124_{\pm 0.847}$
Carbohydrate near-zero	$[-60, -30]$	$[0.5, 8]$	$5.592_{\pm 2.443}$	$289.834_{\pm 130.610}$	$0.925_{\pm 0.105}$	$60.329_{\pm 4.766}$	$\underline{10.070}_{\pm 1.305}$	$15.408_{\pm 1.625}$	$23.357_{\pm 2.265}$
Carbohydrate high	$[-60, -30]$	$[80, 160]$	$8.654_{\pm 1.507}$	$498.454_{\pm 100.500}$	$1.228_{\pm 0.085}$	$85.867_{\pm 8.013}$	$11.104_{\pm 0.968}$	$15.437_{\pm 0.900}$	$24.334_{\pm 1.461}$
JH-PAFM									
Reference	$[-200, -100]$	$[20, 220]$	$6.517_{\pm 0.811}$	$321.678_{\pm 48.010}$	$1.626_{\pm 0.050}$	$113.612_{\pm 3.384}$	$5.493_{\pm 0.287}$	$8.907_{\pm 0.331}$	$14.582_{\pm 0.489}$
Insulin near-zero	$[-20, -10]$	$[20, 220]$	$1.740_{\pm 0.509}$	$88.001_{\pm 29.790}$	$1.372_{\pm 0.080}$	$98.661_{\pm 3.204}$	$\underline{5.378}_{\pm 0.207}$	$\underline{8.649}_{\pm 0.377}$	$\underline{14.271}_{\pm 0.480}$
Insulin high	$[-400, -200]$	$[20, 220]$	$9.438_{\pm 2.552}$	$519.416_{\pm 156.004}$	$1.517_{\pm 0.056}$	$110.383_{\pm 2.490}$	$6.547_{\pm 0.785}$	$10.049_{\pm 0.727}$	$15.544_{\pm 0.482}$
Carbohydrate near-zero	$[-200, -100]$	$[2, 22]$	$4.553_{\pm 0.476}$	$229.819_{\pm 29.854}$	$0.908_{\pm 0.060}$	$61.267_{\pm 3.057}$	$\mathbf{5.262_{\pm 0.090}}$	$\mathbf{8.583_{\pm 0.152}}$	$\mathbf{14.062_{\pm 0.272}}$
Carbohydrate high	$[-200, -100]$	$[220, 440]$	$3.908_{\pm 1.809}$	$192.594_{\pm 95.291}$	$1.450_{\pm 0.071}$	$107.404_{\pm 4.738}$	$6.271_{\pm 0.735}$	$9.497_{\pm 0.726}$	$14.860_{\pm 1.091}$

Table 11: Effect of fidelity weight λ_{fid} and sampling ratio ρ_{fid} on AR-PAFM response magnitudes and forecasting accuracy on Control-IQ. The weight sweep fixes $\rho_{\text{fid}} = 0.1$; the ratio sweep fixes $\lambda_{\text{fid}} = 1$. Response units follow Table 9; RMSE is in mg/dL. Entries are means ± SD over training seeds. **Bold** and underlined values indicate the best and second-best mean RMSEs within each sweep and forecast horizon, respectively.

Setting	$\overline{v}^{\text{peak}}_{\text{ins}}$	$\overline{\text{AUEC}}_{\text{ins}}$	$\overline{v}^{\text{peak}}_{\text{carb}}$	$\overline{\text{AUEC}}_{\text{carb}}$	$\text{RMSE}_{30\text{m}}$	$\text{RMSE}_{60\text{m}}$	$\text{RMSE}_{120\text{m}}$
Fidelity weight ($\rho_{\text{fid}} = 0.1$)							
$\lambda_{\text{fid}} = 0$	$23.556_{\pm 26.508}$	$1071.443_{\pm 1142.133}$	$1.603_{\pm 0.610}$	$82.663_{\pm 29.728}$	$21.431_{\pm 6.167}$	$32.118_{\pm 7.981}$	$52.086_{\pm 7.844}$
$\lambda_{\text{fid}} = 0.5$	$6.967_{\pm 2.382}$	$364.447_{\pm 132.485}$	$1.109_{\pm 0.060}$	$79.108_{\pm 4.066}$	$\underline{10.307}_{\pm 2.145}$	$\mathbf{14.242_{\pm 1.466}}$	$21.870_{\pm 1.384}$
$\lambda_{\text{fid}} = 1$	$7.674_{\pm 2.565}$	$412.047_{\pm 142.446}$	$1.173_{\pm 0.059}$	$81.997_{\pm 2.360}$	$\mathbf{9.836_{\pm 0.887}}$	$\underline{14.257}_{\pm 0.452}$	$\mathbf{21.724_{\pm 1.003}}$
Fidelity ratio ($\lambda_{\text{fid}} = 1$)							
$\rho_{\text{fid}} = 0$	$12.294_{\pm 6.997}$	$630.540_{\pm 392.234}$	$1.313_{\pm 0.089}$	$83.852_{\pm 2.154}$	$13.959_{\pm 0.417}$	$19.974_{\pm 1.852}$	$31.060_{\pm 4.435}$
$\rho_{\text{fid}} = 0.1$	$7.674_{\pm 2.565}$	$412.047_{\pm 142.446}$	$1.173_{\pm 0.059}$	$81.997_{\pm 2.360}$	$9.836_{\pm 0.887}$	$\underline{14.257}_{\pm 0.452}$	$\underline{21.724}_{\pm 1.003}$
$\rho_{\text{fid}} = 1$	$4.919_{\pm 2.592}$	$255.060_{\pm 134.518}$	$1.125_{\pm 0.081}$	$77.718_{\pm 4.806}$	$\mathbf{7.860_{\pm 0.798}}$	$\mathbf{12.155_{\pm 1.023}}$	$\mathbf{18.930_{\pm 1.233}}$

Table 12: Evaluation settings for Control-IQ and PAS. Directional, ordering, and complete-chain thresholds are zero. Control-IQ doses $\delta_{\tilde{a},i}$, the i-th dose level of treatment $\tilde{a}$ are absolute; PAS doses are multiples of each patient's maintenance infusion rate (Appendix C.1).

Setting	Control-IQ	PAS
Decreasing treatment	Insulin	Propofol
Increasing treatment	Carbohydrate	Norepinephrine
Dose levels, decreasing	$\{0.5, 1, 2, 4\}$ U	$\{0.5, 1, 1.5, 2\} \times \mathsf{rate}^{\text{propo,base}}_{i^\dagger}$
Dose levels, increasing	$\{10, 20, 50, 100\}$ g	$\{0.5, 1, 1.5, 2\} \times \mathsf{rate}^{\text{nore,base}}_{i^\dagger}$
Forecast step	5 min	10 s
Decreasing-treatment window	$[20, 120]$ min	$[3, 5]$ min
Increasing-treatment window	$[10, 120]$ min	$[1, 5]$ min
Outcome support $(y_{\min}, y_{\max})$	$(40, 400)$ mg/dL	$(30, 200)$ mmHg
Support buffer $\varrho_{\tilde{a}}$	10 mg/dL	1 mmHg
RMSE lead times	$30/60/120$ min	$1/2.5/5$ min
Held-out patients per seed	10	10
Seeds	$\{1, 10, 42, 50\}$	$\{1, 10, 42, 50\}$
Monte Carlo samples N_{MC}	64	64



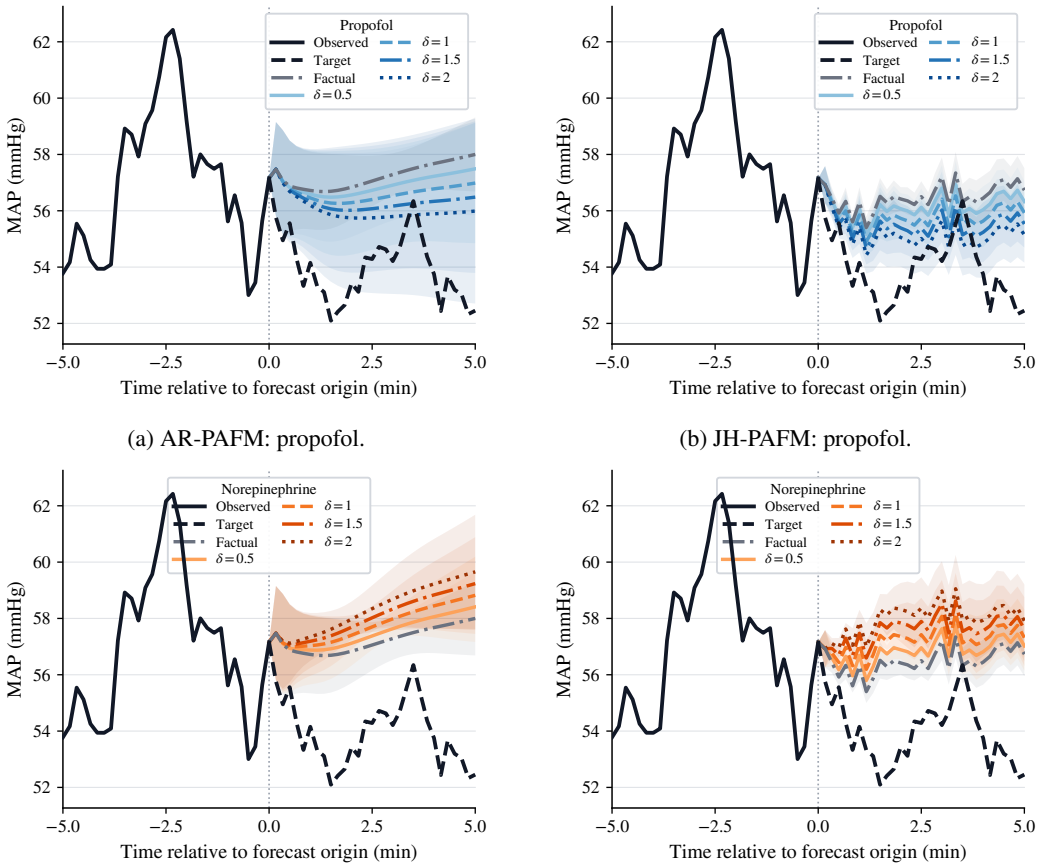
(a) AR-PAFM: insulin.

(b) JH-PAFM: insulin.

(c) AR-PAFM: carbohydrate.

(d) JH-PAFM: carbohydrate.

Figure 3: Example 120-minute Control-IQ forecasts at $\alpha = 0.1$ under insulin (a, b) and carbohydrate (c, d) perturbations. Panels show the observed history and future target, factual forecasts, and forecasts at the indicated additional dose levels. Forecast curves show means over four training seeds; shading indicates ± 1 SD across seeds.



(a) AR-PAFM: propofol.

(b) JH-PAFM: propofol.

(c) AR-PAFM: norepinephrine.

(d) JH-PAFM: norepinephrine.

Figure 4: Example 5-minute PAS forecasts at $\alpha = 0.3$ under propofol (a, b) and norepinephrine (c, d) perturbations. Panels show the observed history and future target, factual forecasts, and forecasts at the indicated additional dose levels. Forecast curves show means over four training seeds; shading indicates ± 1 SD across seeds.

Table 13: Control-IQ architecture search spaces and configurations selected using validation MSE for PAFM and the baseline models.

Model	Hyperparameter	Search space	Selected
BNODE	Latent dimension	$\{32, 64, 128\}$	128
	MLP width	$\{32, 64, 128\}$	32
	MLP hidden layers	$\{1, 2, 3\}$	3
	ODE layers	$\{1, 2\}$	2
Hovorka	Encoder dimension	$\{32, 64, 128\}$	128
	MLP width	$\{32, 64, 128\}$	32
	MLP hidden layers	$\{1, 2, 3\}$	2
	LSTM layers	$\{1, 2\}$	2
LP-Reduced	Latent dimension	$\{32, 64, 128\}$	128
	MLP width	$\{32, 64, 128\}$	32
	MLP hidden layers	$\{1, 2, 3\}$	3
	ODE layers	$\{1, 2\}$	1
LSTM	Latent dimension	$\{32, 64, 128\}$	128
	Recurrent layers	$\{1, 2, 3\}$	3
S4D	Model dimension	$\{32, 64, 128\}$	128
	State dimension	$\{32, 64, 128\}$	32
	S4D layers	$\{1, 2\}$	2
TCN	Channel width	$\{32, 64, 128\}$	128
	Kernel size	$\{2, 3, 5\}$	5
	Number of levels	$\{2, 3, 4\}$	4
AR-PAFM	Flow hidden dimension	$\{32, 64, 128\}$	128
	Model dimension	$\{32, 64, 128\}$	128
	Encoder layers	$\{1, 2, 3\}$	3
	Flow layers	$\{1, 2, 3\}$	1
JH-PAFM	Flow hidden dimension	$\{32, 64, 128\}$	64
	Model dimension	$\{32, 64, 128\}$	128
	Encoder layers	$\{1, 2, 3\}$	3

Table 14: PAS architecture search spaces and configurations selected using validation MSE for PAFM.

Model	Hyperparameter	Candidate values	Selected
AR-PAFM	Flow hidden dimension	$\{32, 64, 128\}$	128
	Model dimension	$\{32, 64, 128\}$	128
	Encoder layers	$\{1, 2, 3\}$	3
	Flow layers	$\{1, 2, 3\}$	1
JH-PAFM	Flow hidden dimension	$\{32, 64, 128\}$	64
	Model dimension	$\{32, 64, 128\}$	128
	Encoder layers	$\{1, 2, 3\}$	3

Table 15: Training and model settings for Control-IQ and PAS. Paired PAFM entries are ordered AR-PAFM/JH-PAFM. Baseline settings apply only to Control-IQ.

Hyperparameter	Control-IQ	PAS
Optimization		
Optimizer	AdamW	AdamW
Learning rate	10^{-3}	10^{-3}
Weight decay	10^{-4}	10^{-4}
Batch size	512	512
Maximum epochs	150	150
Early-stopping patience	15	15
Training seeds	$\{1, 10, 42, 50\}$	$\{1, 10, 42, 50\}$
Data		
Data split	patient-level $80/10/10$	patient-level $80/10/10$
Split seed	42	42
History length L	288 steps (24 h)	600 steps (100 min)
Prediction horizon H	24 steps (120 min)	30 steps (5 min)
PAFM objective and flow		
Jacobian weight α	$\{0, 0.1, 0.3, 0.5, 0.7, 1\}$	$\{0, 0.1, 0.3, 0.5, 0.7, 1\}$
Fidelity weight λ_{fid}	$1/0$	$1/0$
Fidelity sampling ratio ρ_{fid}	$0.1/0$	$0.1/0$
Flow integration steps K_f	20	20
PAFM response transformation and bounds		
Decreasing-channel scale ς_c	1 (insulin)	1 (propofol)
Increasing-channel scale ς_c	25 (carbohydrate)	1 (norepinephrine)
Output range $(y_{\min}, y_{\max})$	$(40, 400)$ mg/dL	$(30, 200)$ mmHg
Decreasing-channel bounds $[b_c^-, b_c^+]$	$[-60, -30]/[-200, -100]$	$[-15, -2]/[-50, -5]$
Increasing-channel bounds $[b_c^-, b_c^+]$	$[5, 80]/[20, 220]$	$[2, 15]/[5, 50]$
Baseline ranking objective		
Ranking weight α	$\{0, 0.1, 0.3, 0.5, 0.7\}$	—
Inverse temperature $\varkappa$	10^4	—