

Safety Generalization Under Distribution Shift in Safe Reinforcement Learning

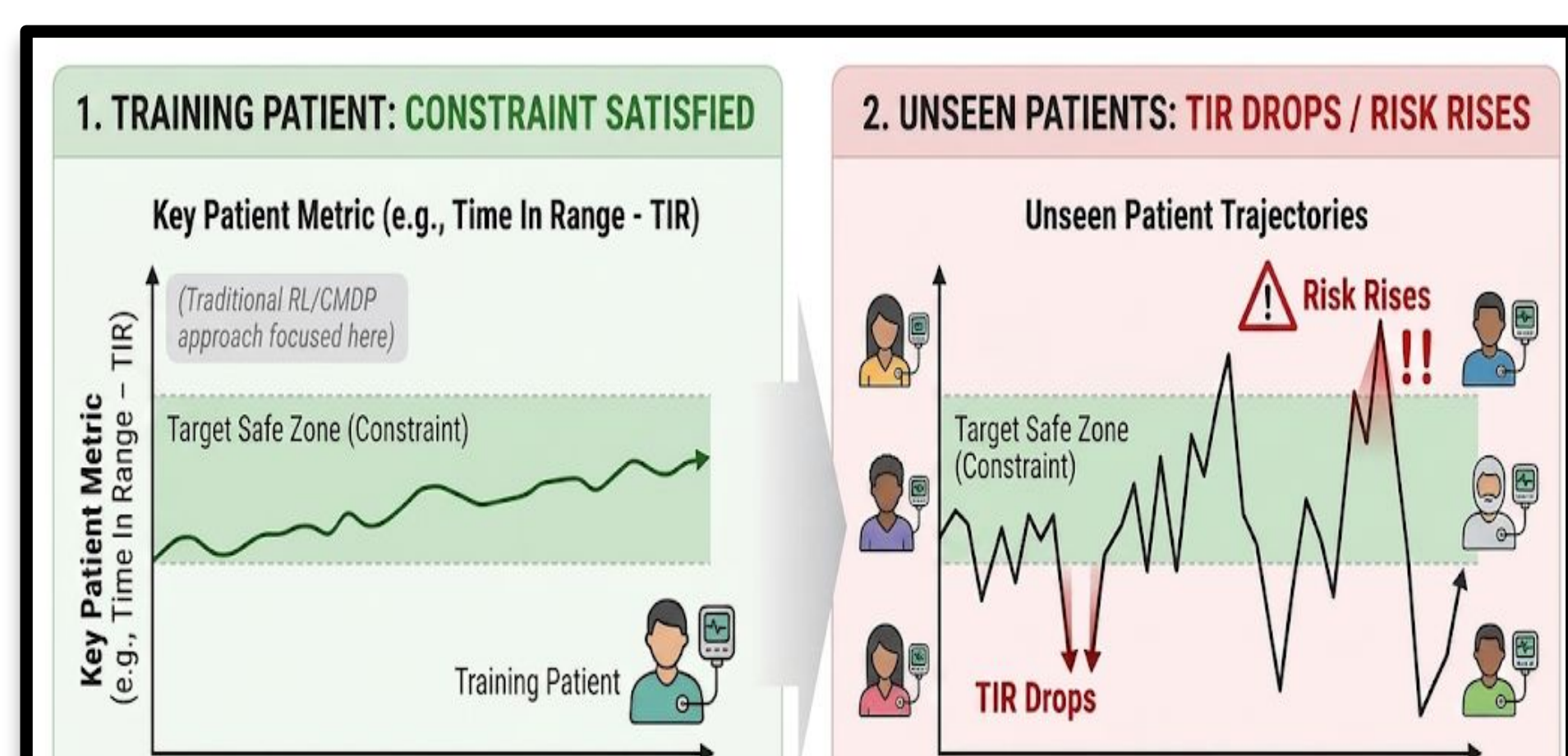


Minjae Kwon, Josephine Lamp, Lu Feng

Motivation

Safe RL optimizes a training-time constrained objective:

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_t \gamma^t R(s_t, a_t) \right] \quad \text{s.t.} \quad \mathbb{E}_{\pi} \left[\sum_t \gamma^t C(s_t, a_t) \right] \leq d$$



Does training-time safety transfer to test-time safety?

Unified Diabetes Simulator

Why Diabetes? Safety-critical sequential decision problem with patient-to-patient dynamics shift.

Regime	Child	Adol.	Adult
T1D	✓	✓	✓
T2D	✓	✓	✓
T2D no pump	✓	✓	✓

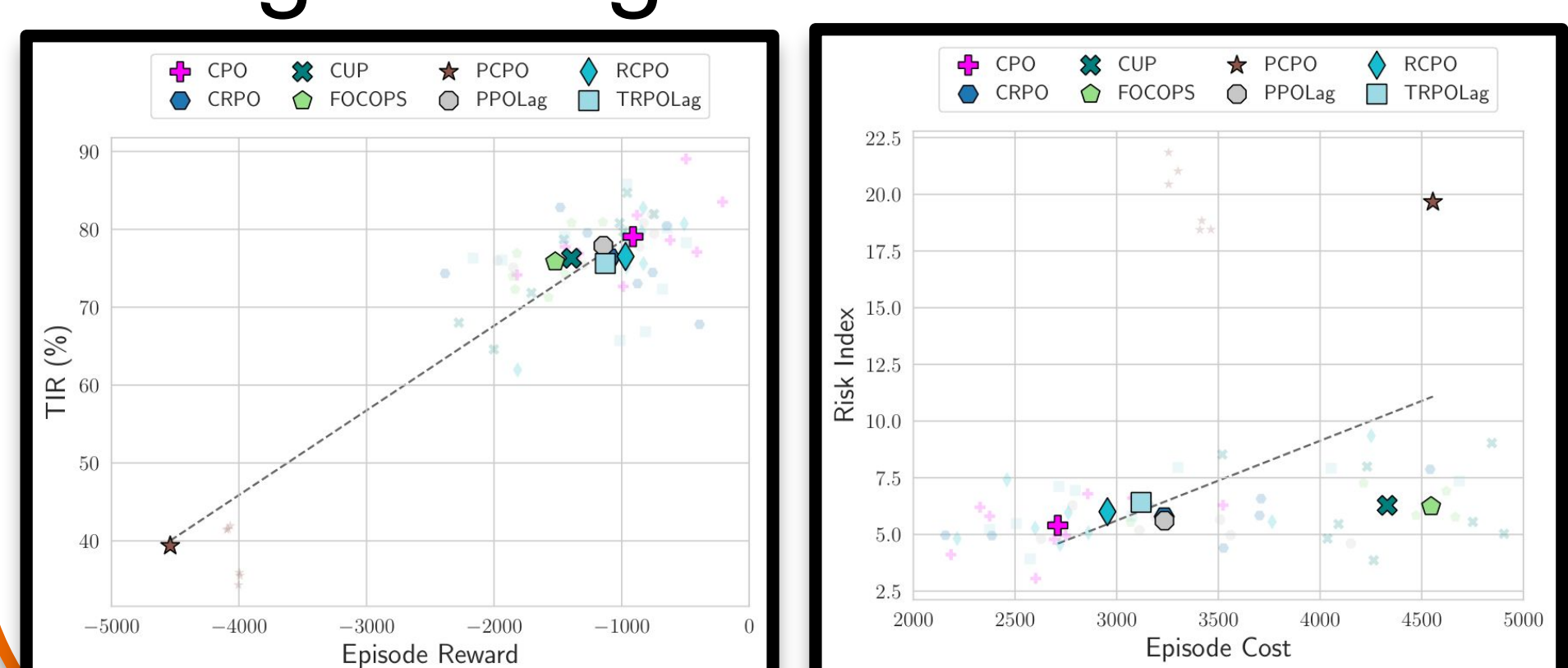
- Observation:** CGM, IOB, COB, Meal Schedule, Time, Trend.
- Action:** Bolus, and Meal Recommendation.
- Train:** Patient#01 per cohort.
- Test:** Unseen Patients#02-10. per cohort.

Clinical Metrics

- TIR** ↑ is the fraction of time blood glucose stays in [70, 180] mg/dL
- Risk Index** ↓ is the mean asymmetric glucose risk $r_t = 15.09(\ln(G)^{1.084} - 5.381)$
- CV** ↓ is glucose variability, $CV = \sigma/\mu$.

Reward & Cost Designs

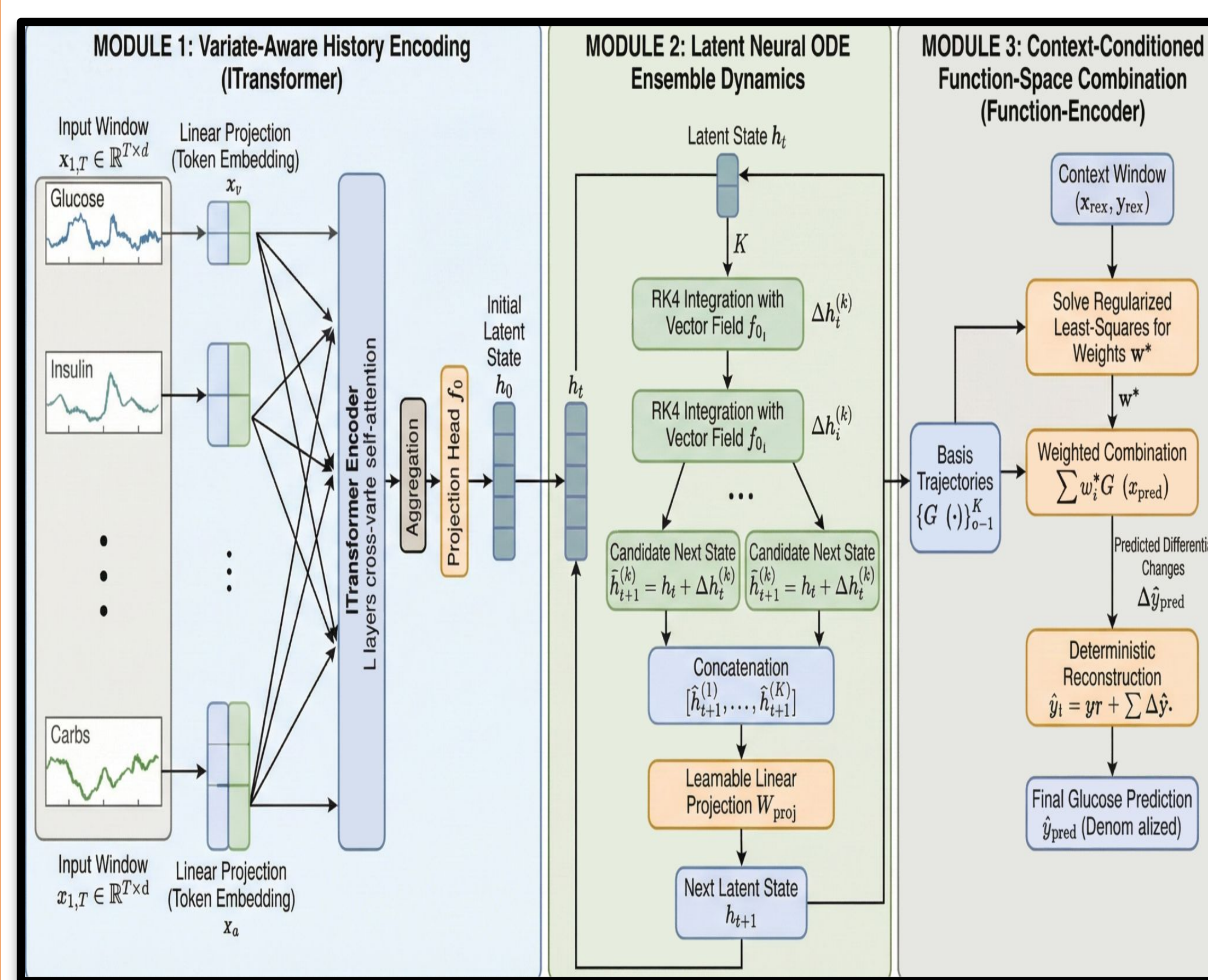
- Reward:** reduce predicted 2-hour clinical risk vs. no action; avoid unnecessary interventions.
- Cost:** penalize clinical risks, threshold violations, and dangerous glucose trends.



- Takeaway:** Unseen simulated patients reveal a safety generalization gap; predictive shielding improves TIR and Risk Index without retraining.

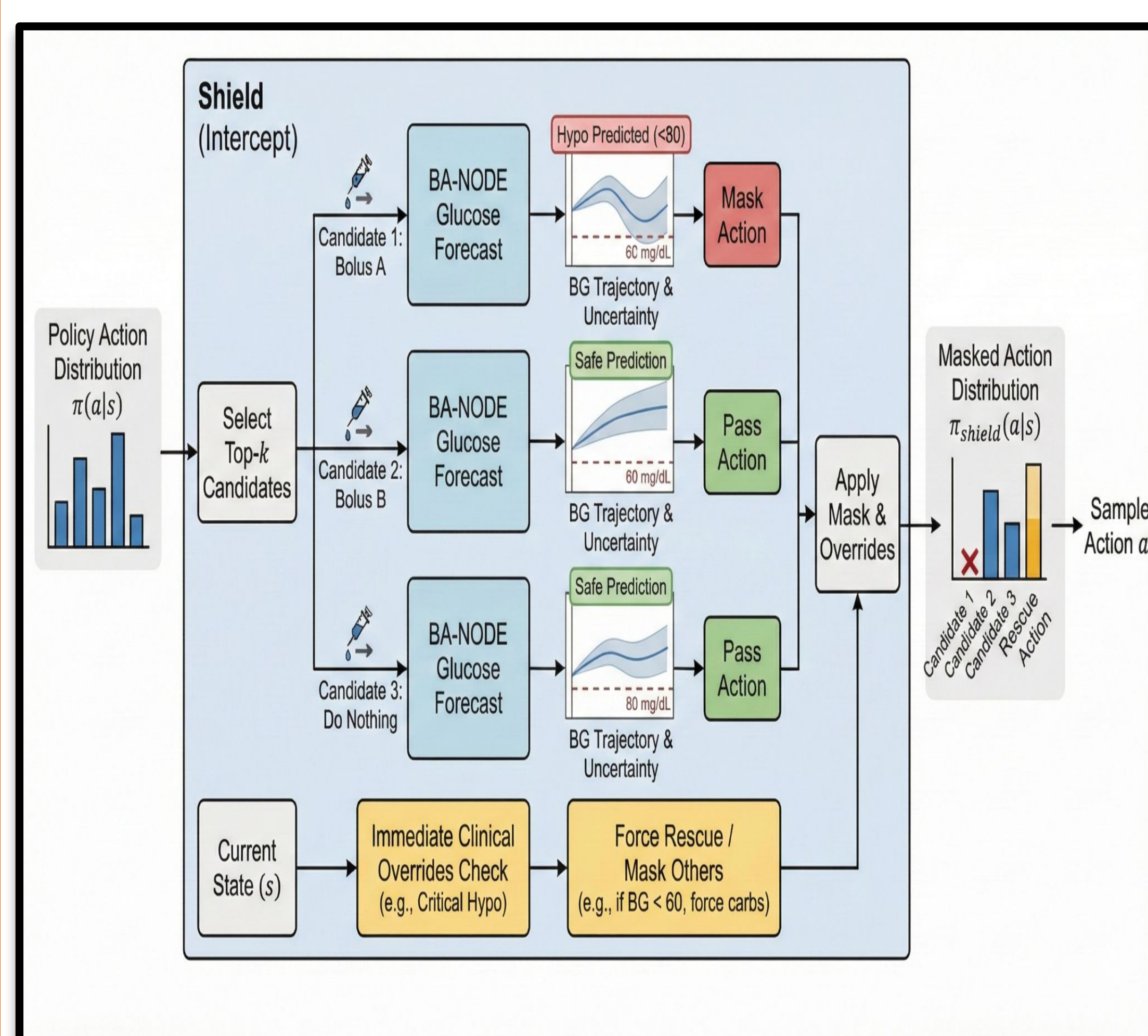
Methods

- BA-NODE:** Personalized blood glucose predictor using function space adaptation mechanism.



(i) Encode time series data with ITransformer, (ii) use latent Neural ODE basis flows, (ii) adapt basis weights for each patient.

- Predictive Shielding:** Runtime Shielding combined with BA-NODE

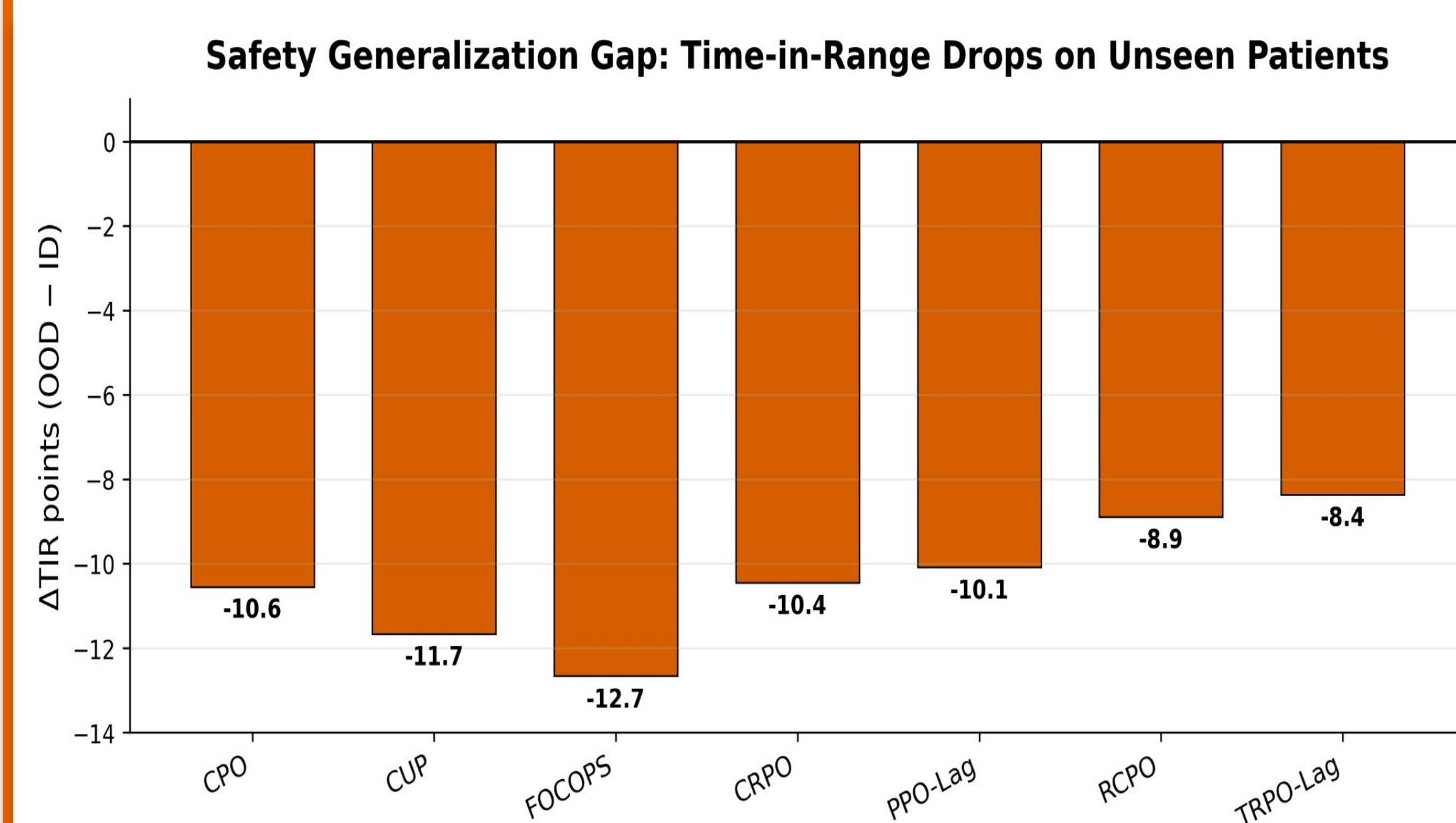
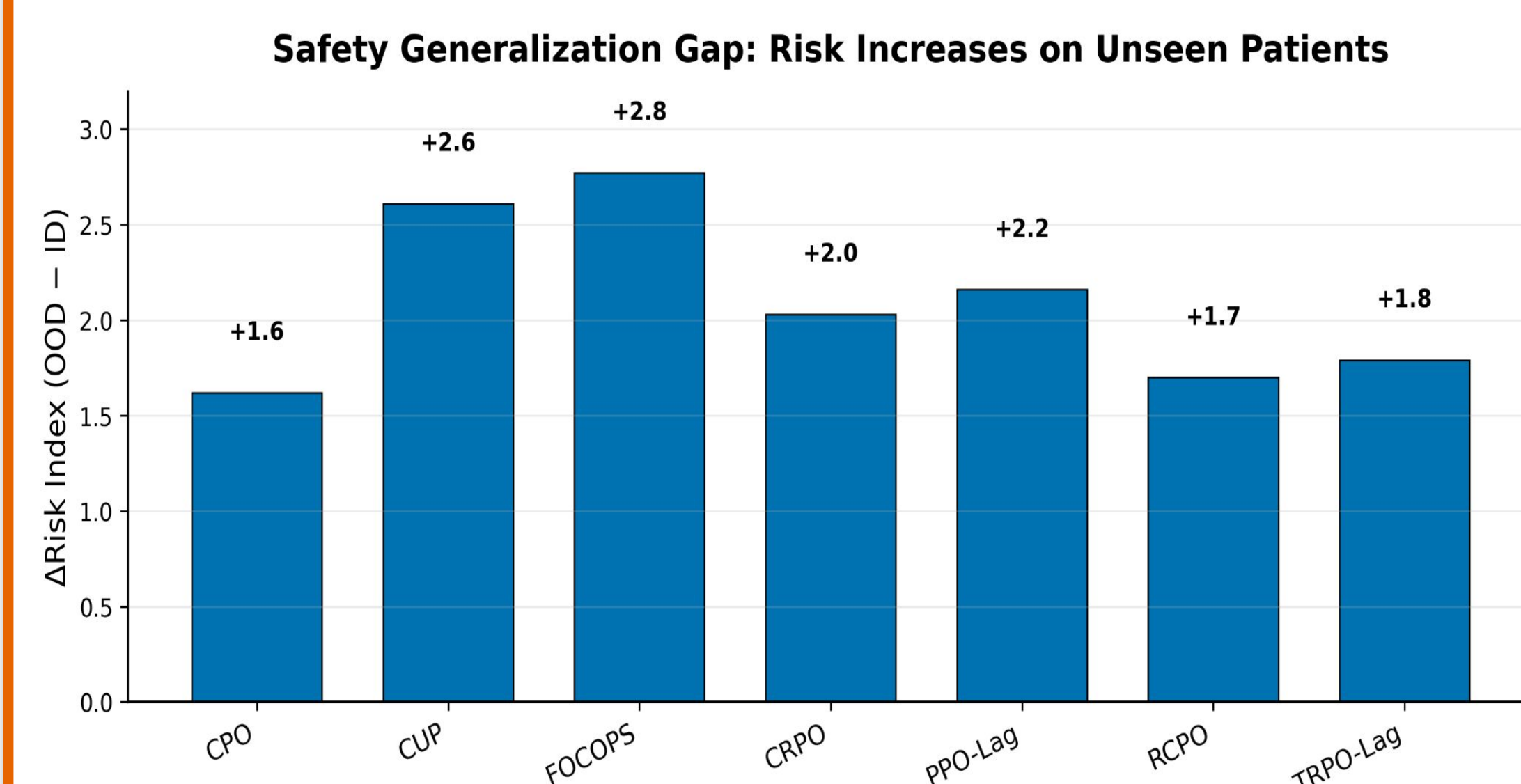


No retraining required. The shield wraps a trained policy at test time and filters risky actions before execution. In theory, better calibrated forecasts shows stronger finite-horizon safety check.

Experiments

We test (i) Parameter Generalization (Unseen Patients), (ii) Horizon Generalization (1 Day during training, 7 Days during evaluation).

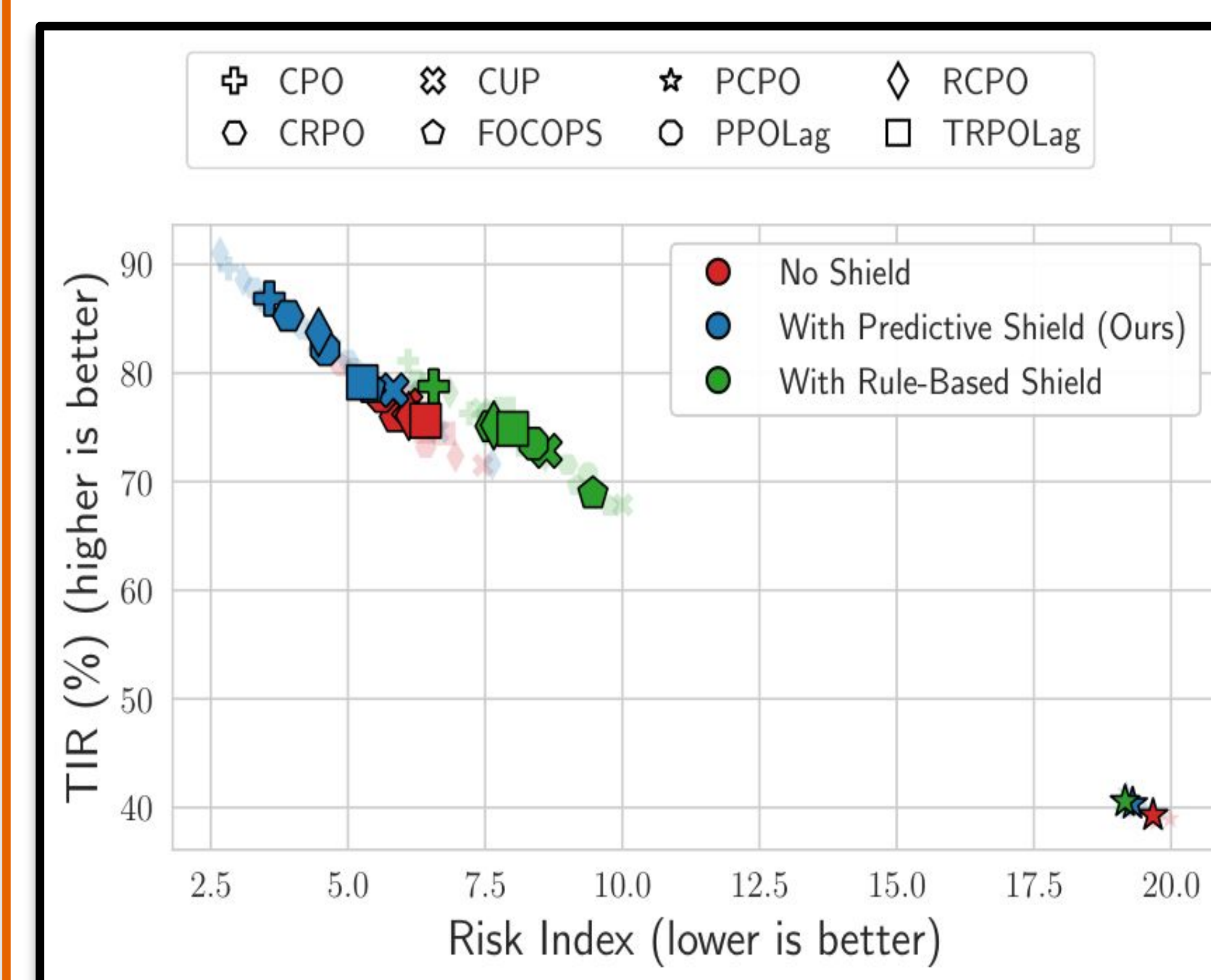
- Training time constrained satisfaction does not hold anymore, as we expected! TIR degrades more than 10%, Risk index increases as well.



- BA-NODE outperforms baselines regarding blood glucose prediction.

Model	MAE ↓	FDE ↓	RMSE ↓
ITransformer	4.11 ± 0.10	5.39 ± 0.16	7.25 ± 0.29
NODE	3.18 ± 0.60	4.11 ± 1.23	5.10 ± 1.45
BA-NODE	2.82 ± 0.01	3.63 ± 0.03	4.42 ± 0.02

- Predictive shielding mitigates the safety generalization gap without retraining.



Safe RL needs test-time safety checks.