SafeCF–SSM: Cognitive-Flexible Control with Explicit Physical Safety Guarantees for Latent-Space MPC

Thanana Nuchkrua Sudchai Boonto, *Senior Member, IEEE*

Department of Control Systems and Instrumentation Engineering
King Mongkut's University of Technology Thonburi

IEEE L-CSS

https://thanana.github.io/SafeCF-SSM.html

Outline

Motivation

Consider a safety-critical system under distributional shift:

Process Drift	Sensor Bias	Actuator Degradation
Model drifts as operating conditions change	Sensor develops a time-varying bias b_t	Rotor efficiency degrades gradually

The controller must **simultaneously**: detect the shift, update its model, and maintain safety — *without any action compromising the others.*

Gap 1 — Unregulated Adaptation

No existing method bounds latent reorganization rate in a way that is triggered by mismatch, coupled to safety, and provably non-destabilizing.

Gap 2 — Physical-Latent Gap

Latent-MPC enforces safety in z-space. No prior work provides an explicit certificate when decoder carries error $\varepsilon_{\text{dec}} > 0$.

SafeCF–SSM: Single Architectural Decision

Define the **surprise signal**:

$$\mathcal{S}_t := -\log p_{\theta_t}(o_{t+1} \mid z_t, u_t)$$

which is large when the latent model is inconsistent with observations.

$\mathcal{S}_t$ serves as a **shared input** to two coupled layers:

Adaptation Layer (Gap 1)

$$\eta_t \leq \frac{\eta_{\max}}{1 + \sqrt{\mathcal{S}_t}}$$

Large $\mathcal{S}_t$ **slows** adaptation, preventing destabilizing updates during high-mismatch transients.

Safety Layer (Gap 1–2)

$$\beta_{i,t} = \max\big(c_i \mathcal{S}_t,\ L_{g,i} L_d\, r_{\delta_i,t}\big)$$

$r_{\delta_i,t} = c_{\delta_i}\sqrt{\lambda_{\max}(\Sigma_t)}$: calibrated confidence radius.
Large $\mathcal{S}_t$ **tightens** the BMPC margin (Lemma 1).

Proposition 1 — Physical Safety (Gap 2)

A Lipschitz decoder argument lifts latent-space safety to physical space:

$$\mathbb{P}\big((x_t, u_t) \in \mathcal{S}\big) \ \geq\ 1 - \delta - f(\varepsilon_{\mathrm{dec}}), \quad \textstyle\sum_i \delta_i \leq \delta, \quad f(0) = 0, \quad \forall t \geq 0$$

Theoretical Guarantees

Theorem 1 — Bounded Representation Drift (G1)

$$\mathbb{E}[\mathrm{CFI}_t] \le 1, \quad \forall t \ge 0, \quad \mathrm{CFI}_t := \|\phi_{\theta_{t+1}} - \phi_{\theta_t}\| / \varepsilon_{\mathrm{cf}}$$

Bounds latent reorganization rate without freezing adaptation.

Theorems 2–3 — Recursive Feasibility + ISS (G2)

Adaptive tightening achieves recursive feasibility and ISS:

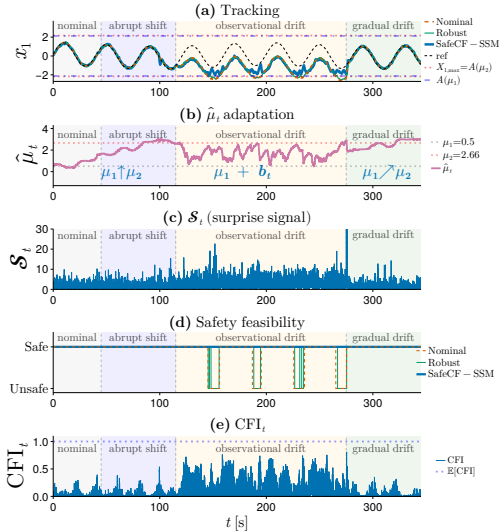
$$\|z_t\| \le \bar{\beta}(\|z_0\|, t) + \gamma(\bar{d}), \quad \bar{\beta} \in \mathcal{KL},\ \gamma \in \mathcal{K}$$

Recovers full control authority when the model is accurate — *strictly improving over fixed-margin robust MPC*.

Corollary 1 — Physical Safety Certificate (G3)

$$\mathbb{P}\big((x_t, u_t) \in \mathcal{S}\big) \ge 1 - \delta - f(\varepsilon_{\mathrm{dec}}), \quad \forall t \ge 0$$

Van der Pol Benchmark



Setup:

- $\mu_1 = 0.5,\ \mu_2 = 2.66$
- 345 s total, 4 regimes
- $N = 10,\ M = 25$ runs

Safety ($|x_1| \le X_{1,\max}$):

- ✓ SafeCF–SSM: $\geq 99.8\%$
- ✗ Nominal: violates
- ✗ Robust: violates

(during observational drift regime)

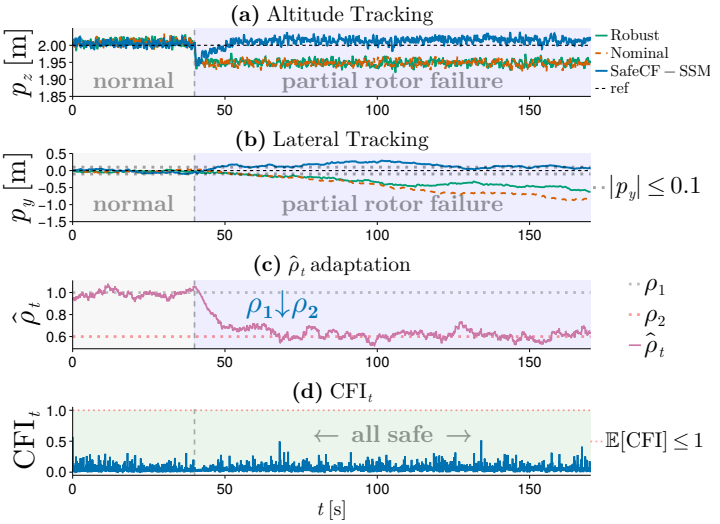
RMSE (x_1):

- SafeCF: 0.876
- Nominal: 0.937
- Robust: 0.924

$\mathbb{E}[\mathrm{CFI}] = 0.088 < 1$ (Thm. 1)

Gap 2: $\geq 99.8\%$ at $\varepsilon_{\mathrm{dec}} = 0.10$

3D Quadrotor Benchmark



Setup:

- 12-state, 4-input
- $\rho_1 = 1.0 \rightarrow \rho_2 = 0.6$ at $t_s = 40\,\mathrm{s}$
- $M = 25$ runs

Safety ($p_z, |\phi|, |\theta|$):

✓ All controllers: $\geq 99.8\%$

Lateral drift p_y:

✓ SafeCF: $|p_y| \leq 0.1\,\mathrm{m}$

✗ Nom/Rob: $> 0.5\,\mathrm{m}$ drift

$\mathbb{E}[\mathrm{CFI}] = 0.055 < 1$ (Thm. 1)
Smooth reorganisation confirms scalability to 12-state system

Quantitative Validation

Table: Physical safety under latent-space MPC ($M = 25$ Monte Carlo runs)

Property / Metric	Thm./Cor.	VdP	VdP ($\varepsilon_{\text{dec}} = 0.1$)	3D-Quad.
$\mathcal{S}_t$ bounded $\mathcal{O}(1\text{–}20)$	—	✓	✓	✓
$\mathbb{E}[\text{CFI}_t] \leq 1$	Thm. 1	✓	✓	✓
Safety $\geq 99.8\%$	Cor. 1	✓	✓	✓
SafeCF–SSM $\neq$ Nominal	Thm. 3	✓*	✓*	✓
$\hat{\theta}_t \to \theta_2$	Thm. 1	✓	✓	✓
SafeCF–SSM RMSE	—	0.876 ± 0.00	0.876 ± 0.00	0.058
Nominal MPC RMSE	—	0.937 ± 0.01	—	0.169
Robust MPC RMSE	—	0.924 ± 0.07	—	0.159

*Partial during observational drift ($t \in [115, 275]$ s): $b_t \neq \mathbf{0}$ limits tracking recovery, consistent with design scope.

Conclusion

SafeCF–SSM closes two structural gaps via a single decision: routing $\mathcal{S}_t$ to both adaptation and safety.

Contributions:

1. Formalize physical-latent gap as certifiable criterion (Prop. 1)
2. $\mathbb{E}[\mathrm{CFI}_t] \leq 1$ — bounded drift (Thm. 1)
3. Recursive feasibility + ISS (Thm. 2–3)
4. **First explicit physical safety guarantee** for latent-MPC (Cor. 1): $\mathbb{P}(\cdot) \geq 1-\delta-f(\varepsilon_{\mathrm{dec}})$

Validated on:

- ✓ VdP: 4 distributional-shift regimes $\geq 99.8\%$ safety, $M = 25$ runs
- ✓ 3D Quadrotor: 12-state system $|p_y| \leq 0.1\,\mathrm{m}$ despite rotor failure

Future work:

- Joint dynamics-observation adaptation
- Nonlinear decoder extensions

`https://thanana.github.io/SafeCF-SSM.html`