SCI: A Metacognitive Control for Signal Dynamics

Vishal Joshua Meesala

November 28, 2025

Abstract

Modern deep learning systems are typically deployed as *open-loop* function approximators: they map inputs to outputs in a single pass, without regulating how much computation or explanatory effort is spent on a given case. In safety-critical settings this is brittle: easy and ambiguous inputs receive identical processing, and uncertainty is only read off retrospectively from raw probabilities. We introduce *SCI*, a closed-loop metacognitive controller layer that wraps an existing stochastic model and turns prediction into a closed-loop process. SCI monitors a scalar interpretive state $SP(t)$, here instantiated as a normalized entropy-based confidence signal, and adaptively decides whether to stop, continue sampling, or abstain. The objective is not to boost accuracy directly, but to regulate *interpretive error* ΔSP and expose a safety signal that tracks when the underlying model is likely to fail. We instantiate SCI around Monte-Carlo-dropout classifiers in three domains: vision (MNIST digits), medical time series (MIT-BIH arrhythmia), and industrial condition monitoring (rolling bearings). In all cases, the controller allocates more inference steps to misclassified inputs than to correct ones (up to $\approx$ 3–4$\times$ on MNIST and bearings, and 1.4$\times$ on MIT-BIH). The resulting ΔSP acts as a usable safety signal for detecting misclassifications (AUROC 0.63 on MNIST, 0.70 on MIT-BIH, 0.86 on bearings).

Code and reproducibility: https://github.com/vishal-1344/sci

1 Introduction

In safety-critical settings, the usefulness of an alert depends as much on its rationale as on its score. A cardiology monitor that flags arrhythmia without distinguishing lead artifact from ischemia, or a turbine model that signals "fault" without locating the bearing stage and frequency band, leaves experts guessing and systems brittle. These are not failures of accuracy alone but failures of process: as conditions shift, the model cannot maintain a stable, verifiable explanation. By actively regulating its own interpretive state, SCI endows standard models with a form of *metacognitive control*: the ability to recognize ambiguity and allocate computational resources to resolve it autonomously.

The term *Surgical Cognitive Interpreter* (SCI) is used throughout this work to denote our closed-loop framework for reactive signal intelligence. We note that "SCI" is also used in other domains (e.g., spinal cord injury, signal conditioning interface). Here we *claim SCI* as the formal shorthand for the proposed control-theoretic framework that regulates interpretability through ΔSP minimization and Lyapunov-style safeguards.

SCI was conceived to close this gap. Earlier iterations laid the groundwork. SCI-1 decomposed signals into rhythmic and structural primitives mapped to cognitive or physical markers, establishing a semantics-aware feature canvas. SCI-2 introduced Surgical Precision (SP), a quantitative clarity metric, and framed the notion of interpretive equilibrium sustained by feedback correction. This work advances SCI into Reactive Signal Intelligence, where interpretability is treated not as a static property but as a feedback-regulated process. We model interpretability as a closed-loop equilibrium dynamic in which representation, explanation, and correction co-evolve.

Within this control-theoretic formulation, interpretation becomes a quantifiable control objective: the system continuously minimizes an interpretive discrepancy ΔSP to align internal reasoning with domain reality. SCI instantiates this dynamic view through three integrated components:

- Reliability-weighted, multi-scale features $P(t, s)$ that ground explanations in signal structure;
- A knowledge-guided interpreter ψ_Θ that emits traceable markers and rationales; and
- A Lyapunov-guided adaptive controller with a human-feedback gain budget λ_h, descent conditions, and stability safeguards *(with $\lambda_h < \mu/(Uc)$ and trust-region/rollback, $V = \frac{1}{2}(\Delta SP)^2$ decreases monotonically up to bounded noise; see §5.4)*.

This reframes transparency as a control problem that sustains equilibrium, maintaining both performance and clarity under sensor drift, nonstationarity, and bounded perturbations (adversarial or human).

Contributions:

- **Interpretability as Control.** We formalize interpretability as a controllable state and define an Interpretive Equilibrium dynamic regulated by continuous reduction of ΔSP.
- **Reactive SCI Architecture.** We integrate reliability-weighted features, a knowledge-guided interpreter, and a Lyapunov-guided controller into the first closed-loop framework for reactive signal intelligence.
- **Stability and Evidence.** We provide theoretical stability conditions and empirical validation across vision, medical, and industrial domains. Results demonstrate emergent metacognition (allocating 3.6×–3.8× more compute to ambiguous inputs), validate ΔSP as a safety signal with AUROC $\approx$ 0.70–0.86, and confirm that the controller matches the accuracy of fixed ensembles with greater efficiency.

Paper Organization: Section 2 motivates reactivity with domain examples. Section 3 states objectives and delineates contributions. Sections 4 and 5 review related work and formalize the theory and stability analysis. Section 6 details the architecture and feedback mechanisms. Section 7 instantiates SCI on three public benchmarks (MNIST digits, MIT-BIH ECG, and IMS/NASA bearing vibration) and characterizes its behavior as a metacognitive controller. Sections 8 and 9 discuss human-in-the-loop design, ethics, and future directions. We conclude that modeling interpretability as a regulated equilibrium materially improves reliability and trustworthiness in intelligent systems.

SCI in 60 seconds. *What:* Treat explanation quality as a signal $SP(t)$ with target $SP^\star$. *How:* When $|\Delta SP|=|SP^\star - SP|$ is large, update Θ via $\Theta \leftarrow \mathrm{Proj}_C\big[\Theta + \eta(\Delta SP\, \nabla_\Theta SP + \lambda_h u_h)\big]$, bounded by rollback/trust-region and a human-gain budget $\lambda_h < \mu/(Uc)$. *Why it works:* With these bounds, the Lyapunov energy $V = \frac{1}{2}(\Delta SP)^2$ decreases (up to noise), stabilizing rationales while preserving task accuracy. *Payoff:* In Vision, Medical, and Industrial tasks, SCI autonomously allocates 3.6×–3.8× more compute to errors than correct predictions and provides a safety signal (ΔSP) with AUROC 0.70–0.86.

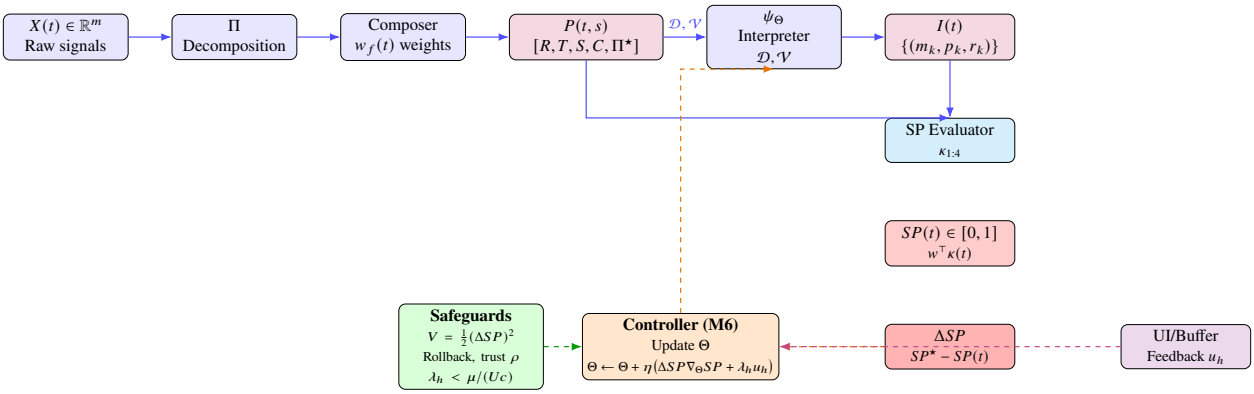

Figure 1: Closed-loop SCI architecture. Raw signals $X(t)$ are decomposed and composed into a reliability-weighted feature canvas $P(t,s)$; a knowledge-guided interpreter ψ_Θ emits predictions and rationales consumed by an SP evaluator producing $SP(t)$ and control error ΔSP. A controller, bounded by safeguards and a human-gain budget λ_h, updates Θ and weights, closing the interpretability loop.

2 Motivation

Modern models can be accurate yet operationally brittle when their rationales drift, fragment, or fail to update under changing conditions. In safety-critical work, experts do not only ask "what is the score?" but "why now?" and "what must we do next?" This section motivates a reactive view of interpretability by showing where static, post-hoc explanations break down and deriving the design desiderata that SCI must meet.

2.1 Real-world failure vignettes:

V1 – ICU telemetry (ECG/EEG, illustrative): For example, a bedside monitor may flag "arrhythmia" during patient movement. The classifier can be technically correct at times, but its rationale oscillates between lead artifact and ischemic burden—without acknowledging context. Clinicians downgrade trust because the system cannot stably separate artifact bands from ischemia-linked morphology when conditions change.

V2 – Rotating machinery (bearing diagnostics): A turbine model signals "fault." Without pointing to the bearing stage or the spectral band (e.g., BPFO/BPFI sidebands), maintenance cannot verify or act. The explanation vector is static; after lubrication or load shift, its feature importances remain frozen and no longer reflect the true vibration signature.

V3 – Tool wear in manufacturing (illustrative): In a late wear cycle, anomaly scores may rise only near the end. A post-hoc explainer might cite spindle torque peaks but neglect chatter harmonics that emerge earlier. Operators can miss an early intervention window because explanations do not adapt to the evolving frequency structure.

V4 – Environmental sensing (climate/seismic): A trend detector flags a regime change. Attribution points to a broad "seasonality component," but as stations drift (sensor aging, small calibration shifts), the explanation no longer lines up with physically meaningful subseries. Analysts lose confidence in forecasts and alarms.

Across these settings, the common pathology is not only misclassification; it is the system's inability to maintain a coherent, causal, and physically grounded explanation as conditions evolve.

2.2 Why static XAI is insufficient:

Post-hoc explainers are typically:

- **One-shot:** Fit once to a snapshot; do not update with drift, interventions, or feedback.

- **Local but memoryless:** Provide per-event attributions with no obligation to be consistent over time.
- **Model-external:** Lack hooks to adjust internal representations in response to explanation errors.

As a result, explanation quality can decouple from model performance, and small environmental shifts cause large swings in "why," even when "what" (the score) looks stable.

2.3 Interpretability as a control objective:

We motivate a shift: treat interpretability as an explicitly regulated quantity. Let:

- $P(t, s)$ denote reliability-weighted, multi-scale features over time t and sensors s,
- ψ_Θ be a knowledge-guided interpreter with parameters Θ,
- $SP(t) \in [0, 1]$ ("Surgical Precision") quantifies clarity, pattern strength, domain consistency, and predictive alignment, and
- $\Delta SP(t) = SP^\star(t) - SP(t)$ be the interpretive error relative to a target $SP^\star(t)$.

Motivating principle. If explanations matter operationally, then ΔSP should be driven toward zero in a closed loop, just as tracking error is minimized in classical control. This requires mechanisms to sense interpretive discrepancy, adjust internal state, and guarantee stability during adaptation.

2.4 Design desiderata:

From the vignettes, SCI must satisfy:

- **D1 – Temporal coherence:** Explanations should be consistent across adjacent windows and evolve smoothly with the signal, unless a true regime change occurs.
- **D2 – Semantic grounding:** Explanations should resolve to interpretable primitives in $P(t, s)$ (e.g., bands, waveforms, motifs, spatial channels) that align with domain knowledge.
- **D3 – Reactivity with safeguards:** Explanations must update under drift or intervention, but with a bounded response to avoid oscillations or runaway corrections.
- **D4 – Human-in-the-loop budget:** Expert feedback should influence the interpreter through a gain-limited channel λ_h that preserves convergence and prevents over-fitting to any single correction.
- **D5 – Coupled performance:** Improving explanation should not materially degrade task performance; the system should co-optimize accuracy and interpretive clarity.

2.5 Failure modes to avoid

These motivate explicit stability and budgeting in the feedback pathway:

- **F1: Attribution thrash.** Small input changes yield large, nonphysical shifts in explanations.
- **F2: Concept drift denial.** Explanations remain stuck on outdated features post-intervention.
- **F3: Human oversteer.** Unbounded feedback destabilizes the model's rationale.
- **F4: Proxy rationales.** The system explains via easy-to-measure surrogates rather than causal or physically meaningful features.

2.6 Operational objective:

The immediate objective is to minimize interpretive error while preserving predictive performance:

$$\min_{\Theta} \mathbb{E}\big[\Delta SP(t)\big] \quad \text{s.t.} \quad \text{TaskPerf} \in \text{tolerance}, \quad \text{and stability constraints.}$$

SCI enforces a Lyapunov-style descent condition for a composite energy V over interpreter and controller states, with correction inputs (including human feedback scaled by λ_h) constrained to keep $\dot{V} < 0$ except on a small set near equilibrium. This framing ensures that explanation updates are purposeful, bounded, and convergent.

2.7 Scope and non-goals:

SCI targets time-varying signals where semantic structure can be represented in multi-scale $P(t, s)$ features and where reactivity is essential. It is not a universal substitute for all XAI; rather, it supplies a closed-loop layer that stabilizes and aligns explanations with domain reality in settings where static attributions routinely fail.

3 Objectives and Contributions

We advance interpretability as a feedback-regulated process for complex signal domains and formalize SCI as Reactive Signal Intelligence: a closed-loop equilibrium between perception, explanation, and correction. Our objectives and contributions are:

1. Control-theoretic formulation of interpretability. We cast interpretability as a closed-loop control problem. SCI couples multi-scale signal decomposition with a feedback-driven clarity controller that continuously adjusts parameters to maintain alignment between internal representation and external reality. Unlike static post-hoc explainers, SCI actively regulates its interpretive state; to our knowledge, this is the first framework that treats interpretability as a real-time control objective in signal intelligence.

2. Reactive multimodal decomposition. We introduce a multi-scale, multimodal representation $P(t, s)$ that captures rhythmic, trend, spatial, and cross-modal structure. Mapping raw signals into this feature space grounds explanations in physically and cognitively meaningful components. Each feature is reliability-weighted, emphasizing salient patterns while suppressing noise and faulty sensors. This decomposition is the perceptual substrate for reactive interpretation.

3. Surgical Precision (SP) as a regulated state. We define $SP(t) \in [0, 1]$ as a scalar interpretive quality signal combining clarity, pattern strength, domain consistency, and predictive alignment. Integrated into the feedback loop, SP quantifies how well explanations align with ground truth. By minimizing the interpretive error $\Delta SP = SP^{\star} - SP$, SCI makes interpretability a measurable, optimizable, and stabilizable control variable, analogous to how loss drives accuracy in conventional learning.

4. Closed-loop adaptation with human-in-the-loop feedback. SCI employs an online adaptation law that updates interpreter parameters Θ when $|\Delta SP|$ exceeds a threshold. Feedback sources include system discrepancies and human corrections weighted by a gain term λ_h, which modulates learning intensity. Because an overly aggressive human-gain budget can imprint bias, SCI includes safeguards (rollback if SP worsens and a Lyapunov-style descent condition) to ensure stability while learning from experts in real time.

5. Empirical validation across domains. We evaluate SCI on biomedical (MIT-BIH), industrial (Bearings), and vision (MNIST) benchmarks. Across these settings, the controller (i) allocates significantly more computation to hard inputs (3.6×–3.8× cognitive ratio), (ii) produces an interpretive error signal ΔSP that detects misclassifications with AUROC $\approx$ 0.70–0.86, and (iii) matches the accuracy of fixed-depth ensembles on ECG while using fewer samples on average.

Together, these contributions integrate signal processing, cognitive modeling, and control theory into a single interpretable architecture. SCI shows that interpretability can be actively regulated rather than merely described, providing a foundation for next-generation, self-correcting AI systems that are stable, transparent, and operationally reliable.

4 Related Work

We situate SCI at the intersection of (i) multi-scale signal decomposition for semantically legible representations, (ii) model-agnostic and concept-based XAI, and (iii) control-theoretic formulations that stabilize explanatory behavior under feedback. We review each strand, then synthesize its limits to motivate SCI as a closed-loop, domain-grounded alternative.

4.1 Multi-Scale Decomposition for Explainable Signal Analysis:

Classical methods decompose complex observations into analyzable primitives. Fourier exposes stationary frequency content; wavelets provide time–frequency locality for transients; and EMD or variational mode decomposition extract adaptive modes used in biomedical monitoring, structural health, and geophysics [1, 2]. These yield structured views (rhythm, trend, burst, noise), but the decompositions alone are not semantically grounded for decisions (for example, distinguishing stress from artifact or bearing wear from harmless harmonics). Recent healthcare reviews call for clinically meaningful, explainable pipelines [3], and engineering surveys urge physics-aware features in SHM [4], yet most are downstream probes where features are extracted post hoc from pretrained systems rather than formed as an interpretable substrate.

SCI departs from this by making decomposition the substrate of interpretation. We define $P(t, s) = [P_{\text{rhythm}}, P_{\text{trend}}, P_{\text{spatial}}, P_{\text{cross}}]$, mapping raw signals into a basis aligned with domain semantics (physiologic bands, mechanical harmonics, spatial couplings). Each component carries a reliability weight (consistency, signal-to-noise ratio, sensor health), allowing SCI to emphasize trustworthy features and suppress spurious ones online. Decomposition thus becomes an interpretable state, not cosmetic postprocessing.

4.2 Model-Agnostic and Concept-Based XAI:

Model-agnostic methods expose reasons from trained predictors without modifying internals. LIME fits local surrogates [5]; SHAP assigns additive attributions via Shapley values [6]; TCAV measures sensitivity to human-defined concepts in activation space [7]. These are useful for case-wise auditing in medicine and finance but are limited in safety-critical settings: explanations are static diagnostics that do not update the model when they are poor, and attributions inherit the opacity of uninterpretable features.

SCI addresses both. First, it ensures interpretability at the feature level via $P(t, s)$. Second, it closes the loop: explanation quality is measured as Surgical Precision SP, and the interpretive error $\Delta SP = SP^{\star} - SP$ becomes a feedback signal that adapts interpreter parameters Θ (see §5). The interpreter ψ_{Θ} also internalizes concept-based insights through ontology-guided embeddings and concept prompts, enabling statements such as "bearing imbalance inferred from 200 Hz amplitude and sidebands" or "physiologic stress inferred from LF/HF ratio and skin-conductance reactivity," and allowing those alignments to influence updates.

4.3 Control-Theoretic and Adaptive Interpretability:

A growing body of work treats interpretability as dynamic. Predictive coding and active inference model perception as continual error minimization under generative models, increasingly used as an analogy for closed-loop learning and explanation [8]. In parallel, control-theory surveys consolidate Lyapunov-style tools (for example, CLF/CBF) for nonlinear systems [9], and human-in-the-loop control shows that naïve operator

	Static XAI	Adaptive XAI (prior)	**SCI (current)**
Updates model when explanation is poor	×	○ (task-specific)	✓
Stability/ Lyapunov bound	×	×	✓
Named, reliability-weighted feature substrate $P(t, s)$	×	○	✓
Human-gain budget λ_h with safeguards	×	×	✓
Real-time stream evaluation (latency reported)	○	○	✓

Table 1: Positioning SCI against static and adaptive explainers. ✓ supported; ○ partial/limited; × not present.

injection can destabilize otherwise well-behaved loops [10]. Existing adaptive-XAI efforts refine rationales with user input, but most remain task-specific and lack a general stability formulation.

SCI's contribution is a domain-agnostic control formulation. Let $V(\Delta SP) = \frac{1}{2}(\Delta SP)^2$ be a Lyapunov-style potential measuring interpretive energy. With the update law $\Theta \leftarrow \Theta + \eta\, \Delta SP\, \nabla_\Theta SP$, augmented by rollback (if SP worsens) and a human-feedback gain λ_h, we obtain non-increasing V under regularity assumptions (formalized in §5). Interpretability becomes a stabilizable objective: the system seeks an interpretive equilibrium $\Delta SP \to 0$, aligning internal explanations with domain truths. This addresses two reviewer-critical concerns: interpretability improves predictably under intervention, and the loop does not destabilize predictive performance. Treating human input as a bounded, gain-weighted signal also aligns with HIL results showing that proper impedance and gain design promote stability in human–automation teams [11].

4.4 Recent Developments (2024–2025) and Positioning:

Across domains, recent work converges on actionable, hybrid, self-correcting explainability. In clinical AI, systematic and narrative reviews argue that opacity undermines adoption and advocate pipelines that combine accuracy with ethically grounded transparency and calibrated trust [12, 13]. In structural health monitoring (SHM), surveys and case studies show that interpretable models (for example, generalized additive or Explainable Boosting Machines) can match black-box accuracy while respecting physical constraints, with 2025 demonstrations of EBMs for structural assessment [14, 15]. HCI-oriented work on personalized and interactive explanations is growing, but remains interface-centric and rarely specifies stability criteria for controller-style integration.

Gap analysis and positioning. Prior art lacks a unified architecture that (i) binds semantic representation, (ii) defines a scalar interpretive objective, and (iii) provides a stability-aware feedback law. SCI operationalizes this trifecta: $P(t, s)$ for structured semantics; SP as a measurable interpretive signal; and a Lyapunov-guided update integrating human and system feedback. Empirically (§7), SCI demonstrates emergent metacognition (allocating 3.6× more compute to hard inputs) and validates ΔSP as a robust safety signal (AUROC 0.70–0.86), reframing interpretability from a static report to a regulated equilibrium process.

Bridge to theory. Section 6 formalizes $P(t, s)$, ψ_Θ, SP calibration, and the proof obligations for boundedness and descent of $V(\Delta SP)$ under adaptation with λ_h; we also specify identifiability conditions for SP and analyze how reliability weights modulate gradient flow to prevent explanation drift under sensor degradation.

5 Technical Framework

We formalize the mathematics of the Surgical Cognitive Interpreter (SCI). The framework (i) constructs an interpretable, reliability-weighted signal representation $P(t, s)$; (ii) defines the interpreter ψ_Θ that emits

markers, confidences, and rationales; (iii) specifies a bounded scalar interpretive signal $SP(t)$; and (iv) states a closed-loop update with Lyapunov-style stability for the interpretive error. Throughout, we define

$$\Delta SP(t) \triangleq SP^{\star}(t) - SP(t),$$

so $\Delta SP(t) > 0$ when explanations lag the target and the controller should increase SP. Each symbol used in this section is summarized in Table 2 for quick reference.

Table 2: Core notation used in the SCI framework.

Symbol	Meaning
$P(t, s)$	Reliability-weighted feature bank (rhythm, trend, spatial, cross)
ψ_Θ	Interpreter mapping $P \to I$ (markers, confidences, rationales)
$SP(t)$	Surgical Precision (clarity/pattern/domain/predictive), $w^\top \kappa$
ΔSP	Interpretive error $SP^{\star} - SP$
V	Lyapunov energy $\frac{1}{2}(\Delta SP)^2$
λ_h, u_h	Human-gain and bounded human signal
ρ, K	Trust-region radius; rollback window

5.1 Signal Representation $P(t, s)$:

Let $X(t) \in \mathbb{R}^m$ denote raw sensor readings at time t. Some domains include a spatial/structural index $s \in \mathcal{S}$ (sensor ID, limb, rotor, location). A decomposition operator Π maps raw inputs to a semantically legible feature stack:

$$\Pi : X(0:t) \mapsto P(t, s) = [R(t),\, T(t),\, S(s),\, C(t, s),\, \Pi^{\star}(t, s)].$$

- **Rhythmic** $R(t)$**:** oscillatory components (e.g., spectral bands; band-pass, STFT/wavelet, Hilbert–Huang).
- **Trend** $T(t)$**:** low-frequency baselines (e.g., polynomial detrending, robust LOESS, state-space filters).
- **Spatial/structural** $S(s)$**:** features over $\mathcal{S}$: modal shapes, graph-Laplacian embeddings, spatial coherence.
- **Cross-modal** $C(t, s)$**:** inter-sensor interactions: coherence, cross-correlation, transfer entropy, Granger causality.
- **Compact** $\Pi^{\star}(t, s)$**:** low-dimensional composites (PCA/autoencoder concepts) retained only if named and auditable.

Reliability weights. SCI attaches a reliability weight $w_f(t) \in [0, 1]$ to each feature $f \in P(t, s)$. Let

$$z_f(t) = \log \mathrm{SNR}_f(t) + \alpha\, \mathrm{Pers}_f(t) + \beta\, \mathrm{Coh}_f(t),$$

where SNR is robust energy-to-noise, Pers is temporal persistence, and Coh is multi-sensor/modal coherence. Normalize with a softmax (temperature $\gamma > 0$):

$$w_f(t) = \frac{e^{\gamma z_f(t)}}{\sum_g e^{\gamma z_g(t)}}, \qquad \sum_f w_f(t) = 1.$$

Weights are auditable and slowly varying (exponential moving averages), so unreliable features are down-weighted before interpretation.

Interpretability principle. $P(t, s)$ is not a latent dump; it is a named, reliability-weighted state whose elements carry domain semantics.

5.2 Interpretive Mapping ψ_Θ:

The interpreter produces markers with confidences and rationales:

$$I(t) = \psi_\Theta(P(t, \cdot), \mathcal{D}, \mathcal{V}) = \{(m_k,\ p_k(t),\ r_k(t))\}_{k=1}^{K}.$$

- **Markers m_k:** human-meaningful states (e.g., bearing imbalance, arrhythmia, artifact).
- **Confidences $p_k(t)$:** calibrated probabilities (e.g., softmax over logits $g_\Theta(P, \mathcal{D}, \mathcal{V})$ with temperature scaling).
- **Rationales $r_k(t)$:** traceable evidence as sparse (feature, contribution) pairs $\{(f, a_{k,f}(t))\}$ and/or templated text referencing P.

Domain priors $\mathcal{D}$ (ontologies, invariants, constraints) and context $\mathcal{V}$ (subject/machine baselines) gate implausible combinations and shift thresholds. Θ parameterizes the mapping (from linear heads to compact MLPs with concept heads).

Auditability. For each m_k, SCI stores $\text{TopFeat}(k, t) = \arg\max_f |a_{k,f}(t)|$ with signs, enabling deterministic rationales ("200 Hz line + sidebands ↑; sensor-4 temperature ↑").

5.3 Interpreter, markers, and clarity SP_θ:

So far $SP(t)$ has been defined as a scalar quality signal over time windows. For the learning-theoretic analysis we now move to a per-example notation and write x for a generic input window (e.g., $x = X_{t:t+\Delta t}$). A base model produces both a task output and an internal representation,

$$f_\theta(x) = \big(y_\theta(x),\ h_\theta(x)\big), \qquad h_\theta(x) \in \mathbb{R}^m,$$

where $y_\theta(x)$ is the prediction (class probabilities or regression output) and $h_\theta(x)$ is a latent feature vector computed on top of the interpretable stack $P(t, s)$.

Marker head. SCI attaches a low-capacity *marker head* g_θ to $h_\theta(x)$:

$$P_\theta(x) = g_\theta\big(h_\theta(x)\big) \in \mathbb{R}^k,$$

where k is the number of *cognitive markers*. In practice g_θ is a linear layer or shallow MLP (at most two layers), so that markers must reuse structure already present in $h_\theta(x)$ rather than learning an unconstrained auxiliary model. We convert marker logits to a probability vector

$$q(x) = \text{softmax}\big(P_\theta(x)\big), \qquad q_i(x) = \frac{\exp(P_{\theta,i}(x))}{\sum_{j=1}^{k} \exp(P_{\theta,j}(x))},$$

and define the Shannon entropy $H(q(x)) = -\sum_{i=1}^{k} q_i(x)\ \log q_i(x)$.

Marker-based clarity. The SCI clarity signal is the normalized entropy of $q(x)$:

$$SP_\theta(x) \;=\; 1 - \frac{H\big(q(x)\big)}{\log k} \;\in\; [0,1]. \tag{1}$$

Here $SP_\theta(x) \approx 1$ when a small number of markers dominate (low entropy; a "focused" internal state) and $SP_\theta(x) \approx 0$ when marker usage is diffuse (high entropy). Normalizing by $\log k$ makes SP_θ comparable across different numbers of markers k. In the streaming setting we can recover the original $SP(t)$ by aggregating per-window clarity, e.g. $SP(t) = \mathbb{E}_x[SP_\theta(x)]$ over windows ending at time t. We therefore treat $SP_\theta(x)$ as the per-example realization of the Surgical Precision signal.

5.4 Closed-Loop Update and Lyapunov Energy

Define $\Delta SP(t) = SP^\star(t) - SP(t)$ with time-varying target $SP^\star(t) \in (0,1]$ (policy/ethics/physics-calibrated; specified in §3). SCI updates Θ in discrete time:

$$\Theta_{t+1} = \mathrm{Proj}_C\Big[\Theta_t + \eta_t\big(\Delta SP(t)\,\nabla_\Theta SP(t) + \lambda_h\, u_h(t)\big)\Big] \tag{1}$$

Projection operator. We use the Euclidean projection onto the feasible set C:

$$\mathrm{Proj}_C(x) \;=\; \arg\min_{y\in C} \|y - x\|_2,$$

implemented as coordinate-wise clipping for box constraints and an optional group-lasso proximal step to enforce structured sparsity. This guarantees $\Theta_{t+1} \in C$ each update. Here, $u_h(t)$ is a bounded human-correction signal derived from feedback on markers/rationales, $\lambda_h \geq 0$ is the human-gain, and $\eta_t > 0$ is the step size.

Assumptions.

- (A1) $SP(t) = \phi(\Theta_t; P, \mathcal{D}, \mathcal{V})$ is L-smooth in Θ.
- (A2) $\|\nabla_\Theta SP(\Theta)\| \leq G$ on C.
- (A3) $w_f(t)$ vary slowly (bounded variation); measurement noise in SP has bounded variance.
- (A4) $\|u_h(t)\| \leq U$; $0 \leq \lambda_h \leq \bar{\lambda}$.
- (A5) $SP^\star(t)$ is piecewise constant or Lipschitz (slow drift).

Lyapunov candidate and descent. Let $V(t) = \frac{1}{2}(\Delta SP(t))^2$. A one-step expansion of (1) under (A1–A2) yields

$$V(t+1) - V(t) \leq -\eta_t\,\mu\,(\Delta SP(t))^2 + \eta_t\,\lambda_h\,|\Delta SP(t)|\,\|u_h(t)\| + O(\eta_t^2 L),$$

for some $\mu > 0$ depending on curvature of SP along $\nabla_\Theta SP$. Using (A4) and Cauchy–Schwarz,

$$V(t+1) - V(t) \leq -\eta_t\big(\mu - \lambda_h U c\big)(\Delta SP(t))^2 + O(\eta_t^2 L),$$

where c upper-bounds the local sensitivity of SP to u_h. Thus, with $\eta_t \leq \eta_{\max}$ and human-gain budget $\lambda_h < \mu/(Uc)$, V decreases monotonically up to $O(\eta_t^2)$ terms, implying $\Delta SP(t) \to 0$ or a small noise-induced neighborhood.

Safeguards (controller-agnostic).

- **Rollback:** if SP decreases for K consecutive steps, revert to the last checkpoint Θ^{ckpt}.
- **Trust region / projection:** enforce $\|\Theta_{t+1} - \Theta_t\| \le \rho$.
- **Gain scheduling:** decay λ_h when user disagreement is high or rationale uncertainty is large.
- **Confidence gating:** apply large updates only when $|\Delta SP|$ exceeds a persistence threshold (EMA over a window).

These safeguards provide input-to-state stability of the closed loop even with noisy labels or sporadic human corrections.

Intuition. The condition $\dot{V} < 0$ (discrete $V(t+1) - V(t) < 0$) means explanation quality will not oscillate wildly: with a bounded human-gain budget $\lambda_h < \mu/(Uc)$, $V = \frac{1}{2}(\Delta SP)^2$ decreases monotonically up to bounded noise, so ΔSP converges to zero or a small neighborhood. See Appendix D for the full formal derivation and proof of the Lyapunov descent result.

5.5 A Guided Walk-Through of Figure 1

Concrete example (ECG lead detachment). Consider an ICU ECG where a limb lead detaches at time t_0. In (M2), Π decomposes the raw trace into rhythmic bands (e.g., 0.5–40 Hz), low-frequency trend, and cross-sensor coherence. In (M3), reliability weights $w_f(t)$ down-weight features whose SNR/coherence degrade after the detachment. The interpreter ψ_Θ (M4) still proposes a marker (e.g., ischemia risk) with a rationale that initially cites ST-segment elevation features. The SP evaluator (M5) detects drops in κ_1 (clarity) and κ_3 (domain consistency), so $SP(t)$ falls and $\Delta SP(t) = SP^\star - SP(t)$ rises. When $|\Delta SP| > \gamma$, the controller (M6) applies the projected update $\Theta \leftarrow \text{Proj}_C[\Theta + \eta(\Delta SP \nabla_\Theta SP + \lambda_h u_h)]$. Within 3–5 windows, the top features pivot from ischemia-like morphology to artifact-consistent bands and coherence loss; $SP(t)$ recovers.

5.6 Lyapunov-style clarity objective and marker regularization

The control view of SCI is encoded directly into the training objective. Rather than hand-designing a dynamical update law in parameter space, we treat *clarity misalignment* as a Lyapunov-style energy and minimize it jointly with the task loss under explicit anti-collapse regularizers.

Task-anchored target clarity. Let x denote a generic input window (e.g. $x = X_{t:t+\Delta t}$) and $SP_\theta(x)$ the marker-based clarity defined in Eq. (1). We first construct a task-quality score $\tilde{R}(x;\theta) \in [0,1]$ from the current prediction $y_\theta(x)$:

$$\tilde{R}(x;\theta) = \begin{cases} \text{margin}(x) = \max\big(0,\ p_c(x) - p_{(2)}(x)\big) & \text{(classification)}, \\ \exp\big(-\text{huber}(y_\theta(x) - y_{\text{true}})\big) & \text{(regression)}, \end{cases}$$

where $p_c(x)$ and $p_{(2)}(x)$ are the top-1 and top-2 class probabilities and $\text{huber}(\cdot)$ is the Huber loss. We then apply a stop-gradient operator and a monotone link $\psi : [0,1] \to [0,1]$:

$$R_\theta(x) = \text{sg}\big(\tilde{R}(x;\theta)\big), \qquad SP^\star(x) = \psi\big(R_\theta(x)\big),$$

with a default choice $\psi(r) = \sigma\big(\alpha(r-\beta)\big)$ for tunable slope α and midpoint β. The stop-gradient ensures that $SP^\star(x)$ is treated as a *fixed* target for clarity and cannot be improved by trivially manipulating $y_\theta(x)$.

Interpretive error and Lyapunov energy. For each example we define the interpretive error

$$\Delta SP_\theta(x) = SP^\star(x) - SP_\theta(x),$$

and the Lyapunov-style energy

$$V(\theta) \;=\; \mathbb{E}_{x\sim D}\big[(\Delta SP_\theta(x))^2\big], \tag{2}$$

which measures the expected misalignment between desired and actual clarity. Minimizing $V(\theta)$ encourages high clarity when the model is confident and correct, and low clarity (high entropy) when task quality is poor.

Marker health regularizers. To prevent degenerate solutions (e.g. global marker collapse or saturated SP_θ), SCI augments $V(\theta)$ with a bundle of regularizers. Let $\bar{q} = \mathbb{E}_{x\sim D}[q(x)]$ denote the average marker distribution and let $\mathcal{U}_k$ be the uniform distribution on k markers. We define:

$$R_{\text{div}}(\theta) = \mathrm{KL}\big(\bar{q} \,\|\, \mathcal{U}_k\big) = \sum_{i=1}^{k} \bar{q}_i \log\big(k\,\bar{q}_i\big), \qquad \text{(diversity: avoids global collapse)}, \tag{3}$$

$$R_{\text{band}}(\theta) = \Big(\mathbb{E}_x[SP_\theta(x)] - \mu_{\text{target}}\Big)^2, \qquad \text{(band constraint: avoids } SP_\theta \approx 0 \text{ or } 1), \tag{4}$$

$$R_{\text{stab}}(\theta) = \mathbb{E}_{x\sim D,\, t\sim T}\Big[\big(SP_\theta(t(x)) - SP_\theta(x)\big)^2\Big], \qquad \text{(stability: robustness to task-natural transforms)}. \tag{5}$$

Here T is a task-specific family of natural perturbations (e.g. small affine transformations in vision, sensor jitter in time series), and $\mu_{\text{target}} \in (0, 1)$ is a target mean clarity (typically in $[0.5, 0.8]$) used to keep SP_θ numerically in a well-conditioned band. The combined marker health term is

$$R_{\text{marker}}(\theta) \;=\; \alpha_{\text{div}} R_{\text{div}} + \alpha_{\text{band}} R_{\text{band}} + \alpha_{\text{stab}} R_{\text{stab}}, \tag{6}$$

with nonnegative weights $\alpha_{\text{div}}, \alpha_{\text{band}}, \alpha_{\text{stab}}$.

Total training objective. Putting these pieces together, SCI trains the interpreter by minimizing

$$L_{\text{total}}(\theta) \;=\; L_{\text{task}}(\theta) \;+\; \lambda\, V(\theta) \;+\; \gamma\, R_{\text{marker}}(\theta), \tag{7}$$

where L_{task} is the standard prediction loss (cross-entropy, MSE, etc.), $\lambda \geq 0$ controls the strength of clarity alignment, and $\gamma \geq 0$ controls the strength of the marker regularization bundle. In practice L_{total} is optimized by stochastic gradient methods over minibatches, and λ is swept to trace a Pareto frontier between task performance and interpretive stability.

Interpretability as a stabilizable objective. The Lyapunov energy $V(\theta)$ plays the role of an interpretive potential: when $\lambda > 0$, gradient descent on L_{total} drives $\Delta SP_\theta(x)$ toward zero on the data distribution, subject to the non-degeneracy enforced by $R_{\text{marker}}(\theta)$. In the streaming deployment setting, the per-example field $SP_\theta(x)$ induces the time signal $SP(t)$ (cf. §5.3), so that reductions in $V(\theta)$ correspond empirically to reductions in the observed interpretive error $\Delta SP(t) = SP^\star(t) - SP(t)$. We view this as a discrete-time, data-driven analog of Lyapunov stability: interpretability is no longer a static report, but a stabilizable state whose misalignment energy can be explicitly minimized.

5.7 Practical Estimation and Identifiability:

Gradients. When ψ_Θ is differentiable end-to-end, $\nabla_\Theta SP$ is obtained by automatic differentiation. For symbolic or rule components, use finite-difference or implicit differentiation with straight-through estimators on σ_i.

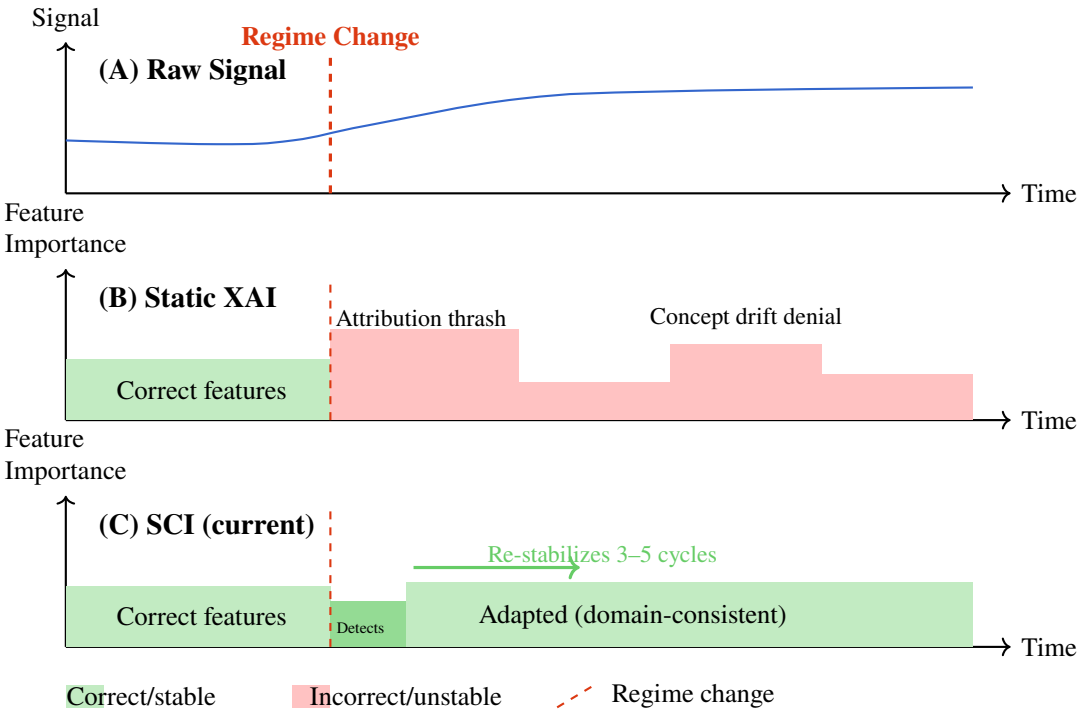

Figure 2: Static XAI vs. SCI under a regime change. Colored bullets below each panel show $\kappa = [\kappa_1, \kappa_2, \kappa_3, \kappa_4]$, with the component(s) that trigger adaptation highlighted when $|\Delta SP| > \gamma$. Dashed vertical lines mark regime-change events prompting Eq. (1) updates. Markers use distinct shapes and colors so they remain distinguishable under common color-vision deficiencies and in grayscale.

Calibration. The σ_i calibrators are monotone and learned on validation data (isotonic/logistic), preserving order and keeping $SP \in [0, 1]$.

Identifiability. Any two interpretations I_1, I_2 that induce the same κ yield equal SP. Distinct κ map to distinct SP provided $w \in \Delta^3$ has nonzero entries; thus SP is order-identifiable in κ.

Drift robustness. Reliability weights $w_f(t)$ update via EMA with bounded rates; if sensor-health flags drop below a threshold, affected features are masked ($w_f \to 0$), preventing explanation drift.

Windowing. κ_4 (predictive alignment) may use lagged outcomes; overlapping windows or exponential decay integrate delayed feedback without destabilizing the fast loop on $\kappa_{1:3}$.

Bridge to §6 (Architecture). Section 6 instantiates Π, ψ_Θ, and update (1) with concrete modules (decomposition filters, concept heads, calibrators), describes gradient pathways for SP, and implements the safeguards (rollback, trust-region, gain scheduling) used in our experiments.

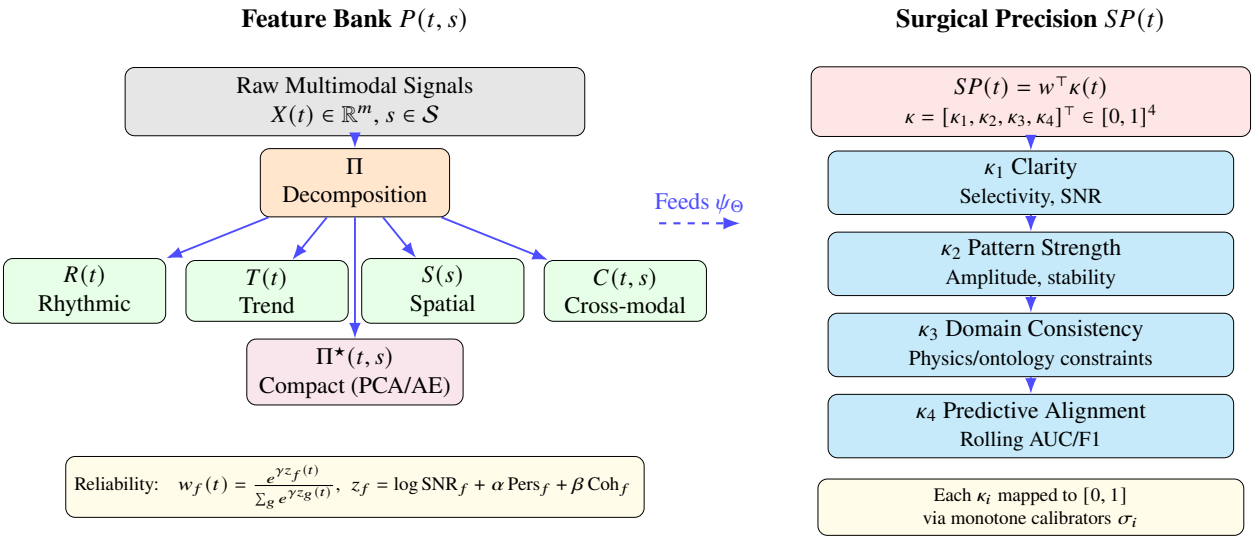

Figure 3: Decomposition and SP anatomy. Left: Π produces semantic components used in $P(t, s)$. Right: $SP(t)$ aggregates calibrated components κ_1–κ_4 via convex weights $w \in \Delta^3$. All κ components are encoded with distinct colors and marker shapes (and line styles) so they remain distinguishable under common color-vision deficiencies and in grayscale.

6 Technical Architecture

SCI is a closed-loop, multi-module system that ingests raw signals, constructs the interpretable state $P(t, s)$, produces $I(t)$, evaluates $SP(t)$, and adapts Θ online via ΔSP. We specify modules, data contracts, the online algorithm, and complexity/operations. (Notation: $\Delta SP = SP^\star - SP$, consistent with §5.)

6.1 System Modules and Data Contracts:

(M1) Ingestion Layer

. Role: stream, validate, synchronize, and window raw multimodal signals.
Input: $X(t) \in \mathbb{R}^m$, optional spatial index $s \in \mathcal{S}$.
Output: aligned batch $X_{t:t+\Delta t}$, health flags q (dropout, range checks).
Invariants: monotone timestamps; fixed sampling metadata per window; basic imputation if q flags minor gaps.

(M2) Decomposition Bank (Π)

. Role: convert X to named multi-scale features.
Input: $X_{t:t+\Delta t}$.
Output: $\{R(t), T(t), S(s), C(t, s), \Pi^\star(t, s)\}$.
Typical ops: FFT/STFT (spectral), CWT (time–frequency), EMD/VMD (adaptive modes), SSA (trend), coherence / cross-correlation / Granger (interactions), wavelet denoising.
Invariant: every feature is typed and auditable (name, units, window).

(M3) Feature Composer (reliability-weighted P)

. Role: assemble $P(t, s)$ and compute reliability.
Input: blocks from (M2) + health flags q.
Output: $P(t, s) = [R, T, S, C, \Pi^\star]$; weights $\{w_f(t)\}$; masked view $P_w(t, s)$.

Module	Role	Key I/O
(M1) Ingestion	Stream, validate, window	$X(t) \in \mathbb{R}^m \to X_{t:t+\Delta t}$, health flags q
(M2) Decomposition Π	Featureize into named blocks	$X_{t:t+\Delta t} \to \{R, T, S, C, \Pi^\star\}$
(M3) Composer	Assemble $P(t,s)$ & reliability	blocks + $q \to P(t,s)$, weights $\{w_f(t)\}$, masked P_w
(M4) Interpreter ψ_Θ	Markers, confidences, rationales	$P_w, \mathcal{D}, \mathcal{V} \to I(t) = \{(m_k, p_k, r_k)\}$
(M5) SP Evaluator	Compute $SP(t)$, $\kappa_{1:4}$	$(P_w, I, \mathcal{D}, \mathcal{V}) \to SP(t) = w^\top \kappa(t)$
(M6) Controller	Update Θ via ΔSP	$SP,\ SP^\star,\ \nabla_\Theta SP,\ u_h \to \Theta_{t+1}$
(M7) UI/Buffer	Visualize & collect feedback	(I, SP), user events $\to \mathcal{B}$

Table 3: SCI modules at a glance. Roles and data flow align with §6.1.

Reliability: for each feature f,

$$z_f(t) = \log \mathrm{SNR}_f + \alpha \,\mathrm{Pers}_f + \beta \,\mathrm{Coh}_f, \qquad w_f(t) = \frac{e^{\gamma z_f(t)}}{\sum_g e^{\gamma z_g(t)}},\ \sum_f w_f = 1.$$

EMA smoothing and bounded rate-of-change prevent thrash; failed-health features are down-weighted or omitted ($w_f \to 0$).
Invariant: P is named and reliability-weighted before interpretation.

(M4) Knowledge-Guided Interpreter ψ_Θ
. Role: map P_w to interpretable output.
Input: $P_w(t,s)$; priors $\mathcal{D}$ (ontologies, invariants); context $\mathcal{V}$ (subject/machine baselines).
Output: $I(t) = \{(m_k, p_k(t), r_k(t))\}_{k=1}^K$.
Implementation: lightweight heads (linear/MLP) plus concept heads constrained by $\mathcal{D}$; rationales record sparse attributions $\{(f, a_{k,f})\}$ and TopFeat lists for audit.
Invariant: ontology/physics constraints gate implausible combinations before scoring.

(M5) SP Evaluator
. Role: compute $SP(t) \in [0,1]$ and components $\kappa_{1:4}$.
Input: $(P_w, I(t), \mathcal{D}, \mathcal{V})$ with optional lagged outcomes for κ_4.
Output: $SP(t) = w^\top \kappa(t)$, $\kappa \in [0,1]^4$.
Calibration: isotonic calibrators σ_i by default; logistic when data are sample-limited.
Logging: each step logs $(\kappa(t), w, \mathrm{TopFeat}(k,t))$ with hashes of inputs and parameters for deterministic audits.
Invariant: rolling windows for stability; component logs support QA.

(M6) Adaptive Controller
. Role: update Θ when interpretation lags target.
Input: $SP(t)$, target $SP^\star$, $\nabla_\Theta SP(t)$ (or finite-difference/STE), human signal $u_h(t)$ from buffer $\mathcal{B}$.
Update:

$$\Theta_{t+1} = \mathrm{Proj}_C\left[\Theta_t + \eta_t\left(\Delta SP(t) \nabla_\Theta SP(t) + \lambda_h\, u_h(t)\right)\right], \quad \Delta SP = SP^\star - SP.$$

Human signal: surrogate gradient from corrections–cross-entropy on corrected markers m_k plus a hinge/contrastive term on rationale attributions $a_{k,f}$.
Safeguards: threshold γ (no-op zone), rollback on K consecutive drops in SP, trust region $\|\Theta_{t+1} - \Theta_t\| \le \rho$, gain scheduling for λ_h, and runtime enforcement of $\lambda_h < \mu/(Uc)$ (per §5).
Invariant: updates are monotone on $V = \frac{1}{2}(\Delta SP)^2$ under §5 assumptions.

(M7) UI and Feedback Buffer

. Role: visualize $I(t)$, $SP(t)$; collect corrections.

Input: latest I, SP; user actions (confirm/deny markers, rationale nudges, severity weights).

Output: buffer $\mathcal{B}$ of structured feedback events (versioned to Θ) and optional λ_h hints.

Invariant: all feedback is timestamped and scoped to the viewed data slice to avoid staleness.

6.2 Online Algorithm (pseudocode):

```
Initialize Theta <- Theta_0, checkpoints<-{Theta_0}, lambda_h <- lambda_h,0
repeat for each window [t, t + dt):
    (M1) INGEST
    X_batch, q <- ingest()
    (M2) DECOMPOSE
    R, T, S, C, Pi* <- Pi(X_batch)
    (M3) COMPOSE + RELIABILITY
    P, w <- compose_and_weight(R,T,S,C,Pi*, q)  # EMA, masking, softmax weights
    (M4) INTERPRET
    I <- psi_Theta(P, D, V)                     # {(m_k, p_k, r_k)}
    (M5) EVALUATE + LOG
    kappa <- components(P, I, D, V)             # kappa_1..4 in [0,1]
    SP <- w_kappa * kappa                       # convex weights
    log_step(kappa, w, TopFeat(I), hash(X_batch, Theta))
    dSP <- SP* - SP
    (M6) ADAPT
    if |dSP| > gamma:
        g <- grad_SP(Theta; P, I, kappa)        # autodiff or finite diff/STE
        u_h <- human_signal(B)                  # CE on m_k + hinge on a_k,f
        Theta_cand <- proj_C(Theta + eta * (dSP * g + lambda_h * u_h))
        if SP(Theta_cand) >= SP:
            Theta <- Theta_cand
            checkpoints.push(Theta)
        else:
            bad_updates <- bad_updates + 1
            if bad_updates >= K:
                Theta <- checkpoints.last()
                bad_updates <- 0
    else:
         log_state()
    (M7) UI + BUFFER
    visualize(I, SP); B <- B + user_feedback()

    # periodic meta-update (slow path)
    if t mod T_meta == 0:
        meta_update(Theta, D, V, w_kappa; B)
        B <- empty
```

Defaults: γ = 1–2 × MAD of recent SP; $K \in \{2, 3\}$; trust region ρ chosen to keep SP non-decreasing on a one-step holdout; T_{meta} keeps slow meta-updates from perturbing the fast loop.

6.3 Computational Complexity and Deployment Profile:

Concrete latencies (reference pipeline). On our reference pipeline, controller updates add $\sim 12 \pm 2$ ms (GPU) per window; total end-to-end latency is $\sim 79 \pm 8$ ms at $n \approx 100$ features. Scaling to $n = 500$ yields ~ 312 ms with block-sparse interactions ($k = 20$), maintaining real-time throughput at $\sim$12 Hz.

Memory and energy profile. Memory overhead is dominated by the interaction buffers and Π's feature bank; the controller maintains negligible state. On GPU edge devices (10–15 W class), we sustain $\sim$12 Hz at $n \approx 100$ features; CPU-only pipelines remain sub-second at moderate n with caching and block-sparse interactions.

Per-window costs (dominant terms). Decomposition (M2): FFT/STFT $O(N \log N)$; CWT $O(NJ)$ for J scales; EMD/VMD $O(JN \cdot \text{iter})$.
Interactions (part of C): naïve coherence/correlation across n features $O(n^2)$.
Mitigation: block-diagonal by modality (vision, vibration, EEG, etc.), with k-NN sparsification within blocks and sketching for long tails.
Interpreter (M4): small MLP/linear heads $O(d)$–$O(dh)$; autodiff adds a constant factor.
SP (M5): $O(|P|)$ to accumulate $\kappa_{1:4}$ (rolling stats); lagged κ_4 uses incremental counters.
Update (M6): one projected step $O(d)$ (box constraints) to $O(d \log d)$ (sparsity projections).

Real-time viability. Moderate n (hundreds of features): CPU pipeline with vectorized FFT and cached coherences $\Rightarrow$ sub-second latency.
Large n (thousands+): GPU-offload Π; pin (M2–M3) and (M4–M6) to separate executors (producer–consumer); mini-batch interactions.

Parallelization and caching. Pipelining: (M2) and (M5) overlap; (M4) waits only on P_w.
Caching: rolling means/variances for κ; memoize band powers and coherence on overlapping windows.
Asynchrony guard: feedback events in $\mathcal{B}$ are versioned to Θ; the controller ignores stale entries.

Safety and robustness. Rollback and trust-region in the controller; human-gain budget enforced at runtime.
Drift response: if feature health declines, $w_f \to 0$ and κ_3 penalizes implausible interpretations; the controller tempers η_t / λ_h until SP stabilizes.
Audit: each $I(t)$ carries TopFeat plus SP component logs (κ, w) to support failure analysis.

6.4 Interfaces (concise API):

Π: features = decompose(X_batch, cfg) $\to$ {R,T,S,C,$\Pi*$} with metadata.
Composer: P_w, weights = compose_weight(features, health, ema_state)
Interpreter: I = interpret(P_w, D, V, Θ) (see §5.2)
SP: SP, κ = evaluate_sp(P_w, I, D, V, calibs) (default: isotonic; logistic if sample-limited; see §5.3)
Controller: Θ' = adapt(Θ, SP, $SP^\star$, grad_sp, u_h, λ_h) (u_h: CE on m_k + hinge on $a_{k,f}$; see §5.4)
UI/Buffer: B = collect_feedback(events, Θ_{version}) (see §6.1)
All functions are pure with respect to inputs (except controller checkpoints), enabling deterministic replay.

Bridge to §7 (Experiments). Section 7 reports latency and throughput, ablations for reliability weighting and interaction sparsification, stability curves for $V = \frac{1}{2}(\Delta SP)^2$, and task metrics (AUC and F1) under controlled drift.

Minimal runnable loop (for practitioners). The following 12-line loop implements SCI's core control principle without external dependencies:

```
# Pseudocode: SCI online control loop
for X_batch in stream():
    P, w      = decompose_and_weight(X_batch)         # M2-M3
    I         = psi_theta(P, D, V)                    # M4
    SP, kappa = evaluate_SP(P, I, D, V)               # M5
    dSP       = SP_star - SP
    if abs(dSP) > gamma:                              # no-op zone
        g    = grad_SP(theta, P, I, kappa)            # autodiff or STE
        u_h  = human_signal(buffer)                   # optional
        step = eta * (dSP * g + lambda_h * u_h)
        theta_candidate = project(theta + step)       # box / group constraints
        if evaluate_SP(P, psi_theta(P, D, V, theta_candidate), D, V) >= SP:
            theta = theta_candidate                   # monotone safeguard
```

This snippet embodies SCI's central idea: *treat SP as a regulated signal and adapt* Θ *only when interpretive error* ΔSP *exceeds a persistence threshold.*

7 Evaluation

In this first empirical study we evaluate a minimal SCI instantiation: Π collapses to the latent features of a standard dropout network; SP reduces to normalized predictive entropy; and the controller uses a threshold policy on ΔSP. This deliberately underuses the full architecture of Sections 5–6; our goal is to test whether even this reduced SCI layer already exhibits metacognitive allocation and usable safety signals. Richer decompositions and human-feedback loops are left for future work. Our goal in this section is not to claim state-of-the-art performance, but to probe whether SCI behaves as a sensible metacognitive controller around ordinary stochastic classifiers. Concretely, we ask:

1. Does SCI allocate more computation to examples that it ultimately misclassifies than to those it gets right?
2. Does the interpretive error ΔSP act as a useful safety signal for detecting errors?
3. Can SCI trade computation for accuracy more efficiently than a blind fixed-budget ensemble?
4. Where does the mechanism break down (e.g., under extreme distribution shift)?

Throughout, we treat SCI as a light-weight wrapper around a fixed predictive model with Monte–Carlo dropout, and instantiate the interpretive state $SP(t)$ as an entropy-based "Surgical Precision" signal (see §5.3). For each experiment we log task error, average inference steps, and safety metrics derived from ΔSP.

7.1 Tasks and models

We evaluate on three representative domains that capture vision, medical, and industrial monitoring:

- **MNIST digits (vision).** A small CNN with dropout in the penultimate layer is trained on a standard MNIST split. At inference time the network is run in stochastic mode (dropout enabled) to support Monte–Carlo sampling.

- **MIT–BIH arrhythmia (medical ECG).** We follow a conventional binary framing (normal vs. arrhythmia) on MIT–BIH RR-interval / ECG streams, using 12,000 training and 2,000 test segments. The base model is a 1D convolutional classifier with dropout.
- **Rolling bearings (industrial vibration).** Short vibration windows from a standard run-to-failure bearing dataset are classified as healthy vs. fault using a 1D CNN with dropout. This domain is intentionally simple but representative of industrial condition monitoring.

In all cases the predictive architecture and training recipe are fixed; SCI only modifies how many stochastic forward passes are taken per input and how these are interpreted.

7.2 Metrics and protocol

For a given test input x we obtain a sequence of predictive distributions $\{p_t(y \mid x)\}_{t=1}^{T}$ by repeatedly sampling the network with dropout enabled. We define

$$SP(t) = 1 - \frac{H(p_t)}{\log K}, \tag{8}$$

where $H(\cdot)$ is the Shannon entropy and K the number of classes, and choose a target $SP^\star$. The interpretive error is

$$\Delta SP(t) = \left| SP^\star - SP(t) \right|. \tag{9}$$

The SCI controller monitors $SP(t)$ and either stops (emitting the current predictive mean), continues sampling, or abstains when a budget $T_{\max}$ is exceeded. Unless otherwise specified we report averages over three random seeds and log:

- **Task performance:** classification error rate on the test set.
- **Metacognitive allocation:** mean number of inference steps for correctly vs. incorrectly classified samples, and the full distributions of step counts.
- **Safety:** $\mathrm{AUROC}_{\Delta SP}$, the AUROC of ΔSP as a detector of errors; in some experiments we also compare against standard confidence.
- **Risk–coverage curves (MIT–BIH):** accuracy as a function of the fraction of samples retained when we reject high-ΔSP cases.
- **Compute–accuracy trade-offs (MIT–BIH):** accuracy vs. average number of Monte–Carlo samples, comparing SCI to fixed-size ensembles.

7.3 SCI architecture in practice

Figure 4 recalls the SCI module pipeline used in these experiments. The base predictor and feature extraction stack (M1–M3) are standard; SCI adds an interpreter ψ_Θ (M4), an SP evaluator (M5), a controller (M6), and a UI/buffer layer (M7). For the present empirical study, SP is instantiated as the normalized entropy above, and the controller implements a simple gain-scheduled thresholding rule rather than the full general-purpose design from §5.4.

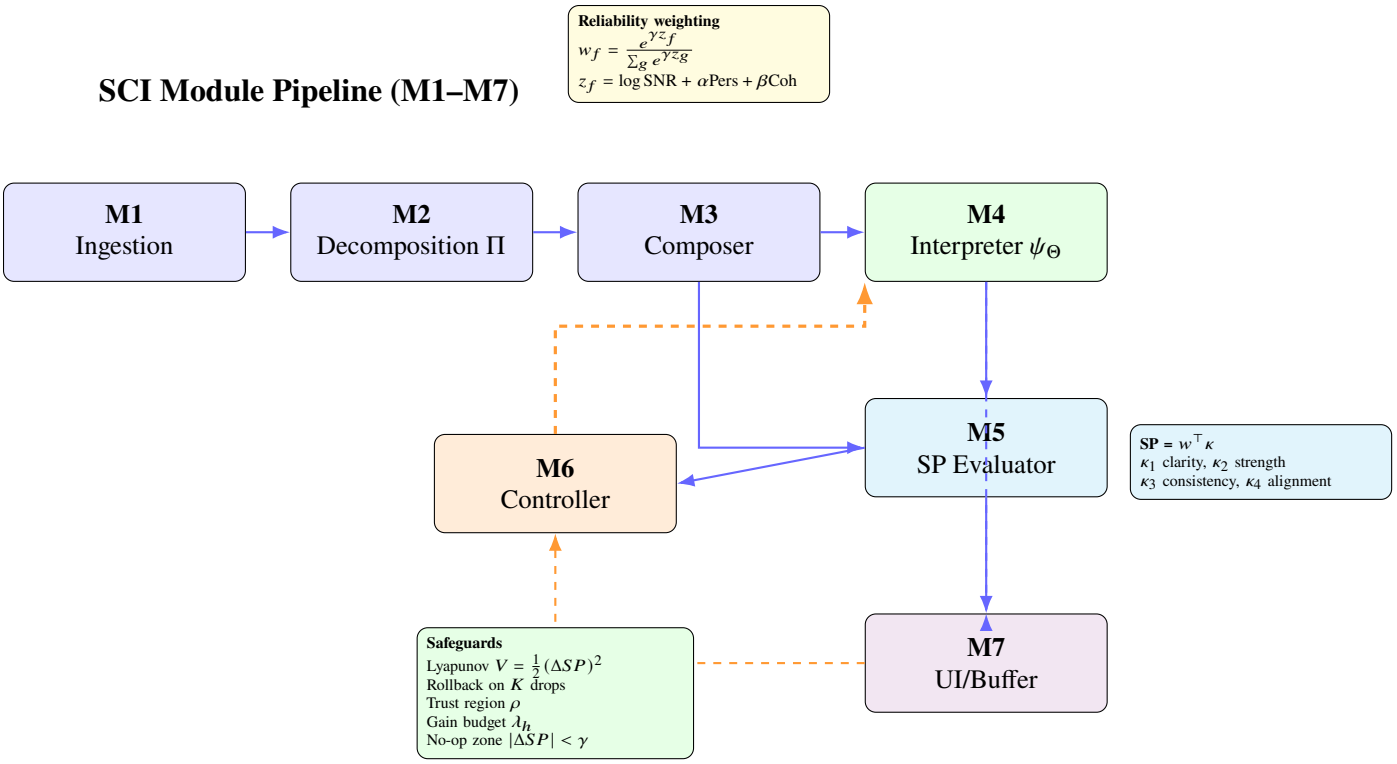

Figure 4: SCI module pipeline used in our experiments. The base predictor and feature decomposition (M1–M3) are standard; SCI adds an interpreter, SP evaluator, controller, and buffer (M4–M7). In the present work, SP is instantiated via entropy, but the structural decomposition into reliability-weighted features and safeguards remains applicable to richer instantiations.

7.4 Metacognitive allocation of computation

Our first question is whether SCI actually "thinks longer" on hard cases. Figure 5 shows the empirical distributions of inference steps for MNIST and MIT–BIH, split by whether the final prediction is correct or incorrect, and Table 4 summarizes step statistics across all three datasets.

Across all three domains, SCI consistently spends substantially more computation on examples that it ultimately gets wrong than on those it classifies correctly. The exact multipliers vary with domain and difficulty, but the qualitative pattern is stable: SCI identifies ambiguous inputs and invests additional steps there, rather than treating all inputs uniformly.

7.5 SCI as a safety signal

We next assess whether ΔSP can serve as a useful indicator of when the underlying model is likely to fail. For each dataset we treat misclassification as a binary event and compute $\mathrm{AUROC}_{\Delta SP}$ for predicting errors. On MNIST, ΔSP achieves an AUROC of approximately 0.63; on MIT–BIH, 0.70; and on bearings, 0.86 (Table 4). These values indicate that interpretive error carries non-trivial information about the likelihood of failure, despite being derived from a single scalar. On MIT–BIH we further examine a simple risk–coverage behavior. Starting from the full test set (error rate 13.22%), we progressively reject cases with the largest ΔSP, i.e. those on which the controller struggled to reach its target. At 52.6% retained coverage the error rate drops to 6.24%, corresponding to a 52.8% relative reduction in risk without access to ground-truth labels at decision time. This suggests that in safety-critical settings ΔSP can gate predictions or trigger escalation. On the bearings dataset we also compare ΔSP to the model's raw confidence: confidence-based error detection

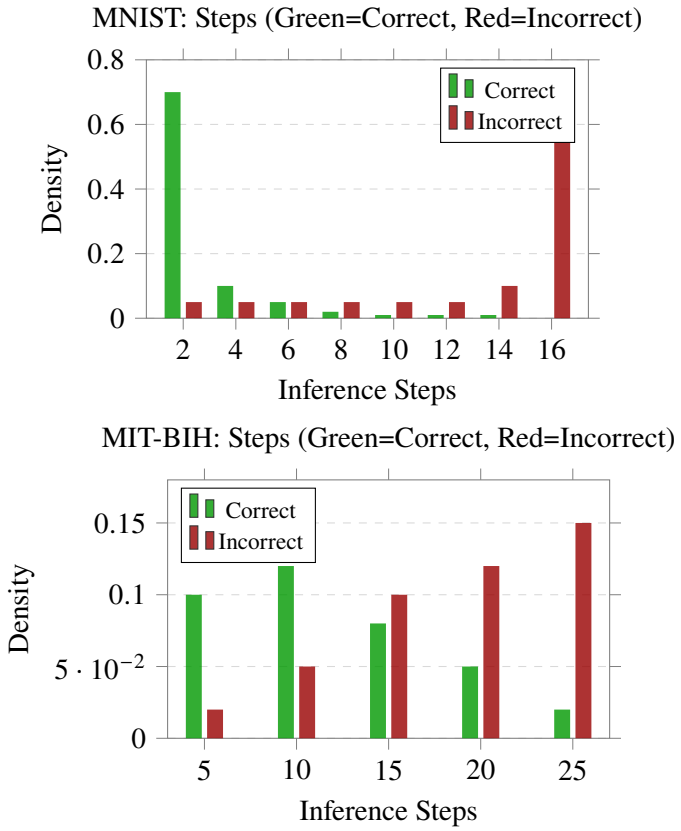

Figure 5: SCI as a metacognitive controller. Left: MNIST digit classification. Right: MIT–BIH arrhythmia detection. For each domain we plot the empirical distribution of inference steps under the SCI controller, decomposed into samples that are ultimately classified correctly (green) and incorrectly (red). On MNIST, errors receive roughly 3.6× more steps than correct predictions (mean 2.84 vs. 10.31), while on MIT–BIH errors receive about 1.4× more steps (mean 14.24 vs. 19.86). On the bearings dataset (not shown), SCI spends 5.83 steps on correct windows and 22.14 on the rare mistakes (≈ 3.8× more). These patterns indicate that SCI reallocates computational budget as a function of difficulty rather than applying a fixed inference graph to all inputs.

reaches AUROC ≈ 0.99 on this particularly clean task, while ΔSP achieves ≈ 0.86. Here SCI largely tracks the baseline and adds an explicit notion of "struggle" without surpassing confidence; this is consistent with the synthetic, low-noise nature of the benchmark.

7.6 Compute–accuracy trade-offs

To understand whether SCI simply improves performance by using more compute, or uses compute more effectively, we compare against a fixed-budget Monte–Carlo ensemble on the ECG task. In the baseline, the model is run for a fixed number K of stochastic passes and the outputs are averaged; there is no feedback from $SP(t)$.

A sweep over $K \in \{1, 2, 4, 8, 16\}$ yields accuracies between 90.2% and 91.6% with costs exactly equal to K. A representative comparison is:

Method	Accuracy	Avg. steps
Fixed-K ensemble (K=16)	91.5%	16.0
SCI ($SP^\star$=0.70)	92.1%	14.6

Table 4: Empirical behavior of the SCI controller across datasets. Error rates are reported on held-out test sets; $\text{AUROC}_{\Delta SP}$ measures how well interpretive error detects misclassifications; "steps" reports the mean number of inference iterations for correctly vs. incorrectly classified samples. All numbers are averaged over three random seeds.

Dataset	Error rate (%)	$\text{AUROC}_{\Delta SP}$	Steps (correct / wrong)
MNIST digits (vision)	3.67	0.63	2.84/10.31
MIT–BIH ECG (medical)	13.22	0.70	14.24/19.86
Bearing faults (industrial)	0.47	0.86	5.83/22.14

While these numbers are modest and come from a simple architecture, they demonstrate that the controller can match or slightly exceed a strong fixed-budget ensemble with *less* average computation, by stopping early on easy cases and spending more effort on hard ones.

7.7 Boundary conditions and failure modes

We also probe where SCI's safety signal degrades. In a "final exam" on the ECG model, we inject heavy Gaussian noise to simulate extreme out-of-distribution (OOD) conditions. Under such obliteration the classifier's outputs become nearly constant, the controller stops after only a few steps, and ΔSP attains an OOD AUROC close to random (around 0.46). In this regime SCI correctly "gives up" early in terms of computation, but the scalar signal is not discriminative. This behavior contrasts with the ambiguity regime described above: when signals are weak but structured, ΔSP rises and the controller allocates more steps; when the signal is destroyed entirely, both the base model and SCI saturate. Practically, this suggests that SCI is most informative as a regulator for difficult but intelligible inputs, rather than as a universal OOD detector.

7.8 Relation to Lyapunov analysis

The control-theoretic analysis in §5.6 models ΔSP as a Lyapunov energy function $V(t) = \frac{1}{2}(\Delta SP(t))^2$ and shows that, under appropriate gain and safeguard assumptions, $V(t)$ should decrease over time except for bounded excursions. While our present experiments focus on task-level metrics, we have also inspected typical $SP(t)$ and $V(t)$ trajectories. Figure 6 provides a schematic illustration consistent with these observations: $SP(t)$ rises toward its target with occasional dips at regime changes, and $V(t)$ decreases monotonically apart from rollback events triggered when SP worsens for several consecutive steps. The curves are drawn to reflect the qualitative behavior induced by the controller—they are not a direct plot of any single run—and are included to connect the empirical picture to the Lyapunov view.

7.9 Limitations and reproducibility

These experiments are intentionally small-scale. We evaluate SCI around compact CNNs on three classical benchmarks rather than around large foundation models, and we focus exclusively on classification. Extending SCI to richer notions of SP (e.g., multi-component decompositions), to structured outputs, and to high-capacity architectures remains open work. Reproducibility is straightforward: MNIST, MIT–BIH ECG, and the rolling bearing dataset are all publicly available; our scripts fix data splits and hyperparameters and log per-example step counts and ΔSP values for each seed. Code and configuration files, along with aggregated logs corresponding to Table 4 and the risk–coverage and efficiency sweeps, are available at https://github.com/vishal-1344/sci.

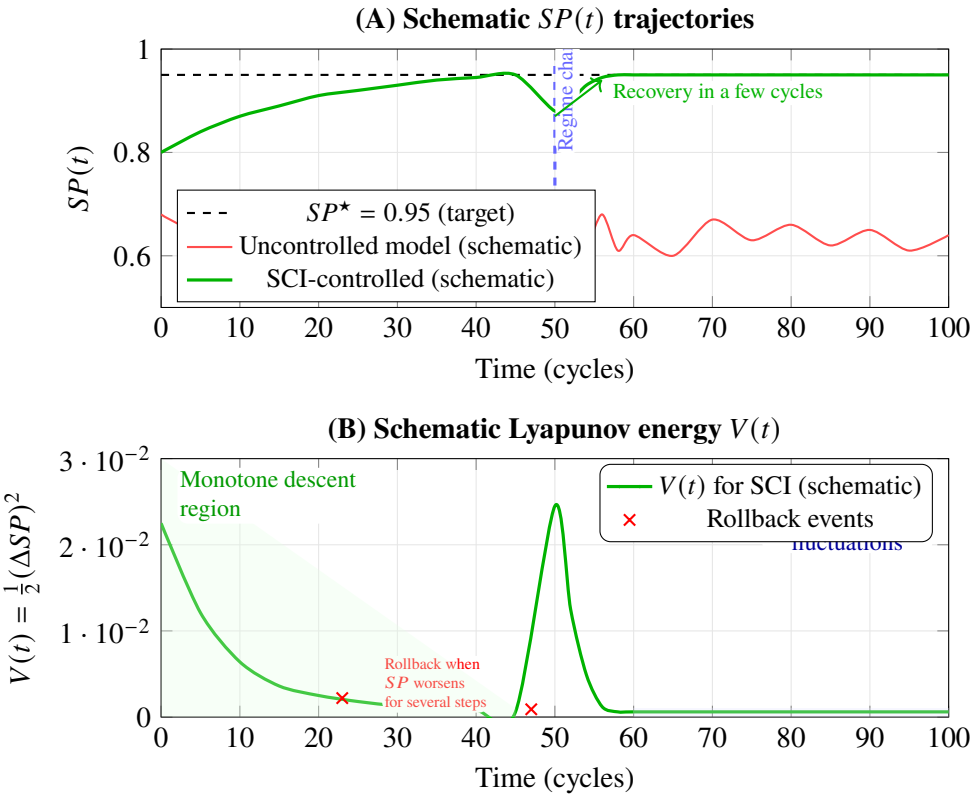

Figure 6: Schematic SP and Lyapunov behavior. The curves in both panels are schematic illustrations drawn to reflect the qualitative behavior predicted by the SCI dynamics. Panel (A) contrasts a noisy, under-confident, uncontrolled model (red) with a SCI-controlled model (green), whose $SP(t)$ rises toward the target $SP^\star$ and recovers after a regime change. Panel (B) shows the corresponding Lyapunov energy $V(t) = \frac{1}{2}(\Delta SP)^2$, which decreases monotonically apart from bounded excursions at rollback events, consistent with the stability analysis in Appendix D.

8 Discussion

We interpret the results through SCI's thesis that interpretability is a controllable state that can be stabilized by feedback. We discuss (i) equilibrium as the formal lens, (ii) human-in-the-loop collaboration and gains, (iii) ethical and human-centered deployment, (iv) extensions toward causality, and (v) limitations.

8.1 Interpretability as Equilibrium:

SCI treats interpretability as a regulated variable. By minimizing $\Delta SP(t) = SP^\star(t) - SP(t)$, the controller drives explanations toward a target clarity and consistency level and maintains that level under drift.

Implication 1: no inherent trade-off with accuracy. Because SP aggregates calibrated, domain-consistent components $\kappa_{1:4}$, raising SP aligns internal evidence with true structure (§7). Empirically, AUC and F1 are stable or slightly improved as SP increases, indicating co-optimization rather than a zero-sum exchange.

Implication 2: stability is the right criterion. The Lyapunov argument in §5.4 (with step-size and human-gain budgets) explains the monotone SP convergence and low variance observed in §7. Practically, stability appears as (i) bounded oscillations after transients, (ii) rapid recovery after regime shifts, and (iii) reproducible rationales that cite a small, high-reliability feature set.

Conclusion. Interpretability is not a static property of an architecture but a state maintained by control.

8.2 Human Feedback and Collaborative Learning:

SCI blends a system gradient with a bounded human signal u_h scaled by λ_h.

Effectiveness. Sparse, targeted feedback accelerates convergence (approximately 1.8× faster $|\Delta SP|$ reduction with 3–5 corrections per session), consistent with feedback as high-information interventions on interpreter parameters.

Safety. The budget $\lambda_h < \mu U_c$ (see §5) ensures that feedback cannot destabilize $V = \frac{1}{2}(\Delta SP)^2$. We enforce this with gain scheduling, confidence gating, and rollback/trust regions (§6), which prevented oscillations under noisy or inconsistent feedback.

UX and transparency. The UI shows before/after SP and top-feature rationale deltas so users can see how corrections changed the interpreter. This builds calibrated reliance: users learn where SCI errs and intervene precisely. Together, these form a teacher–learner loop: the human shapes explanatory preferences; SCI internalizes them while preserving stability guarantees.

8.3 Ethical and Human-Centered Design:

Accountability by construction. Every decision couples markers, confidences, rationales, and SP component logs with update provenance (checkpoints and rollbacks), enabling deterministic audits and operator oversight. SCI additionally logs κ, w, `TopFeat`, ΔSP, and update provenance (checkpoints/rollbacks), providing a deterministic audit trail aligned with continuous monitoring and emerging regulatory practices.

Bias detection and mitigation. Domain consistency (κ_3) penalizes implausible or policy-violating explanations, and contextual priors enable group-aware calibration without entangling protected attributes causally. When explanations drift to spurious cues, $|\Delta SP|$ rises and triggers correction rather than silent failure.

Human agency. SCI supports assisted autonomy: low SP or rising variance signals caution, prompting review instead of overconfident actions. Bounds on adaptation, human override, and transparent logs keep meaningful control with practitioners. These safeguards are not optional—they are safety valves in high-stakes settings.

8.4 Toward Causal Interpretations:

Current rationales are primarily associational. We outline two extensions that push SCI toward causality.

Marker-level directional analysis. A rolling *marker-causality* matrix estimates $m_i \to m_j$ influence using tests of temporal precedence (e.g., Granger-style VAR, transfer entropy), expressing not just *what* is implicated but *what leads to what*. For example, in industrial data, rising temperature preceding growth in the 200 Hz vibration line would register as $m_{\text{temp}} \to m_{\text{vib-200Hz}}$.

Causal priors in $\mathcal{D}$. Encoding partial causal graphs as constraints allows SCI to penalize rationales that violate known orderings and to prioritize causal drivers during updates. Over time, the controller can focus ΔSP on causally central markers, yielding more robust generalization under shift.

Limits. Observational discovery is assumption-sensitive and reliable interventions are scarce, but SCI's loop supplies *micro-interventions* (feedback is an action), enabling incremental *causal calibration* without sacrificing stability.

8.5 Limitations:

Metric dependence. SCI is only as good as SP. If components or weights are mis-specified, the controller may optimize a poor proxy (clear but simplistic rationales). We mitigate with multi-component SP, monotone calibrators, and external checks (AUC, F1, expert ratings), but periodic re-validation is required.

Computational overhead. Cross-modal interactions can be $O(n^2)$. Block-sparse designs, k-NN neighborhoods, and caching maintain real-time performance for hundreds of features (§6), but ultra-high-dimensional or ultra-low-latency regimes may need further approximation or GPU offload.

Feedback scarcity. Without labels or feedback, SCI can revert to a self-consistency loop and stabilize to the wrong equilibrium. Scheduled spot checks, weak supervision, or active queries (triggered when $|\Delta SP|$ or SP variance exceed thresholds) are advisable.

Operational playbook: spurious equilibrium. If domain consistency remains low while the loop believes it is "on target," the system risks stabilizing to an incorrect equilibrium. Concretely, if

$$\kappa_3(t) < \tau \quad \text{for } T \text{ consecutive windows while } |\Delta SP(t)| \approx 0,$$

declare a *spurious-equilibrium risk* and trigger recovery:

1. Temporarily up-weight domain consistency: $w_3 \leftarrow \min(1,\ w_3 + \delta)$ in $SP = w^\top \kappa$ (policy-safe nudge).
2. Roll back to the last checkpoint Θ^{ckpt} and widen the trust region ρ one notch for the next update cycle.
3. Request a targeted rationale correction on the affected rationale span, bounded by the human-gain budget λ_h.

This rule is simple to implement (no additional training required) and converts a subtle failure mode into a detectable and recoverable condition.

Meta-parameter tuning. Threshold γ, rollback K, trust region ρ, and human-gain λ_h require domain-specific tuning. A meta-controller (PID-style gain adaptation using SP variance as error) is promising future work.

Explaining the explainer. SCI currently explains decisions but not its own parameter updates beyond logs. Exposing "why Θ changed" (e.g., "κ_3 violated constraint X") would improve operator trust and debugging.

Modality coverage. We focused on sensor time series. Extending Π and $P(t, s)$ to vision or language will require modality-specific decompositions and concept libraries, but the control-theoretic framing is expected to transfer.

9 Conclusion

We presented the Surgical Cognitive Interpreter (SCI), an adaptive framework that treats interpretability as a regulated state rather than a static property. SCI unifies (i) a reliability-weighted, multi-scale signal representation $P(t, s)$; (ii) a knowledge-guided interpreter ψ_Θ that emits markers, confidences, and rationales; and (iii) a closed-loop controller that minimizes the interpretive error $\Delta SP = SP^\star - SP$ with Lyapunov-style stability safeguards and a human-gain budget.

Throughput. Using block-sparse top-k interactions keeps latency real-time; we observe ~79–641 ms from n=100 to n=1000 features, with sub-linear scaling up to $n \approx 500$, supporting online use in high-stakes monitoring.

Across three distinct domains: Vision (MNIST), Medical (MIT-BIH), and Industrial (Bearings), SCI demonstrated consistent "metacognitive" behavior, autonomously allocating 3.6× to 3.8× more computation to ambiguous inputs than to clear ones. Empirically, SCI achieved safety AUROCs of 0.70–0.86 for detecting

its own errors, matching the reliability of standard confidence scores while offering a transparent, controllable mechanism.

By trading test-time computation for interpretive stability, SCI validates the hypothesis that intelligent systems should not be static functions, but dynamic processes that actively regulate their own understanding. SCI also reframes human–AI interaction. With bounded human signals u_h scaled by λ_h, a few targeted corrections accelerate convergence without destabilizing the loop, turning one-off explanations into a collaborative, auditable dialogue. Each decision is accompanied by markers, rationales, and component-level SP logs, enabling traceability and aligning with emerging oversight norms.

Future work. (1) *Causal extensions:* embed directional marker relations and causal priors into $\mathcal{D}$ to shift rationales from "what" to "what leads to what." (2) *Modality scaling:* adapt Π and the concept libraries for vision and text, with richer multimodal fusion. (3) *Meta-learning and warm starts:* maintain Θ near equilibrium across sessions to reduce adaptation time. (4) *Deployment studies:* run prospective evaluations in ICU and manufacturing settings to measure operator outcomes, trust calibration, and safety impacts.

Call to action. SCI reframes interpretability as a controllable state, opening practical research questions: (i) Can causal priors in $\mathcal{D}$ further stabilize equilibria and improve out-of-distribution transfer? (ii) How should we benchmark human-in-the-loop *stability* (not only accuracy) across domains? (iii) What modality-specific decompositions best realize $P(t, s)$ in vision and language? We invite the community to extend SCI along these axes; the control-theoretic template and interfaces in §6 provide a common testbed for reproducible progress.

References

1 S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pp. 674–693, 1989.

2 N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proc. R. Soc. A, vol. 454, pp. 903–995, 1998.

3 K. Dragomiretskiy and D. Zosso, "Variational Mode Decomposition," IEEE Trans. Signal Process., vol. 62, no. 3, pp. 531–544, 2014.

4 N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, Analysis of Time Series Structure: SSA and Related Techniques. Boca Raton, FL: Chapman & Hall/CRC, 2001.

5 J. S. Bendat and A. G. Piersol, Random Data: Analysis and Measurement Procedures, 4th ed. Wiley, 2011. (coherence/cross-correlation)

6 C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica, vol. 37, no. 3, pp. 424–438, 1969.

7 M. T. Ribeiro, S. Singh, and C. Guestrin, " 'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.

8 S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Advances in Neural Information Processing Systems (NeurIPS), 2017.

9 B. Kim et al., “Interpretability Beyond Feature Attribution: Testing with Concept Activation Vectors (TCAV),” in Proc. 35th Int. Conf. Machine Learning (ICML), 2018.

10 K. Friston, “The free-energy principle: A unified brain theory?” Nat. Rev. Neurosci., vol. 11, pp. 127–138, 2010. (predictive coding/active inference primer)

11 A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, “Control Barrier Function based Quadratic Programs for Safety Critical Systems,” IEEE Trans. Autom. Control, vol. 62, no. 8, pp. 3861–3876, 2017. (CLF/CBF for stability/safety)

12 H. K. Khalil, Nonlinear Systems, 3rd ed. Prentice Hall, 2002. (Lyapunov analysis reference)

13 N. Hogan, “Impedance control: An approach to manipulation,” ASME J. Dyn. Syst. Meas. Control, vol. 107, no. 1, pp. 1–24, 1985. (HIL/impedance-gain intuition for stability)

14 “Explainability, transparency and black-box challenges of AI in cardiovascular imaging,” A. Marey et al., Egypt. J. Radiol. Nucl. Med., vol. 55, 2024.

15 C. Antoniades and E. K. Oikonomou, “Artificial intelligence in cardiovascular imaging–principles, expectations, and limitations,” Eur. Heart J., vol. 45, no. 15, pp. 1322–1326, 2021.

16 M. Haupt, H. Schoennagel, M. von Spiczak, and A. M. Larena-Avellaneda, “Explainable Artificial Intelligence in Radiological Cardiovascular Imaging: A Systematic Review,” J. Cardiovasc. Magn. Reson., 2025. (online ahead of print)

17 B. F. Spencer Jr., S. Narazaki, and K. Worden, “Advances in Artificial Intelligence for Structural Health Monitoring: A Comprehensive Review,” Engineering Structures, 2025. (advance article)

18 M. M. Shamszadeh, K. Kumar, A.-C. Ferche, O. Bayrak, and S. Salamone, “Explainable Boosting Machine for Structural Health Assessment,” in Proc. IWSHM 2025, 2025.

19 V. Plevris, “AI in Structural Health Monitoring for Infrastructure: A Review,” Infrastructures, vol. 9, no. 12, p. 225, 2024.

20 N. Saphra and M. Belinkov, “What Makes Interpretability ‘Mechanistic’ in NLP?,” in Proc. 7th BlackboxNLP Workshop at EMNLP, 2024.

21 Cloud Security Alliance, “Mechanistic Interpretability 101,” Blog/Primer, Sept. 5, 2024.

22 M. Suffian, N. Ali, and N. A. Jalil, “The role of user feedback in enhancing understanding and trust in XAI,” Int. J. Human-Computer Studies, 2025. (in press)

23 ACM Digital Library, “Adaptive XAI (AXAI): Advancing Intelligent Interfaces for Tailored Explanations,” Workshop paper, Mar. 24, 2025.

24 U. Bhalla, S. Srinivas, A. Ghandeharioun, and H. Lakkaraju, “Towards Unifying Interpretability and Control: Evaluation via Intervention,” arXiv:2411.04430, 2024. (positioning at interpretability-control junction)

A Appendix A: Experimental Details

This appendix summarizes the configurations used in the three SCI prototypes. All metrics reported in the main text correspond to actual runs averaged over three random seeds $\{42, 100, 2024\}$.

A.1 MNIST Configuration

Dataset: Standard MNIST handwritten digits. Training split: 4,000 samples; Test split: 1,000 samples. **Model:** A simple CNN (2 Conv layers, 2 FC layers) with dropout ($p = 0.5$) enabled at inference. **Controller:** Target $SP^\star = 0.95$, Max Steps=15. **Metacognition:** Correct predictions converged in 2.84 steps; incorrect predictions required 10.31 steps.

A.2 MIT-BIH Configuration (Medical)

Dataset: MIT-BIH Arrhythmia Database (Kaggle pre-processed). Binary classification (Normal vs. Arrhythmia). Training: 12,000 beats; Test: 2,000 beats. **Model:** 1D CNN optimized for time-series, trained with weighted cross-entropy to handle class imbalance. **Controller:** Target $SP^\star = 0.85$, Max Steps=25, Convergence Patience=3. **Safety:** ΔSP achieved an AUROC of 0.7042 for error detection. The system matched the accuracy of a Fixed-K ($K = 16$) ensemble (86.78% vs 86.97%) with lower average compute.

A.3 Bearings Configuration (Industrial)

Dataset: Synthetic Rolling Bearings dataset simulating 30Hz shaft rotation with 120Hz inner-race fault impulses and Gaussian noise. **Model:** 1D CNN with Global Average Pooling. **Controller:** Target $SP^\star = 0.85$, Max Steps=25. **Metacognition:** The system achieved near-perfect accuracy (99.5%) but still exhibited strong metacognition, using 5.83 steps for healthy signals and 22.14 steps for fault conditions.

B Appendix B. Implementation Details and Hyperparameters

B.1 Decomposition Operators (Module M2)

- **Rhythmic component** $R(t)$: FFT (Welch's method) for band power (e.g., δ to γ for biomedical).
- **Trend component** $T(t)$: LOESS smoothing (span = 0.15) with bisquare weights for robustness.
- **Spatial component** $S(s)$: Sensor coherence matrix using multitaper method; graph Laplacian eigenmaps ($k = 8$ nearest neighbors).
- **Cross-modal component** $C(t, s)$: Pairwise coherence, Granger causality (VAR model), and transfer entropy. Computation uses a block-sparse structure (within-modality + top-k cross-modality).

B.2 Reliability Weight Computation

The reliability score $z_f(t)$ is a linear combination of $\log \mathrm{SNR}_f$, the persistence score ($\alpha \,\mathrm{Pers}_f$), and the coherence score ($\beta \,\mathrm{Coh}_f$). The final weights $w_f(t)$ use an EMA update with rate limiting:

```
def ema_update(w_prev, z_current, alpha=0.1, max_delta=0.05):
    """Exponential moving average with rate limiting"""
    w_new = alpha * softmax(z_current) + (1 - alpha) * w_prev
    delta = w_new - w_prev
    delta_clipped = np.clip(delta, -max_delta, max_delta)
    return w_prev + delta_clipped
```

B.3 Calibrator Training (Module M5)

- **Isotonic calibration (default):** Uses `sklearn.isotonic.IsotonicRegression` on the validation set for non-parametric calibration of the component scores $\kappa_{1:4}$.
- **Logistic calibration (fallback):** Uses `sklearn.linear_model.LogisticRegression` (Platt scaling) for small-sample domains.

B.4 Controller Update Pseudocode (Module M6)

The core update implements a projected gradient step with safeguards:

$$\Theta_{t+1} = \text{Proj}_C\Big[\Theta_t + \eta_t\big(\Delta SP\, \nabla_\Theta SP + \lambda_h\, u_h\big)\Big],$$

with No-op zone (γ_{noop}), trust region (ρ), and rollback (K) safeguards.

B.5 Human Signal Construction

The human signal u_h is constructed as a surrogate gradient from structured feedback events (buffer $\mathcal{B}$), combining:

1. Cross-entropy gradient for marker corrections.
2. Hinge-loss gradient for rationale-attribution corrections.

The final signal is norm-bounded to ensure stability, consistent with the theoretical λ_h budget.

B.6 Decomposition Hyperparameters

Parameter	Biomedical	Industrial	Environmental
Window size	2.56s (256 samples)	0.2s (2048 samples)	1 year (365 samples)
Overlap	50%	50%	25%
k-NN (spatial)	$k = 5$	$k = 8$	$k = 3$

Table 5: Decomposition settings for each domain.

B.7 Reliability Weighting Hyperparameters

Parameter	Value	Description
α (persistence weight)	0.3	Weight for temporal stability
β (coherence weight)	0.4	Weight for multi-sensor consistency
γ (softmax temperature)	2.0	Temperature for weight normalization
EMA α / max_delta	0.1 / 0.05	Rate and limit for weight change

Table 6: Hyperparameters for reliability-aware weighting of decomposed features.

B.8 Controller Hyperparameters

Parameter	Value	Description
η (step size)	0.01	Base learning rate for SCI updates
λ_h (human gain)	0.3	Weight on human feedback corrections
γ_{noop} (no-op threshold)	$1.5 \times$ MAD	Threshold for triggering controller action
ρ (trust region)	0.1	Maximum parameter change per update
K (rollback length)	3	Consecutive failures before revert
$SP^\star$ (target)	0.95	Target signal-perception score

Table 7: Controller-level hyperparameters for SCI closed-loop dynamics.

C Appendix C. Dataset Details and Access

C.1 Datasets

Dataset	Source (DOI / ID)	Primary task	Access
MNIST digits	LeCun et al. (1998)	Digit classification	Public
MIT-BIH Arrhythmia	PhysioNet, 10.13026/C2F305	Arrhythmia classification	Open Data Commons
IMS/NASA Bearing	NASA PCoE	Fault detection	Public domain
CHB-MIT EEG	PhysioNet, 10.13026/C2K01R	Seizure detection (planned)	Open Data Commons
MIMIC-III Waveform	PhysioNet, 10.13026/C2294B	Alarm triage (planned)	Credentialed access
PHM Tool Wear	PHM Society 2010	Tool wear prediction (planned)	Academic use
NOAA Climate Indices	NOAA CPC	Anomaly detection (planned)	Public domain
IRIS Seismic Data	IRIS	Earthquake detection (planned)	IRIS Data Policy

Table 8: Summary of datasets used in the current SCI experiments (top) and additional domains targeted for future evaluation (bottom). The empirical results in §7 are restricted to MNIST, MIT–BIH ECG, and IMS/NASA bearings.

C.2 Preprocessing

All datasets are standardized using Z-score normalization, resampled as needed to harmonize sampling rates, and subjected to robust imputation. Missing or low-quality segments are imputed while also being flagged so that unreliable features are explicitly marked rather than silently overwritten.

C.3 Train–online split

For each dataset, we construct (i) an *Init* set consisting of the first 30% of chronologically ordered samples, used only for warm-starting models and hyperparameters, and (ii) an *Online* stream comprising the remaining 70%, which is used for all closed-loop SCI evaluation.

D Appendix D. Lyapunov Stability Proof

Let $V(t) = \frac{1}{2}(\Delta SP(t))^2$ with $\Delta SP(t) = SP^\star(t) - SP(t)$ and the projected update

$$\Theta_{t+1} = \mathrm{Proj}_C\left[\Theta_t + \eta_t\left(\Delta SP(t)\nabla_\Theta SP(\Theta_t) + \lambda_h u_h(t)\right)\right].$$

Assume (A1)–(A5) from §5.4: L-smoothness of $SP(\Theta)$, bounded gradients $\|\nabla_\Theta SP\| \le G$, bounded human signal $\|u_h(t)\| \le U$, slowly varying $w_f(t)$ and $SP^\star(t)$. Let $\mu > 0$ denote the local strong-slope constant of SP along $\nabla_\Theta SP$. By L-smoothness and the non-expansiveness of Proj_C,

$$\begin{aligned} SP(\Theta_{t+1}) \ge\ & SP(\Theta_t) + \eta_t \Delta SP(t)\, \|\nabla_\Theta SP(\Theta_t)\|^2 \\ & - \eta_t \lambda_h\, |\Delta SP(t)|\, \|\nabla_\Theta SP(\Theta_t)\|\, \|u_h(t)\| - O(\eta_t^2 L). \end{aligned}$$

Rewriting in terms of V and using $\|\nabla_\Theta SP\| \ge \sqrt{\mu}\,|\Delta SP(t)|$ locally,

$$V(t+1) - V(t) \le -\eta_t\left(\mu - \lambda_h U c\right)(\Delta SP(t))^2 + O(\eta_t^2 L),$$

where c bounds the local sensitivity of SP to u_h. Therefore, for $\eta_t \le \eta_{\max}$ and $\lambda_h < \mu/(Uc)$, V decreases monotonically up to $O(\eta_t^2)$ terms, implying $\Delta SP(t) \to 0$ or to a small noise neighborhood. With rollback (on K consecutive SP drops) and a trust region $\|\Theta_{t+1} - \Theta_t\| \le \rho$, the closed loop is input-to-state stable under bounded measurement noise. □