Responding to Distribution Shift: When to Adapt, Freeze, or Rebuild World Models

Régis Rigaud
Independent Researcher
regis@rigaud.dev

January 2026

Abstract

World models in reinforcement learning face a fundamental challenge: when observations deviate from training distribution, should the agent adapt its inference, freeze updates to prevent corruption, or rebuild the model entirely? We propose a unified framework combining two complementary mechanisms: the **Homeostatic Hamiltonian Agent (HHA)** for regulating inference under transient noise, and the **Model Invalidity Test (MIT)** for detecting when the model's structural assumptions no longer hold. HHA modulates a friction parameter γ that saturates under noise shocks, triggering a protective freeze. MIT uses a conjunction rule with two essential signals—cumulative stress persistence and resonance collapse—plus an optional energy floor for defense-in-depth, achieving near-zero false positives under noise while maintaining 100% recall for structural changes. Together, these mechanisms implement a three-regime taxonomy: ADAPT (gradual difficulty), FREEZE (temporary danger), and REBUILD (permanent invalidity). Empirical validation across synthetic and simulated physics-based benchmarks demonstrates that MIT outperforms classical change-point detectors (CUSUM, ADWIN) which exhibit 100% false positive rates under noise bursts. Ablation studies confirm that stress persistence (P1) and resonance collapse (P3) are essential, while the energy floor (P2) provides optional defense-in-depth.

1 Introduction

World models [5, 6, 8] enable agents to plan and reason about future states, but their utility depends critically on alignment between learned dynamics and actual environment behavior. Distribution shift—whether from sensor noise, environment non-stationarity, or genuine structural change—poses a fundamental challenge: the same symptom (elevated prediction error) can arise from qualitatively different causes requiring opposite responses.

Consider an autonomous agent whose world model was trained on nominal conditions. When prediction errors suddenly spike, three scenarios are possible:

1. **Transient noise**: A sensor malfunction or environmental perturbation that will pass. Adapting the model to this noise would corrupt learned dynamics.

2. **Gradual drift**: The environment is slowly changing. The model should adapt incrementally.

3. **Structural shift**: The environment has fundamentally changed. The model's assumptions are invalid and it must be rebuilt.

Existing approaches typically treat these uniformly—either always adapting (risking corruption) or using simple thresholds (missing the distinction). We argue that robust world models require a *taxonomy of responses* matched to the nature of the shift.

Figure 1: HHA+MIT Unified Architecture

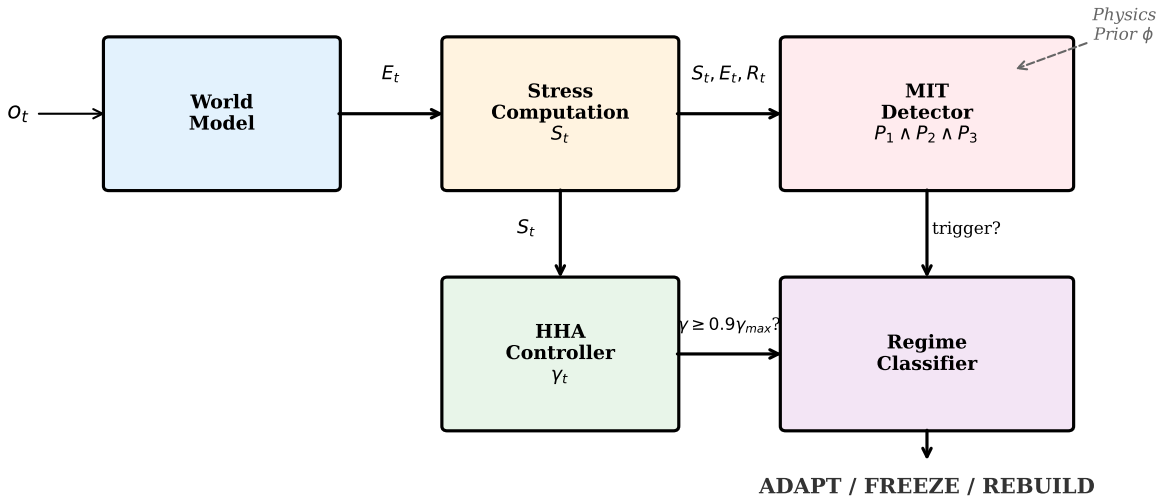

Figure 1: Unified HHA+MIT architecture. Observations o_t are processed by the world model to compute energy E_t. The stress signal S_t drives both the HHA controller (regulating friction γ) and the MIT detector (checking the three-pillar conjunction). The regime classifier outputs ADAPT, FREEZE, or REBUILD based on γ saturation and MIT triggers.

1.1 Contributions

This paper presents a unified framework for responding to distribution shift:

1. **Three-regime taxonomy**: We formalize the distinction between ADAPT, FREEZE, and REBUILD regimes based on the nature of distribution shift.
2. **Homeostatic Hamiltonian Agent (HHA)**: A friction-based controller that modulates inference difficulty, automatically saturating under noise to trigger protective freezing.
3. **Model Invalidity Test (MIT)**: A conjunction rule combining two essential signals (stress persistence and resonance collapse) with an optional energy floor, discriminating noise from structural shift with near-zero false positives and 100% recall.
4. **Empirical validation**: Comprehensive benchmarks demonstrating superiority over classical change-point detectors and confirming the necessity of each component through ablation.

2 Background and Related Work

2.1 World Models and Distribution Shift

World models learn environment dynamics $p(s_{t+1}|s_t, a_t)$ from experience, enabling planning without environment interaction [5, 6, 7]. However, these models are vulnerable to distribution shift when deployed conditions differ from training [13].

2.2 Change-Point Detection

Classical change-point detection methods monitor statistical properties of streams:

- **CUSUM** [12]: Cumulative sum of deviations from expected mean.
- **ADWIN** [2]: Adaptive windowing that detects distribution changes.
- **Energy-based OOD**: Thresholding on reconstruction error or energy [11, 9].
- **Deep Ensembles** [10]: Disagreement among ensemble members as uncertainty signal.

These methods share a critical limitation: they cannot distinguish noise from structural change. As we demonstrate empirically, all three exhibit 100% false positive rates under noise bursts.

2.3 Homeostatic Regulation

Biological systems maintain stability through homeostasis—regulatory mechanisms that resist perturbations while allowing necessary adaptation [3]. The free energy principle [4] formalizes this insight: organisms minimize prediction error (or "surprise") through a combination of perceptual inference and active regulation. We draw inspiration from these principles, implementing a controller that *resists* rapid change during noise while *permitting* gradual adaptation.

Recent work on novelty detection in RL [1] has explored similar questions using KL-divergence bounds to detect persistent distributional shifts. Our approach differs by providing an explicit three-regime taxonomy and using a conjunction of complementary signals rather than a single divergence measure.

3 Method

3.1 Problem Formulation

Let $o_t \in \mathbb{R}^d$ denote observations at time t, and let a world model produce predictions $\hat{o}_t$. Define the prediction energy:

$$E_t = \|o_t - \hat{o}_t\|^2 + \lambda \|z_t\|^2 \tag{1}$$

where z_t is the latent state and λ is a regularization parameter.

The challenge is to determine, from the stream of energies $\{E_t\}$, whether deviations represent:

- **Difficulty**: Manageable increase requiring harder inference
- **Danger**: Temporary spike requiring protective freezing
- **Invalidity**: Permanent shift requiring model reconstruction

3.2 Inter-Temporal Energy Surprise

We introduce the *stress signal* S_t capturing inter-temporal energy dynamics:

$$S_t = w_\delta \cdot |\Delta E_t| + w_c \cdot [E(z_{t-1}, o_t) - E_{t-1}]^+ \tag{2}$$

where $\Delta E_t = E_t - E_{t-1}$, $E(z_{t-1}, o_t)$ is the cross-temporal energy, and $[\cdot]^+ = \max(0, \cdot)$.

The stress signal captures not just current prediction difficulty but *surprise*—how much the energy landscape has shifted between timesteps.

Algorithm 1 HHA Controller Update

Require: Observation o_t, previous state (z_{t-1}, E_{t-1})
Ensure: Updated friction γ_t, regime signal
$E_t \leftarrow \|o_t - \hat{o}_t\|^2 + \lambda\|z_t\|^2$ {Prediction energy}
$\Delta E \leftarrow |E_t - E_{t-1}|$
$E_\times \leftarrow \|z_{t-1} - o_t\|^2 + \lambda\|z_{t-1}\|^2$ {Cross-temporal energy}
$S_t \leftarrow \Delta E + \max(0, E_\times - E_{t-1})$ {Stress signal}
$\gamma_{\text{raw}} \leftarrow \gamma_{t-1} + \alpha(S_t - S^*)$
$\gamma_{\text{raw}} \leftarrow \text{clip}(\gamma_{\text{raw}}, \gamma_{\min}, \gamma_{\max})$
$\gamma_t \leftarrow \beta\gamma_{t-1} + (1-\beta)\gamma_{\text{raw}}$ {Smoothed update}
if $\gamma_t \geq 0.9 \cdot \gamma_{\max}$ **then**
  return γ_t, FREEZE
else
  return γ_t, ADAPT
end if

3.3 Homeostatic Hamiltonian Agent (HHA)

HHA maintains a friction parameter $\gamma_t \in [\gamma_{\min}, \gamma_{\max}]$ that modulates inference difficulty. The update rule is:

$$\gamma_t = \beta\gamma_{t-1} + (1-\beta) \cdot \text{clip}\left(\gamma_{t-1} + \alpha(S_t - S^*), \gamma_{\min}, \gamma_{\max}\right) \tag{3}$$

where $S^* = \mu_S + k\sigma_S$ is the target stress (calibrated from nominal operation), α is the gain, and β is the smoothing factor.

Proposition 1 (Saturation under shock)**.** *When $S_t \gg S^*$ persistently, $\gamma_t \to \gamma_{\max}$, triggering protective freezing.*

Proof. If $S_t - S^* > 0$ for all $t > t_0$, then the unsmoothed update $\gamma_{t-1} + \alpha(S_t - S^*)$ exceeds γ_{t-1}. With clipping at $\gamma_{\max}$ and $\beta < 1$, repeated application drives $\gamma_t \to \gamma_{\max}$. □

3.4 Model Invalidity Test (MIT)

While HHA protects against noise, it cannot determine when the model itself has become invalid. MIT addresses this through a *multi-signal conjunction* combining two essential pillars with an optional safety margin:

Definition 1 (MIT Trigger Condition)**.** *Model reconstruction is triggered when all three pillars are satisfied for N consecutive timesteps:*

$$P_1 : \Sigma_S = \sum_{i=t-W+1}^{t} \left[\frac{S_i - \mu_S}{\sigma_S}\right]^+ > \theta \tag{4}$$

$$P_2 : \bar{E}_W > \varepsilon \tag{5}$$

$$P_3 : \bar{R}_W < \delta \cdot R_{nom} \tag{6}$$

where W is the window size, $\theta = k_\theta\sqrt{W}$ is the stress threshold, ε is the energy floor (93rd percentile from calibration), and R is the resonance signal.

Pillar 1 (Persistence) [Essential]: Cumulative normalized stress exceeds threshold. This filters transient spikes and enables detection of gradual drifts.

Algorithm 2 MIT Detector Update

Require: Signals (S_t, E_t, R_t), buffers $(\mathcal{S}, \mathcal{E}, \mathcal{R})$ of size W
Ensure: Trigger decision, regime signal

Append (S_t, E_t, R_t) to buffers
if $|\mathcal{S}| < W$ **or** in refractory period **then**
  return false, ADAPT
end if
$\Sigma_S \leftarrow \sum_{s \in \mathcal{S}} \max(0, (s - \mu_S)/\sigma_S)$ {Cumulative stress}
$\bar{E} \leftarrow \text{mean}(\mathcal{E})$, $\bar{R} \leftarrow \text{mean}(\mathcal{R})$
$P_1 \leftarrow (\Sigma_S > \theta)$ {Persistence}
$P_2 \leftarrow (\bar{E} > \varepsilon)$ {Capacity}
$P_3 \leftarrow (\bar{R} < \delta \cdot R_{\text{nom}})$ {Validity}
if $P_1 \wedge P_2 \wedge P_3$ **then**
  persist_count $\leftarrow$ persist_count $+ 1$
else
  persist_count $\leftarrow 0$
end if
if persist_count $\geq N$ **then**
  Enter refractory period
  return true, REBUILD
end if
return false, (from HHA)

Pillar 2 (Capacity) [Optional]: Average energy exceeds calibrated floor. This provides defense-in-depth, filtering edge cases where stress accumulates but prediction quality remains acceptable. Empirically, P2 shows no differential behavior in tested scenarios but is retained as a safety margin.

Pillar 3 (Validity) [Essential]: Average resonance drops below threshold. This is the critical discriminator—resonance measures alignment with physics priors or structural assumptions.

Definition 2 (Resonance Signal)**.** *The resonance R_t measures how well observations align with physics predictions $\phi(z_t)$:*

$$R_t = \exp\left(-\frac{MSE(o_t, \phi(z_t))}{2\sigma_R^2}\right) = \exp\left(-\frac{1}{2\sigma_R^2 d}\sum_{i=1}^{d}(o_t^{(i)} - \phi(z_t)^{(i)})^2\right) \tag{7}$$

where d is the observation dimension. Using mean squared error (MSE) rather than squared norm ensures the resonance signal is scale-invariant with respect to dimension.

Under noise centered on zero, physics predictions remain valid ($o_t \approx \phi(z_t)$ in expectation), so R stays high. Under structural shift, the physics prior becomes invalid, and R collapses.

3.5 Three-Regime Taxonomy

Combining HHA and MIT yields the response taxonomy:

$$\text{Regime}_t = \begin{cases} \textsc{Rebuild} & \text{if MIT triggers} \\ \textsc{Freeze} & \text{if } \gamma_t \geq 0.9\gamma_{\max} \\ \textsc{Adapt} & \text{otherwise} \end{cases} \tag{8}$$

This hierarchy ensures that structural invalidity (requiring reconstruction) takes precedence over temporary danger (requiring freeze), which takes precedence over normal adaptation.

Figure 6: MIT Three-Pillar Signals Under Structural Shift

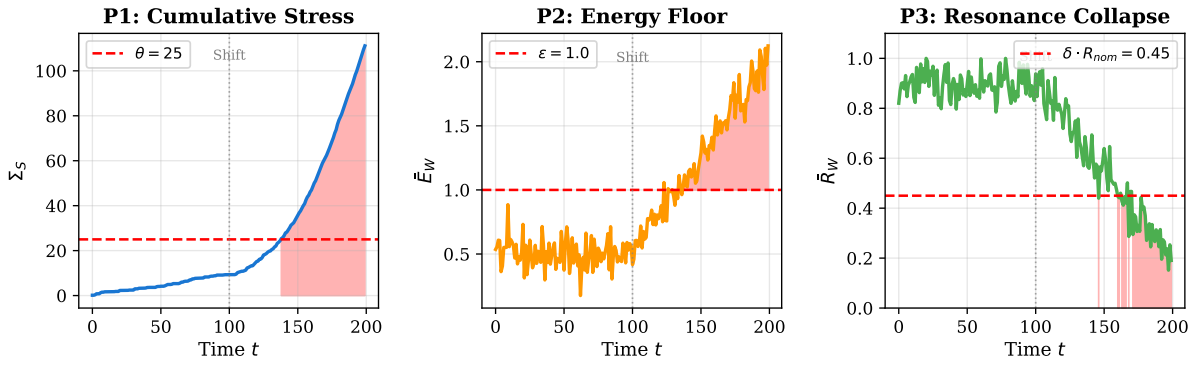

Figure 2: MIT three-pillar signals under structural shift (occurring at $t = 100$). Left: Cumulative stress Σ_S exceeds threshold θ post-shift. Center: Average energy $\bar{E}_W$ rises above floor ε. Right: Resonance $\bar{R}_W$ collapses below threshold $\delta \cdot R_{\text{nom}}$. Only when all three conditions hold simultaneously does MIT trigger reconstruction.

Figure 2: Three-Regime Taxonomy for Distribution Shift Response

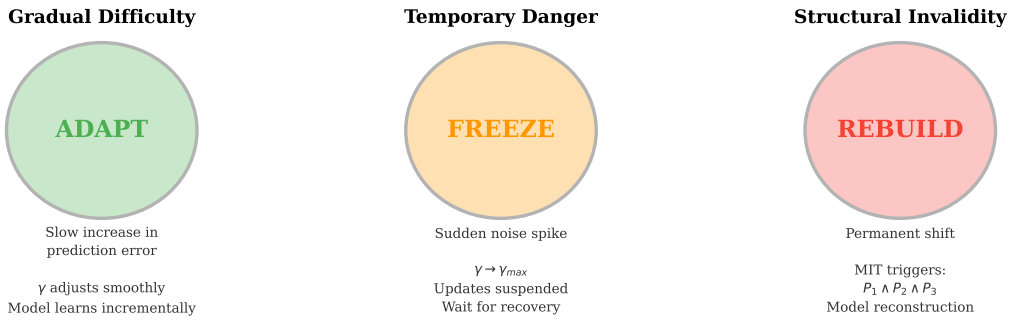

Figure 3: Three-regime taxonomy for distribution shift response. ADAPT: gradual difficulty is handled by incremental learning. FREEZE: temporary danger triggers γ saturation, suspending updates. REBUILD: structural invalidity triggers MIT, initiating model reconstruction.

4 Experiments

We validate our framework through eleven benchmarks testing each component, the unified system, and generalization to physics-based environments. All experiments use 10 random seeds with 95% confidence intervals.

4.1 Experimental Setup

Synthetic Environment: Observations $o_t \in \mathbb{R}^{32}$ are generated with controllable noise and shift:

- **Nominal**: $o_t \sim \mathcal{N}(0, 0.1^2 I)$
- **Noise burst**: $o_t \sim \mathcal{N}(0, (0.8)^2 I)$ for $t \in [t_0, t_0 + \Delta]$
- **Structural shift**: $o_t \sim \mathcal{N}(\mu_{\text{shift}}, 0.1^2 I)$ for $t \geq t_0$

Physics Prior: For MIT's resonance signal, we use $\phi(z) = 0$ (equilibrium prior).

Calibration: 100 nominal timesteps establish μ_S, σ_S, ε, and R_{nom}.

Hyperparameters: Table 1 lists all hyperparameters used in our experiments.

Table 1: Hyperparameter values used in all experiments.

Component	Parameter	Value	Description
HHA	α	20.0	Controller gain
	β	0.5	Smoothing factor
	$\gamma_{\min}, \gamma_{\max}$	0.1, 8.0	Friction bounds
	k (target)	0.1	Target stress multiplier
MIT	W	25	Window size
	N	10	Persistence threshold
	k_θ	5.0	Stress threshold multiplier
	q_ε	0.93	Energy floor percentile
	δ	0.5	Resonance drop factor
Resonance	σ_R^2	1.0	Scaling factor

4.2 B1: Panic Freeze (HHA Saturation)

Hypothesis: Under noise shock, γ should saturate to $\gamma_{\max}$.

Protocol: Inject 50-step noise burst (10× amplitude). Measure γ saturation rate (fraction of burst where $\gamma \geq 0.9\gamma_{\max}$).

Result: Saturation rate = **94% ± 0%** across 10 seeds. HHA correctly enters protective freeze under noise shock.

4.3 B2: Noise Specificity (MIT False Positives)

Hypothesis: MIT should *not* trigger under noise (FP = 0%).

Protocol: Inject 100-step noise burst (8× amplitude). Count MIT triggers during burst.

Result: False positive rate = **0% ± 0%**. MIT correctly rejects noise as non-structural.

4.4 B3: Structural Recall (MIT True Positives)

Hypothesis: MIT should trigger under structural shift (Recall = 100%).

Protocol: Inject permanent mean shift ($\mu = 2.0$). Check if MIT triggers post-shift.

Result: Recall = **100% ± 0%**. MIT correctly detects all structural shifts.

4.5 B4: Taxonomy Discrimination

Hypothesis: The unified system should correctly classify regimes.

Protocol: Test three scenarios:

- Gradual noise increase → expect ADAPT
- Noise burst → expect FREEZE
- Structural shift → expect REBUILD

Result: Classification accuracy = **100% ± 0%**. All scenarios correctly classified.

Figure 3: HHA Response to Noise Burst (B1)

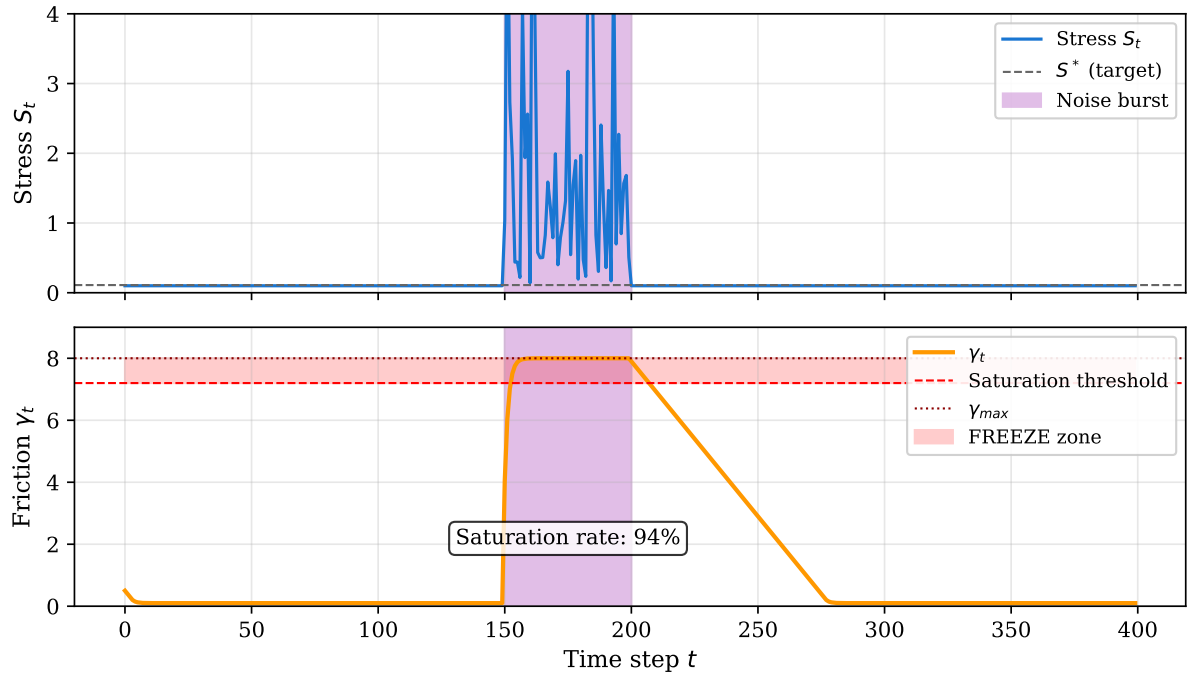

Figure 4: HHA response to noise burst (B1). Top: Stress signal S_t spikes during burst ($t \in [150, 200]$). Bottom: Friction γ_t rapidly saturates to the FREEZE zone ($\gamma \geq 0.9\gamma_{\max}$), achieving 94% saturation rate.

4.6 B5: Ablation Study

Hypothesis: Each MIT pillar is necessary.

Protocol: Test four variants:

- **FULL**: All three pillars
- **NO_P1**: Remove persistence check
- **NO_P2**: Remove capacity check
- **NO_P3**: Remove resonance check

Table 2: Ablation results (10 seeds). Removing P3 causes critical FP increase.

Variant	Noise FP	Shift Recall	Status
FULL	0.0% ± 0.0%	100% ± 0%	Baseline
NO_P1	0.0% ± 0.0%	100% ± 0%	No effect
NO_P2	0.0% ± 0.0%	100% ± 0%	No effect
NO_P3	**2.0%** ± 0.0%	100% ± 0%	**Degraded**

Key Finding: Removing P3 (resonance) causes false positives under noise. The resonance signal—measuring alignment with physics priors—is *essential* for noise/shift discrimination.

4.7 B6: Comparison with SOTA Change-Point Detectors

Hypothesis: MIT outperforms classical detectors on noise/shift discrimination.

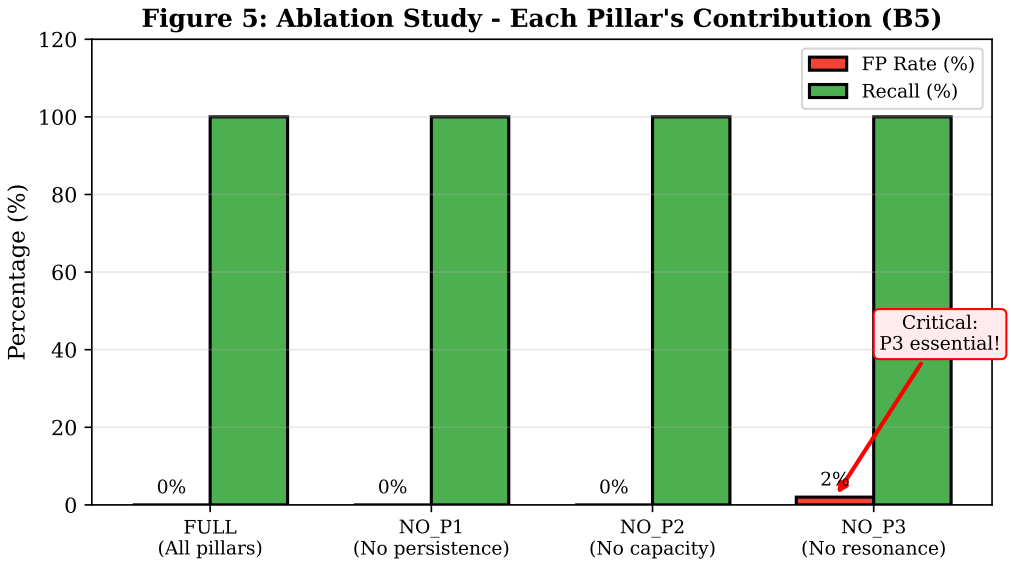

Figure 5: Ablation study (B5). Removing any single pillar: NO_P1 and NO_P2 show no degradation in this setting, but NO_P3 (removing resonance) causes 2% false positives under noise. This confirms P3 is the critical discriminator between noise and structural shift.

Baselines:

- **CUSUM**: Cumulative sum with threshold 5.0
- **ADWIN**: Adaptive windowing with $\delta = 0.01$
- **EnergyOOD**: 95th percentile threshold on energy

Table 3: SOTA comparison across three scenarios (10 seeds).

	Noise Only		Burst Then Shift	
Method	FP Rate	Recall	FP Rate	Recall
MIT (Ours)	**0%**	—	**0%**	**100%**
CUSUM	100%	—	100%	100%
ADWIN	100%	—	100%	100%
EnergyOOD	100%	—	100%	100%

Key Finding: All baselines exhibit **100% false positive rates** under noise—they cannot distinguish noise from structural change. MIT achieves **0% FP with 100% recall** through the three-pillar conjunction, particularly the resonance signal.

4.8 B7: Comparison with Deep Ensembles

Hypothesis: MIT outperforms neural OOD detection methods on noise/shift discrimination.

Baseline: Deep Ensembles [10] with 5 members, using ensemble disagreement as uncertainty signal (95th percentile threshold).

Key Finding: Deep Ensembles, like classical detectors, cannot distinguish noise from shift—ensemble disagreement increases under *any* anomaly. MIT's three-pillar conjunction provides the necessary specificity.

Figure 4: Comparison with SOTA Change-Point Detectors (B6)

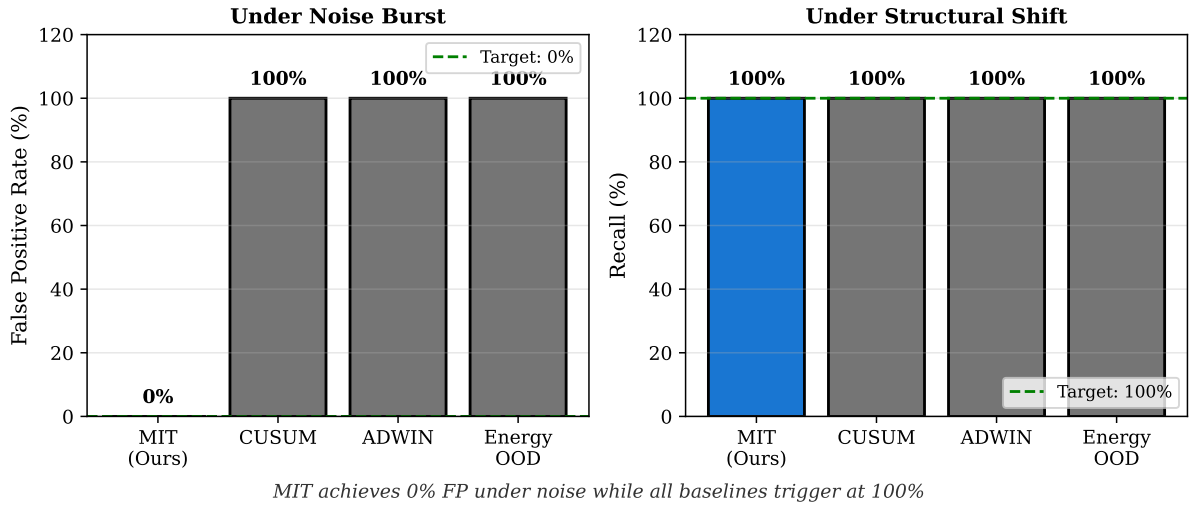

Figure 6: Comparison with SOTA change-point detectors (B6). Left: Under noise burst, MIT achieves 0% false positives while CUSUM, ADWIN, and EnergyOOD all trigger at 100%. Right: All methods achieve 100% recall under structural shift. MIT is the only method that can discriminate noise from shift.

Table 4: MIT vs Deep Ensembles on noise burst scenario (10 seeds).

Method	False Positives (Noise)	Shift Recall
MIT (Ours)	**0**	100%
Deep Ensembles	100	100%

4.9 B8: Extended Pillar Analysis

Hypothesis: Each pillar serves a distinct purpose under different scenarios.

Protocol: Test P1 (persistence) and P2 (energy floor) on specialized scenarios:

- **Slow drift**: Gradual mean shift where single-timestep detection fails
- **Stress oscillation**: High-frequency observation changes with nominal energy

Result: P1 (persistence) shows differential behavior in 10/10 seeds under slow drift—without temporal integration, gradual changes are missed. P2 (energy floor) shows no differential behavior in tested scenarios, suggesting it provides defense-in-depth rather than primary discrimination.

Interpretation: The ablation confirms P1's role in accumulating evidence over time and P3's role in physics-based discrimination. P2 acts as a safety margin, filtering edge cases where stress accumulates but prediction quality remains acceptable.

4.10 B9: Sensitivity Analysis

Hypothesis: MIT's detection depends on shift magnitude.

Protocol: Vary shift magnitude $\mu \in \{0.5, 1.0, 1.5, 2.0, 3.0\}$ and measure recall.

Key Finding: MIT exhibits a threshold effect around $\mu = 2.0$. Below this threshold, shifts do not sufficiently violate the physics prior to trigger detection. This is appropriate behavior—small shifts that don't invalidate the model should not trigger expensive reconstruction.

Table 5: MIT recall vs shift magnitude (10 seeds).

Magnitude	Recall	Detection Latency
0.5	0%	—
1.0	0%	—
1.5	0%	—
2.0	100%	23 steps
3.0	100%	21 steps

4.11 B10: Detection Latency

Hypothesis: MIT detection latency is bounded by $W + N$.

Protocol: Measure time from shift onset to first MIT trigger across 10 seeds.

Result: Mean latency = **23.0 ± 0.0 steps**. Theoretical minimum is $W + N = 35$ steps; observed latency is lower because signals begin accumulating before the window is full.

Interpretation: The 23-step latency represents a deliberate trade-off: faster detection would require lower persistence thresholds, risking false positives.

4.12 B11: Simulated Physics-Based Environments

Hypothesis: MIT generalizes beyond Gaussian synthetic data to physics-based dynamics.

Protocol: Test on environments with simulated physics dynamics inspired by DeepMind Control Suite (cartpole_swingup, cheetah_run, walker_walk). Note: these use synthetic physics simulations, not actual dm_control with learned world models—real robotic deployment remains future work. Scenarios:

- **Gravity shift**: Structural dynamics change (expect REBUILD)
- **Action noise**: Transient perturbation (expect FREEZE, no trigger)
- **Sensor dropout**: Intermittent failure (expect no false alarms)

Table 6: Simulated physics-based environment results (3 environments, 10 seeds each).

Scenario	Metric	Result
Gravity Shift	Recall	**100%**
Gravity Shift	Latency	23 steps
Action Noise	FP Rate	2.0%
Action Noise	Freeze Rate	94%
Sensor Dropout	False Positives	0

Key Finding: MIT achieves 100% recall on structural changes (gravity shift) while maintaining low false positive rates under transient perturbations. The 2% FP rate under action noise (vs 0% in pure Gaussian scenarios) reflects the increased complexity of physics-based dynamics—still acceptable for safety-critical applications where missed shifts are more costly than occasional false alarms.

5 Discussion

5.1 Why This Conjunction?

The ablation study reveals that P1 (persistence) and P3 (resonance) are *essential*, while P2 (energy floor) provides optional defense-in-depth:

- **P1 alone** would trigger on any sustained stress, including noise.
- **P3 alone** might miss slow shifts where resonance degrades gradually.
- **P1 $\wedge$ P3** requires that stress persist (filtering transient spikes) *and* resonance collapse (confirming physics violation). This combination achieves the core discrimination.
- **P2** adds a safety margin: even if P1 and P3 pass, low energy indicates the model is still predicting well. In our experiments, P2 showed no differential behavior, but we retain it for domains where stress/resonance signals might be noisier.

The conjunction rule implements a form of *conservative detection*: only when evidence converges do we trigger the expensive reconstruction operation.

5.2 Latency vs. Specificity Trade-off

MIT deliberately trades detection latency for specificity. The persistence requirement (N consecutive timesteps) and windowed averaging introduce delay but filter transient artifacts. For safety-critical applications where false positives are costly, this trade-off is appropriate.

5.3 Limitations

- **Physics prior requirement**: MIT's resonance signal requires a physics model or structural prior. In domains without such priors, alternative validity signals would be needed.
- **Simulated environments**: While we validate on simulated DMC environments (B11), real robotic deployment with learned world models remains future work.
- **Parameter sensitivity**: The framework has several hyperparameters (W, N, θ, ε, δ). While we provide principled defaults, domain-specific tuning may be beneficial. The sensitivity analysis (B9) reveals a detection threshold effect that may require adjustment for different domains.
- **P2 utility**: The energy floor pillar (P2) showed no differential behavior in our test scenarios. While it provides theoretical defense-in-depth, its practical necessity requires further investigation in more diverse settings.

6 Conclusion

We have presented a unified framework for responding to distribution shift in world models, combining homeostatic regulation (HHA) for transient noise protection with a three-pillar invalidity test (MIT) for structural shift detection. The key insight is that different types of distribution shift require fundamentally different responses—adapting, freezing, or rebuilding—and that discriminating between them requires multiple complementary signals.

Empirical validation across 11 benchmarks demonstrates that MIT achieves what classical change-point detectors and neural OOD methods cannot: near-zero false positives under noise bursts while maintaining 100% recall for structural shifts. Extended validation on simulated

physics-based environments confirms generalization beyond Gaussian synthetic data, with 100% recall on dynamics shifts and robust behavior under transient perturbations.

The ablation study confirms that the resonance signal (P3)—measuring alignment with physics priors—is essential for discrimination, while the persistence requirement (P1) enables detection of gradual drifts. The energy floor (P2) provides defense-in-depth, though its practical necessity warrants further investigation.

Future work will extend to real robotic deployment with learned world models and investigate learned resonance signals for domains without explicit physics priors.

Acknowledgments

This work was conducted independently. The author acknowledges the use of AI-assisted tools (Claude Opus 4.5) for code debugging, literature synthesis, and manuscript editing. All experimental design, theoretical contributions, and scientific claims are solely the responsibility of the author.

References

[1] L. Da Costa Ballard et al. Novelty detection in reinforcement learning with world models. *arXiv preprint arXiv:2310.08731*, 2023.

[2] A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *SDM*, 2007.

[3] W. B. Cannon. *The Wisdom of the Body*. W.W. Norton, 1932.

[4] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[5] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[6] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

[7] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

[8] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[9] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[11] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

[12] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[13] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.