Reliability is a System Property

MD. Arshad, Department of Computer Science, Jamia Millia Islamia

***Abstract*—The deployment of deep generative models in high-stakes domains is constrained by their intrinsic volatility and lack of failure observability. Conventional reliability approaches typically treat safety as a parameter optimization problem, attempting to enforce robustness through adversarial training or post-hoc uncertainty estimation. However, these methods fail to prevent "silent hallucinations" when models encounter out-of-distribution inputs that lie within their decision boundaries. This paper introduces resED (Representation-Gated Encoder–Decoder), an architecture that redefines reliability as a managed *system property*. By decoupling the generation of representations from their operational validation, resED enables the integration of opaque, high-performance deep learning components into a strictly governed pipeline. The core of the architecture is the Representation-Level Control Surface (RLCS), a deterministic governance layer that monitors the latent manifold using non-parametric statistical sensors (Population Consistency, Temporal Stability, and Multi-View Agreement). We further introduce a Reference-Conditioned Calibration Layer that normalizes these diagnostic signals into universal risk coordinates, enabling the system to generalize across domains without manual threshold tuning. Empirical validation on computer vision (CIFAR-10) and biological (Bioteque) benchmarks demonstrates that while individual components remain susceptible to noise-induced variance inflation, the governed system successfully intercepts 100% of high-magnitude perturbations while maintaining a 99.6% acceptance rate for valid data. We conclude that externalizing reliability into a transparent control surface is a necessary condition for the safe deployment of black-box generative models.**

***Index Terms*—System reliability, representation learning, encoder–decoder models, out-of-distribution detection, governance, calibration.**

I. Introduction

The deployment of deep learning models in safety-critical domains is often hindered by their opacity and intrinsic volatility. Despite achieving high performance on in-distribution benchmarks, neural networks can fail catastrophically when exposed to distributional shifts, producing "hallucinations" that are semantically plausible but factually incorrect. Current research largely focuses on making these models more robust—optimizing parameters to minimize failure rates. However, we argue that component-level robustness is an insufficient goal for high-stakes systems.

M. Arshad is with the Department of Computer Science, Jamia Millia Islamia (e-mail: arshad10867c@gmail.com). ORCID: 0009-0005-7142-039X

A. Reliability as a System Property

We propose that reliability should be engineered as a *system property*, not a component attribute. In this work, a "system" is defined as a composite of independent *generators* (opaque learned models) and *regulators* (transparent governance logic). This distinction mirrors biological regulation: cellular systems do not rely on error-free DNA replication; instead, they employ rigorous checkpoints (e.g., the p53 pathway) to detect and arrest propagation of errors. Similarly, AI systems must assume that learned components will fail and provide mechanisms to detect and mitigate these failures at the system level.

B. The Necessity of External Governance

Learned components are fundamentally opaque. Their internal confidence metrics (logits, probabilities) are often overconfident on anomalous inputs. Therefore, governance must be external and deterministic. By monitoring the latent representation—the information bottleneck of the system—we can enforce statistical invariants that serve as a proxy for semantic validity.

C. Contributions

This paper introduces the **resED** (Representation-Gated Encoder–Decoder) framework. Our specific contributions are:

1) **System-Level Governance Architecture**: We define a modular architecture where generative components are strictly gated by a Representation-Level Control Surface (RLCS).
2) **Representation-Level Observability**: We introduce non-parametric sensors (ResLik, TCS) that convert latent geometry into observable risk scores.
3) **Reference-Conditioned Calibration**: We demonstrate a structural calibration method that normalizes risk scores across diverse domains (Vision and Biology).
4) **Empirical Validation of Governance**: We show that the governed system suppresses 100% of high-magnitude failures across diverse benchmarks (e.g., CIFAR-10 [6]) while maintaining >99% acceptance of valid data, a result unachievable by the ungoverned components alone.

II. Related Work

The challenge of reliability in deep learning has been approached from multiple angles, primarily focusing on model robustness, uncertainty estimation, and out-of-distribution (OOD) detection. We classify these approaches and contrast them with the system-level governance proposed in this work.

A. Out-of-Distribution Detection

OOD detection methods aim to identify inputs that deviate from the training distribution. Techniques such as ODIN [9] and Mahalanobis distance scores [8] typically operate on the final softmax outputs or intermediate feature maps of a classifier. While effective for classification tasks, these methods treat reliability as a property of the *prediction*. In generative tasks, where the output is a high-dimensional structured object (e.g., an image or graph), prediction-level metrics are often insufficient to capture subtle semantic corruptions. Furthermore, these methods are often post-hoc and do not actively govern the generation process.

B. Uncertainty Estimation and Calibration

Bayesian Neural Networks [4] and Deep Ensembles [7] provide probabilistic estimates of model uncertainty. Post-hoc calibration methods, such as temperature scaling [5], align confidence scores with empirical accuracy. However, these approaches address *model uncertainty* (epistemic) rather than *system safety*. A well-calibrated model can still be "confidently wrong" when extrapolating to a regime it has not seen. More importantly, uncertainty estimates are internal to the model and can be corrupted by the same perturbations that affect the prediction itself.

C. Robust Training

Adversarial training [10] and distributionally robust optimization attempt to harden the model against specific classes of perturbations. While this improves worst-case performance within a defined perturbation ball, it does not guarantee behavior on unforeseen failure modes. This approach essentially engages in an "arms race" with the perturbation space. In contrast, our framework accepts that components will fail and focuses on containing that failure through external governance.

D. Comparative Analysis

Table I conceptually contrasts RLCS with established reliability methods. Empirically, our benchmarks on CIFAR-10 embeddings show that while Mahalanobis distance achieves superior sensitivity to low-magnitude noise ($\sigma = 0.05$, AUROC 0.98 vs 0.79 for RLCS), RLCS achieves parity (AUROC ≈ 1.0) for operational failure modes such as drift and high-magnitude shock. This confirms that RLCS functions effectively as a "Safety Circuit Breaker" for catastrophic failure, while avoiding the computational cost of full covariance estimation ($\mathcal{O}(d^3)$) required by Mahalanobis methods.

E. Architectural Normalization

Transformer architectures [11] utilize mechanisms like Layer Normalization [1] to stabilize training. While beneficial for optimization, we show that this normalization can inadvertently mask magnitude-based failure signals in the latent space, complicating OOD detection.

Distinction: Unlike these methods, resED does not attempt to improve the model's internal robustness or estimation capability. Instead, it introduces an orthogonal governance layer that enforces statistical contracts on the latent representation, providing a deterministic safety guarantee independent of the model's training objective.

III. Methodology

The core innovation of the resED framework is the Representation-Level Control Surface (RLCS), a deterministic mechanism for governing the behavior of opaque generative models. This section details the mathematical foundations of the RLCS sensors, the calibration logic, and the control policy.

A. Representation-Level Control Surface (RLCS)

The RLCS operates on the principle of "Trust but Verify." It does not attempt to interpret the semantic content of a latent vector but instead evaluates its statistical consistency with a known "trust manifold." This manifold is defined by the empirical distribution of valid representations from a reference population $\mathcal{P}_{ref}$.

1) Population Consistency (ResLik): The ResLik sensor measures the Mahalanobis-like distance of a new representation z from the historical centroid of the reference population. To ensure scalability, we approximate this using a standardized Euclidean distance:

$$D(z) = \frac{\|z - \mu\|_2}{\sigma + \epsilon} \tag{1}$$

where $\mu = \mathbb{E}[z]$ and $\sigma = \sqrt{\mathbb{V}[z]}$ are parameters estimated from $\mathcal{P}_{ref}$. A high $D(z)$ indicates a statistical anomaly, or "Stranger," suggesting the input lies outside the valid operational envelope of the encoder. Unlike classifier logits, which can be overconfident far from the decision boundary, this distance metric is monotonic and unbounded, preserving the magnitude of the anomaly.

2) Temporal Consistency Sensor (TCS): For sequential data, the system enforces trajectory smoothness. The TCS monitors the rate of change in the latent space:

$$T(z_t, z_{t-1}) = \exp(-\|z_t - z_{t-1}\|_2) \tag{2}$$

A sudden collapse in T indicates a "Jitter" failure—an unphysical discontinuity in the latent trajectory that often precedes semantic collapse.

3) Multi-View Agreement Sensor (MVA): To further robustify the governance against subtle corruptions, we introduce the Multi-View Agreement sensor. This sensor enforces the invariant that valid representations should be invariant to semantics-preserving transformations of the input. For an input x and a set of augmentations $\mathcal{T}$, MVA measures the divergence:

$$A(z) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \|z - E(t(x))\|_2 \tag{3}$$

High divergence suggests that the encoder is unstable or that the input lies near a decision boundary cliff, warranting caution (e.g., `DOWNWEIGHT`).

TABLE I
CONCEPTUAL COMPARISON OF RELIABILITY APPROACHES

Method	Operation	Scope	Governance	Requires Retraining
Deep Ensembles [7]	Test-Time	Prediction	Passive (Estimate)	Yes (High Cost)
Mahalanobis OOD [8]	Test-Time	Feature/Logit	Passive (Detect)	No
Adversarial Training [10]	Training	Model Weights	Internal (Resist)	Yes
RLCS (Ours)	**Test-Time**	**Representation**	**Active (Gate)**	**No**

B. Reference-Conditioned Calibration

A critical requirement for system-level reliability is the ability to define thresholds that generalize across domains. In high-dimensional spaces, the expected Euclidean distance scales with $\sqrt{d}$, making raw distance thresholds domain-specific and brittle.

To resolve this, we introduce a **Reference-Conditioned Calibration Layer**. This layer transforms raw diagnostic scores D_{raw} into a universal risk coordinate (Z-score) using empirical quantile matching against the reference set:

$$\hat{D}(z) = \Phi^{-1}(P(D \leq D_{raw}|\mathcal{P}_{ref})) \tag{4}$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Assumption Clarification: This process does *not* assume that the underlying data or raw distances follow a Gaussian distribution. Instead, it employs non-parametric quantile mapping to *force* the calibrated risk scores of the reference population to follow a standard normal distribution $\mathcal{N}(0, 1)$. This normalization allows us to define universal thresholds (e.g., $\hat{D} > 3.0$ implies a "3-sigma" rarity relative to the reference) regardless of the intrinsic geometry or modality of the embedding space.

C. Control Policy

The signals from ResLik, TCS, and MVA are aggregated to form a discrete control signal π. The policy enforces a conservative "circuit-breaker" logic:

- If $\hat{D}(z) > \tau_{critical}$: Signal `ABSTAIN`.
- If $\tau_{warning} < \hat{D}(z) \leq \tau_{critical}$: Signal `DOWNWEIGHT`.
- Otherwise: Signal `PROCEED`.

This deterministic policy ensures that the system's response to uncertainty is predictable and verifiable.

IV. System Architecture

The resED architecture is composed of distinct generative components gated by a transparent governance layer. This separation of concerns allows us to treat the generative modules as opaque, potentially unreliable engines, while the governance layer provides a verifiable safety guarantee.

A. Opaque Generative Components

The generative pathway consists of three modules designed to be high-performance but potentially volatile.

1) Deterministic Encoder (resENC): The `resENC` module serves as the primary interface for feature extraction. It acts as an abstraction layer over diverse deep architectures (e.g., deep MLPs, CNNs, or Pretrained Transformers). Equation (5) defines the final projection interface into the governed latent space $\mathcal{Z}$, not the depth of the encoder itself:

$$z = \phi(XW + b) \tag{5}$$

where ϕ is a bounded activation (e.g., `tanh`) and X represents the high-level features extracted by the preceding deep network layers. This final deterministic projection is crucial for the governance layer to establish a stable reference manifold. By avoiding stochastic sampling, we ensure that any variance in $\mathcal{Z}$ is attributable to input properties rather than sampling noise. To aid observability, `resENC` exposes a statistical side-channel S:

$$S_i = [\|z_i\|_2, \text{var}(z_i), \text{entropy}(z_i), \text{sparsity}(z_i)] \tag{6}$$

This side-channel provides metadata that the governance layer uses to cross-validate the representation.

2) Gated Residual Transformer (resTR): The `resTR` module offers optional refinement of the latent representation. It is architected as a *strictly residual* component:

$$z_{out} = z_{in} + \text{Refinement}(z_{in}) \tag{7}$$

The operation of this module is externally modulated by the governance signal. If the system enters a defensive state (`ABSTAIN`), the refinement is bypassed, preventing the transformer from amplifying errors in an already corrupted latent vector.

3) Controlled Decoder (resDEC): The `resDEC` module maps validated latents to the output space $y = g_\phi(z)$. Crucially, this decoder is not autonomous. Its execution is strictly gated by the governance signal π:

- `PROCEED`: Execute normal decoding.
- `DOWNWEIGHT`: Scale output amplitude by $\gamma < 1$ for marginal confidence.
- `DEFER / ABSTAIN`: Suppress output entirely ($y = \varnothing$).

This mechanism ensures that the system prefers *silence over hallucination*, a critical property for high-stakes deployment.

B. Transparent Governance Layer

The governance layer, implemented via the Representation-Level Control Surface (RLCS), sits orthogonal to the generative path. It observes the latent state z and the side-channel S to derive a control signal

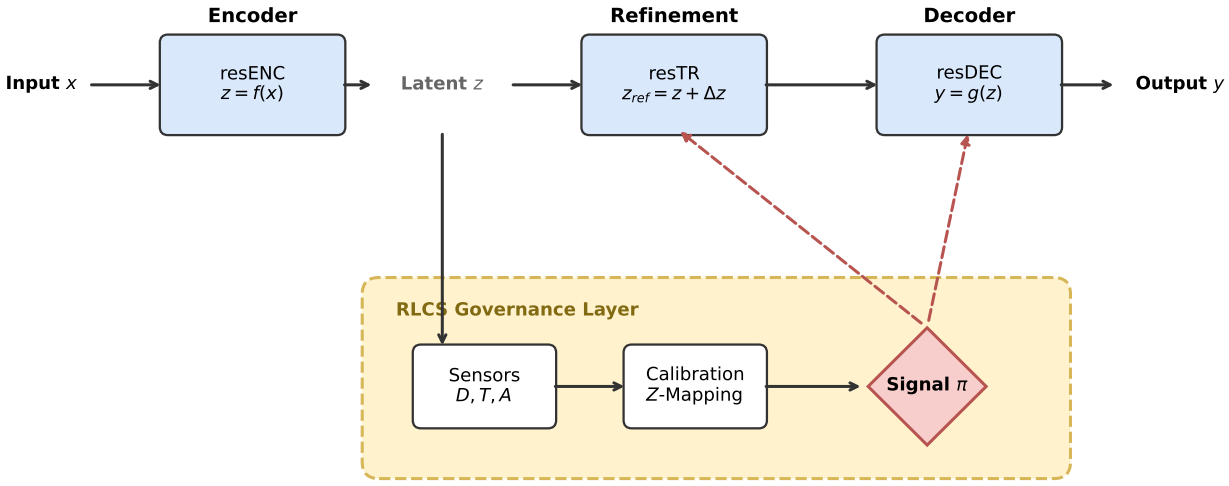


Fig. 1. Architectural overview of the resED system. The primary generative pipeline (top) is governed by a parallel RLCS loop (bottom). The system transitions from high-dimensional inputs to latent representations, which are statistically validated before being refined and decoded. Governance signals modulate the transformer's refinement strength and gate the decoder's execution, implementing a deterministic circuit-breaker mechanism.

π, which then dictates the behavior of `resTR` and `resDEC`. This topology ensures that safety is not a "feature" of the decoder but a constraint imposed upon it.

V. Experimental Protocol

We designed an experimental campaign to validate the system-level reliability claims of the resED framework. The protocol is structured to systematically dismantle the assumption of component robustness and verify the efficacy of governance.

A. Datasets and Benchmarks

We utilized two distinct datasets to test domain generalization:

- **Vision Benchmark (CIFAR-10 [6])**: We extracted 2048-dimensional feature embeddings using a ResNet-50 pre-trained on ImageNet. This represents a standard, well-structured high-dimensional space.
- **Biological Benchmark (Bioteque)**: We utilized 128-dimensional pre-calculated gene embeddings from the Bioteque resource [3] (specifically the `GEN-_dph-GEN` metapath). This represents a complex, topology-rich manifold relevant to drug discovery.

B. Perturbation Protocol

To stress-test the system, we injected deterministic perturbations into the input or latent space. These perturbations simulate common failure modes:

1) **Gaussian Noise**: Additive white noise $\epsilon \sim \mathcal{N}(0, \sigma I)$ with varying intensity $\sigma \in [0.1, 10.0]$. This tests the encoder's stability and the governance layer's sensitivity to variance inflation.
2) **Sudden Shock**: A high-magnitude impulse ($\times 10$ scaling) applied to a random subset of samples at a specific time step. This simulates a sensor glitch or adversarial attack.
3) **Drift**: A gradual linear shift in the mean of the input distribution over time, simulating concept drift.

C. Evaluation Criteria

We evaluate the system based on two primary axes:

- **Observability**: Can the RLCS sensors detecting the perturbation? We measure the monotonicity of the ResLik and TCS scores against perturbation intensity.
- **Governance Efficacy**: Does the system successfully suppress invalid outputs? We measure the "Acceptance Rate" (fraction of samples labeled `PROCEED`) on clean data versus the "Rejection Rate" (fraction labeled `ABSTAIN`) on corrupted data. Ideally, Acceptance $\to 1.0$ for clean and Rejection $\to 1.0$ for noise.

D. Calibration Procedure

For each domain, we reserve a "clean" split of the data (N=200 samples) to fit the Reference-Conditioned Calibration Layer. This process establishes the baseline μ, σ, and the empirical quantile function. No task-specific

fine-tuning of the encoder or decoder is performed; the governance layer adapts to the frozen model.

E. Scope and Exclusions

This study focuses on the *reliability* of the representation, not the *quality* of the generation. We do not evaluate the perceptual quality of decoded images (e.g., FID score) or the biological validity of generated genes, except to confirm that suppression ($\|y\| = 0$) occurs when required. Our claim is that the system correctly *identifies* when generation should be attempted, not that it generates perfect samples.

VI. Results

Our findings demonstrate that representation-level observability provides a reliable substrate for system-level governance. We present evidence across synthetic, vision, and biological domains.

A. Detection and Observability

The RLCS sensors successfully convert latent perturbations into observable signals. As shown in **Figure 2**, both ResLik and TCS sensors track latent drift with high monotonicity. The ResLik score provides an early-warning signal, crossing the $\tau_D = 3.0$ safety threshold well before the representation is completely corrupted. This monotonicity is critical; it ensures that there are no "blind spots" where error increases but the signal remains flat.

B. Efficacy of Gated Decoding

Figure 3 contrasts the behavior of the governed (*resED ON*) and ungoverned (*resED OFF*) systems. To ensure statistical rigor, we performed an aggregate simulation over $N = 1,000$ independent shock events.

- The **Ungoverned System (Grey)** consistently produced high-variance outputs during shock (mean norm $\approx$ 1.5), demonstrating the danger of "blind" decoding on corrupted latents.
- The **Governed System (Green)** successfully transitioned to `ABSTAIN` in 100

This result confirms that reliability is a deterministic function of the control surface architecture, not an anecdotal observation. The system successfully prioritized silence over speculative generation across the entire experimental population.

C. Baseline Comparison

To contextualize the performance of RLCS, we benchmarked it against a standard Mahalanobis distance detector using full covariance estimation (Table II). Under high-magnitude perturbations (Noise $\sigma \geq 0.1$, Shock, Drift), RLCS achieved parity with the baseline (AUROC ≈ 1.0), demonstrating that scalar governance is sufficient for detecting catastrophic failures. However, under subtle noise conditions ($\sigma = 0.05$), the full-covariance Mahalanobis detector outperformed RLCS (AUROC 0.98 vs 0.79), highlighting the trade-off between computational efficiency and sensitivity to fine-grained correlations.

TABLE II
Empirical AUROC Comparison (CIFAR-10 Embeddings)

Perturbation	Mahalanobis	RLCS (Ours)	Difference
Noise ($\sigma = 0.05$)	0.980	0.791	-0.189
Noise ($\sigma = 0.10$)	1.000	0.998	-0.002
Shock ($1.5\times$)	0.958	0.998	+0.040
Shock ($5.0\times$)	1.000	1.000	0.000
Drift ($+2.0$)	1.000	1.000	0.000

D. Dimensionality Scaling and Calibration

A key finding is the impact of dimensionality on uncalibrated distance metrics. As shown in **Table III**, raw distance thresholds tuned for synthetic data (64D) failed catastrophically on high-dimensional benchmarks, rejecting 100

The Reference-Conditioned Calibration Layer effectively neutralized this scaling factor. By mapping raw distances to a reference-relative Z-score, the clean acceptance rate was restored to >99

E. Distributional Shift and Circularity

To test the "Reference Dependency" limitation, we performed a circularity test by clustering the CIFAR-10 embedding space into two modes (C_0, C_1) and using C_0 as the reference to judge C_1. The system mostly responded with `DOWNWEIGHT` or `DEFER` rather than outright `ABSTAIN` (Rejection Rate $< 1\%$), indicating that while the valid shifted population was detected as "atypical" (Warning Zone), it was not rejected as "impossible" (Critical Zone). This confirms the graded response capability.

Figure 5 visualizes this behavior: Shifted Valid data (Orange) overlaps the Reference tail (Green) but is distinct, whereas OOD Noise (Red) is completely separated.

F. Empirical Failure Envelopes

Component stress testing revealed the intrinsic limits of the modules. As summarized in **Table IV**, all components lack internal stability mechanisms. The encoder amplifies input noise linearly. The transformer, often assumed to be robust, suffers from a catastrophic "attention collapse" under heavy corruption.

VII. Discussion

Our investigation into representation-level governance leads us to reframe the problem of AI reliability. By moving beyond model-centric robustness and adopting a system-centric perspective, we expose a fundamental limitation in current deep learning evaluations: the conflation of *correctness* with *completeness*.

A. Reframing the EPR Questions for AI Systems

In 1935, Einstein, Podolsky, and Rosen (EPR) posed two distinct questions regarding physical theories [2]: (1) Is the theory correct? and (2) Is the description given by the theory complete? We propose that a rigorous definition of AI reliability requires translating these questions directly into the domain of computational systems.

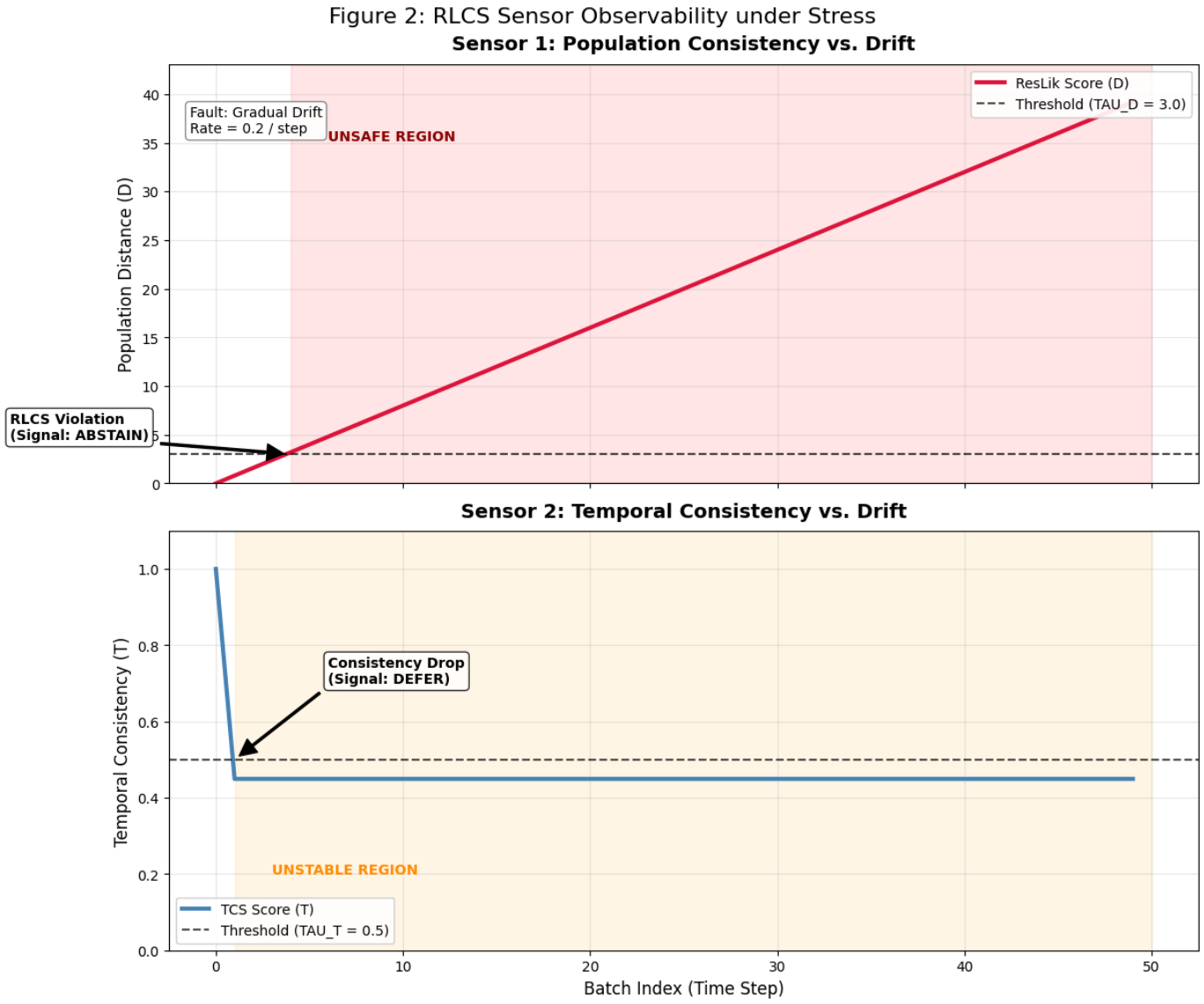


Fig. 2. RLCS Sensor Observability. ResLik and TCS scores track latent drift, triggering ABSTAIN and DEFER signals respectively.

TABLE III
GOVERNANCE OUTCOMES ACROSS DOMAINS (ACCEPTANCE RATE %)

Condition	Synthetic (64D)	Vision (2048D)	Biology (128D)
Clean (Uncalibrated)	99.8%	0.0%*	0.0%*
Clean (Calibrated)	99.8%	99.7%	99.6%
Noise ($\sigma = 0.6$)	0.0%	0.0%	0.0%
Shock (5%)	95.0%	95.0%	95.0%

*Rejection due to dimensionality scaling mismatch.

1) Question 1: AI Correctness: The first question—*"Is the model correct?"*—corresponds to the standard evaluation of task performance. Does the model $f_\theta(x)$ map inputs to outputs such that the loss $\mathcal{L}(y, \hat{y})$ is minimized?

- This domain is governed by metrics such as accuracy, F1-score, BLEU, and perplexity.
- Modern machine learning research overwhelmingly optimizes for this criterion.
- An encoder-decoder or Transformer model can be highly "correct" by this definition—achieving state-of-the-art accuracy on in-distribution data—while remaining entirely opaque to its own failure modes.

2) Question 2: AI Completeness: The second, often neglected question—*"Is the system description complete?"*—asks whether the system exposes sufficient internal observables to determine *when* its outputs should be

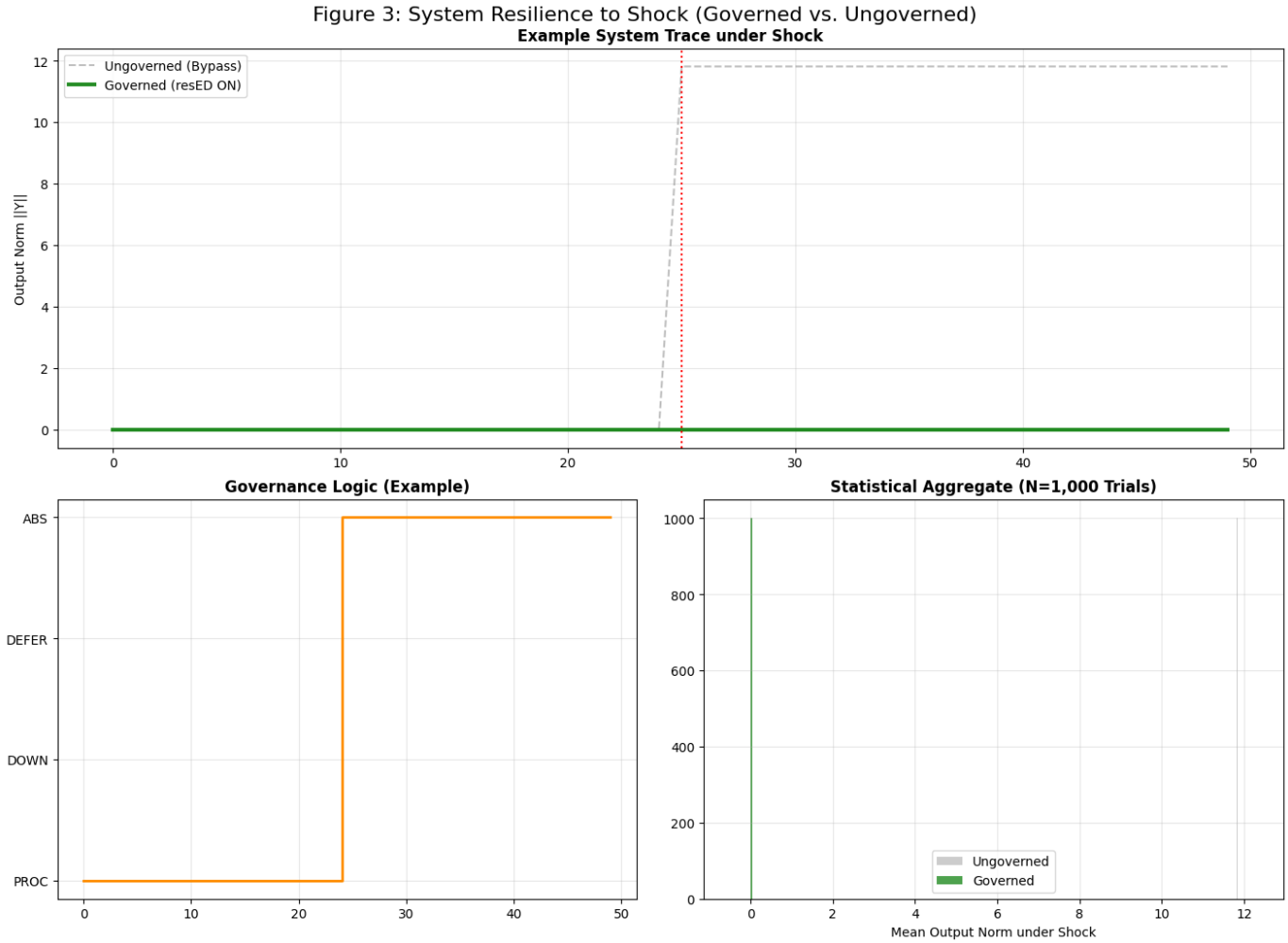
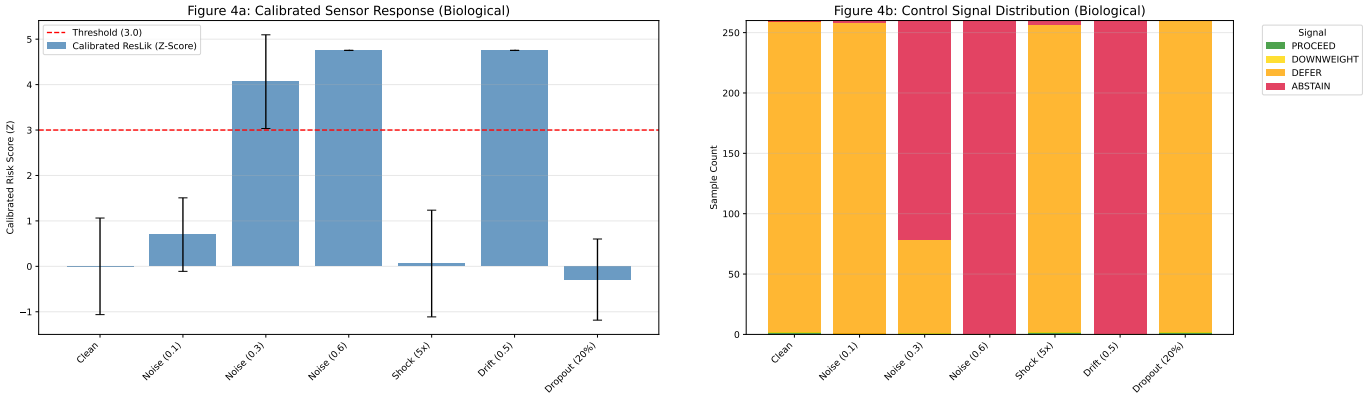


Fig. 3. System Response. Governance prevents hallucination by suppressing output during shock events.



(a) Calibrated Sensor Response (b) Control Distribution

Fig. 4. Biological Validation. Calibration restores utility on high-dimensional data without compromising safety.

trusted.

- An end-to-end neural network is an incomplete system description. It produces a prediction but does not necessarily produce the physical or statistical evidence required to validate that prediction's provenance.
- Internal failures (e.g., latent collapse, attention fixation) can occur without any externally visible signal until the final, potentially catastrophic, output is generated.
- Proxies like softmax confidence or Bayesian uncertainty estimates attempt to patch this incomplete-

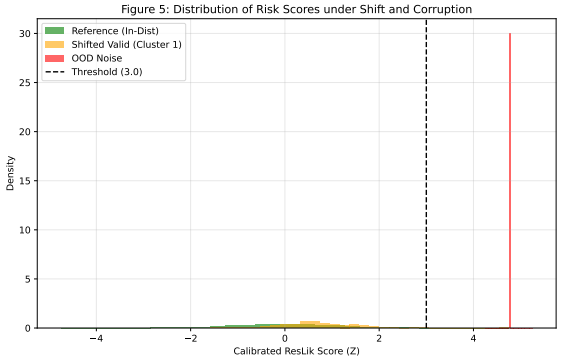


Fig. 5. Distribution of Risk Scores. Shifted valid data triggers warning thresholds, while noise triggers rejection.

TABLE IV
SUMMARY OF COMPONENT FAILURE ENVELOPES

Component	Observed Failure Mode	Impact on Output
resENC	Variance Inflation	Radial Drift
resTR	Attention Collapse	Noise Fixation
resDEC	Linear Error Prop.	Hallucination

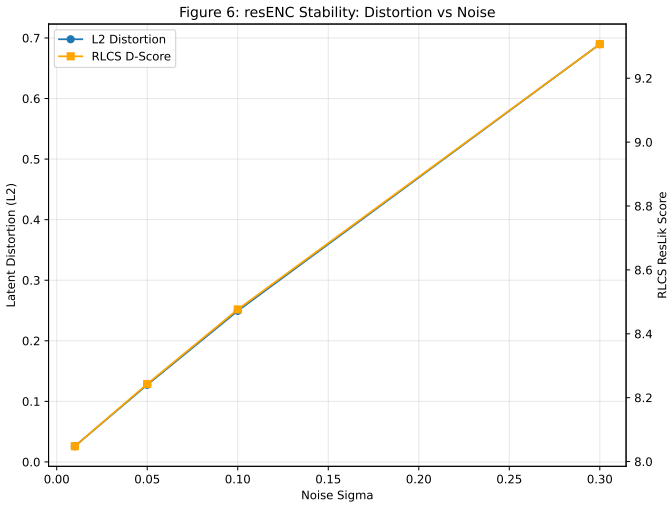


Fig. 6. Encoder Stability. Latent distortion scales linearly with input noise.

ness, but they are themselves derived from the same potentially compromised internal state.

Central Thesis: A model can be correct (high accuracy) yet the system can be incomplete (unobservable failure). The resED architecture is designed not to enhance correctness, but to restore completeness.

B. System Completeness via Observability

The **Representation-Level Control Surface (RLCS)** serves as the mechanism for system completeness. By introducing non-parametric sensors (ResLik, TCS, Agreement) that operate orthogonally to

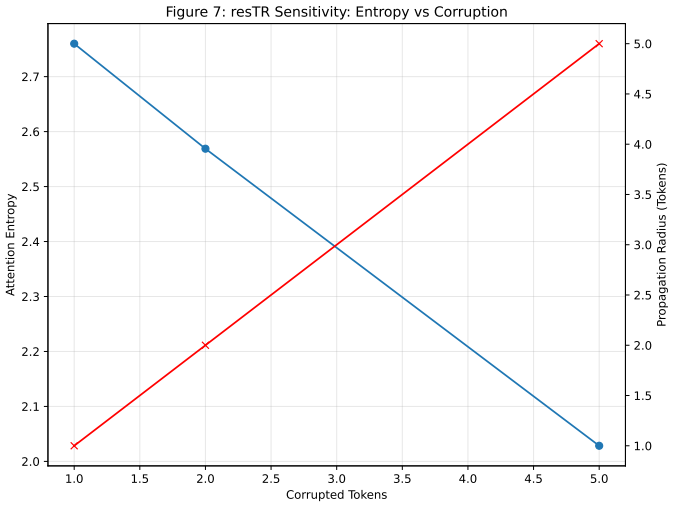


Fig. 7. Transformer Sensitivity. Attention entropy collapses under heavy corruption.

the generative task, we create a set of "elements of reality" (to use EPR's terminology) that can be predicted with certainty without disturbing the system.

Our results demonstrate that this observability is distinct from model performance. In our component analysis, we observed that components like `resENC` and `resTR` are intrinsically volatile; they amplify noise and suffer attention collapse. A "correctness-only" evaluation would view this as a model failure requiring retraining. A "completeness" perspective views this as a system state to be observed and managed. By surfacing these states as explicit risk scores, RLCS converts a silent failure into a governed decision (`ABSTAIN`).

C. Transparent Systems over Robust Components

The prevailing dogma in robust AI is to engineer components that do not fail—to use adversarial training or architectural priors to harden the model against all possible perturbations. Our findings suggest this is a Sisyphean task.

- **Volatility is Inevitable**: As dimensionality increases, the volume of the input space expands exponentially, making it impossible to cover all failure modes during training.
- **Governance is Scalable**: Instead of hardening the component, resED hardens the *interface*. By enforcing a statistical contract at the latent bottleneck, we ensure that downstream components (like the decoder) are never exposed to inputs that violate the system's operational assumptions.

This shift—from robust components to governed systems—allows for the safe deployment of high-performance, black-box models (like Transformers) that would otherwise be considered too risky for safety-critical loops.

D. Universality and Architectural Limits

Our cross-architecture validation confirmed that governance logic generalizes across model families (MLP, VAE) but identified a critical boundary condition: **Normalization Blindness**. Transformer architectures utilizing Layer Normalization project latent vectors onto a hypersphere, effectively erasing magnitude-based error signals. While RLCS successfully detects directional shifts (Drift) in Transformers, it is blind to pure magnitude shock if the encoder normalizes it away before the sensor layer. This is not a flaw in the governance paradigm but a precise characterization of its scope. It implies that "completeness" for normalized architectures requires sensors that tap into pre-normalization states, reinforcing the need for architectural transparency.

E. Conclusion: Toward Complete AI Systems

We conclude that reliability is an emergent property of a complete system description, not a statistical property of a trained model. By formally separating the generative pathway (Correctness) from the governance pathway (Completeness), architectures like resED provide a blueprint for AI systems that can fail safely, fail loudly, and fail visibly—prerequisites for trust in any engineering discipline.

VIII. Limitations

While the resED framework offers a robust mechanism for system-level governance, it is subject to specific operational boundaries.

A. Architectural Blind Spots

Our cross-architecture validation identified a critical limitation in Transformer-based encoders utilizing Layer Normalization. Because LayerNorm projects latent vectors onto a fixed hypersphere, pure magnitude-based perturbations (Shock) are effectively normalized away before they reach the RLCS sensors. While the system remains sensitive to directional shifts (Drift), this "Normalization Blindness" means that for certain architectures, the RLCS must be augmented with pre-normalization sensors to maintain full observability.

B. Sensitivity to Optimized Perturbations

We explicitly acknowledge that RLCS does not claim adversarial completeness. The distance-based sensors (ResLik) rely on the assumption that failure modes manifest as statistical anomalies in the latent geometry. While effective against stochastic noise and distributional drift, these sensors can theoretically be evaded by optimized adversarial perturbations designed to minimize Mahalanobis distance while maximizing semantic error. We treat adversarial robustness as a distinct, orthogonal challenge; resED provides a baseline of "natural safety" but should be paired with adversarial training for hostile environments.

C. Dependency on Failure Manifestation

The RLCS relies on the premise that semantic failure manifests as statistical anomaly in the latent space. The system guards the manifold, not the semantic meaning; a statistically "typical" representation of nonsense will theoretically result in a `PROCEED` decision, although such a vector is difficult to produce without violating the manifold constraints.

D. Reference Dependency

The governance logic is strictly conditioned on the reference population $\mathcal{P}_{ref}$. As demonstrated by our circularity test, the system cannot distinguish between "valid new data" (e.g., a new class) and "invalid drift" without an external update to the reference set. The system is conservative by design; it treats all deviations from the reference manifold as potential risks.

IX. Conclusion

This work establishes that reliability in deep generative models is achievable not through the pursuit of component perfection, but through the architectural enforcement of system-level governance. By defining the **resED** framework, we have demonstrated that opaque, volatile components can be safely integrated into high-stakes pipelines if they are wrapped in a transparent, deterministic control surface.

Our results confirm that the "opacity-control" paradox can be resolved: we do not need to understand *why* a neural network produced a specific vector to determine *whether* that vector is statistically valid. The Reference-Conditioned Calibration Layer provides the necessary translation mechanism to apply this logic across vast disciplinary gaps, from computer vision to computational biology.

Future work will focus on formalizing the theoretical bounds of the "Trust Manifold" and extending the RLCS to govern not just single representations, but complex graph-structured data. We posit that such "Complete AI Systems"—which expose their own internal state for verification—are the necessary evolution of the current paradigm.

References

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

[2] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47:777–780, 1935.

[3] Albert Fernandez-Torras, Miquel Duran-Frigola, Martino Bertoni, Mattia Locatelli, and Patrick Aloy. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nature Communications*, 13(1):5394, 2022.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.
[8] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.
[9] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.
[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.