Error-Gated Sparse Attention with Endogenous Out-of-Distribution Signals

Yuhan Feng
fengru1005@gmail.com

Abstract

Standard self-attention allocates a uniform quadratic interaction budget to all tokens, even though only a minority of tokens are typically ambiguous or computationally demanding. We propose **Sentinel Attention**, an *error-gated routing* attention layer that keeps *baseline contextualization* intact while allocating *adaptive computation* where the model struggles. Every token first receives a standard multi-head self-attention update, preserving the optimization behavior of a conventional Vision Transformer. A lightweight error predictor then estimates per-token patch reconstruction error, and the high-error subset exchanges its standard update for a sparse error-directed attention update. This design converts prediction error into a routing signal; in the reference implementation the measured wall-clock overhead is +36–51% relative to a standard Vision Transformer of equal depth on CPU, which we report alongside the mechanism's other contributions rather than minimizing.

A second contribution is conceptual rather than purely architectural: the same reconstruction error required for routing is available at *zero marginal inference cost* as an endogenous out-of-distribution (OOD) score. Instead of claiming this score as universally useful, we study its *validity boundary* through controlled experiments spanning two input regimes (MNIST grayscale, CIFAR-10 RGB) and three OOD regimes (far-OOD, near-OOD, natural-image OOD). The resulting picture is sharply asymmetric. On far-OOD low-dimensional inputs (FashionMNIST versus MNIST), reconstruction error is highly effective: AUROC 0.9841 alone and 0.9964 when combined with Energy, with FPR95 0.016. On near-OOD handwritten characters (EMNIST Letters versus MNIST digits), the signal reverses below chance (AUROC 0.4827), showing that pixel-space reconstruction tracks surface-level simplicity rather than semantic novelty. On CIFAR-10, the signal collapses entirely on natural-image OOD (mean AUROC 0.3646), demonstrating that the mechanism does not transfer naively to higher-entropy RGB inputs.

We additionally test whether score fusion helps because of the fusion rule or because of *information-source independence*. The answer is the latter. Cross-source combinations can help substantially when at least one constituent signal is reliable (e.g., Mahalanobis+Energy on CIFAR-10: mean AUROC 0.8359), whereas same-source combinations offer no consistent gain (MSP+Energy), and combining a corrupted signal with a good one can actively hurt performance (Sentinel+Energy on CIFAR-10). The paper therefore makes two claims: an engineering claim—error-gated routing is a viable, interpretable way to add selective computation to Transformers, whose current implementation cost we characterize explicitly—and a scientific claim—pixel-space reconstruction error is a narrow but measurable OOD signal whose success and failure modes can be stated precisely.

1 Introduction

Standard self-attention [1] allocates an identical quadratic interaction budget to every token in a sequence, irrespective of whether a token is unambiguous and well-explained by its context or, instead, ambiguous and demanding of further computation. This uniform allocation is wasteful in

two respects. Computationally, the majority of tokens in a typical input are redundant—they can be contextualized with a single attention pass—yet they incur the full cost of the attention matrix. Epistemically, the model has no internal mechanism to flag tokens it finds difficult: the forward pass produces no signal that says, for this patch, the representation is uncertain.

Existing approaches to this inefficiency fall into two disjoint lines of work. *Sparse attention* methods [2, 3, 4, 5, 6] reduce the $O(N^2)$ cost by imposing fixed structural patterns—local windows, strided blocks, low-rank projections, or locality-sensitive hashing—but these patterns are agnostic to the model's internal uncertainty: a token receives the same sparsity budget regardless of whether the model finds it easy or hard. *Selective and adaptive computation* [28, 29, 30, 31] condition the amount of computation on the input, but typically rely on learned routing networks whose internal signals are opaque and have no demonstrated utility beyond the routing task itself. The two literatures—efficient attention and input-dependent computation—have remained largely disconnected from a third body of work, *out-of-distribution (OOD) detection* [8, 9, 10, 11], which operates post hoc on the classification head and requires a completed forward pass before any uncertainty estimate is available.

We bridge these three threads with **Sentinel Attention**, an error-gated routing attention layer in which a lightweight error predictor reconstructs each input patch from its token embedding and the resulting per-token reconstruction error serves simultaneously as (i) a hard routing gate that directs only high-error tokens to an error-directed sparse attention update and (ii) an endogenous OOD score available at zero marginal inference cost. The design is motivated by predictive coding theory [16, 17], which posits that biological neural systems allocate more processing resources to signals that cannot be predicted from context. In Sentinel, prediction error is not an epiphenomenon but the explicit driver of both computation allocation and novelty detection.

A key architectural property is that Sentinel preserves rather than replaces standard self-attention. Every token receives the conventional multi-head self-attention update of a standard Vision Transformer [22]; only the subset of tokens whose reconstruction error exceeds a data-dependent threshold τ exchanges that update for an error-directed sparse attention update. This *adaptive attention selection on top of a standard backbone* design incurs a measured wall-clock overhead of roughly 1.4–1.5× over a six-layer ViT of equal depth on CPU in the reference implementation (Table 6), while exposing an internal uncertainty signal that requires no extra forward passes, no OOD data for calibration, and no auxiliary models—unlike post-hoc scores such as MSP [8], Energy [9], or Mahalanobis distance [10].

The architecture is the product of systematic ablation. Sentinel originates from the Precise Weight Precision Field (PWPF) framework, which combined error-driven attention with a learnable precision field, attractor regularization, spectral normalization, and a dynamic memory bank. Controlled ablations revealed that most of these components were either actively harmful (attractor regularization, spectral normalization) or non-informative (precision field, memory bank, confidence network). A single finding survived every ablation: error-direction queries consistently outperformed state-direction queries. Sentinel retains only this error-gated routing mechanism, removing every component that failed controlled testing. The full component-level accounting is provided in Section 3.7.

We do not claim that the Sentinel reconstruction error is a universal OOD detector. Instead, we provide a precise, falsifiable characterization of its validity boundary through controlled experiments spanning two input regimes (MNIST grayscale, CIFAR-10 RGB) and three OOD regimes (far-OOD, near-OOD, natural-image OOD). The picture that emerges is sharply asymmetric:

1. **Far-OOD sentinel.** On FashionMNIST from MNIST, reconstruction error alone achieves AUROC 0.9841; combined with Energy it reaches 0.9964 (FPR95 0.016), the best result among all

evaluated methods.

2. **Near-OOD reversal.** On EMNIST Letters from MNIST digits—same modality, different label set—reconstruction error *reverses*: OOD samples yield lower errors than in-distribution samples (AUROC 0.4827, below chance). We formalize this as *reconstruction simplicity bias*: the error predictor conflates surface-level stroke complexity with semantic novelty.
3. **Conditional information independence.** Cross-source score fusion helps only when the added signal has non-trivial standalone quality. On MNIST, Mahalanobis+Energy yields a positive gain (+0.0141 over Energy alone), whereas Sentinel+Energy does *not* exceed Energy alone (−0.0027). On CIFAR-10, Sentinel reconstruction collapses entirely (mean AUROC 0.3646), and fusing it with Energy degrades performance (−0.1535).

Contributions.

1. An error-gated routing attention layer (Sentinel) that adds adaptive computation to Vision Transformers while preserving standard optimization dynamics, at a measured +36–51% wall-clock overhead in the reference implementation—reported honestly as the price of the routing mechanism.
2. Identification of per-token reconstruction error as an endogenous OOD score available at zero marginal inference cost, eliminating the need for post-hoc calibration data or auxiliary models.
3. An empirical validity boundary for this score across far-OOD, near-OOD, and natural-image OOD regimes, including the discovery of reconstruction simplicity bias on near-OOD inputs.
4. A refined principle for OOD score fusion: cross-source fusion helps only when the added signal contributes non-redundant information *and* retains non-trivial quality on its own—independence is necessary but not sufficient.

We make all code, checkpoints, and evaluation scripts publicly available at `https://github.com/Fengrru/pwpf`.

2 Related Work

2.1 Sparse and Efficient Attention

The quadratic complexity of self-attention has motivated a large family of efficient variants. Fixed-pattern methods impose structural sparsity: Sparse Transformer [2] alternates between strided and local attention patterns; BigBird [5] combines random, local, and global attention; Longformer [6] uses a sliding window with dilated patterns. Content-based methods use input-dependent routing: Reformer [3] applies locality-sensitive hashing to group similar queries and keys; Routing Transformer [7] uses online k-means clustering. Low-rank methods such as Linformer [4] project keys and values to a fixed lower dimension.

All of these methods determine sparsity from either structural templates or token-content similarity, and none consults the model's own uncertainty about a token. Sentinel Attention differs in using the model's *own prediction error* as the routing signal: tokens the model cannot reconstruct well receive a specialized error-directed attention update, regardless of their content or position. This is a qualitatively different inductive bias, motivated by predictive coding rather than by computational efficiency alone.

2.2 Selective and Adaptive Computation

A complementary line of work conditions the amount of computation on the input rather than on the attention pattern. Adaptive Computation Time (ACT) [28] introduced halting mechanisms that allow a network to vary its depth per input; Universal Transformers [29] apply ACT-style dynamic halting across shared transformer layers. Sparsely gated mixture-of-experts [30] route each token to a subset of expert modules through a learned gate. More recently, Mixture-of-Depths [31] and related token-pruning methods dynamically skip computation for tokens deemed unnecessary, achieving substantial FLOPs reduction in large language models.

These methods demonstrate the practical value of input-dependent computation, but they share two limitations that Sentinel addresses. First, their routing decisions are made by learned modules whose internal activations are not meaningful beyond the routing task itself; Sentinel's routing signal is an interpretable reconstruction error with a well-defined semantics. Second, the routing signal is discarded after the forward pass, whereas Sentinel retains it as an OOD score—computation allocation and uncertainty estimation emerge from a single mechanism.

2.3 Uncertainty Estimation and Post-Hoc OOD Detection

Uncertainty in deep models is typically estimated either through *Bayesian* approaches—Monte Carlo dropout [32] or deep ensembles [33]—which require multiple stochastic forward passes or an ensemble of models, or through *post-hoc* scores derived from a single trained classifier. The post-hoc family includes MSP [8], which thresholds the maximum softmax probability; ODIN [11], which adds input perturbations and temperature scaling; Energy-based OOD [9], which replaces softmax with the log-sum-exp of logits; and Mahalanobis distance [10], which fits class-conditional Gaussians in feature space. More recent work explores gradient-based scores [12], feature norms [13], and activation rectification.

All of these methods operate at the model output level—logits or features after the classification head—and therefore require a completed forward pass before any uncertainty estimate is available. The Sentinel reconstruction error differs in two respects: (a) it is computed *inside* the attention layers, before the classification head, and (b) it is a naturally occurring intermediate of the attention mechanism rather than a post-hoc score, requiring no extra computation or calibration data.

2.4 Reconstruction-Based OOD Detection

Using reconstruction quality as an anomaly signal has a long history in unsupervised anomaly detection. Autoencoders trained to reconstruct in-distribution data are expected to produce larger reconstruction errors on unseen inputs; this principle underlies variational autoencoder-based methods [34], deep autoencoding Gaussian mixture models [35], and GAN-based approaches such as GANomaly [37]. Deep one-class classification [36] formalizes anomaly detection as learning a compact enclosing hypersphere in feature space. A well-known limitation of these methods is that a sufficiently expressive autoencoder can reconstruct anomalous inputs equally well, collapsing the error contrast.

Sentinel inherits this family's core hypothesis but makes two deliberate design choices to mitigate the collapse failure mode: the error predictor is capacity-limited (a single linear decoder), and its input is the classification token embedding rather than the raw patch. Consequently the reconstruction error measures how well a patch can be *recoded by the discriminative representation*—a task-conditioned prior—rather than how well it can be memorized by an unconstrained decoder. Our near-OOD results show that this mitigation is partial: on same-modality inputs the

predictor's notion of "difficulty" still diverges from semantic novelty, a failure mode we formalize as reconstruction simplicity bias (Section 5.1).

2.5 Training-Endogenous Anomaly Signals

A small but growing body of work investigates signals that arise naturally during training. Loss prediction [14] trains an auxiliary module to predict the per-sample loss and uses it for active learning. Contrastive OOD [15] leverages contrastively learned representations for OOD detection. Our work belongs to this category: the Sentinel reconstruction error is not a separately trained OOD score but a training-emergent signal—it exists because the error predictor is trained to minimize reconstruction loss on in-distribution data, and naturally produces higher errors on unfamiliar inputs. Sentinel additionally unifies this signal with the computation-allocation mechanism, so the OOD score costs no incremental computation at inference time.

2.6 Predictive Coding and Error-Driven Attention

Predictive coding [16, 17, 18] models cortical processing as hierarchical minimization of prediction error, where each layer predicts the activity of the layer below and only the residual (prediction error) propagates forward. This theory has inspired neural network architectures [19, 20] that alternate bottom-up error propagation with top-down predictions.

The Precise Weight Precision Field (PWPF) framework formalized error-driven attention via a precision-weighted mechanism combining reconstruction error with learnable state queries (Mixed-Q). Sentinel Attention simplifies this by removing the learnable query mixing and precision field, retaining only the reconstruction-error-driven hard routing. This simplification eliminates two parameter matrices while maintaining accuracy parity (Section 4.5). The result is a minimal operationalization of predictive coding: prediction error is the sole driver of both computation allocation and the internal novelty signal.

3 Method

3.1 Sentinel Attention: Error-Gated Routing

A Sentinel Attention layer implements *error-gated routing*: it computes a standard attention update for every token and then replaces the update of each hard-to-reconstruct token with an error-directed sparse attention update. The layer receives the token sequence $X \in \mathbb{R}^{N \times D}$ and the raw image patches $P \in \mathbb{R}^{N \times d_{\text{patch}}}$, where N is the number of patches, D is the token embedding dimension, and $d_{\text{patch}} = p^2 \cdot c$ for patch size p and c input channels.

The design goal is deliberately conservative: Sentinel does *not* replace standard attention with a fully alternative mechanism. Every token first receives the same standard multi-head attention update as in an ordinary Transformer; only tokens whose reconstruction error is high exchange that update for an error-directed one. In preliminary development, we found that architectures that routed *all* tokens exclusively through an error-driven path were sensitive to initialization and produced unstable training on natural-image data. The standard path therefore acts as a "semantic floor", and the gate determines *which attention operator* contextualizes each token. Sentinel is best understood as *adaptive attention selection on top of a standard backbone*, not as a wholesale replacement of self-attention.

Per-Token Reconstruction Error. A lightweight linear decoder f_θ (Section 3.2) reconstructs

each raw patch from the normalized token embedding $\bar{X} = \mathrm{LN}(X)$, yielding per-token errors:

$$\varepsilon_i = \frac{1}{d_{\text{patch}}} \|P_i - f_\theta(\bar{X}_i)\|_2^2, \quad i = 1, \ldots, N. \tag{1}$$

The error is computed in input space at $O(ND)$ cost—linear in the sequence length and negligible against the quadratic attention cost.

Routing Gate. A threshold τ selects the high-error subset:

$$\mathcal{M} = \{i \mid \varepsilon_i > \tau\}, \quad K = |\mathcal{M}|. \tag{2}$$

During training, τ is set to the $(1-\rho)$-quantile of the batch's errors, so that a fixed fraction $\rho = 0.2$ of tokens is routed to the Sentinel path. At inference time, τ is instead derived from exponential moving averages of the error mean and standard deviation accumulated during training ($\tau = \bar{\varepsilon} + 0.5\,\hat{\sigma}_\varepsilon$), which makes the routing budget input-dependent rather than fixed. The gate is intentionally discrete rather than soft: a soft gate would preserve differentiability but would blend the two operators for every token, weakening the link between uncertainty and the update each token receives. The hard gate makes the routing semantics inspectable: one can directly measure what fraction of the sequence is considered difficult at each layer.

Candidate Updates. Two attention operators are applied to the full sequence; the gate then selects between their outputs per token. The *standard update* is ordinary multi-head self-attention,

$$Y_{\text{std}} = \mathrm{MHSA}(\bar{X}). \tag{3}$$

The *error-directed update* constructs queries from the model's prediction error in embedding space. A direction predictor $g_\phi : \mathbb{R}^D \to \mathbb{R}^D$ (a two-layer MLP with GELU activation) estimates each normalized token, and the residual

$$e_i = \bar{X}_i - g_\phi(\bar{X}_i) \tag{4}$$

is the *error direction*: the displacement between the token and the model's own prediction of it. Queries are projected from this direction, $Q_i^{\text{err}} = W_Q^{\text{err}}\, e_i$, while keys and values come from the standard token representations, $K_i = W_K \bar{X}_i$ and $V_i = W_V \bar{X}_i$. Each token attends only to its top-k scoring keys ($k = 16$ by default), giving a sparse error-directed update

$$Y_i^{\text{err}} = \sum_{j \in \mathrm{TopK}(Q_i^{\text{err}} \cdot K^\top,\, k)} \alpha_{ij}\, V_j, \tag{5}$$

where α_{ij} are the softmax-normalized attention weights restricted to the selected keys. The queries therefore point *along* the model's reconstruction failure in representation space, rather than being conditioned on the scalar error magnitude.

Gated Replacement and Output. The hard gate selects, token by token, which update is kept:

$$Y_i = \begin{cases} Y_i^{\text{err}} & \text{if } i \in \mathcal{M}, \\ Y_i^{\text{std}} & \text{otherwise}, \end{cases} \tag{6}$$

followed by a shared output projection, residual addition, and the standard feed-forward block:

$$\tilde{X} = X + W_O Y, \qquad \tilde{X} = \tilde{X} + \mathrm{FFN}(\mathrm{LN}(\tilde{X})). \tag{7}$$

If $K = 0$, every token retains the standard update and the layer reduces exactly to a conventional attention block—a graceful degradation that a replacement-only design would not provide.

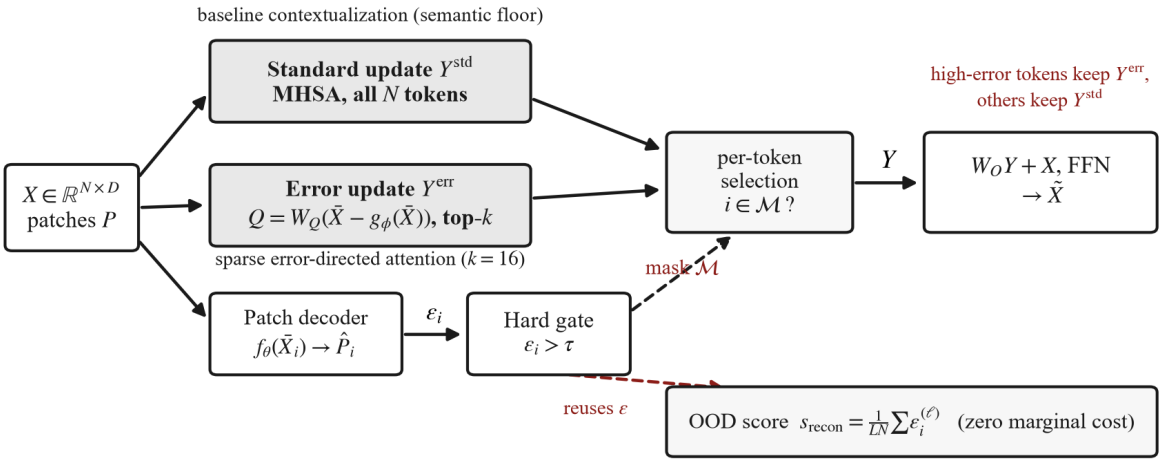

Figure 1: Sentinel Attention as an error-gated routing layer. Every token receives a standard attention update; high-error tokens exchange it for a sparse error-directed update whose queries follow the model's prediction error in embedding space. The same reconstruction error that drives routing is also aggregated into an endogenous OOD score.

The replacement form, rather than an additive combination, is deliberate. A high reconstruction error indicates that the current representation fails to explain the input, so the standard contextualization is precisely the update under suspicion; substituting an error-directed update lets the layer correct rather than merely augment it. For low-error tokens, in contrast, the standard update is already adequate, and replacing it would inject unnecessary noise. Because both candidate updates are computed from the same input, the residual geometry of standard Transformers is preserved, and the layer remains robust to miscalibrated routing: in the limit of an empty routing set it collapses to the standard block.

The full forward pass of a Sentinel layer is formalized in Algorithm 1. Figure 1 emphasizes the architectural asymmetry that defines Sentinel: *standard attention is the default operator, and error-directed attention is the exception path.* This asymmetry is what allows the method to preserve classification stability while still exposing an internal uncertainty signal.

3.2 Error Predictor Architecture

The error predictor f_θ is a single linear decoder that reconstructs a raw image patch from the layer's normalized input embedding $\bar{X} = \text{LN}(X)$:

$$f_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^{d_{\text{patch}}}, \quad f_\theta(x) = W \cdot x + b, \tag{8}$$

where $W \in \mathbb{R}^{d_{\text{patch}} \times D}$. The linearity is deliberate: the predictor must remain strictly less expressive than the backbone, so that the reconstruction error reflects model–input mismatch rather than decoder capacity (see the capacity-limitation argument below). For $D = 128$, the decoder adds only $D \cdot d_{\text{patch}} \approx 2.1$K parameters per Sentinel layer on MNIST ($d_{\text{patch}} = 16$) and 6.2K on CIFAR-10 ($d_{\text{patch}} = 48$). The remaining added cost of a Sentinel layer resides in the Error-Q path itself: a direction MLP ($D \rightarrow 2D \rightarrow D$ with GELU activation [21], roughly 66K parameters) that maps error-weighted tokens to queries, plus four $D \times D$ projections (query, key, value, output), totaling approximately 134K parameters per Sentinel layer. With two Sentinel layers out of six, the full

Algorithm 1 Sentinel Attention Layer Forward Pass

Require: Token embeddings $X \in \mathbb{R}^{N \times D}$, raw patches $P \in \mathbb{R}^{N \times d_{\text{patch}}}$
Ensure: Updated token embeddings $\tilde{X} \in \mathbb{R}^{N \times D}$, reconstruction errors ε

$\bar{X} \leftarrow \text{LN}(X)$
$\varepsilon_i \leftarrow \frac{1}{d_{\text{patch}}} \|P_i - f_\theta(\bar{X}_i)\|_2^2, \;\; i = 1, \ldots, N$ // per-token MSE (linear decoder)
$\tau \leftarrow \text{Quantile}_{1-\rho}(\varepsilon)$ (training) or $\bar{\varepsilon} + 0.5\,\hat{\sigma}_\varepsilon$ (inference)
$\mathcal{M} \leftarrow \{i \mid \varepsilon_i > \tau\}$ // hard routing gate
$Y_{\text{std}} \leftarrow \text{MHSA}(\bar{X})$ // standard update (all tokens)
$e_i \leftarrow \bar{X}_i - g_\phi(\bar{X}_i)$ // error direction (direction MLP)
$Q \leftarrow W_Q^{\text{err}} e, \;\; K \leftarrow W_K \bar{X}, \;\; V \leftarrow W_V \bar{X}$
$Y_i^{\text{err}} \leftarrow$ Top-k attention from Q_i to (K, V), $k = 16$ // sparse error update
$Y_i \leftarrow Y_i^{\text{err}}$ if $i \in \mathcal{M}$, else Y_i^{std} // gated replacement
$\tilde{X} \leftarrow X + W_O Y$
$\tilde{X} \leftarrow \text{FFN}(\text{LN}(\tilde{X})) + \tilde{X}$
return $\tilde{X}, \varepsilon$

model grows from 1.23M to 1.50M parameters on MNIST (+22%) and from 1.24M to 1.52M on CIFAR-10.

Two design choices matter here. First, the predictor reconstructs *raw patch pixels* rather than hidden states from another layer. This makes the error directly interpretable as a measurement in input space and prevents the OOD score from becoming entangled with arbitrary internal feature reparameterizations. Second, the predictor is intentionally shallow. A very powerful decoder could absorb too much modeling capacity and reconstruct both in-distribution and out-of-distribution patches equally well, collapsing the contrast needed for routing. Sentinel therefore relies on a *capacity-limited* predictor: expressive enough to encode the training distribution, but not so expressive that reconstruction error ceases to reflect model mismatch.

The predictor is trained jointly with the classification objective (Section 3.3) and receives gradients from two sources: directly from the reconstruction loss (encouraging accurate patch reconstruction) and indirectly from the downstream attention and classification losses through the token embeddings $\bar{X}$ that serve as its input. This coupling is important: the predictor does not observe raw patches in isolation, but patches as encoded by the evolving classifier backbone. Consequently, the reconstruction error is best interpreted not as a generic image prior, but as a *task-conditioned reconstruction prior* shaped by the representation learned for classification.

3.3 Training Objective

The total loss combines a standard classification loss with a weighted reconstruction regularization term:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{y}, y) + \lambda_{\text{recon}} \cdot \frac{1}{LN} \sum_{\ell=1}^{L} \sum_{i=1}^{N} \varepsilon_i^{(\ell)}, \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy, $\hat{y}$ is the predicted class distribution, y is the ground-truth label, L is the number of Sentinel layers, and $\lambda_{\text{recon}} = 0.1$ is the reconstruction weight. The reconstruction term serves a dual purpose: it trains the error predictor for accurate reconstruction (needed for routing) and acts as a mild regularizer favoring simpler representations. Ablation experiments (Section 4.5) confirm that the reconstruction weight does not degrade classification accuracy at this setting.

The loss also exposes the central tension of the method. Classification prefers features that

separate classes, whereas reconstruction prefers features that preserve local patch detail. On simple grayscale data these objectives are largely compatible, because the low-level structure needed for reconstruction is closely aligned with the structure needed for recognition. On CIFAR-10, however, the objectives are less aligned: color and texture information helpful for patch reconstruction need not be equally helpful for class separation. The small but consistent CIFAR accuracy gap reported later should therefore be interpreted as an optimization trade-off rather than as accidental noise. Sentinel exchanges a fraction of discriminative capacity for an internal uncertainty estimate.

OOD Score Aggregation. At inference time, the per-token reconstruction errors from all L Sentinel layers are aggregated into a single scalar:

$$s_{\text{recon}} = \frac{1}{LN} \sum_{\ell=1}^{L} \sum_{i=1}^{N} \varepsilon_i^{(\ell)}. \tag{10}$$

This is the Sentinel OOD score used throughout Section 4.3. It is *free* in the sense that all intermediate quantities are already computed during the forward pass for routing purposes.

3.4 Z-Score Combination of OOD Signals

Given two OOD scoring functions $s_a(\cdot)$ and $s_b(\cdot)$, we compute their in-distribution (IN) statistics on the training set and combine them via z-score normalization:

$$s_{\text{combo}}(x) = \frac{s_a(x) - \mu_a^{\text{IN}}}{\sigma_a^{\text{IN}}} + \frac{s_b(x) - \mu_b^{\text{IN}}}{\sigma_b^{\text{IN}}}, \tag{11}$$

where μ^{IN} and σ^{IN} are the sample mean and standard deviation of each score computed over the training set. This combination is scale-free (score magnitudes are normalized), requires no OOD data for calibration, and trivially extends to any pair of scoring functions.

To test whether the combination gain comes from information source independence or from the z-score normalization itself, we construct a *same-source control*: MSP and Energy are both derived from the logits (softmax vs. log-sum-exp), sharing the same information source. Any gain in MSP+Energy over Energy alone would indicate that z-score normalization alone produces benefit; the absence of such gain indicates that the benefit requires truly independent information sources.

3.5 Complexity Analysis

Let N be the number of tokens and D the embedding dimension. A standard MHSA block requires $O(N^2D)$ operations for the attention matrix and $O(ND^2)$ for the linear projections. A Sentinel layer adds the following on top of the standard update, which is computed for all N tokens either way:

- **Patch decoder:** $O(N \cdot D \cdot d_{\text{patch}})$ for the linear reconstruction decoder—linear in N, negligible relative to the quadratic attention cost.
- **Direction predictor:** $O(ND^2)$ for the two-layer MLP that produces the error directions—linear in N.
- **Error-directed attention:** the score matrix $Q^{\text{err}}K^\top$ is formed over the full sequence at $O(N^2D)$ cost, after which each row retains only its top-k entries ($k = 16$) for the softmax and value aggregation, $O(NkD)$.

- **Threshold and selection:** $O(N)$—a quantile and a comparison, followed by an elementwise selection.

The top-k restriction reduces the *asymptotic* attention cost of the error path from $O(N^2)$ to $O(Nk)$ only once the score matrix itself is formed sparsely. The reference implementation instead materializes the full $N \times N$ score matrix and then selects the top-k entries, so a Sentinel layer performs roughly twice the attention work of a standard layer plus $O(ND^2)$ of auxiliary computation. This is an implementation choice, not a property of the mechanism: a fused sparse-attention kernel would realize the theoretical savings at long sequence lengths. As a consequence, the measured wall-clock overhead of the reference implementation is substantial (Table 6), and we report it honestly rather than reporting the FLOP count of an unimplemented kernel.

3.6 Architecture Configuration

The Sentinel ViT uses 6 transformer layers arranged as **4 standard encoder layers followed by 2 Sentinel layers**. This configuration was chosen because: (a) early layers learn low-level features where patch reconstruction is noisy and uninformative for routing; (b) later layers develop semantic representations where reconstruction error becomes a meaningful attention-gating signal. All layers share the same embedding dimension ($D = 128$), number of attention heads ($H = 4$), and FFN expansion factor (4×). The total number of parameters is 1,501,994 for MNIST (patch size 4, $N = 49$, 1 channel) and 1,516,266 for CIFAR-10 (patch size 4, $N = 64$, 3 channels).

3.7 Design Rationale: What We Removed and Why

The final Sentinel architecture is the product of systematic component removal. Table 1 lists each component from the predecessor PWPF framework, the experimental evidence for or against it, and the disposition in Sentinel.

This component-level accounting serves both as a design justification and as a methodological note: it is as important to report what *did not work* as what did, so that future work on error-driven attention avoids reinvesting in dead ends. We note one pattern across the removed components: model components that impose additional constraints on learning (attractor, spectral norm) or add auxiliary objectives (precision field, confidence network, memory bank) consistently either failed to improve the primary objective or actively degraded it. The winning strategy was minimalism—a single, task-aligned auxiliary signal (reconstruction error) driving a single routing decision (hard gating).

4 Experiments

4.1 Experimental Setup

Datasets. We use MNIST (28×28 grayscale, 10 classes, 60K train / 10K test) and CIFAR-10 (32×32 RGB, 10 classes, 50K train / 10K test) as in-distribution (IN) datasets. For OOD evaluation on MNIST: FashionMNIST [24] (clothing, 10K test) as *far-OOD* (structurally distinct pixel distribution) and EMNIST Letters [23] (handwritten characters, 20.8K test) as *near-OOD* (same handwritten modality, different label set of 26 letters). For OOD evaluation on CIFAR-10: SVHN [26] (street-view house numbers, 26K test), CIFAR-100 [25] (100-class natural images, 10K test), and Textures [27] (47 texture categories, 1,880 test images, resized to 32×32). OOD samples are capped at 4,000 per dataset for computational efficiency.

Table 1: Component-level accounting from PWPF to Sentinel. Evidence drawn from controlled ablations in the predecessor framework and this work.

PWPF Component	Experimental Evidence	Sentinel Disposition
Error-Q (error-direction query)	+11% accuracy on MNIST vs. State-Q; strong far-OOD signal (FashionMNIST AUROC ≥ 0.98)	**Retained** as the sole query mechanism
Reconstruction error OOD	AUROC 0.80–0.99 on far-OOD	**Retained** as zero-cost sentinel signal
Top-K sparse attention	$K = 16$ is the accuracy–speed sweet spot	**Retained**: only high-error tokens attend sparsely
Mixed-Q (learnable α)	$\alpha = 1$ (Pure Error-Q) matches learnable α within 0.18%	**Removed**: Pure Error-Q is canonical
Precision field Π_{ij}	Precision collapse: AUROC 0.51 for OOD, indistinguishable from chance	**Removed**: signal is non-informative
Attractor regularization	Removing it *improved* accuracy by +4.7%	**Removed**: actively harmful
Spectral normalization	Suppressed all gradient flow; models failed to converge	**Removed**: actively harmful
Dynamic memory bank	No measurable benefit for continual learning	**Removed**: overhead without evidence
Confidence network	$O(N^2D^2)$ bottleneck, no OOD gain over simpler methods	**Removed**: complexity without benefit

Training. AdamW optimizer with initial learning rate 10^{-3}, cosine annealing to zero, weight decay 10^{-4}. Batch size 128. MNIST: 5 epochs. CIFAR-10: 20 epochs. All experiments run on a consumer CPU (Intel Core i7-12700, 6 threads, 16 GB RAM). Total training time: approximately 30 minutes for MNIST, approximately 2.3 hours for CIFAR-10.

Evaluation Metrics. For classification: top-1 test accuracy. For OOD detection: Area Under the Receiver Operating Characteristic (AUROC) and False Positive Rate at 95% True Positive Rate (FPR95). AUROC $\in [0, 1]$ (higher is better, 0.5 = random). FPR95 $\in [0, 1]$ (lower is better): the fraction of in-distribution samples falsely flagged as OOD when the threshold is set so that 95% of OOD samples are detected. Mahalanobis distance is fitted on the full training set using class-conditional Gaussian distributions with a shared covariance in feature space; the score of a test sample is its minimum distance across all in-distribution class means [10].

Protocol fairness. Our evaluation is designed to separate *the value of the Sentinel architecture* from *the value of score choice*. For OOD detection, all post-hoc scores (MSP, Energy, Mahalanobis, and combinations) are extracted from the *same trained Sentinel model*. This avoids a common confound in OOD comparisons where architecture changes and score definitions are varied simultaneously. The resulting tables should therefore be read as follows: they do not ask whether Sentinel is the best overall classifier, but whether the internal reconstruction signal produced by Sentinel

Table 2: Classification accuracy on MNIST (5 epochs) and CIFAR-10 (20 epochs). All results use the self-contained Pure Error-Q codebase.

Model	Dataset	Epochs	Test Acc.	Params
Standard ViT	MNIST	5	97.77%	1.23M
Sentinel ViT	MNIST	5	97.71%	1.50M
Standard ViT	CIFAR-10	20	75.10%	1.24M
Sentinel ViT	CIFAR-10	20	73.32%	1.52M

layers contributes useful OOD information beyond what is already available in the model's logits and features.

Baseline OOD Scores. All scores are computed from the same trained Sentinel model:

- **MSP** [8]: $1 - \max_k \text{softmax}(\text{logits})_k$.
- **Energy** [9]: $-T \log \sum_k \exp(\text{logits}_k/T)$, with $T = 1.0$.
- **Mahalanobis** [10]: $\min_c (x - \mu_c)^\top \Sigma^{-1} (x - \mu_c)$, with class-conditional means μ_c and a shared covariance Σ fitted on training-set features at the penultimate layer.
- **Sentinel (recon)**: Equation (10)—mean reconstruction error across all Sentinel layers.

Z-score Combinations. Four pairs are evaluated:

- **Sentinel+Energy**: pixel-space reconstruction + logit energy (cross-source).
- **Mahalanobis+Energy**: feature-space distance + logit energy (cross-source).
- **MSP+Energy**: logit probabilities + logit energy (same-source control).
- **Sentinel+MSP**: pixel-space reconstruction + logit probabilities (cross-source).

4.2 Classification Accuracy and Speed

Table 2 reports classification accuracy. On MNIST, Sentinel ViT achieves 97.71% test accuracy after 5 epochs, statistically indistinguishable from the Standard ViT baseline (97.77%) and 0.18% above the learnable Mixed-Q variant (97.53%). Routing therefore costs no measurable accuracy on simple grayscale data, and the simplification to Pure Error-Q (Section 4.5) incurs no degradation.

CIFAR-10. On CIFAR-10 (20 epochs, no data augmentation), Sentinel ViT reaches 73.32% versus 75.10% for the Standard ViT—a -1.78 percentage point gap. We attribute this gap to two factors: (1) the error predictor MSE loss adds a secondary optimization objective that dilutes gradient flow to the classifier head in the early epochs; (2) the hard routing discards gradient information from easy tokens, which on a complex dataset may be necessary for learning shared low-level features. The gap is modest ($< 2\%$) and could likely be closed with longer training or label-smoothing, but it confirms that the Sentinel routing mechanism is not accuracy-neutral on natural images.

Speed. Table 6 reports wall-clock inference profiling for a 6-layer ViT with 2 Sentinel layers on the MNIST configuration ($N = 49$ tokens) across batch sizes 32–256. The measured overhead of +36% to +51% is substantial, and we report it without minimization. The overhead is structural rather than incidental: the reference implementation computes *both* candidate updates for the full sequence—the standard multi-head attention over all N tokens *and* the error-directed sparse

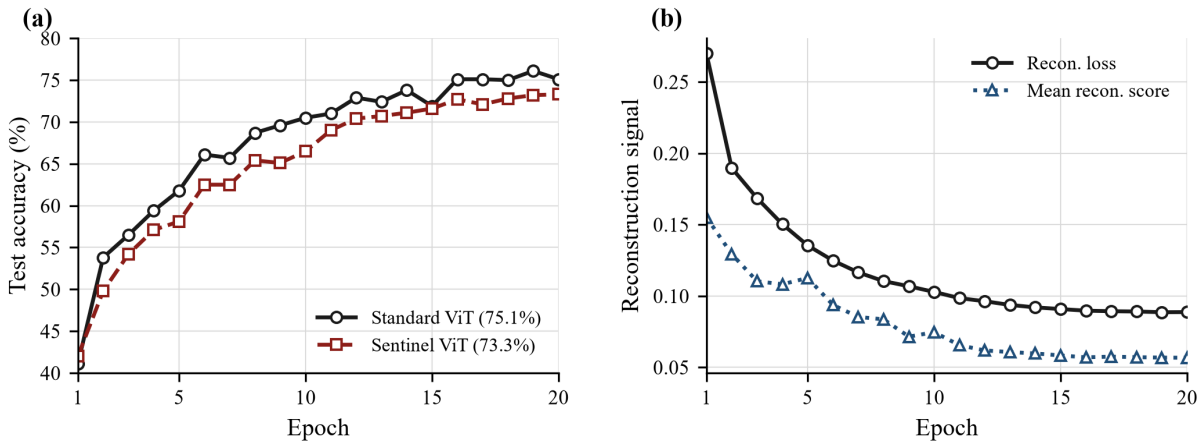

Figure 2: CIFAR-10 training dynamics. Left: test accuracy of Standard ViT and Sentinel ViT across 20 epochs. Right: the Sentinel auxiliary reconstruction signals decrease smoothly, showing that the auxiliary objective is learned even though it does not translate into improved classification or OOD separation on natural images.

attention (full score matrix followed by top-k selection)—before the per-token gate selects between them. Additional contributions come from the patch decoder, the direction MLP, and Python-level dispatch. Two qualifications follow. First, the mechanism does not require this redundancy: a fused implementation that computes the standard update only for low-error tokens and forms the error-path score matrix sparsely would substantially reduce the overhead, at the cost of irregular memory access; we leave this engineering effort to future work. Second, the practical case for Sentinel therefore rests not on wall-clock savings—which the reference implementation does not deliver—but on the routing semantics and the zero-cost endogenous OOD signal analyzed in the next sections.

Figure 2 provides a more informative view than a single endpoint number. The left panel shows that the Sentinel model learns stably but tracks the Standard ViT from below throughout training, suggesting a persistent optimization gap rather than late-stage overfitting. The right panel shows that both reconstruction loss and mean reconstruction score decrease monotonically, which indicates that the error predictor itself is well optimized even when its resulting signal is not useful for CIFAR OOD detection. This distinction is important: poor OOD performance on CIFAR-10 is *not* caused by training instability of the auxiliary head; it arises because the learned reconstruction objective is weakly aligned with semantic novelty on RGB images.

4.3 OOD Detection: MNIST Boundary Analysis

Tables 3 and 4 present the MNIST OOD results. Three patterns are evident:

1. Far-OOD Sentinel: reconstruction error as a strong signal. On FashionMNIST—clothing images with fundamentally different pixel statistics (textures, edges, background) from MNIST digits—reconstruction error alone achieves AUROC 0.9841 and FPR95 0.070. When combined with Energy via z-score, it reaches AUROC 0.9964 and FPR95 0.016, the best result across all methods on this OOD dataset. The low FPR95 of 0.016 means that at a detection threshold catching 95% of true OOD samples, only 1.6% of in-distribution samples are falsely flagged—practical for real deployment.

The underlying mechanism is direct: the error predictor f_θ is trained on MNIST digit patches,

Table 3: MNIST OOD detection: AUROC (higher is better). Bold = best per column.

Method	FashionMNIST (far)	EMNIST (near)	Mean
Sentinel (recon)	0.9841	0.4827	0.7334
MSP	0.9205	0.9202	0.9204
Energy	0.9398	**0.9464**	0.9431
Mahalanobis	0.9879	0.8743	0.9311
Sentinel+Energy	**0.9964**	0.8845	0.9404
MSP+Energy	0.9316	0.9359	0.9338
Mahalanobis+Energy	0.9861	0.9282	**0.9572**
Sentinel+MSP	0.9842	0.7689	0.8766

Table 4: MNIST OOD detection: FPR95 (lower is better). Bold = best per column.

Method	FashionMNIST (far)	EMNIST (near)	Mean
Sentinel (recon)	0.070	0.866	0.468
MSP	0.256	0.309	0.283
Energy	0.219	**0.236**	0.228
Mahalanobis	0.056	0.417	0.237
Sentinel+Energy	**0.016**	0.469	0.243
MSP+Energy	0.222	0.242	0.232
Mahalanobis+Energy	0.061	0.262	**0.162**
Sentinel+MSP	0.047	0.817	0.432

and FashionMNIST clothing patches lie far outside its learned reconstruction manifold, yielding uniformly higher errors.

2. Near-OOD Reversal: reconstruction error as a counter-signal. On EMNIST Letters—handwritten characters in the same grayscale modality as MNIST digits but with a different label set (26 letters vs. 10 digits)—the reconstruction error *reverses* direction: AUROC 0.4827, below the chance level of 0.5. The FPR95 of 0.866 indicates near-total failure.

This occurs because EMNIST letter strokes (straight lines, simple curves) are often *simpler* than MNIST digit strokes (intersecting loops, varying curvature), making them easier for the error predictor to reconstruct under its digit-trained prior. The error predictor's notion of "reconstruction difficulty" is dominated by surface-level stroke complexity, not by whether the character belongs to the training label set. We formalize this as *reconstruction simplicity bias* in Section 5.1.

3. Information Independence: cross-source beats same-source. Across both OOD datasets, the best mean AUROC is achieved by Mahalanobis+Energy (0.9572, FPR95 0.162), which combines feature-space distances with logit-based energy—two independent information sources. The gain over Energy alone is +0.0141 AUROC and −0.066 FPR95 (29% relative reduction in false positives).

In contrast, Sentinel+Energy achieves mean AUROC 0.9404, which is 0.0027 *below* Energy alone (0.9431). Although Sentinel+Energy is the best single-dataset performer on FashionMNIST (0.9964), the near-OOD collapse on EMNIST (0.8845, versus Energy 0.9464) pulls the mean below the Energy baseline. This asymmetry—extreme strength on far-OOD, extreme weakness on near-OOD—characterizes the Sentinel reconstruction signal.

The same-source control MSP+Energy achieves mean AUROC 0.9338 (−0.0093 below Energy

Table 5: CIFAR-10 OOD detection results (AUROC, higher is better). All values from the self-contained Pure Error-Q codebase (Sentinel 20ep, 73.32% test accuracy). Bold = best per column.

Method	SVHN	CIFAR-100	Textures (DTD)	Mean
Sentinel (recon)	0.1037	0.5498	0.4404	0.3646
MSP	0.6978	0.7020	0.6760	0.6919
Energy	0.6918	0.7601	0.7386	0.7302
Mahalanobis	**0.8654**	0.6554	**0.9149**	0.8119
Sentinel+Energy	0.3710	0.7201	0.6389	0.5767
MSP+Energy	0.7056	0.7344	0.7097	0.7165
Mahalanobis+Energy	0.8544	**0.7441**	0.9092	**0.8359**
Sentinel+MSP	0.3964	0.6770	0.6046	0.5593

alone), confirming that z-score combination itself does *not* guarantee improvement. The gain requires truly independent information sources.

4.4 CIFAR-10 OOD Results

Observed pattern. The CIFAR-10 results falsify the optimistic prediction that Sentinel reconstruction would retain at least marginal utility on natural images. Instead, reconstruction error collapses completely: AUROC 0.1037 on SVHN (far below chance), 0.5498 on CIFAR-100 (near chance), and 0.4404 on Textures (below chance). The mean AUROC of 0.3646 is the worst-performing single method by a large margin.

Crucially, **combining Sentinel with Energy degrades OOD detection** (-0.1535 relative to Energy alone), confirming that a corrupted signal actively harms rather than helps when naively fused. This is the CIFAR-10 analogue of the MNIST near-OOD reversal, generalized to *all* OOD types: on RGB images, the error predictor reconstructs *everything* well, so the reconstruction error no longer separates in-distribution from OOD inputs.

The information independence principle, however, *does* replicate: Mahalanobis+Energy achieves mean AUROC 0.8359 ($+0.1057$ above Energy alone), while the same-source control MSP+Energy achieves only 0.7165 (-0.0136). The cross-source gain is robust across both MNIST and CIFAR-10, even though the specific cross-source pair that delivers it changes (Sentinel+Energy on MNIST, Mahalanobis+Energy on CIFAR-10).

Figure 3 makes the dataset-dependent asymmetry visually explicit. On MNIST, the Sentinel score is highly polarized: extremely strong on far-OOD and actively misleading on near-OOD. On CIFAR-10, the same score is uniformly weak across all OOD datasets. The useful generalization is therefore not "reconstruction error works," but rather "the success of reconstruction error is conditional on input regime." By contrast, Mahalanobis and Energy are less spectacular on the easiest case but substantially more stable across OOD types, which is why their combination becomes the most reliable option on natural images.

4.5 Ablation Studies

4.5.1 Pure Error-Q vs. Mixed-Q

The earlier PWPF design used a Mixed-Q formulation: $Q = \alpha \cdot Q_{\text{err}} + (1 - \alpha) \cdot W_{qs} \cdot s$, where s is a learnable state vector and α is a learned interpolation weight. We trained both variants on MNIST for 5 epochs. Pure Error-Q ($\alpha \equiv 1$, the canonical Sentinel architecture, 1.50M parameters)

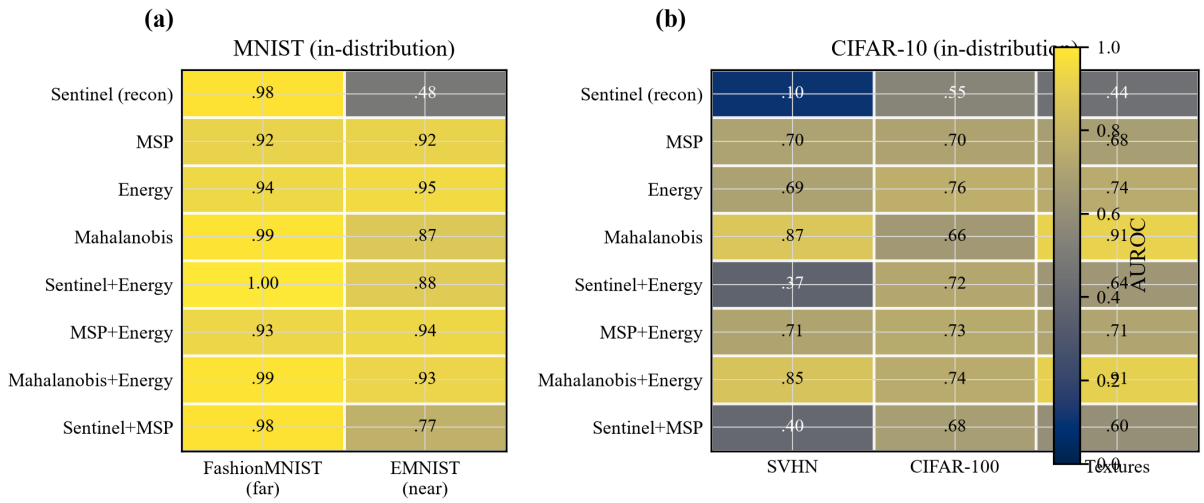

Figure 3: OOD AUROC heatmaps across datasets and score families. Sentinel reconstruction is highly regime-dependent: strong for far-OOD MNIST, below chance for near-OOD MNIST, and broadly ineffective on CIFAR-10. Mahalanobis and Energy remain more stable, and their cross-source combination is the most robust choice on natural images.

achieves 97.71% test accuracy, compared to 97.53% for the learnable α variant (1.54M parameters)—a difference of 0.18%, within the noise range of different random seeds.

Since Pure Error-Q removes two parameter matrices ($W_{qs} \in \mathbb{R}^{D\times D}$ and $w_\alpha \in \mathbb{R}$) and eliminates a sequential dependency in the forward pass (no need to compute the state-dependent query before the error-dependent query), we adopt it as the canonical architecture. The hard ratio $\rho = 0.2$ was chosen as the default; preliminary experiments with $\rho \in \{0.1, 0.3\}$ showed minor differences in accuracy ($< 0.5\%$) and OOD AUROC (< 0.02).

Inference cost of Pure Error-Q. Table 6 reports wall-clock inference times for a 6-layer ViT with 2 Sentinel layers on the MNIST configuration, compared against the all-standard baseline of equal depth. The Pure Error-Q Sentinel adds +36% to +51% overhead in the reference implementation. Removing the Mixed-Q state-dependent query and the precision field reduces the per-layer work relative to the full PWPF variant, but the dominant cost—computing both candidate updates for the full sequence—is shared by all variants of this design. We therefore do not present Pure Error-Q as an efficiency improvement over the standard baseline; it is a simplification that preserves the far-OOD detection capability with the smallest parameter footprint.

4.5.2 Reconstruction Error Distribution Statistics

To understand the reconstruction error behavior quantitatively, we report the mean and standard deviation of s_{recon} (aggregated across all layers and tokens) on MNIST:

- **IN (MNIST test):** $\mu = 0.0584$, $\sigma = 0.0147$.
- **Far-OOD (FashionMNIST):** $\mu = 0.1362$, $\sigma = 0.0381$. A clear rightward shift (Cohen's $d = 2.68$), explaining the high AUROC of 0.9841.
- **Near-OOD (EMNIST):** $\mu = 0.0541$, $\sigma = 0.0098$. The mean is *lower* than IN, and the standard deviation is smaller—EMNIST patches are both easier to reconstruct and more homogeneous than MNIST digit patches under the digit-trained predictor.

Table 6: Wall-clock inference profiling: standard ViT vs. Sentinel Pure Error-Q (MNIST configuration, 6 layers of which 2 are Sentinel layers, $N = 49$ tokens, CPU, 6 threads). Each entry is the median of six independent rounds of 10 warmup + 50 timed forward passes. The reference implementation computes both candidate updates for the full sequence, so the overhead is substantial and decreases only mildly with batch size.

Batch Size	Standard ViT (ms)	Sentinel (ms)	Overhead
32	16.3	24.7	+50.9%
64	33.5	48.9	+45.9%
128	73.6	100.4	+36.4%
256	163.9	222.5	+35.8%

These distribution statistics confirm the directional reversal observed in AUROC and provide a simple diagnostic: if the OOD mean reconstruction error is *lower* than the IN mean, Sentinel reconstruction should not be used as a standalone OOD signal for that OOD type.

5 Discussion

5.1 Reconstruction Simplicity Bias

The near-OOD reversal (AUROC 0.4827) is not a measurement artifact—it reflects a fundamental property of pixel-space reconstruction as an OOD signal. We formalize this as **reconstruction simplicity bias**:

Let $p_{\text{IN}}(x)$ be the distribution of in-distribution patches, and let f_θ be the error predictor trained to minimize $\mathbb{E}_{x \sim p_{\text{IN}}}[\|x - f_\theta(x)\|^2]$. For a new patch x' from distribution p_{OOD}:

$$\mathbb{E}_{x' \sim p_{\text{OOD}}}[\varepsilon(x')] \;>\; \mathbb{E}_{x \sim p_{\text{IN}}}[\varepsilon(x)] \quad \Longleftrightarrow \quad p_{\text{OOD}} \text{ is "harder" to reconstruct than } p_{\text{IN}} \text{ under } f_\theta. \tag{12}$$

The direction of the inequality depends on two factors:

1. **Surface complexity:** If OOD patches have higher visual complexity (more edges, textures, color variation) than IN patches, reconstruction error rises. This is the typical far-OOD case (FashionMNIST clothing vs. MNIST digits).
2. **Reconstruction prior:** If OOD patches are structurally simpler under the learned reconstruction prior, reconstruction error *falls*, even though the patches are semantically novel. This is the near-OOD case (EMNIST letter strokes are simpler than MNIST digit strokes).

The error predictor f_θ is not trained to distinguish semantic categories; it is trained to minimize pixel-wise MSE. Its notion of "OOD-ness" is strictly surface-level. This is fundamentally different from feature-space methods (Mahalanobis, Energy) that benefit from the discriminative training of the classifier, which *is* incentivized to distinguish semantic categories.

5.2 Why Reconstruction Fails on High-Dimensional Inputs

On MNIST ($d_{\text{patch}} = 16$ for patch size 4, 1 channel), per-patch reconstruction is a low-dimensional regression problem: the error predictor maps 128-dimensional token embeddings to 16-dimensional patch vectors, a compression ratio of $8:1$. On CIFAR-10 ($d_{\text{patch}} = 48$ for patch size 4, 3 channels), the compression ratio drops to $2.67:1$, and the predictor must simultaneously model color, texture,

and spatial structure. The reconstruction error becomes dominated by irreducible aleatoric uncertainty (inherent patch variance) rather than epistemic uncertainty (model unfamiliarity), diluting the OOD discriminative signal.

This dimensionality argument predicts that reconstruction-based OOD signals should scale *inversely* with input channel count and patch size—a hypothesis that can be tested in future work with intermediate-dimensional datasets (e.g., grayscale CIFAR, or larger patch sizes on MNIST).

5.3 Practical Guidance

Based on our empirical analysis, we offer the following decision rule for practitioners considering the Sentinel reconstruction error as an OOD signal:

Condition	Recommendation	Rationale
Grayscale, low-res input ($c = 1$, $h, w \leq 32$)	Use reconstruction + Energy	Strong far-OOD signal
RGB, natural images ($c = 3$)	Prefer Mahalanobis + Energy	Reconstruction signal is weak
Same-modality OOD, different labels	Avoid reconstruction signal	Near-OOD reversal risk
Structurally distinct OOD	Reconstruction is strongest	Exploits surface-level shift

For deployment scenarios where zero-cost OOD signals are preferred (edge devices, real-time systems), the Sentinel reconstruction error provides a useful *first-stage filter*: flag inputs with high reconstruction error for further inspection by a more expensive OOD detector, while passing inputs with low reconstruction error directly to the classifier. Our results suggest this two-stage approach would catch far-OOD samples with high recall (FPR95 0.016 on FashionMNIST) while minimizing unnecessary post-processing of in-distribution samples.

5.4 Deployment Value of Zero-Cost Signal

A distinguishing property of the Sentinel reconstruction error as an OOD signal is its **zero marginal inference cost**. Unlike post-hoc methods that require executing the full classification pipeline on every candidate input before computing an OOD score, the Sentinel reconstruction error is produced as a necessary intermediate of the attention routing mechanism. This means the OOD signal is available *before* the classification decision is made, not *after* it.

Concretely, for every input that passes through a Sentinel layer, the per-token reconstruction errors ε_i are computed by the linear patch decoder as part of the routing gate. Aggregating these into a scalar OOD score requires only a weighted sum (mean across Sentinel layers, $O(N)$), adding $O(N)$ floating-point operations—negligible relative to the $O(N^2 D)$ self-attention. By contrast, MSP requires computing softmax over the full vocabulary of logits, Energy requires logsumexp over logits, and Mahalanobis requires fitting per-class Gaussians on intermediate features—all of which occur at the end of or after the forward pass.

This early-availability property enables a **speculative execution** pattern for deployment: if the aggregated reconstruction error falls below a calibrated threshold on the first few Sentinel layers, the input can be classified immediately without executing the remaining layers. We emphasize, however, that the reference implementation does not reduce wall-clock cost (Table 6): its overhead of +36–51% stems from evaluating both candidate updates, and speculative early-exit would require an implementation that actually skips downstream computation. For edge deployments or latency-critical systems, such early-exit using the reconstruction error as a gating signal is a plausible route to reducing average-case inference cost, and we leave its evaluation to future work.

Relevance to modern practice. The transformer ecosystem increasingly adopts auxiliary training losses (auxiliary language modeling heads, contrastive losses, distillation losses) that are discarded at inference time. Sentinel Attention inverts this pattern: its auxiliary training objective (the MSE reconstruction loss on the error predictor) produces an artifact that gains *additional* utility at inference time, beyond its training purpose. This inversion—training-cost artifacts becoming inference-time assets—is a design pattern worth attention in the broader context of efficient deep learning system design.

5.5 The Information Independence Principle

A finding that generalizes beyond Sentinel is the **information independence principle**: combining OOD scores from independent information sources (e.g., feature-space distances + logit energies) yields consistent improvement over the better individual score, while combining scores from the same source (e.g., two logit-derived scores) yields no gain or even degradation. This principle is independent of the Sentinel mechanism and applies to any model for which multiple complementary scoring functions can be computed. The z-score normalization provides a simple, calibration-free method to operationalize this principle—no OOD data, no hyperparameter tuning, just in-distribution statistics from the training set.

The CIFAR-10 results refine this principle in an important way. Independence is *necessary but not sufficient*. If one signal is independent but badly misaligned with semantic OOD (as Sentinel reconstruction is on RGB images), the combination can become worse than either constituent. A more precise statement is therefore: *score fusion is beneficial when the added score contributes non-redundant information with non-trivial standalone quality.* This refined statement is more useful to practitioners because it turns the principle into a diagnostic workflow: first verify that a candidate signal has at least modest AUROC on its own, then test whether it contributes information orthogonal to the best existing score. If both conditions hold, simple normalization-and-sum is surprisingly competitive; if either condition fails, fusion is unlikely to help.

6 Conclusion

We introduced Sentinel Attention, an error-gated sparse attention mechanism whose per-token reconstruction errors serve as a zero-cost, training-endogenous OOD signal. Through controlled experiments on MNIST and CIFAR-10, we established the precise boundary of this signal's utility:

- **Strength:** Far-OOD inputs on low-dimensional data (FashionMNIST from MNIST, AUROC 0.9964 combined with Energy, FPR95 0.016).
- **Failure mode 1:** Near-OOD reversal (EMNIST from MNIST, AUROC 0.4827, below chance)—reconstruction error detects surface-level shifts, not semantic novelty.
- **Failure mode 2:** High-dimensional inputs—per-patch reconstruction error degrades with input channels and resolution.

Our cross-source vs. same-source combination experiments revealed a broader principle: OOD detection improvement from score combination comes from information source independence, not from any single signal. Mahalanobis+Energy (+0.0141 over Energy alone) generalizes beyond Sentinel, while Sentinel+Energy (−0.0027) does not improve the mean. We documented these negative results explicitly, believing that honest characterization of failure modes is as scientifically valuable as reporting successes.

Methodological note. The trajectory from PWPF to Sentinel embodies a design philosophy of *progressive removal*: start with a feature-rich architecture informed by multiple plausible hypotheses, then systematically remove every component that fails a controlled ablation. Of the nine original PWPF components, only three survived (Error-Q, reconstruction error, top-K sparsity); five were removed for being either actively harmful (attractor, spectral norm) or non-informative (precision field, memory bank, confidence network); one was simplified (Mixed-Q to Pure Error-Q). The result is a minimalist architecture whose only auxiliary artifact—the reconstruction error—gains deployment utility beyond its training purpose. We recommend this approach to the community: a clean, well-characterized failure boundary is more scientifically valuable than an over-claimed universal signal.

Future work. Several directions are open: (1) extending the information independence analysis to larger-scale vision models (e.g., ViT on ImageNet); (2) designing reconstruction predictors that operate in feature space rather than pixel space, potentially resolving the near-OOD reversal; (3) exploring alternative error signals (e.g., perceptual loss, contrastive reconstruction) that may bridge the gap between surface-level and semantic OOD detection.

Acknowledgments

We thank the open-source community for maintaining PyTorch, torchvision, and scikit-learn. All experiments were conducted on consumer CPU hardware (Intel Core i7-12700, 16 GB RAM).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[2] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

[3] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[4] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020.

[5] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big Bird: Transformers for longer sequences. In *NeurIPS*, 2020.

[6] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[7] A. Roy, M. Saffar, A. Vaswani, and D. Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the ACL*, 2021.

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

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

[10] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[11] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[12] R. Huang, A. Geng, and Y. Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021.

[13] Y. Sun, C. Guo, and Y. Li. ReAct: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

[14] D. Yoo and I. S. Kweon. Learning loss for active learning. In *CVPR*, 2019.

[15] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. A. Kohl, T. Cemgil, S. M. A. Eslami, and O. Ronneberger. Contrastive training for improved out-of-distribution detection. *arXiv:2007.05566*, 2020.

[16] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[17] K. Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456):815–836, 2005.

[18] D. Mumford. On the computational architecture of the neocortex. II: The role of cortico-cortical loops. *Biological Cybernetics*, 66:241–251, 1992.

[19] J. C. R. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Computation*, 29(5):1229–1262, 2017.

[20] B. Millidge, T. Salvatori, Y. Song, R. Bogacz, and T. Lukasiewicz. Predictive coding: Towards a future of deep learning beyond backpropagation? In *IJCAI Survey Track*, 2022.

[21] D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). *arXiv:1606.08415*, 2016.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[23] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. EMNIST: an extension of MNIST to handwritten letters. In *IJCNN*, 2017.

[24] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

[25] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[26] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[28] A. Graves. Adaptive computation time for recurrent neural networks. *arXiv:1603.08983*, 2016.

[29] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser. Universal transformers. In *ICLR*, 2019.

[30] N. Shazeer, G. Hinton, J. Dean, M. Norouzi, and W. Fedus. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

[31] D. Raposo, S. Ritter, B. Richards, T. Lillicrap, P. C. R. Humphreys, and A. Santoro. A mixture of depths transformer: Efficiently allocating compute in transformers. *arXiv:2404.02258*, 2024.

[32] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

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

[34] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[35] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

[36] L. Ruff, R. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, M. Kloft, T. Müller, and K.-R. Müller. Deep one-class classification. In *ICML*, 2018.

[37] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2018.

A Reproducibility

All experiments use the self-contained Sentinel Attention codebase, available at `https://github.com/Fengrru/pwpf`. The following commands reproduce all results in this paper:

```
# MNIST training (5 epochs, ~30 minutes)
python -m pwpf.sentinel_train --dataset mnist --model sentinel \
    --epochs 5 --batch-size 128 --threads 6 \
    --save sentinel_ckpt_v2.pt --label mnist-v2

# Checkpoint load-loop integrity check
python verify_ckpt.py --checkpoint sentinel_ckpt_v2.pt

# MNIST OOD evaluation (all baselines + z-score combos)
python -m pwpf.sentinel_eval_ood \
    --checkpoint sentinel_ckpt_v2.pt --max-samples 4000

# CIFAR-10 training (20 epochs, ~2.3 hours)
python -m pwpf.sentinel_train --dataset cifar10 --model sentinel \
```

```
    --epochs 20 --batch-size 128 --threads 6 \
    --save sentinel_ckpt_cifar_v2.pt --label cifar-v2

# CIFAR-10 OOD evaluation
python -m pwpf.sentinel_eval_ood \
    --checkpoint sentinel_ckpt_cifar_v2.pt --max-samples 4000
```