TIDE: Asymmetric Neural Circuits for Stabilized Temporal Inhibitory-Excitatory Dynamics

Alexander Kyuroson
Luleå University of Technology,
Luleå, Sweden
akyuroson@gmail.com

Denis Kleyko
Örebro University &
RISE Research
Institutes of Sweden
denis.kleyko@oru.se

Marcus Liwicki
Luleå University of Technology,
Luleå, Sweden
marcus.liwicki@ltu.se

Abstract

Recent Continuous Thought Machine (CTM) architecture decouples internal computation from external inputs via neural dynamics, but relies on unconstrained multi-layer perceptrons for its recurrent computations. We propose to model neural dynamics using asymmetric Excitatory-Inhibitory (E-I) networks, which can be stabilized via principles from network theory, while obeying Dale's principle and can be expressed as energy-based systems optimized through a game-theoretic loss. Building on this perspective, we introduce Temporal Inhibitory-Excitatory Dynamic Engine (TIDE), a neuro-inspired architecture that computes internal representations through neural dynamics stabilized by incorporating the Wilson-Cowan dynamics and lateral inhibition. TIDE balances biological realism by, for instance, using Hierarchical Receptive Fields and enforcing Dale's principle to ensure a realistic $80:20$ E-I balance ratio with an end-to-end trainable architecture. The aim of this paper is to introduce a new architecture that brings neuro-inspired learning and architectural principles to the forefront. We present proofs of convergence, stability, and complexity bounds, along with empirical ablation studies. Overall, TIDE achieves higher `top-1` accuracy than CTM on the majority of the investigated datasets while requiring a smaller training budget. Furthermore, across varying training budgets, TIDE improves `top-1` accuracy on ImageNet-C by an average of $+1.65\%$ to $+6.42\%$ under different perturbation severity levels.

1 Introduction

Recent advances in machine learning driven by Convolutional Neural Networks (CNNs) [1–3], Transformers [4], and Vision Transformers (ViTs) [5] have significantly improved representation learning, scalability, and cross-domain generalization. At the same time, these architectures rely on computational principles that differ widely from those of biological neural systems, which are commonly characterized by sparsity, event-driven activation, population-level dynamics, and asymmetric Excitatory-Inhibitory (E-I) networks [6]. The absence of such principles could also be associated with shortcomings of modern deep learning models, like their brittleness under distribution shift, limited cross-modal generalization, and sample inefficiency relative to their biological counterparts [7–9].

It has been suggested that structure and asymmetrical connections can facilitate more efficient learning in neural systems [10, 11]. Recent work [12, 13] has demonstrated that Daleian-based networks are more robust against perturbations and are capable of efficient learning via single-neuron feature optimization. This indicates that biological mechanisms may play a significant role in designing novel architectures. Furthermore, it has been shown that the $80:20$ ratio for E-I population is vital not only for energy efficiency but also for robustness in continual learning and dynamic stability [14].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

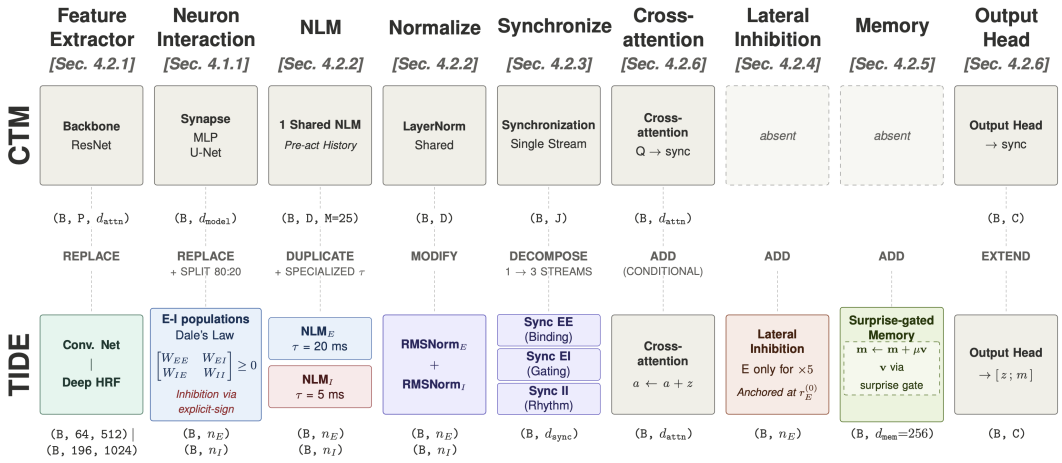

Figure 1: Schematic architectural comparison between CTM and TIDE. TIDE's architectural components are further described in Section 4.2; Appendix E provides additional technical details.

By introducing neural dynamics as a main primitive, Continuous Thought Machine (CTM) architecture (Figure 1) attempts to decouple an internal reasoning process from the external input data, thus introducing recurrence via internal computation steps that replace static activations with per-neuron Neuron-level Models (NLMs), which elevates synchronization from an emergent statistic to an explicit latent representation for continuous reasoning [15]. Although CTM uses recurrence for internal computation, it does not leverage E-I dynamics and only relies on standard Multilayer Perceptron (MLP) to represent its recurrent connections, which typically do not obey Dale's principle [16, 17].

We introduce Temporal Inhibitory-Excitatory Dynamic Engine (TIDE) architecture that evolves the ideas of CTM by using the Wilson-Cowan model for neural dynamics [18], leading to a recurrent model in which every connection is Dale-constrained, and neuron populations interact based on an asynchronous asymmetric model with distinct time constants for different types of neurons. Our main contributions within TIDE can be summarized as follows:

- A novel neuro-inspired recurrent architecture with non-negative weights via the enforcement of the Wilson-Cowan dynamics.
- Hierarchical Receptive Field (HRF) implementation and its adoption as a feature extractor for end-to-end training and inference based on a neuro-inspired model.
- Integration of a differentiable spectral stabilizer based on the Perron-eigenvalue sum-ratio to enable online Lyapunov Diagonal Stability (LDS) verification during training.
- Inclusion of population-specific NLMs with distinct exponential temporal kernels for excitatory and inhibitory neurons given by τ_E=20 ms, and τ_I=5 ms, respectively.
- Comprehensive analysis and ablation studies to investigate TIDE's performance against CTM in the presence of corrupted or Out-of-Distribution (OOD) data. Experimental results demonstrate that TIDE consistently outperforms CTM in robustness.

The remainder of this paper is structured as follows. Section 2 presents related work. Thereafter, Section 3 introduces the theoretical background and notation adopted throughout the paper, covering the CTM formulation and the Dale-constrained E-I dynamics. Section 4 presents the proposed architecture, its components, and their overall theoretical properties. Section 5 introduces the evaluation setup, presents the results, and provides both qualitative and quantitative assessments of the proposed architecture. Section 6 concludes by discussing results and future work.

2 Related Work

Continuous reasoning with recurrency: In contrast to existing sequence- and iteration-based architectures such as recurrent-depth networks [19], deep equilibrium models [20], neural ordinary differential equations [21], and adaptive compute [22, 23], CTM [15] adopts recurrence for internal computation steps and leverages it for reasoning via certainty-based halting. The proposed TIDE

retains the internal computation steps and synchronization read-out blocks proposed in CTM but replaces the unconstrained recurrent weights with the Wilson-Cowan dynamics under Dale constraints.

Wilson-Cowan dynamics & E-I balance: TIDE incorporates a neuro-inspired circuit with separate excitatory and inhibitory neuron populations, each with its own time constants and connectivity, following the Wilson-Cowan model [24, 25]. This allows operating in the asynchronous-irregular regime, where excitatory and inhibitory activations can cancel while the E-I network exhibits irregular yet dynamically stable patterns [26, 27]. Stability is further promoted through homeostasis via inhibitory plasticity [28, 29] and balanced-network mechanisms [30], acting as direct regularizers on network activity.

Dale's principle: Dale's principle [16, 17] states that each neuron releases a single neurotransmitter at every given outgoing connection, implying that the recurrent connectivity matrix is sign-indefinite and weights for E-I network must satisfy $W_{EI} \neq W_{IE}^{\top}$. In contrast to [13], the existing models ignore this principle. By parameterizing recurrent weights as $W = \Sigma \odot |M|$, where $\Sigma \in \{+1, -1\}$ is the fixed per-neuron sign vector assigned via Dale's principle, TIDE allows direct usage of the existing optimization algorithms for a neuro-inspired architecture. At each optimization step, M is projected onto the non-negative orthant via the clipping $M \leftarrow \max(M, 0)$ to preserve the assigned sign.

Energy-based & game-theoretic loss: The use of an energy-based optimization method that leverages Nash equilibria to replace the single scalar loss in homogeneous networks, with generalization to the reinterpretation of dynamics in heterogeneous networks based on E-I networks, is proposed in [18]. Therefore, by using the per-neuron gradient and its optimization via the energy function, it is feasible to address the main shortcoming of Hopfield networks [31], thereby providing a principled objective that is consistent, while ensuring the convergence of internal dynamics. Furthermore, this reformulation provides stability under strong convex-concavity of the local energies in the consensus regime during the recurrent updates, replacing the implicit symmetry-based stability argument in [31].

Cortex-aligned backbone & memory model: It has been demonstrated that an untrained CNNs with multi-scale center-surround filters and subsequent spatial compression resembles the primate ventral-stream responses [32]. Based on hierarchical models [33] and data-driven cortex-alignment studies [34, 35], TIDE uses HRF as its backbone for feature extraction. Moreover, given our adoption of a surprise-gate (Section 4.2.5), which differs in its memory formulation from Titans [36] and MIRAS [37], and is represented as a state buffer updated iteratively during each internal computation step, without reliance on gradient updates, the timing of memory updates can be coupled to the homeostatic states of the network. TIDE's inductive bias, achieved via the E-I network and its temporal mixing operation, is later conditioned by constraints from Dale's principle, thereby removing the need for internal state-space memory and hidden states as used in, e.g., Mamba [38] and xLSTM [39].

Normalization & optimization: Since the mean-subtraction of LayerNorm (LN) would inhibit the absolute E-I balance signaling, which must be preserved, `RMSNorm` [40] per-population is used. Moreover, by using AdamW [41] with a linear warm-up followed by cosine decay [42], we stabilize any further issues raised by the mixed-precision training [43] to improve the computational efficiency of TIDE.

3 Preliminaries

This section presents the notation and background concepts used in the paper. We use superscript $^{(t)}$ to represent internal computation steps. Populations in E-I networks are denoted by E and I, respectively. The activations are $r_E \in \mathbb{R}^{n_E}$, $r_I \in \mathbb{R}^{n_I}$, with $n_E = \lfloor 0.8 \cdot d_{\text{model}} \rfloor$ and $n_I = d_{\text{model}} - n_E$ while population time constants and their corresponding Euler coefficients are τ_E, τ_I and α_E, α_I, respectively. $W_{EE}, W_{EI}, W_{IE}, W_{II}$ are recurrent connectivity matrices whose components are non-negative, in accordance with Dale's principle. P denotes the number of spatial backbone tokens, d_{model} is model dimension while $d_{\text{attn}} = n_{\text{heads}}\, d_{\text{head}}$ and $d_{\text{sync}} = d_{EE} + d_{EI} + d_{II}$ are attention and synchronization dimensions, respectively. Remaining notations are introduced throughout the text.

Given a fixed input, CTM processes it over T internal computation steps, representing thinking by a recurrence whose state consists of a collection of per-neuron activation histories [15]. Two of the main pillars of CTM, namely NLM and latent representation via synchronization, are adopted in TIDE. Each neuron i has a small MLP $g_{\theta_i} : \mathbb{R}^M \to \mathbb{R}$ that maps the last M pre-activations of that neuron as a temporal buffer with a window size M, to its next post-activation $r_i^{\tau} = g_{\theta_i}(a_i^{(\tau - M + 1):(\tau)})$.

The synchronization, where the vector of post-activations is fed to the readout, is the second-order statistic of the post-activation history: $S_{ij}^{(t)} = \frac{\sum_{\tau \le t} \lambda^{t-\tau}\, r_i^\tau\, r_j^\tau}{\sqrt{\sum_{\tau \le t} \lambda^{t-\tau}\,(r_i^\tau)^2 \;\sum_{\tau \le t} \lambda^{t-\tau}\,(r_j^\tau)^2}}$, where $\lambda \in (0, 1]$ denotes the decay factor that down-weights distant history [15]. Note that CTM does not follow Dale's principle; therefore, its formulation and weights are fully unconstrained.

Wilson-Cowan dynamics: The two-population continuous-time rate E-I networks are described by

$$\tau_E\, \dot{r}_E = -r_E + \varphi\big(W_{EE}\, r_E - W_{EI}\, r_I + u_E\big); \quad \tau_I\, \dot{r}_I = -r_I + \varphi\big(W_{IE}\, r_E - W_{II}\, r_I + u_I\big), \quad (1)$$

where φ, u_E, and u_I denote the activation function, and external excitatory and inhibitory inputs, respectively [24]. Excitation decays by $\tau_E \approx 20$ ms while inhibition relaxes with $\tau_I \approx 5$ ms, thus the receptor kinetics is imposed such that $\tau_I < \tau_E$. A first-order forward-Euler discretization with time step $\Delta t = 1$ ms yields the recurrence:

$$r_E^{(t)} = (1 - \alpha_E)\, r_E^{(t-1)} + \alpha_E\, \varphi(h_E^{(t)}); \quad r_I^{(t)} = (1 - \alpha_I)\, r_I^{(t-1)} + \alpha_I\, \varphi(h_I^{(t)}), \quad (2)$$

with $\alpha_\bullet = \Delta t / \tau_\bullet$ and $h_\bullet^{(t)}$ representing the pre-activations that are computed as

$$h_E^{(t)} = W_{EE}\, r_E^{(t-1)} - W_{EI}\, r_I^{(t-1)} + W_E^{\text{in}}\, a^{(t)}; \quad h_I^{(t)} = W_{IE}\, r_E^{(t-1)} - W_{II}\, r_I^{(t-1)} + W_I^{\text{in}}\, a^{(t)}, \quad (3)$$

where $a^{(t)} \in \mathbb{R}^{d_{\text{sync}}}$ denotes the cross-attention among latent vector $z^{(t)}$, see Section 4.2.3.

Dale's principle: Under Dale's principle [16, 17], a neuron's effect on its post-synaptic neighbors has a fixed sign determined by its neurotransmitter. In an artificial network, this is formulated as recurrent connectivity matrices with non-negative components $W_{EE}, W_{EI}, W_{IE}, W_{II}$. The inhibitory population's contribution to downstream activity is given an explicit minus sign as in (3).

Game-theoretic formulation: Since an E-I network violates the assumption of Hopfield networks, $W = W^\top$, per-neuron energies $\{E_i\}_{i=1}^{d_{model}}$ are defined such that a single scalar Lyapunov function can be adopted for asymmetric populations. Following [18], we assign a per-neuron local energy, E_i, with neurons partitioned into excitatory subset, n_E, and inhibitory subset, n_I. Thus, the local energy is defined as $E_i(x, u_i)$, where the joint network state, $x \in \mathbb{R}^{d_{model}}$, is the stacked vector of post-Euler firing rates, $x^{(t)} = [\, r_E^{(t)};\ r_I^{(t)}\,]$. The firing rate of neuron i for $r_E^{(t)}$ and $r_I^{(t)}$ at computation step, t, is denoted by $x_i^{(t)} \in \mathbb{R}$, while $u_i \in \mathbb{R}$ denotes external input driving neuron i, for a given state of neighboring neurons, and is given by $W_\bullet^{\text{in}} a^{(t)}$; x^* is the joint state that denotes the Nash equilibrium in a Zero-sum game where the stationary conditions are satisfied iff: $\nabla_{x_i} E_i(x^*, u_i) = 0 \quad \forall i$.

4 TIDE Architecture & Its Foundations

4.1 Theoretical Foundations

Here we provide theoretical foundations of TIDE; complete proofs are presented in Appendices A–C.

4.1.1 Dale-constrained Gradient Optimization

A projection $W \leftarrow \Pi_{\text{Dale}}(W) = \max(W, 0)$ is used to enforce Dale's principle during optimization updates. It is component-wise, idempotent, and 1-Lipschitz in the Frobenius norm (Proposition A.2), and under Distributed Data Parallel (DDP) produces identical weights for gradients across the model.

Theorem 4.1 (Dale-compatible gradient, informal). Let $\mathcal{L}$ be L-smooth and μ-strongly convex over the non-negative orthant $\mathcal{W}_{\text{Dale}} = \{W \ge 0\}$, while considering the iteration $W^{(k+1)} = \Pi_{\text{Dale}}(W^{(k)} - \eta\, \nabla \mathcal{L}(W^{(k)}))$ with $\eta = 1/L$. Then $W^{(k)} \in \mathcal{W}_{\text{Dale}}$ for every $k \ge 0$, and $\|W^{(k)} - W^*\|_F^2 \le (1 - \mu/L)^k \|W^{(0)} - W^*\|_F^2$ for the every unique is projected minimizer W^* over $\mathcal{W}_{\text{Dale}}$.

The training objective defined in Section 4.2.7 is non-convex. Thus, the theorem applies only at a strict local minimum at which the standard condition, i.e., Lipschitz-continuous gradient, is met. Moreover, Backpropagation Through Time (BPTT) through the constraint is addressed in Remark A.4.

4.1.2 Game-theoretic Formulation & Lyapunov Diagonal Stability

Let $x = [\, r_E;\ r_I\,] \in \mathbb{R}^{d_{model}}$, and $W_{\text{eff}} \in \mathbb{R}^{d_{model} \times d_{model}}$ be the signed effective recurrent matrix obtained by composing the four Dale-constrained matrices as defined in (A.42) while defining the

per-neuron energies following [18] as

$$E_i(x, u_i) = -x_i \sum_{j=1}^{d_{model}} \left(1 - \tfrac{1}{2}\delta_{ij}\right) W_{\text{eff},ij}\, x_j \;-\; x_i u_i \;+\; \int_0^{x_i} \varphi^{-1}(s)\, ds, \tag{4}$$

where δ_{ij}, and $\varphi^{-1} : \text{range}(\varphi) \to \mathbb{R}$ denote the Kronecker delta and the inverse of the neural activation function φ introduced in (1), respectively. Moreover, u_i is the per-neuron component of the cross-attention $a^{(t)}$, and the factor $(1 - \frac{1}{2}\delta_{ij})$ halves the diagonal contribution $W_{\text{eff},ii}\, x_i^2$ so that ∂_{x_i} recovers the full self-coupling term. Differentiating (4) with respect to x_i and using $(d/dx_i)\int_0^{x_i} \varphi^{-1}(s)\, ds = \varphi^{-1}(x_i)$ yields $\varphi^{-1}(x_i) - \partial_{x_i} E_i = (W_{\text{eff}}\, x)_i + u_i$, which is the right-hand side of the Wilson-Cowan dynamics in (1). This provides the following benefits: *i*) A principled soft penalty on the gradient as residual $\sum_i \|\partial_{x_i} E_i\|^2$, which is instantiated as the game loss in Section 4.2.7; and *ii*) A saturation-based path to guaranteed existence of a unique equilibrium under mild contractivity conditions on W_{eff} and φ, formalized in Theorem A.5.

Theorem 4.2 (LDS implies global convergence). If $W_{\text{eff}} - I$ is LDS with $D \succ 0$ such that $D(W_{\text{eff}} - I) + (W_{\text{eff}} - I)^\top D \prec 0$, the linearized dynamics $\dot{x} = W_{\text{eff}} x - x + b$ admits a unique Nash equilibrium $x^* = (I - W_{\text{eff}})^{-1} b$, to which every trajectory converges [18].

Therefore, LDS is monitored at runtime, but does not backpropagate through $\lambda_{\max}(W_{\text{eff}} + W_{\text{eff}}^\top)/2$ given the eigen-decomposition is non-differentiable at crossings. Instead, gradient-based stability is achieved using the differentiable Perron surrogate provided below.

4.1.3 Spectral Stability of Wilson-Cowan Dynamics

The Perron-based spectral regularizer is defined below, while its functionality and limitations are discussed in detail. The analysis is carried out under the assumption of fully isolated E and I populations. Furthermore, the coupled E-I stability is provided under the LDS Theorem 4.2.

Isolated-E bound: Linearizing (2) around $r_E^* = 0$ without inhibitory signals yields $r_E^{(t+1)} = M_E\, r_E^{(t)}$ with $M_E = (1 - \alpha_E) I + \alpha_E W_{EE}$. Given that the eigenvalues of M_E are $\mu_i = (1 - \alpha_E) + \alpha_E \lambda_i(W_{EE})$, for the real non-negative Perron eigenvalue $\lambda_{\text{P}}(W_{EE}) \geq 0$, Schur stability $|\mu_i| < 1$ for every i is reduced to: $\lambda_{\text{P}}(W_{EE}) < 1$ (isolated-E Schur bound). That is, the *E-only* sub-dynamics is contracting iff W_{EE}'s Perron eigenvalue is strictly below 1, for every Euler step $\alpha_E \in (0, 1]$.

Isolated-I bound: With the explicit minus sign, the linearization on the inhibitory population results in $\mu_j = (1 - \alpha_I) - \alpha_I \lambda_j(W_{II})$. Thus, Schur stability $|\mu_j| < 1$ on the Perron eigenvalue reduces to $\lambda_{\text{P}}(W_{II}) < 2/\alpha_I - 1 = 9$ at $\alpha_I = 0.20$. The upper threshold is binding in the I population, unlike the E population, where $(1 - \alpha_E)$ leakage guarantees the lower $|\mu| > -1$ bound trivially as the minus sign flips μ_j past -1 for large λ_j.

The coupled system & limiters: In practice, with an activated inhibitory pathway, TIDE operates above the isolated-E bound, e.g., on ImageNet-1K [44] $\widehat{\lambda}_{\text{P}}(W_{EE}) \approx 14.7$ at convergence. The E-I network is stable as the effective-matrix linearization, $M_{\text{eff}} = I - \text{diag}(\alpha) + \text{diag}(\alpha) W_{\text{eff}}$ with sign-indefinite W_{eff} ((A.8) in the Appendix), has inhibitory signals that suppress the unstable excitatory modes. The Perron values of the isolated populations W_{EE}, W_{II} are, therefore, *not* Schur conditions for the coupled E-I network; they act as a limiting factor for each non-negative population. Empirically, we find that $\tau_{EE} = 15$, and $\tau_{II} = 7$ keep gradient norms bounded and prevent instability during training on the ImageNet-1K dataset.

Differentiable Perron surrogate: Because $W_{..} \geq 0$, the Perron-Frobenius theorem guarantees a real non-negative dominant eigenvalue. Therefore, it can be estimated by the sum-ratio power iteration

$$v_0 = \mathbb{1}/n, \qquad v_{k+1} = W_{..}\, v_k / \|W_{..}\, v_k\|_2, \qquad \widehat{\lambda}_{\text{P}}(W_{..}) = \mathbb{1}^\top W_{..}\, v_K / \mathbb{1}^\top v_K, \quad K = 10, \tag{5}$$

which is differentiable in $W_{..}$ (Proposition C.7). On trained W_{EE}, we observe $\|W_{EE}\|_F / \widehat{\lambda}_{\text{P}}(W_{EE}) \approx 2.5$; thus the sum-ratio penalty is tighter than an operator-norm regularizer, which would collapse W_{EE} to near-diagonal matrices, removing recurrent interactions. For the off-diagonal blocks W_{EI} and W_{IE}, we monitor the maximum singular value as a detached diagnostic, without backpropagation through it, since sustained growth above its rolling median indicates impending population collapse.

4.2 TIDE Architecture

In this section, we present TIDE architecture's components, including its feature extractor, E-I population-based NLM, population-based synchronization mechanism, lateral inhibition, cross-attention, and surprise-gated associative memory, see Figure 1. Each of these components is described in detail below, along with its inputs, outputs, and guiding equations. Figure 1 not only illustrates each of these components and their role but also shows the divergence from the CTM framework. Moreover, the complete training process of TIDE is summarized in Algorithm A.1 in Appendix B.4.

4.2.1 Hierarchical Receptive Field

Given the varying complexity of datasets used in this paper, where different backbones are required to extract and encode features from the input data, we designed two backbone variants for TIDE. A simple shallow HRF backbone that maps 32×32 or 28×28 input images into an 8×8 grid of d_{attn}-token positions, $P=64$, using multi-stage filters, while the second neuro-inspired deep HRF uses a ResNet-style stem, followed by four hierarchical residual stages for extracting features from ImageNet-1K [44]. The initial stage zero is a bank of learnable center-surround filters defined as:

$$\phi^{(s)}(x) \;=\; \mathrm{ReLU}\Big(\mathrm{BN}\big(w_c^{(s)}\,C^{(s)}(x) - w_s^{(s)}\,S^{(s)}(x)\big)\Big), \quad s\in\{1,2,4,8\},$$

with independent learnable center- and surround-convolutions $C^{(s)}$, $S^{(s)}$, and scalar gains $(w_c^{(s)}, w_s^{(s)})$ initialized at $(1,\,0.5)$, respectively, and BN denotes BatchNorm. The above expression subsumes the fixed-form Difference-of-Gaussian (DoG) operator of [45] as the special case where $C^{(s)}=G_s,\ S^{(s)}=G_{\kappa s}$ are frozen isotropic Gaussian distributions; see Remark D.1 in Appendix D. Furthermore, stages one to four apply standard residual blocks with $2\times$ spatial down-sampling and channel expansion, while the final stage produces a 14×14 grid of 1024-dimensional tokens, $P=196$.

Relation to CTM's backbone: In contrast to CTM, which uses the pre-trained ResNet-152 as its backbone, HRF variants use the identity map in their stages one to four as a parameter-space point (cf. Appendix E). This results in differences in stage widths, stem, and block shape. Therefore, we treat them as distinct architectures rather than members of a single parameterized family, and we mark the head-to-head comparisons in Section 5 accordingly.

4.2.2 Population-specific NLM

Each neuron has a two-layer gated MLP with hidden dimension $H=4$ for small images or $H=32$ for ImageNet-1K, implemented as a batched `SuperLinear` operator that maintains an independent $M\to2H$ linear map per neuron. The temporal weighting in the first layer uses an exponential kernel, $w_m=\mathrm{softmax}_m\left(-(M-m)/\tau\right)$ with $\tau=\tau_E=20$ ms and $\tau=\tau_I=5$ ms for excitatory and inhibitory NLM, matching AMPA and GABA_A kinetics, respectively. Since using pre-activations leads to a positive feedback loop that destabilizes the dynamics, post-activations are stored in the FIFO buffer.

4.2.3 Synchronization as Latent Representation

For each pair type $XY\in\{EE, EI, II\}$, a deterministic number of neuron pairs denoted by p_{XY}, is sampled from index tensors $\mathbf{ij}$, to maintain recurrent sums of pair-wise product $\pi_k^{(t)}=r_{X,i_k}^{(t)}\,r_{Y,j_k}^{(t)}$:

$$\nu_{XY}^{(t)} = e^{-\delta_{XY}}\,\nu_{XY}^{(t-1)} + \pi^{(t)}, \qquad \xi_{XY}^{(t)} = e^{-\delta_{XY}}\,\xi_{XY}^{(t-1)} + 1,$$

with learnable decays $\delta_{XY}\in\mathbb{R}^{p_{XY}}$ clipped to $[0,15]$. The step-t synchronization vector is expressed as $z_{XY}^{(t)}=\texttt{Proj}_{XY}\big(\nu_{XY}^{(t)}/\sqrt{\xi_{XY}^{(t)}+\varepsilon}\big)$, where $\texttt{Proj}_{XY}$ is a linear projection from $\mathbb{R}^{p_{XY}}$ to $\mathbb{R}^{d_{XY}}$, which represents per-stream latent space. The full latent space is defined as their concatenation, $z^{(t)}=\texttt{LN}\big[z_{EE}^{(t)};\,z_{EI}^{(t)};\,z_{II}^{(t)}\big]\in\mathbb{R}^{d_{\text{sync}}}$. The individual streams are interpreted as *binding* for E-E phase-lock, *gating* for E-I coherence, and *rhythm* for I-I coherence.

4.2.4 Lateral Inhibition

Lateral inhibition is based on Winner-Take-All (WTA) and leverages a nested E-I network within the excitatory population prior to the sync readout (Algorithm A.1). A pool of auxiliary inhibitory neurons reads the current excitatory activity and sends back an inhibitory signal that is subtracted from the excitatory neurons as follows:

$$r_I^{(k)} = \mathrm{ReLU}(W_{EI}^{\text{lat}}\,r_E^{(k-1)}), \quad r_E^{(k)} = \mathrm{ReLU}(r_E^{(0)} - \gamma\,W_{IE}^{\text{lat}}\,r_I^{(k)}), \quad k=1,\dots,K_{\text{WTA}},$$

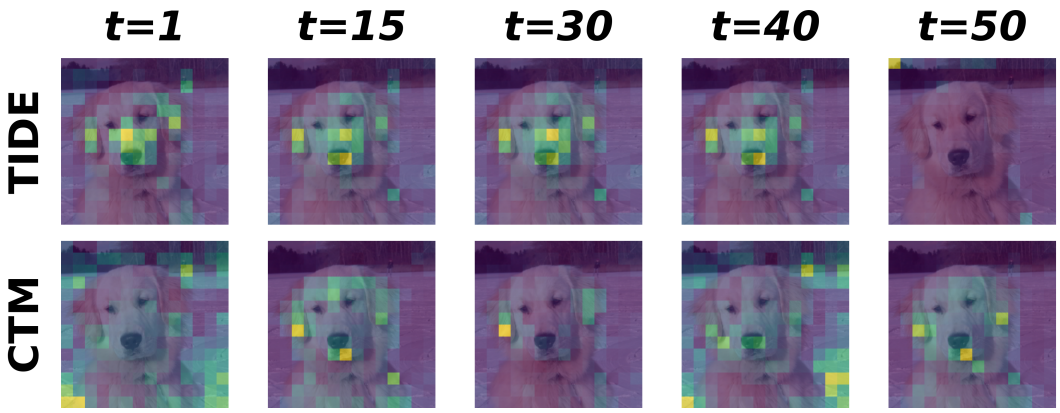

Figure 2: Temporal evolution of mean attention as saliency per computation step for TIDE and CTM.

where $^{(k)}$ are WTA iterations, $\gamma{>}0$ is a learnable gain, $W^{\text{lat}}_{\cdot\cdot}{\geq}0$ are lateral weights, and $K_{\text{WTA}}{=}5$. Following [18], the limit of the expression is given by a Nash equilibrium of the Zero-sum game in which each excitatory neuron maximizes its own activity subject to a shared inhibitory budget. This results in a sparse representation in which only the most strongly driven neurons remain active.

4.2.5 Surprise-gated Associative Memory with an E-I Retention Gate

Inspired by the Titans [36] and MIRAS [37] architectures, TIDE uses an internal-state module as its memory. Unlike CTM, which has no memory module (Figure 1), a combination of a persistent buffer $m{\in}\mathbb{R}^{d_{\text{memory}}}$, a reconstruction head f_{rec}, and a projection head f_{proj}, enables the formation of surprise signal $s^{(t)}{=}\|f_{\text{rec}}(m){-}z^{(t)}\|_2^2$. The E-I retention gate, $\iota^{(t)}{=}\sigma(-\kappa\,|\rho^{(t)}_{EI}{-}\rho^{*}_{EI}|)$ with learnable $\kappa{>}0$, is high when the population-normalized E-I activity ratio, $\rho^{(t)}_{EI}{=}(\|r^{\text{pre}}_E\|_1/n_E)/(\|r_I\|_1/n_I)$, is near TIDE's target $\rho^{*}_{EI}\,{=}\,4$ and low otherwise. The buffer is updated by a momentum-smoothed, surprise-gated rule $m^{(t)}\,{=}\,m^{(t-1)}\,{+}\,v^{(t)}$ with $v^{(t)}\,{=}\,\mu v^{(t-1)}\,{+}\,g^{(t)}f_{\text{proj}}(z^{(t)})$ and $g^{(t)}{=}\,\mathbb{1}[s^{(t)}{>}\theta_s](1\,{-}\,\iota^{(t)})$, for $\theta_s{=}0.5$, $\mu{=}0.9$. The memory state does not receive the task loss gradient, while the heads and κ are optimized via backpropagation. The E-I retention gate is the only component of the architecture that *couples* the memory update with the network's homeostatic state: when the excitation dominates and the E-I ratio is unbalanced, i.e., the network is within an unreliable regime, the memory does not absorb new information.

4.2.6 Cross-attention & Output Head

A multi-headed attention with query $\texttt{Proj}_Q(z^{(t)})$ and keys/values supplied by the backbone is used to compute $a^{(t)}$ that is fed back into (3). For ImageNet-1K [44], we use an additional residual stream $a^{(t)}{\leftarrow}a^{(t)}_{\text{attn}}{+}z^{(t)}$ to enable adaptive attention via joint attention and latent synchronization to improve the attention mechanism, as exemplified in Figure 2, where CTM's attention is comparatively less stable. The logits are computed as $o^{(t)}{=}W_{\text{out}}\,\texttt{LN}\left[\texttt{GLU}(W_{\text{hidden}}[z^{(t)};m^{(t)}])\right]$ with the final prediction denoted by $o^{(T)}$, and the intermediate $\{o^{(t)}\}_{t<T}$, which drive the task loss and the certainty curve used for adaptive compute at inference. To preserve the sign of the gradient, GLU is used as its smooth gate keeps the gradient non-zero while operating on a normalized latent space that is sign-indefinite.

4.2.7 Training Objective

The total loss is a weighted sum of five terms, each capturing a distinct design goal:

$$\mathcal{L}(\theta)\;=\;\mathcal{L}_{\text{task}}\;+\;w(\texttt{step})\cdot\big(\lambda_{\text{EI}}\mathcal{L}_{\text{EI}}{+}\lambda_{\text{game}}\mathcal{L}_{\text{game}}{+}\lambda_{\text{sync}}\mathcal{L}_{\text{sync}}\big)\;+\;\lambda_{\text{spec}}\mathcal{L}_{\text{spec}},\qquad(6)$$

where $w(\texttt{step})\,{\in}\,[0,1]$ is the curriculum warm-up coefficient to allow warm-up training steps. Additionally, the weights for each loss are $(\lambda_{\text{EI}},\lambda_{\text{game}},\lambda_{\text{sync}},\lambda_{\text{spec}}){=}(10^{-2},10^{-3},10^{-4},10^{-1})$.

Task loss: TIDE directly adopts CTM's task loss that is expressed as $\mathcal{L}_{\text{task}}\,{=}\,\frac{1}{2}\,\texttt{CE}\,(o^{(t_{\text{min}})},y)\,{+}\,\frac{1}{2}\,\texttt{CE}\,(o^{(t_{\text{cert}})},y)$, where $t_{\text{min}}{=}\arg\min_t\,\texttt{CE}\,(o^{(t)},y)$ and $t_{\text{cert}}{=}\arg\max_t\;c^{(t)}$ for the entropy-based

certainty $c^{(t)} = 1 - H(p^{(t)})/\log C$. This loss encourages the model not only to provide a correct prediction at a given internal computation step but also to increase its confidence within the recurrent computation, thereby inducing adaptive behavior during internal computation steps.

E-I balance: To enforce E-I population balance, we penalize deviations of the activity ratio ρ_{EI} from its designated target $\rho^*_{EI} = 4$ via $\mathcal{L}_{\text{EI}} = \texttt{clip}\ (\rho_{EI} - \rho^*_{EI}, -50, 50)^2$, which acts as a soft homeostatic regularizer on the population-normalized E-I activity ratio and is set at an 80:20 population split. The target ρ^*_{EI} is a design choice that does not directly reflect the biological constant.

Game loss: We use a scalar, clipped, size-normalized surrogate of the per-neuron gradient residual, $\sum_i \|\partial_{x_i} E_i\|^2$, under mean-field approximation and (4): $\mathcal{L}_{\text{game}} = \frac{1}{d_{\text{model}}} \min(\left[n_E^{-1}\sum_{i=1}^{n_E}\mathcal{E}_E^{(i)} + n_I^{-1}\sum_{j=1}^{n_I}\mathcal{E}_I^{(j)}\right], 100)$, where $\mathcal{E}_E = \frac{[\,(\bar{w}_{EE}-d_E)\, r_E^{(t)} - \bar{w}_{EI}\,\bar{r}_I + u_E\,]^2}{2\,(d_E - \bar{w}_{EE})}$, $\mathcal{E}_I = \frac{[\,\bar{w}_{IE}\,\bar{r}_E - (\bar{w}_{II}+d_I)\, r_I^{(t)} + u_I\,]^2}{2\,(d_I + \bar{w}_{II})}$, $\bar{r}_\bullet = n_\bullet^{-1}\sum_i (r_\bullet)_i$, $d_E > 0$, and $d_I > 0$ denote population dissipation constants, and $\bar{w}_\bullet = \texttt{mean}(\text{diag}(W_\bullet))$ as population-effective scalar weights. Dividing by d_{model} makes the term scale-invariant, while clipping stabilizes transient spikes early in training.

Synchronization & spectral regularizers: $\mathcal{L}_{\text{sync}} = \|z^{(T)}\|_2^2/d_{\text{sync}}$ is used to bound the accumulators and latent space. The spectral regularizer is formulated as $\mathcal{L}_{\text{spec}} = \text{ReLU}\,(\widehat{\lambda}_{\text{P}}(W_{EE}) - \tau_{EE})^2 + \text{ReLU}\,(\widehat{\lambda}_{\text{P}}(W_{II}) - \tau_{II})^2$, where $(\tau_{EE},\, \tau_{II}) = (15, 7)$, while using the sum-ratio Perron estimator of (5). As discussed in Section 4.1.3, we use these values to achieve stability during training.

5 Experimental Evaluation

Setup: We evaluate TIDE on five image classification datasets: MNIST [1], Fashion-MNIST [46], CIFAR-10 and CIFAR-100 [47], and ImageNet-1K [44]; Detailed hyperparameter configurations are specified in Appendix F.2.

Main benchmark: Table 1 provides a detailed comparison between TIDE and CTM under multi-seeded training and evaluation under identical training budgets. Note that we retrain CTM following [15] using the same seeds as the TIDE model to ensure reproducible results. In general, TIDE is more sample efficient and requires fewer steps to reach higher `top-1` accuracy. As shown in Table 1, TIDE without reliance on deep feature extractors as its backbone can outperform CTM in the majority of image classification tasks, while requiring a smaller training budget.

Hyperparameter ablation study: We use MNIST and Fashion-MNIST to analyze how various hyperparameter choices affect `top-1` accuracy, learning outcomes, and the stability of TIDE. All

Table 1: Comparison between TIDE and CTM on five image classification datasets at equal training budget. The *Best Seed* column reports the highest-performing run with its best/final `top-1` accuracy, Acc@1 (%), and `mean` $\pm$ `std` across the multi seeded training for both architectures. CTM is retrained based on [15] for all the ablations with identical seeds. Note that † denotes the mean `top-1` accuracy drop of more than 1%, and ‡ indicates an effective batch size of 64. Moreover, the highest `top-1` accuracy for each dataset is in bold. Additional details on the experiments are provided in Appendix G.

			TIDE				**CTM**			
			`mean` $\pm$ `std`		**Best Seed**		`mean` $\pm$ `std`		**Best Seed**	
Task	**Steps**	d_{model}	best	final	best	final	best	final	best	final
MNIST	50K	256	**$99.64_{\pm 0.03}$**	**$99.61_{\pm 0.04}$**	**99.67**	**99.63**	$99.54_{\pm 0.03}$	$99.52_{\pm 0.02}$	99.56	99.54
	200K	256	**$99.60_{\pm 0.05}$**	**$99.53_{\pm 0.07}$**	**99.64**	**99.60**	$99.59_{\pm 0.06}$	$99.49_{\pm 0.08}$	**99.64**	99.57
F-MNIST	50K	256	**$94.21_{\pm 0.21}$**	**$94.15_{\pm 0.25}$**	**94.42**	**94.42**	$92.51_{\pm 0.11}$	$92.45_{\pm 0.19}$	92.63	92.63
	200K	256	**$94.12_{\pm 0.19}$**	**$93.85_{\pm 0.10}$**	**94.23**	**93.96**	$92.53_{\pm 0.24}$	$92.43_{\pm 0.27}$	92.80	92.73
CIFAR-10	600K	512	**$90.62_{\pm 0.06}$**	**$90.54_{\pm 0.07}$**	**90.68**	**90.62**	$86.23_{\pm 0.34}$	$85.98_{\pm 0.35}$	86.59	86.31
CIFAR-100	600K	718	$62.23_{\pm 0.16}$	$61.39_{\pm 0.82}$	62.38	61.97	$65.19_{\pm 0.48}$	**$65.01_{\pm 0.39}$**	65.70	**65.44**
	600K‡	718	**$65.68_{\pm 0.18}$**	$64.50_{\pm 0.60}{}^\dagger$	**65.79**	65.08				
ImageNet-1K	100K	4096	**$67.22_{\pm 1.52}$**	**$67.01_{\pm 1.58}$**	**68.74**	**68.59**	$52.67_{\pm 1.79}$	$51.47_{\pm 1.62}$	54.46	53.09
	200K	4096	**$68.40_{\pm 1.91}$**	**$68.18_{\pm 2.25}$**	**69.55**	**69.50**	$62.78_{\pm 1.31}$	$61.81_{\pm 1.50}$	64.06	$62.88^\dagger$
	500K	4096	—	—	69.90	69.87	—	—	**71.78**	**$69.88^\dagger$**

Table 2: Multi-seeded hyperparameter ablation study performed on MNIST and Fashion-MNIST (F-MNIST) datasets. The table provides the mean best and final `top-1` accuracy, Acc@1 (%), as the `mean` $\pm$ `std` across the seeds for evaluation during training. Default hyperparameters for the reported experiments are in bold, while the headers indicate which hyperparameter is being studied. We denote training instability, where the mean `top-1` accuracy drops by more than 1%, with †. Appendix G provides additional details on the ablation studies.

Setting	MNIST Acc@1 (%) best / final	F-MNIST Acc@1 (%) best / final	Setting	MNIST Acc@1 (%) best / final	F-MNIST Acc@1 (%) best / final	Setting	MNIST Acc@1 (%) best / final	F-MNIST Acc@1 (%) best / final
(i) Excitatory fraction n_E/d_{model}			***(ii) Game-loss weight*** λ_{game}			***(iii)*** τ_I ***(ms),*** $\tau_E{=}20$ ***fixed***		
0.6	$99.57_{\pm 0.07}$/$99.39_{\pm 0.30}$	$94.01_{\pm 0.36}$/$92.91_{\pm 2.17}$ †	0	$99.61_{\pm 0.04}$/$99.53_{\pm 0.09}$	$92.62_{\pm 2.56}$/$92.19_{\pm 2.96}$	3	$99.61_{\pm 0.03}$/$99.58_{\pm 0.02}$	$93.62_{\pm 0.69}$/$92.01_{\pm 3.09}$ †
0.7	$99.53_{\pm 0.08}$/$99.41_{\pm 0.19}$	$93.83_{\pm 0.16}$/$92.92_{\pm 1.75}$	$\mathbf{10^{-3}}$	**$99.61_{\pm 0.04}$/$99.55_{\pm 0.05}$**	**$93.80_{\pm 0.58}$/$93.70_{\pm 0.60}$**	**5**	**$99.61_{\pm 0.05}$/$99.55_{\pm 0.05}$**	**$93.69_{\pm 0.67}$/$93.48_{\pm 0.84}$**
0.8	**$99.60_{\pm 0.06}$/$99.55_{\pm 0.05}$**	**$93.78_{\pm 0.59}$/$91.88_{\pm 3.84}$** †	10^{-2}	$99.59_{\pm 0.09}$/$99.57_{\pm 0.09}$	$93.90_{\pm 0.33}$/$93.75_{\pm 0.38}$	7	$99.56_{\pm 0.09}$/$98.54_{\pm 2.02}$ †	$93.84_{\pm 0.68}$/$93.77_{\pm 0.68}$
0.9	$99.59_{\pm 0.05}$/$99.51_{\pm 0.16}$	$94.01_{\pm 0.35}$/$93.92_{\pm 0.39}$	10^{-1}	$99.34_{\pm 0.36}$/$94.17_{\pm 6.23}$ †	$93.39_{\pm 1.05}$/$92.95_{\pm 1.61}$	10	$99.59_{\pm 0.04}$/$99.57_{\pm 0.04}$	$93.95_{\pm 0.31}$/$93.91_{\pm 0.30}$
(iv) τ_E ***(ms),*** $\tau_I{=}5$ ***fixed***			***(v) Internal computation depth*** T			***(vi) Lateral inhibition iterations*** K_{WTA}		
10	$99.38_{\pm 0.46}$/$98.66_{\pm 1.79}$	$94.15_{\pm 0.20}$/$92.95_{\pm 2.34}$ †	10	$99.63_{\pm 0.04}$/$99.61_{\pm 0.05}$	$94.12_{\pm 0.18}$/$94.07_{\pm 0.18}$	1	$99.59_{\pm 0.08}$/$99.55_{\pm 0.10}$	$94.09_{\pm 0.06}$/$93.99_{\pm 0.09}$
15	$99.61_{\pm 0.02}$/$99.56_{\pm 0.07}$	$94.13_{\pm 0.23}$/$94.04_{\pm 0.29}$	25	$99.64_{\pm 0.04}$/$99.59_{\pm 0.06}$	$93.73_{\pm 0.62}$/$93.34_{\pm 1.29}$	3	$99.62_{\pm 0.04}$/$99.54_{\pm 0.04}$	$94.02_{\pm 0.29}$/$94.00_{\pm 0.29}$
20	**$99.62_{\pm 0.04}$/$99.57_{\pm 0.05}$**	**$93.96_{\pm 0.59}$/$93.77_{\pm 0.58}$**	**50**	**$99.62_{\pm 0.04}$/$99.56_{\pm 0.04}$**	**$93.86_{\pm 0.56}$/$93.78_{\pm 0.58}$**	**5**	**$99.57_{\pm 0.11}$/$99.42_{\pm 0.29}$**	**$93.66_{\pm 0.69}$/$93.47_{\pm 0.85}$**
25	$99.63_{\pm 0.03}$/$99.59_{\pm 0.06}$	$93.92_{\pm 0.66}$/$93.21_{\pm 1.80}$	75	$99.54_{\pm 0.12}$/$84.20_{\pm 29.80}$ †	$93.90_{\pm 0.15}$/$87.94_{\pm 11.68}$ †	10	$99.61_{\pm 0.08}$/$99.58_{\pm 0.08}$	$94.16_{\pm 0.27}$/$94.07_{\pm 0.33}$
30	$99.49_{\pm 0.17}$/$96.91_{\pm 5.21}$ †	$93.89_{\pm 0.64}$/$93.79_{\pm 0.64}$	100	$99.31_{\pm 0.36}$/$96.03_{\pm 5.49}$ †	$93.33_{\pm 0.97}$/$75.67_{\pm 34.19}$ †			

models are trained for 50 K steps, with an identical simple HRF backbone with $d_{model} = 256$ across all ablation studies. Table 2 reports multi-seeded experiments on TIDE, while Appendix G provides full breakdowns for each experiment. It has been observed that internal computation steps $T \in [10, 50]$ provides stable learning, while $\tau_E \in [15, 25]$ and $\tau_I \in [5, 7]$ provide the highest accuracy. Moreover, the choice of lateral inhibition iterations, K_{WTA}, depends on both the internal computation steps and the dataset. As illustrated in Table 2, $K_{\text{WTA}} = \{1, 3\}$ outperforms the selected default value and results in higher mean `top-1` accuracy for both MNIST and Fashion-MNIST.

Architectural component study: To further analyze each component of the proposed architecture, we perform a leave-one-out ablation study across the four image classification tasks (cf. Table 3) to compare TIDE and its components against CTM under a fixed training budget. Both TIDE and CTM are trained for 50 K steps across MNIST and Fashion-MNIST datasets with identical backbone and seeds similar to Table 1. Additionally, the training budgets for CIFAR-10 and CIFAR-100 follow a similar pattern and are set to 600 K steps. Table 3 demonstrates that TIDE with all components achieves the highest performance across all the datasets compared to CTM. In contrast, removing any individual component of TIDE results in losing `top-1` accuracy. The largest drop in best `top-1` accuracy occurs in MNIST and Fashion-MNIST when removing the memory component, while the choice of backbone has a more detrimental effect on CIFAR-10 and CIFAR-100. Additional ablation studies on the backbone variants are provided in Appendix G.

Table 3: Multi-seeded leave-one-out ablation performed on four image classification datasets to analyze TIDE and its components. The table provides the mean best and final `top-1` accuracy, Acc@1 (%), as the `mean` $\pm$ `std` across the seeds for evaluation during training. Moreover, the highest `top-1` accuracy for each dataset is in bold, and the mean top-1 accuracy drop by more than 1% in the final stages of training is marked with †.

	MNIST		F-MNIST		CIFAR-10		CIFAR-100	
Configuration	best	final	best	final	best	final	best	final
Reference								
CTM	$99.54_{\pm 0.03}$	$99.52_{\pm 0.02}$	$92.51_{\pm 0.11}$	$92.45_{\pm 0.19}$	$86.23_{\pm 0.34}$	$85.98_{\pm 0.35}$	$65.19_{\pm 0.48}$	**$65.01_{\pm 0.39}$**
TIDE (full, $\lambda_{\text{spec}} = 0.0$**)**	**$99.64_{\pm 0.02}$**	$99.57_{\pm 0.02}$	$94.21_{\pm 0.21}$	$94.15_{\pm 0.25}$	**$90.62_{\pm 0.06}$**	**$90.54_{\pm 0.07}$**	**$65.68_{\pm 0.18}$**	$64.50_{\pm 0.60}$ †
Component removed								
− Wilson-Cowan update	$99.61_{\pm 0.02}$	$99.56_{\pm 0.01}$	$93.70_{\pm 0.12}$	$93.65_{\pm 0.15}$	$90.18_{\pm 0.26}$	$90.14_{\pm 0.27}$	$64.52_{\pm 0.67}$	$63.65_{\pm 0.40}$
− Dale's law constraint	$99.59_{\pm 0.03}$	$99.57_{\pm 0.04}$	$93.55_{\pm 0.18}$	$93.51_{\pm 0.16}$	$89.01_{\pm 1.74}$	$85.46_{\pm 6.75}$ †	$65.39_{\pm 0.39}$	$64.53_{\pm 0.91}$
− Lateral inhibition (WTA)	$99.61_{\pm 0.02}$	$99.58_{\pm 0.04}$	$93.80_{\pm 0.17}$	$93.77_{\pm 0.19}$	$89.13_{\pm 1.20}$	$82.90_{\pm 8.69}$ †	$64.94_{\pm 0.30}$	$63.65_{\pm 0.24}$ †
− Surprise-gated memory	$99.59_{\pm 0.01}$	$99.51_{\pm 0.01}$	$92.95_{\pm 0.16}$	$92.76_{\pm 0.16}$	$88.97_{\pm 1.41}$	$85.68_{\pm 4.58}$ †	$64.32_{\pm 0.30}$	$62.03_{\pm 0.97}$ †
− ResNet-18 inst. of HRF backbone	$99.60_{\pm 0.03}$	$99.51_{\pm 0.05}$	$93.03_{\pm 0.12}$	$92.83_{\pm 0.20}$	$83.02_{\pm 2.02}$	$48.82_{\pm 30.94}$ †	$54.49_{\pm 0.76}$	$53.25_{\pm 1.02}$ †
Component added								
+ Spectral regularizer ($\lambda_{\text{spec}} = 0.1$)	$99.64_{\pm 0.03}$	**$99.61_{\pm 0.05}$**	**$94.31_{\pm 0.06}$**	**$94.27_{\pm 0.08}$**	$89.02_{\pm 2.93}$	$88.14_{\pm 4.94}$	$65.38_{\pm 0.35}$	$64.60_{\pm 0.32}$

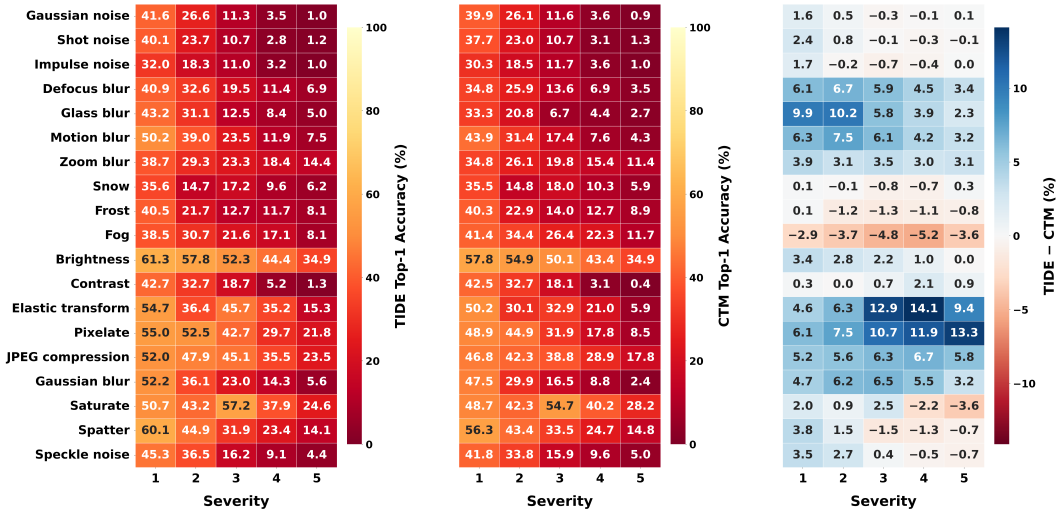

Figure 3: Robustness analysis using ImageNet-C [48] for corrupted images with perturbations. *Left* panel presents results for TIDE, *center* panel corresponds to CTM, and *right* panel reports differences.

Robustness study: Tiny-ImageNet [49], ImageNet-C [48], and ImageNet-R [50] were used to assess the OOD robustness of TIDE and CTM that are solely trained on the original ImageNet-1K (cf. Table 1). Figure 3 summarizes the performance degradation of CTM and TIDE on ImageNet-C that are trained for $500\,\text{K}$ steps. TIDE outperforms CTM in the presence of most perturbations. Note that under aerosol-particle perturbations such as fog, frost, and snow, CTM is more robust. Nevertheless, under all considered perturbations, TIDE improves CTM's `top-1` accuracy by an average of $+2.68\%$ as shown in Figure 3. We also provide additional robustness evaluations for various severities, datasets, and training budgets in Appendix G.10.

6 Discussion & Conclusions

In this study, we presented TIDE, a new architecture that extends CTM with Wilson-Cowan dynamics and Dale's principle, enforcing non-negative weights. This design enables integration of an asymmetric neural circuit to stabilize temporal dynamics while running a two-population recurrence for internal computation steps, as in CTM. Moreover, we used homeostatic E-I ratio activity to introduce a novel retention gate that serves as a persistent memory update.

TIDE can achieve high `top-1` accuracy, even with the population collapse, where the full neuron population converts to excitatory, indicating that CTM can be considered as a special case of TIDE, without inhibitory neurons, $n_I = 0$ (cf. Table 2). Furthermore, Appendix G.10 reports that TIDE trained with a smaller training budget between $100\,\text{K}$ to $200\,\text{K}$ steps can outperform CTM under varied severity of perturbations. This supports the hypothesis about TIDE's stability and sample efficiency while justifying the design with neuro-inspired primitives, which require fewer computational resources to achieve generalization and better OOD performance.

Further performance gains can be achieved by replacing the scalar-weighted game loss with an augmented-Lagrangian formulation that couples the per-neuron Nash residual to $\rho_{EI} \approx \rho^{\star}_{EI}$, enabling manifold-based optimization instead of a task-specific gradient approach. Moreover, imposing stricter biophysical constraints and moving toward more bio-plausible modeling may improve stability under extensive inhibitory feedback, helping address the training drift and computational time instability observed in Table 2. Additional coupling between neural circuits and the input via anchoring [20] can further stabilize the architecture and its temporal dynamics beyond the trained internal computation with recurrence (cf. Appendix G). Finally, further investigation and evaluation are needed to scale TIDE to sequence-based modeling, enabling spike-based implementations for greater energy efficiency.

Acknowledgments and Disclosure of Funding

We thank three anonymous reviewers for their comments that helped to improve the presentation style and experimental evaluation reported in the study. The authors thank RISE ICE Data Center for providing access to the computing infrastructure. DK acknowledges funding from the Swedish Strategic Research Foundation (Grant No. FFL24-0111) and the Swedish Research Council (Grant No. 2025-05421). We thank the Kempe-foundation for providing us with the Lotty Bruzilius cluster used for compute. The authors declare no competing interests.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] ChiYan Lee, Hideyuki Hasegawa, and Shangce Gao. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1406–1426, 2022.

[7] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7:13276, 2016.

[8] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24, 2017.

[9] Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A. Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature Neuroscience*, 24:1010–1019, 2021.

[10] Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1):3770, 2019.

[11] Jeffry S. Isaacson and Massimo Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.

[12] Adam Haber and Elad Schneidman. The computational and learning benefits of Daleian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[13] Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri M. Kullmann, and Blake Richards. Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] Bilal Haider, Alvaro Duque, Andrea R. Hasenstaub, and David A. McCormick. Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *The Journal of Neuroscience*, 26(17):4535–4545, 2006.

[15] Luke Darlow, Ciaran Regan, Sebastian Risi, Jeffrey Seely, and Llion Jones. Continuous thought machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[16] Henry Dale. Pharmacology and nerve-endings. *Proceedings of the Royal Society of Medicine*, 28(3):319–332, 1935.

[17] John C. Eccles, Paul Fatt, and Kyozo Koketsu. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. *The Journal of Physiology*, 126(3):524–562, 1954.

[18] Simone Betteti, William Retnaraj, Alexander Davydov, Jorge Cortés, and Francesco Bullo. Competition, stability, and functionality in excitatory-inhibitory neural circuits. *arXiv:2512.05252*, 2025.

[19] Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. Can you learn an algorithm? Generalizing from easy to hard problems with recurrent networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 6695–6706, 2021.

[20] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[21] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[22] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[23] Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to ponder. In *ICML Workshop on Automated Machine Learning*, 2021.

[24] Hugh R. Wilson and Jack D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1–24, 1972.

[25] Hugh R. Wilson and Jack D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80, 1973.

[26] Carl van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996.

[27] Nicolas Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of Computational Neuroscience*, 8(3):183–208, 2000.

[28] Tim P. Vogels, Henning Sprekeler, Friedemann Zenke, Claudia Clopath, and Wulfram Gerstner. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062):1569–1573, 2011.

[29] Gina G. Turrigiano. The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435, 2008.

[30] Kenneth D. Harris and Thomas D. Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, 2013.

[31] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[32] Atlas Kazemian, Eric Elmoznino, and Michael F. Bonner. Convolutional architectures are cortex-aligned de novo. *Nature Machine Intelligence*, 7:1834–1844, 2025.

[33] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[34] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[35] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

[36] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 113506–113543. Curran Associates, Inc., 2025.

[37] Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. *arXiv:2504.13173*, 2025.

[38] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[39] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[40] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[42] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[43] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[45] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 207(1167):187–217, 1980.

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

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

[48] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[49] Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

[50] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of- distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021.

[51] Andrzej Granas and James Dugundji. *Fixed Point Theory*. Springer Monographs in Mathematics. Springer, 2003.

[52] Carl van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Computation*, 10(6):1321–1371, 1998.

[53] Yashar Ahmadian and Kenneth D. Miller. What is the dynamical regime of cerebral cortex? *Neuron*, 109(21):3373–3391, 2021.

[54] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[55] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002.

[56] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2013.

[57] Nicolas Brunel and Vincent Hakim. Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Computation*, 11(7):1621–1671, 1999.

[58] Stephen W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

[59] Dennis M. Dacey, Beth B. Peterson, Farrel R. Robinson, and Paul D. Gamlin. Fireworks in the primate retina: In vitro photodynamics reveals diverse LGN-projecting ganglion cell types. *Neuron*, 37(1):15–27, 2003.

[60] David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[61] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Boston, MA, 1994.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main theoretical contribution of the paper is described in Sections 3-4, while its empirical validation is provided in Section 5 with additional details following in Appendix F.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6, we discuss several limitations, including the addition of an augmented-Lagrangian formulation to couple the per-neuron Nash residual to $\rho_{EI} \approx \rho^*_{EI}$, enabling manifold-based optimization instead of a task-specific gradient approach. Moreover, imposing strict biophysical constraints to further improve the current architecture should be considered. Additional limitations and open questions for future work are provided in Appendix H.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced. We state the main theoretical results in the main text and provide proofs and additional details in the Appendix, e.g., in Appendices A-C.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information needed to reproduce the main experimental results of the paper is provided in the Appendix. Appendix E provides the pseudocode of the proposed architecture. Appendix F provides all hyperparameters and experimental setup necessary to reproduce the experiments. To streamline reproducibility, we are attaching the source code implementing the architecture.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used for the experiments reported in this study are open-source and are available from Hugging Face. Along with this submission, we provide the source code that implements the architecture. The experimental protocol, preprocessing steps, and hyperparameters used in simulations are provided in Appendix F. The scripts necessary to rerun the simulations will be released prior to publication of the paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details are provided in Appendix F.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Whenever possible, we reported the results of several simulations with different seeds for PRG. The corresponding figures/tables report the mean and standard deviation. However, the overall suite of experiments was computationally demanding (e.g., some runs took up to two weeks, see Appendix F.3 for more details), and taking into account the constraints of our compute infrastructure, some of the experiments and ablation studies we ran used only one seed. This was the choice we made to increase the breadth of the experiments being conducted.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is provided in Appendix F.3.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics while working on this study.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The presented study is foundational research. We do not see any immediate path to negative applications and, therefore, the paper does not discuss potential societal impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for the responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not foresee that at the current stage of research, the proposed architecture requires any safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Everything that was used in this study is open source, and where applicable, the credit was given via academic references.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Appendix E provides the pseudocode of the proposed architecture. We also attach the source code implementing the architecture.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Crowdsourcing was not used in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Human subjects were not involved in this study.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used in the ways that might impact the core methodology, scientific rigor, or originality of the research.

Appendices

Contents

1 Introduction **1**

2 Related Work **2**

3 Preliminaries **3**

4 TIDE Architecture & Its Foundations **4**

4.1 Theoretical Foundations 4

4.1.1 Dale-constrained Gradient Optimization 4

4.1.2 Game-theoretic Formulation & Lyapunov Diagonal Stability 4

4.1.3 Spectral Stability of Wilson-Cowan Dynamics 5

4.2 TIDE Architecture 6

4.2.1 Hierarchical Receptive Field 6

4.2.2 Population-specific NLM 6

4.2.3 Synchronization as Latent Representation 6

4.2.4 Lateral Inhibition 6

4.2.5 Surprise-gated Associative Memory with an E-I Retention Gate 7

4.2.6 Cross-attention & Output Head 7

4.2.7 Training Objective 7

5 Experimental Evaluation **8**

6 Discussion & Conclusions **10**

Appendices **18**

A Discrete-Time E-I Dynamics **20**

A.1 Notation 20

A.2 Forward Euler Recurrence 20

A.2.1 Regime 1: Unconstrained Weights 20

A.2.2 Regime 2: Non-negative Weights 21

A.3 E-I Balance & Fixed-point Equation 21

B TIDE Objective & Training **22**

B.1 Loss Constituents 22

B.2 Curriculum Design 23

B.3 Dual Loss Conditioning 24

B.4 Training Algorithm 24

C Continuous-time Limit, Stability, & Game-theoretic Fixed Points **24**

C.1 Continuous-time E-I ODE 24

C.2 Linear Stability Around a Fixed Point 25

C.3 Lyapunov Diagonal Stability & Game-theoretic Nash Equilibria . . . 25
C.4 Spectral Perron Surrogate . . . 26
C.5 Relationship to Wilson-Cowan & Brunel-Hakim Limits . . . 28

D HRF Backbone **28**
D.1 Learnable Filter Bank . . . 28
D.2 Equivariance Properties . . . 28

E TIDE Components **29**
E.1 Deep-HRF Backbone . . . 29
E.2 Shallow HRF Backbone . . . 29
E.3 Wilson-Cowan E-I Update . . . 29
E.4 Population-specific NLM . . . 30
E.5 Three-type Synchronization . . . 30
E.6 Cross-attention Readout . . . 31
E.7 Lateral Inhibition . . . 31
E.8 Persistent Memory with Retention Gate . . . 31
E.9 Output Head . . . 32
E.10 Perron Sum-ratio Estimator . . . 32

F Experimental Details **32**
F.1 Datasets & Preprocessing . . . 33
F.2 Architecture & Training Hyperparameters . . . 33
F.3 Compute Resources . . . 34
F.4 Low-variance Estimator . . . 35

G Additional Experiments & Ablations **36**
G.1 Main Benchmark . . . 36
G.2 E-I Ratio Analysis . . . 38
G.3 Time-constant Analysis . . . 38
G.4 Iteration-depth Analysis . . . 40
G.5 Game-theoretic Loss Weight Analysis . . . 40
G.6 Lateral Inhibition Iterations . . . 41
G.7 Training-stability Diagnostics . . . 42
G.8 Architectural Component Analysis . . . 44
G.9 Backbone Analysis & Robustness . . . 45
G.10 Robustness & OOD Evaluation . . . 46

H Open Questions & Limitations **50**

A Discrete-Time E-I Dynamics

This section is organized as follows. Initially, notations are introduced in Appendix A.1, and the forward Euler recurrence for the discrete-time Wilson-Cowan E-I dynamics is presented in Appendix A.2, together with its exact form under the parameterization following Dale's principle, while considering the two canonical cases: i) signed weights and ii) Dale-constrained weights. Finally, we derive the fixed-point equation and characterize the loose- versus tight-balance regimes in Appendix A.3.

A.1 Notation

The excitatory and inhibitory populations are denoted by sets $\mathcal{E}=\{1,\ldots,n_E\}$ and $\mathcal{I}=\{1,\ldots,n_I\}$, respectively, with $n_E=\lfloor\rho\cdot\, d_{\text{model}}\rfloor$, $n_I=d_{\text{model}}-n_E$, where $\rho=0.8$ unless mentioned otherwise. Population firing rates are given by vectors $r_E^{(t)}\in\mathbb{R}_{\geq 0}^{n_E}$ and $r_I^{(t)}\in\mathbb{R}_{\geq 0}^{n_I}$ for every internal computation step $t\in\{0,1,\ldots,T\}$, while their pre-activations are denoted by vectors $h_E^{(t)}$ and $h_I^{(t)}$. The four matrices specifying recurrent weights are $W_{EE}\in\mathbb{R}^{n_E\times n_E}$, $W_{EI}\in\mathbb{R}^{n_E\times n_I}$, $W_{IE}\in\mathbb{R}^{n_I\times n_E}$, and $W_{II}\in\mathbb{R}^{n_I\times n_I}$. Note that Dale's principle restricts all four matrices to be positive, while the time constants for the excitatory and inhibitory populations are $\tau_E=20$ ms and $\tau_I=5$ ms, respectively. Moreover, the Euler step is set to $\Delta t=1$ ms, resulting in the following integration coefficients: $\alpha_E=\Delta t/\tau_E=0.05$ and $\alpha_I=\Delta t/\tau_I=0.20$. Activation function ReLU is denoted by $\varphi(\cdot)$, while $\odot$ denotes the Hadamard product, and $\langle a,b\rangle=a^\top b$ denotes the inner product. The population-normalized E-I activity ratio is given by:

$$\rho_{EI}(r_E,r_I) \;=\; \frac{\|r_E\|_1/n_E}{\|r_I\|_1/n_I+\varepsilon},\qquad \varepsilon=10^{-8}, \tag{A.1}$$

where ε is only required during backpropagation to prevent division by zero, and the biological target $\rho_{EI}^*=4$ is defined to enforce reaching $80:20$ population-based optimization and is used for the regularizer of Appendix B. For a matrix M, $\lambda_i(M)$ is used to denote its eigenvalues, $\sigma_i(M)$ for its singular values, $\rho_{\text{sp}}(M)=\max_i|\lambda_i(M)|$ for the spectral radius, and $\lambda_{\text{P}}(M)$ for the Perron eigenvalue whenever M is component-wise non-negative. Lastly, $\mathbb{1}\in\mathbb{R}^n$ denotes the all-ones vector.

A.2 Forward Euler Recurrence

The continuous-time Wilson-Cowan rate model of [24] is expressed as follows:

$$\tau_E\,\dot{r}_E=-r_E+\varphi\big(W_{EE}\,r_E-W_{EI}\,r_I+u_E\big), \tag{A.2}$$

$$\tau_I\,\dot{r}_I=-r_I+\varphi\big(W_{IE}\,r_E-W_{II}\,r_I+u_I\big), \tag{A.3}$$

where u_E,u_I denote external inputs; the parameterization of [13] enforces Dale's principle by fixing the signs of W_{EI},W_{II} through explicit subtraction rather than through negative components. The first-order forward Euler discretizations from (A.2)–(A.3) with step Δt yields:

$$r_E^{(t)}=(1-\alpha_E)\,r_E^{(t-1)}+\alpha_E\,\varphi\big(h_E^{(t)}\big), \tag{A.4}$$

$$r_I^{(t)}=(1-\alpha_I)\,r_I^{(t-1)}+\alpha_I\,\varphi\big(h_I^{(t)}\big), \tag{A.5}$$

with pre-activations computed as:

$$h_E^{(t)}=W_{EE}\,r_E^{(t-1)}-W_{EI}\,r_I^{(t-1)}+W_E^{\text{in}}a^{(t)}, \tag{A.6}$$

$$h_I^{(t)}=W_{IE}\,r_E^{(t-1)}-W_{II}\,r_I^{(t-1)}+W_I^{\text{in}}a^{(t)}. \tag{A.7}$$

Here $a^{(t)}\in\mathbb{R}^{d_{\text{sync}}}$ is the cross-attention output described in Section 4.2.6 and $W_E^{\text{in}}\in\mathbb{R}^{n_E\times d_{\text{sync}}}$, $W_I^{\text{in}}\in\mathbb{R}^{n_I\times d_{\text{sync}}}$ are unconstrained input projections. Note that (A.6)–(A.7) are followed by population-local `RMSNorm` in the actual implementation. However, we omit this mapping in the derivation since it is a smooth, homogeneity-preserving transformation of each population and does not affect the fixed-point analysis. Thus, the formulation can be expressed in the concatenated form and is as follows:

$$x=\begin{bmatrix}r_E\\ r_I\end{bmatrix}\in\mathbb{R}^n,\qquad W_{\text{eff}}=\begin{bmatrix}W_{EE} & -W_{EI}\\ W_{IE} & -W_{II}\end{bmatrix}\in\mathbb{R}^{d_{\text{model}}\times d_{\text{model}}},\qquad d_{\text{model}}=n_E+n_I. \tag{A.8}$$

Next, the analysis of (A.4) and (A.5) in their two canonical regimes is provided.

A.2.1 Regime 1: Unconstrained Weights

Consider the case in which all four connectivity matrices are fully unconstrained, i.e., $W_{\bullet\bullet}\in\mathbb{R}^{n_\bullet\times n_\bullet}$. By applying the triangle inequality and the Lipschitz property of φ with the constant L_φ to (A.4)–(A.5), any two states $x,\tilde{x}$, satisfy:

$$\|\Psi(x)-\Psi(\tilde{x})\|_2\leq(1-\alpha)\,\|x-\tilde{x}\|_2+\alpha\,L_\varphi\,\|W_{\text{eff}}\|_2\,\|x-\tilde{x}\|_2 \quad \because \text{triangle + Lipschitz} \tag{A.9}$$

$$=\gamma\,\|x-\tilde{x}\|_2 \quad \text{with} \quad \gamma\equiv(1-\alpha)+\alpha L_\varphi\,\|W_{\text{eff}}\|_2, \tag{A.10}$$

where the population-specific α_E, α_I is replaced by $\alpha = \max(\alpha_E, \alpha_I)$ for the purpose of this upper bound. Hence, Ψ is a contraction iff $\|W_{\text{eff}}\|_2 < 1/L_\varphi$, which for ReLU ($L_\varphi = 1$) reduces to $\|W_{\text{eff}}\|_2 < 1$. In the continuous-time, given limit $\alpha \to 0$, the contraction factor satisfies $\gamma \to 1$ for any W_{eff}, thus recovering the neutrally stable behavior of the underlying Wilson-Cowan dynamics [24]. This confirms that the continuous-time limit removes the discretization-induced instability, consistent with [24].

A.2.2 Regime 2: Non-negative Weights

Consider the case in which every component of $W_{..}$ is non-negative. In such a case, the signs of the inhibitory currents appear explicitly in (A.6)-(A.7); the block matrix W_{eff} of (A.8) must therefore be sign-indefinite with a block-wise structure as follows:

$$W_{\text{eff}} = \Sigma \odot |W_{\text{eff}}|, \qquad \Sigma = \begin{bmatrix} +\mathbb{1}_{n_E \times n_E} & -\mathbb{1}_{n_E \times n_I} \\ +\mathbb{1}_{n_I \times n_E} & -\mathbb{1}_{n_I \times n_I} \end{bmatrix}, \tag{A.11}$$

with $|W_{\text{eff}}| \geq 0$, that is a non-negative matrix whose components are W_{EE}, W_{EI}, W_{IE}, and W_{II} concatenated without signs. Furthermore, since the sign mask Σ is fixed and only the magnitudes $|W_{\text{eff}}|$ are learned directly, (A.11) is considered as the central Dale parameterization. We formalize this as follows.

Definition A.1 (Dale parameterization). Let $\mathcal{W}_{\text{Dale}} = \{W \geq 0\} \subset \mathbb{R}^{m \times n}$ be the non-negative orthant. A *Dale-parameterized* effective matrix is $W_{\text{eff}}(\theta) = \Sigma \odot |W_{\text{eff}}(\theta)|$ with $|W_{\text{eff}}| \in \mathcal{W}_{\text{Dale}}$ and Σ constructed according to (A.11).

Additionally, we define the Dale projection: $\Pi_{\text{Dale}}(W)_{ij} = \max(W_{ij}, 0)$.

Proposition A.2 (Properties of Π_{Dale}). For every $W, \tilde{W} \in \mathbb{R}^{m \times n}$, *i)* $\Pi_{\text{Dale}}\big(\Pi_{\text{Dale}}(W)\big) = \Pi_{\text{Dale}}(W)$ *(idempotence)*; *ii)* $\|\Pi_{\text{Dale}}(W) - \Pi_{\text{Dale}}(\tilde{W})\|_F \leq \|W - \tilde{W}\|_F$ *(non-expansiveness)*; *iii)* $\Pi_{\text{Dale}}(W) = \arg\min_{W' \in \mathcal{W}_{\text{Dale}}} \|W - W'\|_F^2$ *(Euclidean projection).*

Proof. *i)* $\max(\max(W_{ij}, 0), 0) = \max(W_{ij}, 0)$ by idempotence of the scalar function $\max(\cdot, 0)$. *ii)* Since $\max(\cdot, 0)$ is 1-Lipschitz, $|\max(W_{ij}, 0) - \max(\tilde{W}_{ij}, 0)| \leq |W_{ij} - \tilde{W}_{ij}|$; squaring and summing over (i, j) yields the non-expansiveness in Frobenius norm. *iii)* The minimization $\min_{W' \geq 0} \|W - W'\|_F^2$ is separable over components, and for each scalar w the solution of $\min_{w' \geq 0}(w - w')^2$ is $w' = \max(w, 0)$. □

Theorem A.3 (Convergence of Dale-constrained gradient descent). Let $\mathcal{L}$ be differentiable in every $W_{..}$, and let $W^{(k+1)} = \Pi_{\text{Dale}}\big(W^{(k)} - \eta \nabla_W \mathcal{L}(W^{(k)})\big)$ be the projected-gradient update with learning rate $\eta > 0$. Then: *i)* $W^{(k)} \in \mathcal{W}_{\text{Dale}}$ for every $k \geq 0$ if $W^{(0)} \in \mathcal{W}_{\text{Dale}}$; *ii)* For $\mathcal{L}$ that is L-smooth and μ-strongly convex over $\mathcal{W}_{\text{Dale}}$, the iterates should satisfy $\|W^{(k)} - W^*\|_F^2 \leq (1 - \mu/L)^k \|W^{(0)} - W^*\|_F^2$ for the given unique $W^* = \arg\min_{W \in \mathcal{W}_{\text{Dale}}} \mathcal{L}(W)$.

Proof. *i)* Following Proposition A.2.*i* and *iii*, Π_{Dale} is idempotent, and given it is the Euclidean projection onto $\mathcal{W}_{\text{Dale}}$, its range satisfies $\text{range}(\Pi_{\text{Dale}}) = \mathcal{W}_{\text{Dale}}$. Hence, $W^{(k+1)} = \Pi_{\text{Dale}}(\cdot) \in \mathcal{W}_{\text{Dale}}$ for any $W^{(k)}$ and in particular for any $W^{(0)} \in \mathcal{W}_{\text{Dale}}$. *ii)* By the descent lemma for L-smooth $\mathcal{L}$, the unconstrained step $\tilde{W}^{(k+1)} = W^{(k)} - \eta \nabla_W \mathcal{L}$ with $\eta = 1/L$ obeys $\|\tilde{W}^{(k+1)} - W^*\|_F^2 \leq \|W^{(k)} - W^*\|_F^2 - 2\eta \langle \nabla_W \mathcal{L}(W^{(k)}), W^{(k)} - W^* \rangle + \eta^2 \|\nabla_W \mathcal{L}(W^{(k)})\|_F^2$; μ-strong convexity upper-bounds the middle term by $-\mu \|W^{(k)} - W^*\|_F^2$, while L-smoothness bounds the last by $L^2 \|W^{(k)} - W^*\|_F^2$. Assembling, $\|\tilde{W}^{(k+1)} - W^*\|_F^2 \leq (1 - 2\mu/L + \mu^2/L^2) \|W^{(k)} - W^*\|_F^2 = (1 - \mu/L)^2 \|W^{(k)} - W^*\|_F^2$. Applying Π_{Dale} cannot expand the distance to $W^* \in \mathcal{W}_{\text{Dale}}$ (Proposition A.2.*ii*), giving $\|W^{(k+1)} - W^*\|_F^2 \leq (1 - \mu/L)^2 \|W^{(k)} - W^*\|_F^2 \leq (1 - \mu/L) \|W^{(k)} - W^*\|_F^2$, where we used $(1 - x)^2 \leq 1 - x$ for $x \in [0, 1]$. □

Remark A.4 (Gradient compatibility across internal computation steps). Although Theorem A.3 is stated for a single outer-loop optimizer step, TIDE's forward pass invokes each $W_{..}$ at every internal computation step $t = 1, \ldots, T$. Therefore, as Π_{Dale} is applied once after each optimizer step and not inside the recurrence, the gradient through-time $\partial \mathcal{L} / \partial W_{..} = \sum_{t=1}^{T} (\partial \mathcal{L} / \partial W_{..})^{(t)}$ is still the BPTT sum of (A.6)–(A.7) across all internal computation steps, and the conclusion of the theorem carries over verbatim.

A.3 E-I Balance & Fixed-point Equation

With the cross-attention frozen at $a^{(t)} \equiv a^*$, and with lateral inhibition removed, any fixed point (r_E^*, r_I^*) of the Euler recurrence from (A.4)-(A.5) must satisfy the self-consistency relation

$$r_\bullet^* = \varphi\big(W_{\bullet E}\, r_E^* - W_{\bullet I}\, r_I^* + W_\bullet^{\text{in}}\, a^*\big), \tag{A.12}$$

where $\bullet \in \{E, I\}$. Existence of such a pair follows from Brouwer's theorem:

Theorem A.5 (Fixed-point existence). Assume $\varphi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is continuous and saturating, with $\sup_{u \in \mathbb{R}} \varphi(u) = M_\varphi < \infty$, e.g., $\varphi = \mathrm{ReLU}(\cdot)$ composed with post-LayerNorm clipping, or any bounded continuous activation. Then (A.12) has at least one solution in the compact cube $[0, M_\varphi]^{n_E+n_I}$, if φ is L_φ-Lipschitz and $L_\varphi \|W_{\text{eff}}\|_2 < 1$.

Proof. Define $\Phi : \mathbb{R}^{n_E+n_I} \to [0, M_\varphi]^{n_E+n_I}$ by $\Phi(r_E, r_I) = (\varphi(h_E), \varphi(h_I))$ with $h_\cdot$ as in (A.6)-(A.7), where each coordinate of $\varphi(\cdot)$ is taken component-wise. The hypercube $K = [0, M_\varphi]^{n_E+n_I}$ is compact and convex. The map Φ is continuous, given φ is continuous and the linear map $(r_E, r_I) \mapsto (h_E, h_I)$ is continuous. Finally, for any $r \in \mathbb{R}^{n_E+n_I}$, $\Phi(r) \in [0, M_\varphi]^{n_E+n_I} = K$ by the saturation hypothesis on φ; in particular $\Phi(K) \subseteq K$. Hence, Brouwer's fixed-point theorem [51, Ch. II, Theorem 2.1.5] yields a fixed point of Φ in K, which solves (A.12). □

Remark A.6 (Saturation hypothesis). Given that the saturation hypothesis on φ cannot be satisfied by the standard ReLU used in the TIDE implementation, the forward pass includes `RMSNorm` on both populations before the readout, which bounds the mean-square of the activations. In practice, we observe bounded per-neuron activities at convergence (Appendix G.7); the compact-set hypothesis of Theorem A.5 is therefore satisfied a posteriori at convergence, though the pre-normalization pre-activations are unbounded.

Definition A.7 (Loose and tight E-I balance). The effective excitatory and inhibitory currents at a fixed point are denoted as $I_E^{\text{exc}} = W_{EE} r_E^*$, $I_E^{\text{inh}} = W_{EI} r_I^*$, and analogously for the inhibitory population. It can be stated that the E-I network is in the tight-balance regime iff $I_\bullet^{\text{exc}} - I_\bullet^{\text{inh}} = \mathcal{O}(1)$, while $I_\bullet^{\text{exc}}$, $I_\bullet^{\text{inh}} = \mathcal{O}(\sqrt{K})$ for in-degree K. Note that $\mathcal{O}(\cdot)$ denotes asymptotic scaling, which represents the strong-coupling limit first introduced by [26, 52] and formalized as the asynchronous-irregular balance regime by [27, 53]. Otherwise, the network is in the loose-balance regime.

Furthermore, the target $\rho_{EI}^* = 4$ used in (A.1) is chosen such that the tight-balance regime at $\rho = 0.8$ is achieved, given that normalizing by population size results in a per-neuron excitation-to-inhibition rate ratio close to unity while the current balance is $\mathcal{O}(1)$. The ρ_{EI} regularizer of Appendix B encourages TIDE to operate close to this regime without hard-coding it.

B TIDE Objective & Training

In this section, we first decompose the loss function used as the training objective of TIDE into its five constituent terms in Appendix B.1. Afterward, the curriculum design for the warm-up phase is presented (Appendix B.2), and then the CTM dual loss as a conditional-expectation reduction of per internal step based on cross-entropy is discussed (Appendix B.3). Finally, we provide the training algorithm outline in Appendix B.4.

B.1 Loss Constituents

Let x denote an input image, $y \in \{1, \ldots, C\}$ its label, $o^{(t)} = f_\theta^{(t)}(x)$ the logits at internal computation step t, and $p^{(t)} = \mathrm{softmax}(o^{(t)})$. The TIDE training loss is computed as follows:

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{task}} \;+\; w(\texttt{step}) \cdot \Big[\lambda_{\text{EI}} \mathcal{L}_{\text{EI}} \;+\; \lambda_{\text{game}} \mathcal{L}_{\text{game}} \;+\; \lambda_{\text{sync}} \mathcal{L}_{\text{sync}}\Big] \;+\; \lambda_{\text{spec}} \mathcal{L}_{\text{spec}}, \tag{A.13}$$

where $w(\texttt{step}) \in [0, 1]$ is the curriculum warm-up coefficient described below. The task loss follows CTM's dual cross-entropy formulation [15],

$$\mathcal{L}_{\text{task}} \;=\; \tfrac{1}{2}\, \texttt{CE}\big(o^{(t_{\min})}, y\big) + \tfrac{1}{2}\, \texttt{CE}\big(o^{(t_{\text{cert}})}, y\big), \tag{A.14}$$

with $t_{\min} = \arg\min_t \texttt{CE}(o^{(t)}, y)$ and $t_{\text{cert}} = \arg\max_t c^{(t)}$, where $c^{(t)} = 1 - H(p^{(t)})/\log C$ is the entropy-based certainty. The E-I balance regularizer is

$$\mathcal{L}_{\text{EI}} \;=\; \Big[\texttt{clip}\big(\rho_{EI} - \rho_{EI}^*, -50, 50\big)\Big]^2, \qquad \rho_{EI}^* = 4, \tag{A.15}$$

with ρ_{EI} evaluated on the pre-WTA state r_E^{pre}, since WTA sparsifies r_E and would artificially depress the ratio, thereby interfering with the loss magnitude.

The game loss measures the Nash residual of per-neuron dynamics as stated in Appendix C.3 by reducing four Dale-constrained recurrent populations to population-effective scalar weights, which is expressed as

$$\bar{w}_{EE} \;=\; \tfrac{1}{n_E} \sum_{i=1}^{n_E} (W_{EE})_{ii}, \quad \bar{w}_{II} \;=\; \tfrac{1}{n_I} \sum_{i=1}^{n_I} (W_{II})_{ii}, \tag{A.16}$$

$$\bar{w}_{EI} \;=\; \tfrac{1}{n_E n_I} \sum_{i,j} (W_{EI})_{ij}, \quad \bar{w}_{IE} \;=\; \tfrac{1}{n_I n_E} \sum_{i,j} (W_{IE})_{ij}, \tag{A.17}$$

where $\bar{w}_{\cdot\cdot}$ denote the $\texttt{mean}(\mathrm{diag}(W_{\cdot\cdot}))$, which are paired with per-population dissipation constant $d_E, d_I > 0$ inherited from the continuous-time formulation [18]. Thus, the per-population quadratic energies in the consensual regime are given by

$$\mathcal{E}_E \;=\; \frac{\big[\,(\bar{w}_{EE} - d_E)\, r_E \;-\; \bar{w}_{EI}\, \bar{r}_I \;+\; u_E\,\big]^2}{2\,(d_E - \bar{w}_{EE})}, \tag{A.18}$$

$$\mathcal{E}_I \;=\; \frac{\big[\,\bar{w}_{IE}\, \bar{r}_E \;-\; (\bar{w}_{II} + d_I)\, r_I \;+\; u_I\,\big]^2}{2\,(d_I + \bar{w}_{II})}, \tag{A.19}$$

where $\bar{r}_E := \frac{1}{n_E}\sum_i (r_E)_i$ and $\bar{r}_I := \frac{1}{n_I}\sum_i (r_I)_i$ denote the population-mean activations; thus the game loss can be expressed by:

$$\mathcal{L}_{\text{game}} = \tfrac{1}{d_{\text{model}}}\ \min\Big(\Big[\frac{1}{n_E}\sum_{i=1}^{n_E}\mathcal{E}_E^{(i)} \;+\; \frac{1}{n_I}\sum_{j=1}^{n_I}\mathcal{E}_I^{(j)}\Big],\, 100\Big). \tag{A.20}$$

The stability at the regime boundary is enforced by $|d_E - \bar{w}_{EE}| \geq 0.1$, while clipping suppresses transitional spikes during the warm-up phase. Appendix C.3 provides the full stability proof and derivation of Nash equilibria for the recoverability under the consensual regime.

During the training of ImageNet-1K with $d_{\text{model}} = 4096$, and $n_E = 3277$, it has been observed that the full matrix product in (A.20) had a prohibitive computational cost given our limited computational resources. Therefore, we modified $\mathcal{L}_{\text{game}}$ by dropping the per-player Hessian normalization $1/[2\,(d_\bullet \pm \bar{w}_{\bullet\bullet})]$. Thus, the $\mathcal{L}_{\text{game}}$ is given by:

$$\mathcal{L}^{\star}_{\text{game}} \;=\; \min\Big(\frac{\mathbb{E}_{\text{batch}}\big(\|r_E - \mathrm{ReLU}(h_E)\|_2^2 + \|r_I - \mathrm{ReLU}(h_I)\|_2^2\big)}{d_{\text{model}}},\; 100\Big), \tag{A.21}$$

where $r_E \in \mathbb{R}^{n_E}$ and $r_I \in \mathbb{R}^{n_I}$ are the post-activation excitatory and inhibitory population vectors at the last computation time-step, $t = T$, while h_E and h_I denote the matching pre-activations assembled by the Wilson-Cowan integrator given by (A.6)-(A.7). Furthermore, $\mathrm{ReLU}\,(\cdot) = \max(\cdot, 0)$ is calculated component-wise and $\|\cdot\|_2$ denotes the Euclidean norm over the population index, and $\mathbb{E}_{\text{batch}}\,[\cdot]$ is the mean over per-sample residual norms in the current mini-batch.

The synchronization loss is given by:

$$\mathcal{L}_{\text{sync}} \;=\; \frac{1}{d_{\text{sync}}}\,\|z^{(T)}\|_2^2, \tag{A.22}$$

with $z^{(T)} = \texttt{LN}\big[z_{EE}^{(T)}; z_{EI}^{(T)}; z_{II}^{(T)}\big]$ which is composed of the three-type synchronization readouts. The spectral regularizer, derived in Appendix C.4, is expressed as:

$$\mathcal{L}_{\text{spec}} \;=\; \mathrm{ReLU}\big(\widehat{\lambda}_{\text{P}}(W_{EE}) - \tau_{EE}\big)^2 \;+\; \mathrm{ReLU}\big(\widehat{\lambda}_{\text{P}}(W_{II}) - \tau_{II}\big)^2, \tag{A.23}$$

with $(\tau_{EE}, \tau_{II}) = (15, 7)$ and $\widehat{\lambda}_{\text{P}}$ the sum-ratio estimate defined in Appendix C.4. The weights in (A.13) are $(\lambda_{\text{EI}}, \lambda_{\text{game}}, \lambda_{\text{sync}}, \lambda_{\text{spec}}) = (10^{-2}, 10^{-3}, 10^{-4}, 10^{-1})$. The cosine curriculum coefficient is given by:

$$w(\texttt{step}) \;=\; \begin{cases} 0 & \texttt{step} < t_s, \\ \frac{1}{2}\big(1 - \cos\frac{\pi(\texttt{step} - t_s)}{T_w}\big) & t_s \leq \texttt{step} \leq t_s + T_w, \\ 1 & \texttt{step} > t_s + T_w, \end{cases} \tag{A.24}$$

with $(t_s, T_w) = (1K, 5K)$ for `CIFAR-10`, `CIFAR-100`, `MNIST`, and `Fashion-MNIST` datasets, while $(10K, 10K)$ is used for `ImageNet-1K` (cf. Table A.1).

B.2 Curriculum Design

By treating the $w(\texttt{step})$ as a scalar multiplier on a single-batch gradient for each stochastic estimator step, the expected gradient satisfies

$$\mathbb{E}_{\texttt{step}\sim U\{1,\ldots,S\}}\big[\,w(\texttt{step})\,\nabla_\theta \mathcal{L}_{\text{aux}}\big] = \mathbb{E}_{\texttt{step}}\big[w(\texttt{step})\big]\,\nabla_\theta \mathbb{E}\big[\mathcal{L}_{\text{aux}}\big] \quad \because \mathcal{L}_{\text{aux}} \text{ depends on step through } x, y \tag{A.25}$$

$$= \bar{w}\cdot \nabla_\theta \mathbb{E}\big[\mathcal{L}_{\text{aux}}\big], \tag{A.26}$$

where $\bar{w} = S^{-1}\sum_{s=1}^{S} w(s)$ and $\mathcal{L}_{\text{aux}} = \lambda_{\text{EI}}\,\mathcal{L}_{\text{EI}} + \lambda_{\text{game}}\,\mathcal{L}_{\text{game}} + \lambda_{\text{sync}}\,\mathcal{L}_{\text{sync}}$. Analogously to the Rao-Blackwellised ELBO of [54], the per-step estimator obtained by conditioning on step has a strictly smaller variance. Therefore, the variance reduction is exploited by retaining the per-step form but bounding its variance contribution of $w(\texttt{step})$ by $w(\texttt{step})^2 \leq 1$, to form a tight uniform bound.

B.3 Dual Loss Conditioning

The dual cross-entropy loss in its unrolled form is expressed as $L_{\text{task}} = \frac{1}{T}\sum_{t=1}^{T} \texttt{CE}\left(o^{(t)}, y\right)$. Under temporal discretizations of (A.14), the estimator is effectively conditioned on the temporally dependent distribution, where the categorical distribution places $1/2$ weight on each of the two argmin and argmax indices. Thus, the dual loss is a Rao-Blackwellised estimator of the uniform per-time-step average, obtained by conditioning on the per-sample statistic $(t_{\min}, t_{\text{cert}})$, and can be expressed as

$$\mathcal{L}_{\text{task}} = \tfrac{1}{2}\,\texttt{CE}\left(o^{(t_{\min})}, y\right) + \tfrac{1}{2}\,\texttt{CE}\left(o^{(t_{\text{cert}})}, y\right) \tag{A.27}$$

$$= \mathbb{E}_{t\,\sim\,\text{Cat}((\pi+\xi)/2)}\left[\texttt{CE}\left(o^{(t)}, y\right)\right], \quad \because \text{using (A.27)} \tag{A.28}$$

where $\pi_t = \mathbb{1}[t = t_{\min}]$ and $\xi_t = \mathbb{1}[t = t_{\text{cert}}]$. Moreover, this can be validated by checking the certainty curve $c^{(t)}$, which must be monotone non-decreasing, and CE should also be monotone non-increasing in t. Hence, providing a regime in which $t_{\min} = t_{\text{cert}} = T$ (A.28) converges to $\texttt{CE}\left(o^{(T)}, y\right)$.

B.4 Training Algorithm

The overall training pipeline for TIDE is provided below in Algorithm A.1. The forward pass described in Appendix A.2 corresponds to lines 7-13, while the loss calculation in Appendix B.1 is outlined in lines 14-16. The diagnostic framework for NaN-detection and rejection is outlined in lines 17-20, and the backpropagation and optimizations based on Theorem A.3 are provided in lines 21-24.

Algorithm A.1 Training TIDE

Require: image batch sampler, Dale sign mask Σ, internal computation steps $T = 50$

1: Initialize θ, $W_{..} \geq 0$, $m \in \mathbb{R}^{d_{\text{memory}}}$ ▷ Memory buffer
2: **repeat**
3: $(x, y) \sim q(x, y)$ ▷ Sample a mini-batch
4: $r_E^{(0)} \leftarrow \mathbf{x}_E^{\text{init}},\ r_I^{(0)} \leftarrow \mathbf{x}_I^{\text{init}}$ ▷ Learnable init
5: $u_E^{(0)} \leftarrow \mathbf{s}_E,\ u_I^{(0)} \leftarrow \mathbf{s}_I$ ▷ FIFO traces
6: $K, V \leftarrow \text{Backbone}(x)$
7: **for** $t = 1, \ldots, T$ **do**
8: $z^{(t)} \leftarrow \text{Sync}\left(r_E^{(t-1)}, r_I^{(t-1)}; \alpha, \beta\right)$ ▷ Three-type sync readout
9: $a^{(t)} \leftarrow \text{CrossAttn}\left(z^{(t)}, K, V\right)$
10: $h_{\bullet}^{(t)} \leftarrow$ (A.6)–(A.7) using $a^{(t)}$
11: $r_{\bullet}^{(t)} \leftarrow$ (A.4)–(A.5) with NLM-corrected pre-acts
12: $r_E^{(t)} \leftarrow \text{LateralInhibition}\left(r_E^{(t)}\right)$ ▷ $K_{\text{WTA}} = 5$
13: **end for**
14: $\mathcal{L}_{\text{task}} \leftarrow$ (A.14); $\mathcal{L}_{\text{EI}}, \mathcal{L}_{\text{game}}, \mathcal{L}_{\text{sync}} \leftarrow$ (A.15)–(A.22)
15: $\widehat{\lambda}_{\text{P}}(W_{EE}), \widehat{\lambda}_{\text{P}}(W_{II}) \leftarrow$ PERRONSUMRATIO $(W_{..}, K = 10)$ ▷ Algorithm A.11
16: $\mathcal{L} \leftarrow$ (A.13)
17: **if** $\neg\,\text{isfinite}(\mathcal{L})$ **then**
18: $\mathcal{L}' \leftarrow 0 \cdot \mathbf{1}^{\top}\,\sigma\left(o^{(T)}\right)$ ▷ σ replaces NaN / $\pm\infty$ with 0
19: all_reduce (1.0, op = MAX)
20: **else**
21: $g \leftarrow \nabla_{\theta}\mathcal{L}$; clip g (per-component on `ImageNet-1K`)
22: $\theta \leftarrow \text{AdamW}(\theta, g)$
23: **end if**
24: $W_{..} \leftarrow \Pi_{\text{Dale}}(W_{..})$ ▷ Proposition A.2; clip ≥ 0
25: **until** converged

C Continuous-time Limit, Stability, & Game-theoretic Fixed Points

This section is organized as follows. Initially, the continuous-time E-I system and its Ordinary Differential Equation (ODE) formulation are derived (Appendix C.1) as $\Delta t \to 0$ limit of (2). We then linearize around fixed points and give eigenvalue conditions for stability (Appendix C.2). Afterward, we prove LDS implies global asymptotic stability of the game-theoretic Nash equilibrium (Appendix C.3) and derive the differentiable spectral surrogate used in (A.23) (Appendix C.4). Finally, we compare the provided formulation with the Wilson-Cowan and Brunel-Hakim limits (Appendix C.5).

C.1 Continuous-time E-I ODE

Let $r(t) = [r_E(t)^{\top}, r_I(t)^{\top}]^{\top} \in \mathbb{R}^n_{\geq 0}$ and define $\tau = \text{diag}(\tau_E \mathbb{1}_{n_E}, \tau_I \mathbb{1}_{n_I})$. Equation (2) can be rewritten as $r^{(t)} - r^{(t-1)} = \Delta t\,\tau^{-1}\left(-r^{(t-1)} + \varphi\left(W_{\text{eff}} r^{(t-1)} + b\right)\right)$; taking $\Delta t \to 0$ gives

$$\tau\,\dot{r} \;=\; -r \;+\; \varphi\big(W_{\text{eff}}\, r + b\big), \tag{A.29}$$

with $b=[W_E^{\text{in}}a;\, W_I^{\text{in}}a]$. The Jacobian of the right-hand side of (A.29) at a fixed point r^*, where the ReLU is active, is

$$J = -\tau^{-1}\big(I - W_{\text{eff}}\big). \tag{A.30}$$

Lemma C.1 (Convergence of the forward Euler to (A.29)). Let $r^{(t)}$ be the solution of (2) and $r^{\text{cont}}(t)$ the solution of (A.29), both with the same initial condition. If φ is L_φ-Lipschitz and $\dot{r}^{\text{cont}}$ is bounded on $[0, T\Delta t]$, then for all $k \le \lfloor T/\Delta t \rfloor$, $\|r^{(k)} - r^{\text{cont}}(k\Delta t)\|_2 \le C\Delta t \cdot e^{Lk\Delta t}$ for constants C, L depending only on $L_\varphi, \|W_{\text{eff}}\|_2, \tau_E, \tau_I$.

Proof. Taylor expansion at $t = k\Delta t$ gives $r^{\text{cont}}((k+1)\Delta t) - r^{\text{cont}}(k\Delta t) = \Delta t\, \tau^{-1}(-r^{\text{cont}} + \varphi(\cdot))\big|_{k\Delta t} + \mathcal{O}(\Delta t^2)$; subtracting the discrete update (2) and using Lipschitz continuity of φ gives $\|e^{(k+1)}\|_2 \le (1 + L\Delta t)\,\|e^{(k)}\|_2 + C_0\Delta t^2$ with $L = L_\varphi\|W_{\text{eff}}\|_2 \max(1/\tau_E, 1/\tau_I)$. By discrete Gronwall, $\|e^{(k)}\|_2 \le \frac{e^{Lk\Delta t}-1}{L} C_0\Delta t = \mathcal{O}(\Delta t)$. □

C.2 Linear Stability Around a Fixed Point

Let r^* be a fixed point of (A.29). Assume φ is differentiable at $h^* = W_{\text{eff}}r^* + b$, for instance, ReLU is almost-everywhere differentiable and is differentiable at strictly interior h^*. Given $D^* = \text{diag}(\varphi'(h^*))$, the linearized dynamics obeys $\dot{\delta r} = J^*\delta r$ with

$$J^* = \tau^{-1}\big(-I + D^*W_{\text{eff}}\big). \tag{A.31}$$

For the discrete-time Euler iteration $r^{(t+1)} = r^{(t)} + \Delta t\, \tau^{-1}(-r + \varphi(\cdot))$, the corresponding tangent map is $M^* = I + \Delta t\, J^*$.

Theorem C.2 (Schur stability and the step-size bound). The discrete-time fixed point r^* is asymptotically stable iff $|\lambda_i(M^*)| < 1$ for every i. If J^* has spectrum $\{\mu_i\}_{i=1}^n$ with $\text{Re}(\mu_i) < 0$ for all i, indicating continuous-time stability, then

$$\Delta t \;<\; \min_i \frac{2\,\text{Re}(-\mu_i)}{|\mu_i|^2} \tag{A.32}$$

is necessary and sufficient for Schur stability of M^*.

Proof. Eigenvalues of M^* are $\{1+\Delta t\,\mu_i\}$, and $|1+\Delta t\,\mu_i| < 1$ iff $\Delta t\,\mu_i \in \{z \in \mathbb{C} : |1+z| < 1\}$; for μ_i with $\text{Re}(\mu_i) < 0$, and expanding $|1+\Delta t\,\mu_i|^2 = 1 + 2\Delta t\,\text{Re}(\mu_i) + \Delta t^2|\mu_i|^2 < 1$ reduces to $\Delta t < 2\text{Re}(-\mu_i)/|\mu_i|^2$. Requiring this for every i and taking the most restrictive, i.e., smallest bound, gives (A.32). □

Furthermore, for $\tau_E = 20\,\text{ms}$, $\tau_I = 5\,\text{ms}$, an L_φ-Lipschitz activation φ, and $\|W_{\text{eff}}\|_2 \le c$, every eigenvalue of the continuous-time Jacobian J^* satisfies $|\mu_i| \le (1 + L_\varphi \cdot c)/\min(\tau_E, \tau_I) = (1 + L_\varphi \cdot c)/\tau_I = (1 + L_\varphi \cdot c)/5$, where the binding scale is τ_I since $\tau_I < \tau_E$, and $L_\varphi = 1$ for ReLU. In the real-spectrum regime $\text{Re}(-\mu_i) = |\mu_i|$, thus the loose Schur bound (A.32) admits $\Delta t \le 10/(1 + c)$. Given $\Delta t = 1$ is used in the forward pass, this constraint is satisfied whenever $c \le 9$. The bound on c aligns with the isolated-block Perron bound $\widehat{\lambda}_{\text{P}}(W_{II}) \le 9$ derived in Appendix C.4, motivating the spectral target $\tau_{II} = 7$ in (A.23). The differentiable Perron estimate $\widehat{\lambda}_{\text{P}}(W_{II})$ is constrained rather than $\|W_{II}\|_2$ because the former is differentiable and tractable to penalize, even though the latter can be considered a tighter constraint. The analogous isolated-block Schur bound for the excitatory neurons is bounded as $\widehat{\lambda}_{\text{P}}(W_{EE}) < 1$ (A.40). Note that the spectral target $\tau_{EE} = 15$ is well above the defined bound, and a LDS analysis of such a system is provided in Appendix C.3, which establishes that W_{EE}'s isolated instability is stabilized by the inhibitory feedback through the cross-population blocks.

C.3 Lyapunov Diagonal Stability & Game-theoretic Nash Equilibria

It has been shown that asymmetric firing-rate networks admit a per-neuron energy interpretation [18] that replaces the classical scalar Lyapunov function of [31] with a family of energies $\{E_i\}_{i=1}^n$, one per neuron. For the Wilson-Cowan rate system with ReLU denoted by φ, these energies take the form

$$E_i(x, u_i) \;=\; -x_i \sum_{j=1}^{n} \Big(1 - \tfrac{1}{2}\delta_{ij}\Big) W_{\text{eff},ij} x_j \;-\; x_i\, u_i \;+\; \int_0^{x_i} \varphi^{-1}(s)\,\mathrm{d}s, \tag{A.33}$$

whose partial derivatives satisfy $\partial_{x_i} E_i = -(W_{\text{eff}}\, x)_i - u_i + \varphi^{-1}(x_i)$. Setting $\partial_{x_i} E_i = 0$ gives $x_i = \varphi\big((W_{\text{eff}}x)_i + u_i\big)$, which is precisely the fixed-point condition for (A.29). Consequently, the stationary points of (A.29) are the *Nash equilibria* of the non-cooperative game with per-player payoffs $-E_i(x, u_i)$.

Definition C.3 (LDS). A matrix $M \in \mathbb{R}^{n\times n}$ is Lyapunov diagonally stable (LDS) iff there exists a positive diagonal $D \succ 0$ such that $DM + M^\top D \prec 0$. A sufficient, easily checkable condition is $\lambda_{\max}\big((M + M^\top)/2\big) < 0$, which corresponds to $D = I$.

Theorem C.4 (LDS $\Rightarrow$ global asymptotic stability). Suppose $W_{\text{eff}} - I$ is LDS with Lyapunov diagonal $D \succ 0$. Then the linearized gradient-play dynamics $\dot{x} = W_{\text{eff}} x - x + b$ admits a unique Nash equilibrium $x^* = (I - W_{\text{eff}})^{-1} b$ and every trajectory converges to x^*.

Proof. Existence and uniqueness: LDS implies every eigenvalue of $W_{\text{eff}} - I$ has a negative real part, so $I - W_{\text{eff}}$ is invertible. Global convergence: define the weighted quadratic

$$V(x) = \tfrac{1}{2} (x - x^*)^\top D (x - x^*) \geq 0, \qquad V(x) = 0 \Leftrightarrow x = x^*. \tag{A.34}$$

Then

$$\begin{aligned}
\dot{V} &= (x - x^*)^\top D \, \dot{x} = (x - x^*)^\top D \left(W_{\text{eff}} x - x + b \right) && \text{(A.35)} \\
&= (x - x^*)^\top D (W_{\text{eff}} - I)(x - x^*) + (x - x^*)^\top D \big((W_{\text{eff}} - I) x^* + b \big) && \text{(A.36)} \\
&= (x - x^*)^\top D (W_{\text{eff}} - I)(x - x^*) \quad \because \text{since } (W_{\text{eff}} - I) x^* + b = 0 && \text{(A.37)} \\
&= \tfrac{1}{2} (x - x^*)^\top \big(D (W_{\text{eff}} - I) + (W_{\text{eff}} - I)^\top D \big) (x - x^*) < 0 \quad \text{for } x \neq x^*. && \text{(A.38)}
\end{aligned}$$

By LaSalle's invariance principle [55], every trajectory converges to the largest invariant set inside $\{\dot{V} = 0\}$, which by (A.38) is the singleton $\{x^*\}$. □

Corollary C.5 (Practical LDS test). If $\lambda_{\max}\big(\frac{1}{2}(W_{\text{eff}} + W_{\text{eff}}^\top) - I\big) < 0$, Theorem C.4 holds with $D = I$. For E-I Dale-parameterized W_{eff} given by (A.11), the condition is equivalent to $\lambda_{\max}\big(W_{\text{eff}}^{\text{sym}}\big) < 1$ where $W_{\text{eff}}^{\text{sym}} = \frac{1}{2}(W_{\text{eff}} + W_{\text{eff}}^\top)$.

Remark C.6 (LDS is monitored, not gradient-carrying). The eigen-decomposition of $W_{\text{eff}}^{\text{sym}}$ is non-differentiable at eigenvalue crossings, so it is not used as an optimization signal during backpropagation through $\lambda_{\max}(\cdot)$ for training TIDE; Corollary C.5 is evaluated every 100 optimizer steps as a monitoring signal, while the differentiable surrogate (A.23) is used as part of optimization signaling and carries gradients. Prior work [13, 28] similarly treats matrix stability as a diagnostic rather than a regularizer.

C.4 Spectral Perron Surrogate

A Dale matrix $W \geq 0$ admits a real, non-negative dominant eigenvalue $\lambda_{\text{P}}(W)$ by the Perron–Frobenius theorem [56]. We use the sum-ratio power iteration to estimate it:

$$v_0 = \mathbb{1}/n, \qquad v_{k+1} = \frac{W v_k}{\|W v_k\|_2}, \qquad \widehat{\lambda}_{\text{P}}(W) = \frac{\mathbb{1}^\top W v_K}{\mathbb{1}^\top v_K}. \tag{A.39}$$

The estimator $\widehat{\lambda}_{\text{P}}$ is differentiable in W whenever v_K does not lie on the null space of $\mathbb{1}$.

Proposition C.7 (Convergence of $\widehat{\lambda}_{\text{P}}$). Let $W \geq 0$ and let $\lambda_{\text{P}}(W)$ denote its Perron eigenvalue. i*)* If W is irreducible with a strictly dominant real eigenvalue $\lambda_{\text{P}}(W) > |\lambda_i(W)|$ for all $i \geq 2$ and associated positive Perron eigenvector $v_* \in \mathbb{R}^n_{>0}$, then $v_K \to v_* / \|v_*\|_2$ and $\widehat{\lambda}_{\text{P}}(W) \to \lambda_{\text{P}}(W)$ as $K \to \infty$. *ii*) If $W \geq 0$ is reducible, the same conclusion holds under the generic perturbation $W_\epsilon = W + \epsilon \mathbb{1}\mathbb{1}^\top$ for a rank-one non-negative perturbation, hence $W_\epsilon > 0$ and irreducible and after taking $\epsilon \to 0^+$; the limit estimator $\widehat{\lambda}_{\text{P}}(W_\epsilon) \to \widehat{\lambda}_{\text{P}}(W)$ by continuity of eigenvalues in matrix entries. In both cases, the limiting value satisfies $\lambda_{\text{P}}(W) \leq \|W\|_2$ with equality iff W is normal.

Proof. Case (i) *Irreducible* W. The sequence v_K is the normalized power iteration applied to W; under a strictly dominant real eigenvalue, $v_K \to v_* / \|v_*\|_2$ at geometric rate $|\lambda_2 / \lambda_{\text{P}}|^K$ [56, Theorem 8.5.1]. Since $W \geq 0$ is irreducible and $\lambda_{\text{P}}(W)$ is the unique eigenvalue of maximum modulus, W, is primitive [56, Definition 8.5.0]. Hence, by the Perron-Frobenius limit theorem for primitive matrices [56, Theorem 8.5.1], if the initial vector has a nonzero component in the Perron direction, $v_0 > 0$, the normalized power iterates should satisfy $v_K = W^K v_0 / \|W^K v_0\|_2 \longrightarrow v_* / \|v_*\|_2$. Moreover, Perron-Frobenius for irreducible nonnegative matrices gives $v_* > 0$ component-wise [56, Theorem 8.4.4]. Substituting and using $W v_* = \lambda_{\text{P}} v_*$:

$$\begin{aligned}
\widehat{\lambda}_{\text{P}}(W) &= \frac{\mathbb{1}^\top W v_K}{\mathbb{1}^\top v_K} && \because \text{definition (A.39)} \\
&\longrightarrow \frac{\mathbb{1}^\top W v_*}{\mathbb{1}^\top v_*} = \frac{\mathbb{1}^\top (\lambda_{\text{P}} v_*)}{\mathbb{1}^\top v_*} = \lambda_{\text{P}} && \because v_* > 0 \text{ implies } \mathbb{1}^\top v_* > 0.
\end{aligned}$$

Case (ii) *Reducible* W. For any $\epsilon > 0$, the perturbed matrix $W_\epsilon = W + \epsilon \mathbb{1}\mathbb{1}^\top$ has strictly positive entries and is therefore irreducible. Case *(i)* gives $\widehat{\lambda}_{\text{P}}(W_\epsilon) \to \lambda_{\text{P}}(W_\epsilon)$ as $K \to \infty$. The Perron eigenvalue $\lambda_{\text{P}}(\cdot)$ is continuous in matrix entries by [56, Appendix D], hence $\lim_{\epsilon \to 0^+} \lambda_{\text{P}}(W_\epsilon) = \lambda_{\text{P}}(W)$. Choosing any diagonal sequence $(\epsilon_k, K_k) \to (0^+, \infty)$ and applying Case *(i)* along the sequence W_{ϵ_k} at iteration K_k gives

$\widehat{\lambda}_{\mathrm{P}}(W_{\epsilon_k}; K_k) \to \lambda_{\mathrm{P}}(W)$ by a standard Moore-Osgood argument (uniform convergence of power-iteration on the compact set $\{W_\epsilon : \epsilon \in [0, \epsilon_0]\}$).

Operator-norm bound. $\lambda_{\mathrm{P}}(W)$ is an eigenvalue of W, hence $|\lambda_{\mathrm{P}}(W)| \leq \rho(W) \leq \sigma_1(W) = \|W\|_2$, where $\rho(W)$ is the spectral radius; the second inequality is standard with equality iff W is normal [56, Theorem 5.6.9]. For $W \geq 0$, the Perron-Frobenius theorem gives $\lambda_{\mathrm{P}}(W) = \rho(W)$. □

Remark C.8 (Practical relevance of Case (*ii*)). The Dale projection Π_{Dale} clamps entries of $W_{..}$ to $\mathbb{R}_{\geq 0}$ and can therefore produce sparse matrices with some zero rows or columns, i.e., reducible. Case *(ii)* ensures that Proposition C.7 remains valid at such configurations. We apply the estimator to the raw $W_{..}$ without any perturbation and report $\widehat{\lambda}_{\mathrm{P}}$ as-is, relying on Case *(ii)* as a theoretical guarantee rather than a runtime preprocessor.

For W_{EE}, which is symmetric and positive, the operator norm $\|W\|_2 = \lambda_{\mathrm{P}}(W)$ coincides with the Perron eigenvalue; the sum-ratio estimate can then be considered tight. For asymmetric or non-symmetric non-negative W, $\widehat{\lambda}_{\mathrm{P}}(W) \leq \|W\|_2$ can be substantially smaller, which is the regime in which the sum-ratio surrogate can be considered a tighter regularizer than spectral-norm regularization. Empirically, we observe $\|W_{EE}\|_2 / \widehat{\lambda}_{\mathrm{P}}(W_{EE}) \approx 2.5$ at convergence on `ImageNet-1K`, consistent with the analytical gap in Proposition C.7.

Isolated-block Schur analysis: We now derive the correct Schur bounds for the isolated excitatory and inhibitory recurrences, and clarify how the TIDE targets $(\tau_{EE}, \tau_{II}) = (15, 7)$ relate to those bounds.

Consider first the excitatory population in the absence of inhibitory drive. The recurrence (A.4) linearized at $r_E^* = 0$ gives $r_E^{(t+1)} = M_E\, r_E^{(t)}$ with $M_E = (1 - \alpha_E) I + \alpha_E W_{EE}$. Its eigenvalues are $\mu_i = (1 - \alpha_E) + \alpha_E \lambda_i(W_{EE})$. For the Perron eigenvalue $\lambda_{\mathrm{P}}(W_{EE}) \geq 0$:

$$\begin{aligned} |\mu_i| < 1 &\iff -1 < (1 - \alpha_E) + \alpha_E \lambda_{\mathrm{P}}(W_{EE}) < 1 \\ &\iff \lambda_{\mathrm{P}}(W_{EE}) < 1 \ \text{ and } \ \lambda_{\mathrm{P}}(W_{EE}) > 1 - 2/\alpha_E. \end{aligned} \tag{A.40}$$

For $\alpha_E = 0.05$, the lower bound is define as $\lambda_{\mathrm{P}} > -39$, so the binding condition is $\lambda_{\mathrm{P}}(W_{EE}) < 1$ given that the isolated-E Schur bound is 1, not $1/\alpha_E$, which indicates that the step-size α_E does not play any role.

Now consider the inhibitory population in the absence of excitatory drive. By linearizing (2) at $r_I^* = 0$, $r_I^{(t+1)} = M_I\, r_I^{(t)}$ is given with $M_I = (1 - \alpha_I) I - \alpha_I W_{II}$. Its eigenvalues are $\mu_j = (1 - \alpha_I) - \alpha_I \lambda_j(W_{II})$. For the Perron eigenvalue $\lambda_{\mathrm{P}}(W_{II}) \geq 0$:

$$\begin{aligned} |\mu_j| < 1 &\iff -1 < (1 - \alpha_I) - \alpha_I \lambda_{\mathrm{P}}(W_{II}) < 1 \\ &\iff \lambda_{\mathrm{P}}(W_{II}) < 2/\alpha_I - 1 \ \text{ and } \ \lambda_{\mathrm{P}}(W_{II}) > -1. \end{aligned} \tag{A.41}$$

At $\alpha_I = 0.20$ the binding condition is $\lambda_{\mathrm{P}}(W_{II}) < 9$. Unlike the excitatory case, the minus sign in M_I makes the upper bound on μ_j trivial and the lower bound $\mu_j > -1$ binding; thus, α_I enters the threshold through this inequality.

Spectral targets for TIDE: Notice that the asymmetry between (A.40) and (A.41) is vital. The target $\tau_{II} = 7$ sits below the isolated-I Schur threshold 9 and can reasonably be considered as an upper safe boundary enabling variation during the optimization. The target $\tau_{EE} = 15$, however, is $15\times$ above the isolated-E Schur threshold of 1. The TIDE runtime indeed operates above the isolated-E threshold on `ImageNet-1K`, as shown in Appendix G.7, and remains stable due to the fully coupled E-I system that has an effective recurrent matrix

$$W_{\text{eff}} = \begin{bmatrix} W_{EE} & -W_{EI} \\ W_{IE} & -W_{II} \end{bmatrix}, \tag{A.42}$$

whose spectrum is not the union of the spectra of W_{EE} and W_{II} but depends on all four blocks. The inhibitory feedback adds $-W_{EI} r_I$ to the excitatory drive at every internal computation step; thus, minimizing the effective excitatory gain via the feedback results in a Schur-stable system even when W_{EE} alone can not be considered stable. This is the classical balanced-network phenomenon characterized by [26] and [27].

We therefore can state the following:

- ***i)*** The *isolated-I bound* $\lambda_{\mathrm{P}}(W_{II}) < 9$ of (A.41) is a *sufficient* condition for the stability of the inhibitory sub-dynamics, independent of the excitatory state. The TIDE target $\tau_{II} = 7$ is a safety margin below this bound.
- ***ii)*** The *isolated-E bound* $\lambda_{\mathrm{P}}(W_{EE}) < 1$ of (A.40) is a *sufficient* condition in absence of inhibition, however, TIDE does not operate in this regime. The target $\tau_{EE} = 15$ can be considered a limiter on the non-negative recurrent block, tuned empirically to keep gradient norms bounded, and the coupled system stable under LDS.

C.5 Relationship to Wilson-Cowan & Brunel-Hakim Limits

TIDE's continuous-time system (A.29) can be formulated as the Wilson-Cowan mean-field rate equation of [24, 25]; the Dale parameterization of (A.11) pins down the sign structure that [27] and [57] identify as necessary for the asynchronous-irregular regime. Specifically, [27] derives four regimes as a function of the excitation–inhibition balance parameter $g = \|W_{EI}\|/\|W_{EE}\|$; TIDE operates in the asynchronous-irregular regime in which $g > 1$ and the balance ratio ρ_{EI} of (A.1) is close to $\rho^*_{EI} = 4$.

D HRF Backbone

This section provides a detailed description of TIDE backbone, which uses learnable center-surround filters rather than a fixed-form parametric DoG kernel. Note that this study utilizes two variants of the HRF: *i*) a shallow HRF where a multi-scale filter bank is used, and *ii*) a deep HRF with a single center-surround per-stage filter of kernel size 5. Appendix D.1 provides the overall mapping of stage zero and defines the formulation for the shared center-surround filter bank. Moreover, the translational and rotational equivariance properties are proven under a restricted DoG initialization in Appendix D.2.

D.1 Learnable Filter Bank

Let $C^{(s)},\ S^{(s)}$ denote learnable 2-D convolution filters with kernel sizes $k_c^{(s)}$ denoting center and $k_s^{(s)}$ denoting surround, with $k_s^{(s)} > k_c^{(s)}$. Therefore, the center kernel can be defined as either a 1×1 or 3×3 kernel, and the surround is given by a kernel with a one- to three-pixel wider shape, with zero-padding to preserve spatial dimensions, and $s \in \{1, 2, 4, 8\}$ denotes the indexed operator scale. Hence, the center-surround operator for stage zero at scale s is given by:

$$\phi^{(s)}(x) \;=\; \mathrm{ReLU}\Big(\mathrm{BN}\big(w_c^{(s)}\, C^{(s)}(x) \;-\; w_s^{(s)}\, S^{(s)}(x)\big)\Big), \tag{A.43}$$

where $w_c^{(s)}, w_s^{(s)} \in \mathbb{R}$ are learnable scalar weights initialized to $(1.0, 0.5)$, and BN denotes the 2-D BatchNorm. The center and surround convolution of (A.43) have independent learnable weights, while the center-surround ratio is factorized into a pair of scalar gains, $(w_c^{(s)}, w_s^{(s)})$, instead of a parametric DoG kernel with its optimizable width, κ. This design choice trades the explicit DoG scale-space interpretation for expressivity, with the biologically motivated motif based on ON/OFF-center retinal ganglion cells [58], LGN relay cells [59], and cortical simple cells [60]. In contrast to the shallow HRF, the deep variant does not use the multi-scale bank directly, as its layout is based on ResNet with four hierarchical residual stages with channel widths $(128, 256, 512, 2048)$ and $(2, 2, 3, 3)$ basic residual blocks per stage.

Remark D.1 (DoG as a special case). The fixed-form DoG kernel $\mathrm{DoG}_{s,\kappa}(x, y) = G_s(x, y) - \kappa\, G_{\kappa s}(x, y)$ with $G_\sigma(x, y) = (2\pi\sigma^2)^{-1} \exp\big(-(x^2 + y^2)/(2\sigma^2)\big)$ can be approximately recovered from (A.43) by freezing $C^{(s)}, S^{(s)}$ to the discretized Gaussians $G_s, G_{\kappa s}$ and tying $(w_c^{(s)},\ w_s^{(s)}) = (1,\ \kappa)$. Under this restriction, $\phi^{(s)}$ is reduced to a classical DoG operator, which is half-wave-rectified and batch-normalized, similar to the edge-detection framework of [45] and the scale-space theory of [61]; the latter also gives the differentiability of G_σ in σ and therefore of $\mathrm{DoG}_{s,\kappa}$ in $(s,\ \kappa)$. Due to the direct optimization of the convolution weights, in place of the (s, κ) parameters, the regularity is not used in our implementation.

D.2 Equivariance Properties

Let T_v denote a planar translation $T_v f(x, y) = f(x - v_x, y - v_y)$ and R_θ the in-plane rotation by angle θ. Let $\mathrm{conv}_\mathcal{K}(f)$ denote convolution of f with kernel $\mathcal{K}$ on $\mathbb{R}^2$.

Proposition D.2 (Translation equivariance). For any discretely-supported kernel $\mathcal{K}$ and translation T_v by an integer-pixel vector v, $\mathrm{conv}_\mathcal{K}(T_v f) = T_v\, \mathrm{conv}_\mathcal{K}(f)$, while far from the boundaries to prevent interaction between zero-padding and T_v. In particular, the center-surround bank (A.43) is translation-equivariant, since each of $C^{(s)}, S^{(s)}$ is a standard convolution, BN is a channel-wise operator and translation-equivariant on spatial dimensions, and ReLU is component-wise. Thus, both the shallow and deep HRF are inherently translation equivariant under the same proviso.

Proposition D.3 (Rotation equivariance under DoG). Consider (A.43) restricted to the DoG kernels of Remark D.1: $C^{(s)} = G_s,\ S^{(s)} = G_{\kappa s}$ with isotropic Gaussian kernels, and batch-normalization replaced by a channel-wise affine. Then, in the continuous domain and prior to spatial discretizations, $\phi^{(s)}(R_\theta f) = R_\theta\, \phi^{(s)}(f)$ for every continuous rotation R_θ.

Proof. Proposition D.2 is the standard translation equivariance of planar convolution. The ReLU and channel-wise affine preserve this by component-wise and channel-wise action, respectively. Therefore, for Proposition D.3, the isotropic Gaussian is rotation-invariant as a continuous function of (x, y): $G_\sigma(R_\theta(x, y)) =$

$G_\sigma(x, y)$. Hence $G_s - \kappa\, G_{\kappa s}$ is a rotation-invariant kernel. Convolution with a rotation-invariant kernel commutes with rotation of the continuous input, and component-wise operations commute with rotation trivially. Therefore, the exact equivariance holds only for $\theta \in \{0,\ \pi/2,\ \pi,\ 3\pi/2\}$. □

Remark D.4 (Loss of rotation equivariance in later layers). Apart from Stage zero center-surround filters, the learnable convolutions of both backbones, the per-scale expansion convolution in the shallow HRF, and the residual blocks of deep HRF are randomly initialized, have anisotropic kernels, and are not rotation-equivariant. Thus, TIDE inherits Stage zero rotation equivariance under DoG but not global rotation equivariance; this matches the biological observation that simple cells are orientation-selective, and that orientation equivariance is broken in the extrastriate cortex.

E TIDE Components

This section presents detailed algorithms for individual components of TIDE, describes their functionalities, and clarifies any remaining details.

E.1 Deep-HRF Backbone

The deep HRF backbone provided in the Algorithm A.2 is solely used as a feature extractor for `ImageNet-1K`. Four receptive field modules with channel widths $(128, 256, 512, 2048)$ and an additional basic residual block are used, resulting in a final $14 \times 14 \times 2048$ feature map. Moreover, stages one to three use a single fixed-kernel $k_s = 5$, which is applied as a residual $x \mapsto x + \phi(x)$ at the stage entry. The final outputs are passed through `AdaptiveAvgPool2d(14)` and flattened to provide $P = 196$ tokens, with the additional 2-D sinusoidal positional encoding, resulting in $d_{\text{attn}} = 1024$.

Algorithm A.2 BACKBONE: Deep-HRF

Require: Image batch $x \in \mathbb{R}^{B\times 3\times 224\times 224}$; a learnable 7×7 ResNet-style stem; four receptive blocks $\mathcal{G}_1, \ldots, \mathcal{G}_4$. Key and value projections $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{feat}}\times d_{attn}}$.
Ensure: Keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times P\times d_{attn}}$ with $P = 196$ and $d_{attn} = 1024$.
$y \leftarrow \mathrm{MaxPool}\left(\mathrm{RELU}\left(\mathrm{BN}\left(\mathrm{Conv}_{7\times 7,\ \mathrm{stride}2}(x)\right)\right)\right)$ ▷ ResNet-style stem: $56\times 56\times 64$
for $\ell \leftarrow 1$ **to** 4 **do**
  $y \leftarrow \mathcal{G}_\ell(y)$
end for
$y \leftarrow \mathrm{Flatten}_{\text{spatial}}(\mathrm{AdaptivePool}_{14\times 14}(y)) + \mathrm{PE}_{\text{2D}}$ ▷ $(B, 14\times 14, d_{\text{feat}}) = (B, 196, d_{\text{feat}})$
$\mathbf{K} \leftarrow y\,\mathbf{W}_K, \quad \mathbf{V} \leftarrow y\,\mathbf{W}_V$
return $(\mathbf{K}, \mathbf{V})$

E.2 Shallow HRF Backbone

The shallow HRF backbone is provided in Algorithm A.3, where a two-layer 3×3 $\mathrm{Conv}-\mathrm{BN}-\mathrm{ReLU}$ stem denoted by H_{stem} expands the input to feature channels. The multi-scale center-surround uses four parallel branches at scales $s \in \{1, 2, 4, 8\}$ with surround kernel sizes $k_s = 2s + 1 \in \{3, 5, 9, 17\}$. The four branches are `AdaptiveAvgPool2d(8)`-pooled and concatenated across channels. A subsequent aggregation block $\mathrm{Conv}_1 \circ \mathrm{BN} \circ \mathrm{ReLU} \circ \mathrm{Conv}_3 \circ \mathrm{BN} \circ \mathrm{ReLU}$ denoted by H_{agg} mixes the multi-scale features, followed by another `AdaptiveAvgPool2d(8)`, and a 2-D positional encoding to produce $P = 64$ tokens.

Algorithm A.3 BACKBONE: Shallow HRF

Require: Image batch $x \in \mathbb{R}^{B\times C\times H\times W}$ with $(C, H, W) \in \{(1, 28, 28), (3, 32, 32)\}$ and key/value projections $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{feat}}\times d_{attn}}$.
Ensure: $(\mathbf{K}, \mathbf{V})$ with $P = 64$ (8×8 grid) and $d_{attn} = 512$.
$y \leftarrow \mathcal{H}_{\text{stem}}(x)$
$y \leftarrow \bigoplus_{s\in\mathcal{S}} \mathrm{AdaptivePool}_{8\times 8}\big(\psi^{(s)}(y)\big)$ ▷ Multi-scale filter bank, channel-concatenated
$y \leftarrow \mathcal{H}_{\text{agg}}(y)$
$y \leftarrow \mathrm{Flatten}_{\text{spatial}}(\mathrm{AdaptivePool}_{8\times 8}(y)) + \mathrm{PE}_{\text{2D}}$
$\mathbf{K} \leftarrow y\,\mathbf{W}_K, \quad \mathbf{V} \leftarrow y\,\mathbf{W}_V$
return $(\mathbf{K}, \mathbf{V})$

E.3 Wilson-Cowan E-I Update

Algorithm A.4 presents the Dale-constrained, $\mathrm{RMSNorm}$ stabilized Wilson-Cowan update, which is a discrete analog of (A.2)–(A.3). Given the post-activation state r_E, r_I from the previous internal computation step and

the cross-attention given by a, the pre-activations are formed following (A.6)–(A.7), and then per-population $\mathrm{RMSNorm}$ is applied to yield $h_E,\ h_I$.

Algorithm A.4 Dale-constrained Wilson-Cowan

Require: Previous post-activations $r_E \in \mathbb{R}^{B\times n_E}$, $r_I \in \mathbb{R}^{B\times n_I}$; attention drive $a \in \mathbb{R}^{B\times d_{\mathrm{sync}}}$; Dale-constrained weights $W_{EE}, W_{EI}, W_{IE}, W_{II} \geq 0$; input projections $W_E^{\mathrm{in}}, W_I^{\mathrm{in}}$; per-population $\mathrm{RMSNorm}_E, \mathrm{RMSNorm}_I$.
Ensure: Pre-activations h_E, h_I
$\tilde{h}_E \leftarrow W_{EE}\, r_E \ - \ W_{EI}\, r_I \ + \ W_E^{\mathrm{in}}\, a$ ▷ Inhibition is the **minus** sign, not $W < 0$
$\tilde{h}_I \leftarrow W_{IE}\, r_E \ - \ W_{II}\, r_I \ + \ W_I^{\mathrm{in}}\, a$
$h_E \leftarrow \mathrm{RMSNorm}_E(\tilde{h}_E)$ ▷ Per-population scale
$h_I \leftarrow \mathrm{RMSNorm}_I(\tilde{h}_I)$
return (h_E, h_I)

E.4 Population-specific NLM

Algorithm A.5 presents the NLMs readout of the FIFO with length M for given r_E and r_I, which yields a neuron-specific scalar correction. We use a temporally weighted `SuperLinear` and normalize it given $Z = \sum_{m=1}^{M} \exp(-(M-m)/\tau)$. Moreover, a subsequent `SuperLinear` layer is used to project a dual-channel logit output, which is squeezed post `GLU` gate. Note that for smaller datasets $H=4$, while `ImageNet-1K` requires $H=32$.

Algorithm A.5 Population-specific NLM

Require: Post-activation FIFO $\mathbf{u}_\bullet \in \mathbb{R}^{B\times n_\bullet \times M}$ with $M = 25$; per-neuron super-linear weights $W_1^{(0)} \in \mathbb{R}^{M\times 2H\times n_\bullet}$, $W_1^{(1)} \in \mathbb{R}^{H\times 2\times n_\bullet}$; biases $b_1^{(0)}, b_1^{(1)}$; time constant $\tau_\bullet$ (τ_E=20, τ_I=5).
Ensure: Per-neuron additive correction $n_\bullet \in \mathbb{R}^{B\times n_\bullet}$.
$w_m \leftarrow \exp\big(-(M-m)/\tau_\bullet\big)/Z$ **for** $m = 1, \ldots, M$ ▷ Normalize exponential temporal weights
$\tilde{\mathbf{u}} \leftarrow \mathbf{u}_\bullet \odot w_m$ ▷ Broadcast along the memory axis
$\tilde{\mathbf{u}} \leftarrow \mathrm{LayerNorm}(\tilde{\mathbf{u}})$
$y_0 \leftarrow \texttt{einsum}\big(\texttt{"BNM,MHN->BNH"},\ \tilde{\mathbf{u}},\ W_1^{(0)}\big) + b_1^{(0)}$ ▷ Output $\rightarrow$ BNH
$y_0 \leftarrow y_0/T$ ▷ Per-neuron learnable temperature scaling
$y_0 \leftarrow \mathrm{GLU}(y_0)$ ▷ Halves H to H_{NLM}
$y_1 \leftarrow \texttt{einsum}\big(\texttt{"BNH,HOn->BN2"},\ y_0,\ W_1^{(1)}\big) + b_1^{(1)}$
$y_1 \leftarrow y_1/T$
$y_1 \leftarrow \mathrm{GLU}(y_1)$
$n_\bullet \leftarrow \mathrm{squeeze}(y_1, \mathrm{dim}{=}{-1})$ ▷ $(B, n_\bullet)$
return $n_\bullet$

E.5 Three-type Synchronization

Algorithm A.6 shows the synchronization module that maintains the three exponentially decaying covariance accumulators between neuron pairs. For each accumulator, a per-pair learnable decay $\delta \in [0, 15]$ produces $y = e^{-\delta}$ while pair-wise products $\pi = x_a[:, I_a] \odot x_b[:, I_b]$ update pair-product sum ν and effective-count ξ via $\nu \leftarrow y \cdot \nu + \pi$ and $\xi \leftarrow y \cdot \xi + 1$. Note that the three sub-latent representations corresponding to each of the neuron pairs are concatenated and passed through a final `LN` of width d_{sync} to produce z.

Algorithm A.6 Synchronization Accumulator

Require: Post-activations $x_a \in \mathbb{R}^{B\times n_a}$, $x_b \in \mathbb{R}^{B\times n_b}$; running accumulators $\nu, \xi \in \mathbb{R}^{B\times P_{XY}}$; fixed pair indices $I_a, I_b \in \mathbb{N}^{P_{XY}}$; learnable decay $\delta \in \mathbb{R}^{P_{XY}}$; projection $\mathcal{P}_{XY} : \mathbb{R}^{P_{XY}} \rightarrow \mathbb{R}^{d_{XY}}$ (Linear $\rightarrow 2d_{XY} \rightarrow$ GLU $\rightarrow$ LayerNorm); accumulator clamp C.
Ensure: Sync vector $z_{XY} \in \mathbb{R}^{B\times d_{XY}}$; updated (ν, ξ).
$r \leftarrow \mathrm{clamp}\big(\exp(-\delta),\ e^{-15},\ 1\big)$
$\pi \leftarrow x_a[:, I_a] \odot x_b[:, I_b]$ ▷ (B, P_{XY}) pairwise product
$\nu \leftarrow r \odot \nu + \pi$
$\xi \leftarrow r \odot \xi + \mathbf{1}$
if $C > 0$ **then**
  $\nu \leftarrow \mathrm{clamp}\,(\nu, -C, +C)$
end if
$s \leftarrow \nu \,/\, \sqrt{\xi + \varepsilon}$ ▷ Normalized sync signal
$z_{XY} \leftarrow \mathcal{P}_{XY}(s)$ ▷ (B, d_{XY})
return $(z_{XY},\ \nu,\ \xi)$

E.6 Cross-attention Readout

Algorithm A.7 provides an overview of the cross-attention mechanism that processes the synchronization latent z as a query. The backbone key/value channel is used directly or in cases where the width d_{KV} differs from $d_{\mathrm{attn}} = n_{\mathrm{heads}} \cdot d_{\mathrm{head}}$, separate W_K, W_V projections are used to reshape the keys and values. In the current implementation, the backbone projects to d_{attn} channels upstream, while the standard scaled dot-product attention is applied to the $\mathrm{softmax}$ of the weights. During `ImageNet-1K` training, an additive residual $+z$ is added inside the LN rather than externally to enhance context.

Algorithm A.7 Multi-head Cross-attention

Require: Query $Q \in \mathbb{R}^{B \times d_{\mathrm{sync}}}$; keys and values $K, V \in \mathbb{R}^{B \times P \times d_{\mathrm{attn}}}$; projections $\mathbf{W}_Q \in \mathbb{R}^{d_{\mathrm{sync}} \times d_{\mathrm{attn}}}$, $\mathbf{W}_O \in \mathbb{R}^{d_{\mathrm{attn}} \times d_{\mathrm{sync}}}$; head count H_{attn}, head dim $d_{\mathrm{head}} = d_{\mathrm{attn}}/H_{\mathrm{attn}}$.
Ensure: Attention output $a \in \mathbb{R}^{B \times d_{\mathrm{sync}}}$.
$q \leftarrow Q\,\mathbf{W}_Q$ ▷ (B, d_{attn})
$q \leftarrow \mathrm{reshape}\,(q, B, H_{\mathrm{attn}}, d_{\mathrm{head}})$ ▷ Split heads
$K' \leftarrow \mathrm{reshape}\,(K, B, P, H_{\mathrm{attn}}, d_{\mathrm{head}})$
$V' \leftarrow \mathrm{reshape}\,(V, B, P, H_{\mathrm{attn}}, d_{\mathrm{head}})$
$A \leftarrow \mathrm{softmax}\big(q\,K'^{\top}/\sqrt{d_{\mathrm{head}}}\big)$ ▷ $(B, H_{\mathrm{attn}}, P)$
$A \leftarrow \mathrm{Dropout}_p(A)$ ▷ Dropout on attention probabilities
$\tilde{a} \leftarrow A\,V'$ ▷ $(B, H_{\mathrm{attn}}, d_{\mathrm{head}})$
$a \leftarrow \mathrm{reshape}\,(\tilde{a}, B, d_{\mathrm{attn}})\,\mathbf{W}_O$
$a \leftarrow \mathtt{LN}_{d_{\mathrm{sync}}}\big(a + Q\big)$
return a

E.7 Lateral Inhibition

The lateral inhibition is implemented based on WTA and is used solely on the excitatory population, as shown in Algorithm A.8. We use a post-Euler $r_E^{(0)}$ across $K_{\mathrm{WTA}} = 5$ iterations of an inhibitory feedback loop given by $x_I^{(k)} = \mathrm{ReLU}(W_{EI}^{\mathrm{lat}} r_E^{(k-1)})$, and then $r_E^{(k)} = \mathrm{ReLU}(r_E^{(0)} - \gamma\, W_{IE}^{\mathrm{lat}} x_I^{(k)})$. Furthermore, early termination at various per-computation steps is leveraged to indirectly optimize the number of iterations required during deployment, thus enabling an adaptive halting driven by the variation threshold indicated in line 5.

Algorithm A.8 Lateral Inhibition

Require: Post-Euler excitatory state $r_E \in \mathbb{R}^{B \times n_E}$; Dale-constrained lateral weights $W_{EI}^{\mathrm{lat}} \geq 0 \in \mathbb{R}^{n_{I,\mathrm{lat}} \times n_E}$, $W_{IE}^{\mathrm{lat}} \geq 0 \in \mathbb{R}^{n_E \times n_{I,\mathrm{lat}}}$; learnable gain $\gamma \geq 0.01$; iteration count $K_{\mathrm{WTA}} = 5$.
Ensure: Sparsified excitatory state $r_E^{(K_{\mathrm{WTA}})}$.
$r_E^{(0)} \leftarrow r_E$
for $k \leftarrow 1$ **to** K_{WTA} **do**
  $x_I^{(k)} \leftarrow \mathrm{ReLU}\big(W_{EI}^{\mathrm{lat}}\, r_E^{(k-1)}\big)$
  $r_E^{(k)} \leftarrow \mathrm{ReLU}\big(r_E^{(0)} - \gamma\, W_{IE}^{\mathrm{lat}}\, x_I^{(k)}\big)$ ▷ Anchored E update
  if $\max |r_E^{(k)} - r_E^{(k-1)}| < 10^{-4}$ **then**
    break ▷ Early termination once converged
  end if
end for
return $r_E^{(K_{\mathrm{WTA}})}$

E.8 Persistent Memory with Retention Gate

The memory, along with its implementation, is provided in Algorithm A.9. A single persistent buffer m and a momentum buffer v across all computation steps and batches are maintained, while augmented by a batch-broadcast readout. Moreover, the surprise signal is constructed via an MLP embedding, thereby enabling reconstruction-based analysis of new information, without direct reliance on its gradient. The resulting surprise is defined as the per-sample squared ℓ_2 distance $s = \|\hat{z} - z\|_2^2$ between the reconstruction $\hat{z}$ and the input z.

Algorithm A.9 Persistent Memory

Require: Sync latent $z \in \mathbb{R}^{B \times d_{\text{sync}}}$; pre-WTA excitatory state r_E^{pre}; inhibitory state r_I; persistent buffers $m, v \in \mathbb{R}^{d_{\text{mem}}}$ ($d_{\text{mem}} = 256$); learnable heads $f_{\text{rec}} : \mathbb{R}^{d_{\text{mem}}} \to \mathbb{R}^{d_{\text{sync}}}$, $f_{\text{proj}} : \mathbb{R}^{d_{\text{sync}}} \to \mathbb{R}^{d_{\text{mem}}}$, $f_{\text{read}} : \mathbb{R}^{d_{\text{mem}}+d_{\text{sync}}} \to \mathbb{R}^{d_{\text{mem}}}$; retention sharpness κ; surprise threshold $\theta_s = 0.5$, momentum $\mu = 0.9$, target ratio $\rho^\star = 4$.
Ensure: Memory readout $m \in \mathbb{R}^{B \times d_{\text{mem}}}$.

$\hat{z} \leftarrow f_{\text{rec}}(m)$ ▷ Reconstruct z from persistent state
$s \leftarrow \|\hat{z} - z\|_2^2$ ▷ Per-sample surprise scalar
$\rho \leftarrow \text{mean}\left(|r_E^{\text{pre}}|\right) / \text{mean}\left(|r_I|\right)$ ▷ E-I ratio
$\gamma \leftarrow \sigma\big(-\kappa\,|\rho - \rho^\star|\big)$ ▷ Retention gate
$u \leftarrow \mathbb{1}\left[s > \theta_s\right] \cdot (1 - \gamma)$ ▷ Surprise-gated write signal
$v \leftarrow \mu\, v + u \cdot f_{\text{proj}}(z)$ ▷ Momentum buffer
$m \leftarrow m + v$ ▷ In-place buffer registration
$m \leftarrow f_{\text{read}}\big([\,m;\, z\,]\big)$
return m

E.9 Output Head

The output head consists of a two-layer `GLU`-based MLP applied to the concatenation $[z;\, m^{(t)}]$, as shown in Algorithm A.10. By initially expanding the channels for the `GLU` gate and then halving them back, we create a hidden representation that contains both the memory and the current latent space. Stabilization of the hidden representation is achieved by using LayerNorm_H across a wide range of magnitudes seen across internal computation steps.

Algorithm A.10 OutputHead

Require: Concatenated latent $[z; m] \in \mathbb{R}^{B \times (d_{\text{sync}}+d_{\text{mem}})}$; weights $W_{h_1} \in \mathbb{R}^{(d_{\text{sync}}+d_{\text{mem}}) \times 2H}$, b_{h_1}, $W_{h_2} \in \mathbb{R}^{H \times C}$, b_{h_2}; hidden dim $H = 256$ (CIFAR/`MNIST` default; $H = 2048$ for `ImageNet-1K`); dropout rate p.
Ensure: Class logits $o \in \mathbb{R}^{B \times C}$.

$y_1 \leftarrow [z;\, m]\, W_{h_1} + b_{h_1}$ ▷ $\in \mathbb{R}^{B \times 2H}$
$y_2 \leftarrow \text{GLU}\,(y_1)$ ▷ $\in \mathbb{R}^{B \times H}$; gated linear unit halves the channel dimension
$y_3 \leftarrow \text{LayerNorm}_H(y_2)$
$y_4 \leftarrow \text{Dropout}_p(y_3)$
$o \leftarrow y_4\, W_{h_2} + b_{h_2}$
return o

E.10 Perron Sum-ratio Estimator

The Perron eigenvalue estimator provided in Algorithm A.11 is the differentiable spectral primitive used to regularize W_{EE} and W_{II}. The n_{iter} denoted by K in Appendix C and (5) representing the power-iteration steps is set to $n_{\text{iter}} = 10$ with ℓ_2-normalization produce a vector v_K aligned with the Perron eigenvector. The eigenvalue is then estimated by the sum-ratio $\widehat{\lambda}_{\text{P}} = \mathbb{1}^\top W v_K / \mathbb{1}^\top v_K$, which is the true Perron eigenvalue, and not Single Value Decomposition (SVD) norm $\sigma_{\max}(W)$ that an ℓ_2-norm-ratio $\|Wv\|/\|v\|$ would yield [56]. Furthermore, as shown in line 1, we initialize v_0 from a uniform start, $v_0 = \mathbb{1}/n$, without using a warm start.

Algorithm A.11 PerronSumRatio

Require: Non-negative square matrix $W \in \mathbb{R}_{\geq 0}^{n \times n}$; iteration count n_{iter}.
Ensure: Estimate $\hat{\lambda}_{\text{Perron}}(W)$.

$v_0 \leftarrow \mathbb{1} \,/\, n$
for $k \leftarrow 1$ **to** n_{iter} **do**
  $v_k \leftarrow W\, v_{k-1} \,/\, \|W\, v_{k-1}\|_2$ ▷ Normalized power iteration step
end for
$\hat{\lambda} \leftarrow \dfrac{\mathbb{1}^\top W\, v_{n_{\text{iter}}}}{\mathbb{1}^\top v_{n_{\text{iter}}}}$ ▷ Sum ratio, (A.39)
return $\hat{\lambda}$

F Experimental Details

In this section, we present the preprocessing and the associated datasets used in this paper (Appendix F.1). Appendix F.2 presents the architecture and training hyperparameters used during training. Finally, the compute resources and the low-variance estimators for TIDE are provided in Appendices F.3 and F.4.

Table A.1: Per-dataset training hyperparameters. Empty cells indicate the option is inactive. Shared constants are provided in A.2.

Parameter	`MNIST`	`Fashion-MNIST`	`CIFAR-10`	`CIFAR-100`	`ImageNet-1K`
d_{model}	256	256	512	718	4096
(n_E, n_I)	(205,51)	(205,51)	(410,102)	(574,144)	(3277,819)
Backbone	HRF	HRF	HRF	HRF	Deep-HRF
Positions P	64	64	64	64	196
(d_{EE}, d_{EI}, d_{II})	(256,128,64)	(256,128,64)	(256,128,64)	(256,128,64)	(4096,2048,2048)
n_{heads}	8	8	8	8	16
NLM hidden H	4	4	4	4	32
Output-head hidden	256	256	256	256	2048
Cross-attn residual	—	—	—	—	on
Sync decay r	1	1	1	1	$e^{-0.5}$
Sync accum. clamp	—	—	—	—	100
Local batch	64	64	256	256	64
Grad-accum	1	1	1	1	4
DDP size	1	1	1	1	4
Effective batch	64	64	256	256	1024
Learning rate	10^{-3}	10^{-3}	10^{-4}	10^{-4}	$3{\cdot}10^{-4}$
LR warmup	2K	2K	10K	10K	10K
Total steps	50K-200K	50K-200K	600K	600K	100K-500K
Weight decay	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Dropout	0.1	0.1	0.0	0.0	0.1
Mixed precision	—	—	—	—	AMP (Backbone)
Grad clip (global)	1.0	1.0	1.0	1.0	20.0
Per-group clip	—	—	—	—	(2, 2, 50, 5, 5, 20)
TBPTT K	0	0	0	0	25
λ_{spec}	0	0	0	0	0.1
(τ_{EE}, τ_{II})	—	—	—	—	(15, 7)
Curriculum (t_s, T_w)	(1K,5K)	(1K,5K)	(1K,5K)	(1K,5K)	(10K,10K)
Seed	42-46	42-46	42-46	42-46	42-46

F.1 Datasets & Preprocessing

TIDE is evaluated on five image-classification datasets, namely: `MNIST` [1], `Fashion-MNIST` [46], `CIFAR-10` and `CIFAR-100` [47], and `ImageNet-1K` [44]. The standard recommended train and validation splits for each dataset are used, and evaluation during training is performed on the provided test sets. The best evaluation checkpoints are used for all empirical studies, including the OOD analysis. Moreover, the preprocessing per dataset is as follows:

- `MNIST`: $\text{ToTensor} + \text{Normalize}((0.1307),\ (0.3081))$
- `Fashion-MNIST`: $\text{ToTensor} + \text{Normalize}((0.2860),\ (0.3530))$
- `CIFAR-10/100`: $\text{RandomCrop}(32, 4) + \text{RandomHorizontalFlip} + \text{ToTensor} + \text{Normalize}$, where for `CIFAR-10`, we used $(0.4914, 0.4822, 0.4465)/(0.2023, 0.1994, 0.2010)$ for per channel normalization, while $(0.5071, 0.4867, 0.4408)/(0.2675, 0.2565, 0.2761)$ are used for `CIFAR-100`. At test time for both `CIFAR-10` and `CIFAR-100`, only $\text{ToTensor} + \text{Normalize}$ is applied.
- `ImageNet-1K`: At training time, we used $\text{RandomResizedCrop}(224) + \text{RandomHorizontalFlip} + \text{Normalize}\big((0.485, 0.456, 0.406), (0.229, 0.224, 0.225)\big)$ and at test time, $\text{Resize}(256) + \text{CenterCrop}(224) + \text{Normalize}$ is used with the same normalization values as training time.

F.2 Architecture & Training Hyperparameters

Table A.1 lists the dataset-specific hyperparameters used across all studies in this paper. The provided values are identical across all random seeds, and the recommended dataset split is used for each dataset. Moreover, two backbone variants are used: small datasets use the shallow HRF, while `ImageNet-1K` is trained and evaluated on the deep HRF due to its feature complexity. As motivated by [15], `ImageNet-1K` uses larger NLM hidden dimension, $H = 32$, in contrast to other datasets where $H = 4$. Additionally, to stabilize the training and minimize the computational resources used for ImageNet-1K, four additional strategies, including truncated BPTT with $K = 25$, per-component gradient clipping, spectral regularization with $(\tau_{EE}, \tau_{II}) = (15, 7)$, and mixed-precision AMP on the backbone, are used. Note that the Wilson-Cowan dynamics are kept in `float32` precision to avoid overflow [43].

Table A.2: Hyperparameters shared across all five datasets. The Euler coefficients $\alpha_\bullet$ are derived from Δt and the time constants $\tau_\bullet$.

Group	Parameter	Value
Recurrent Dynamics	Internal computation steps T	50
	E-population fraction f_E	0.8
	Population time constants (τ_E, τ_I)	$(20,\ 5)$ ms
	Step size Δt $(\alpha_E = 0.05,\ \alpha_I = 0.20)$	1 ms
Neuron-level Model	NLM memory length M	25
	NLM layers	2
Lateral Inhibition	WTA iterations K_{WTA}	5
	Lateral gain γ (learnable; initial value)	$1.0;\ 0.01$
Memory	Buffer dimension d_{memory}	256
	Surprise threshold θ_s	0.5
	Write step size η_m	10^{-2}
	Write momentum μ	0.9
	Target E-I ratio ρ^*_{EI}	4
Loss Weights	Task loss weight λ_{task}	1.0
	E-I ratio weight λ_{EI}	10^{-2}
	Game-theoretic weight λ_{game}	10^{-3}
	Sync regularizer weight λ_{sync}	10^{-4}

F.3 Compute Resources

Training and evaluation on the smaller datasets were run on a single NVIDIA V100-SXM2 GPU with 32GB VRAM, while `ImageNet-1K` required four 32GB V100 GPUs with PyTorch DDP. Note that all training and evaluation were performed using CUDA 12.4 and PyTorch 2.6.0. Total wall-clock time required for training of each of the datasets per single seed is as follows: a) `MNIST`: 0.85 steps/s, 65 h for 50K steps. b) `Fashion-MNIST` 0.83 steps/s, 67 h for 50K steps. c) `CIFAR-10` 0.68 steps/s, 246 h for 600K steps. d) `CIFAR-100` 0.69 steps/s, 207 h for 300 K steps. e) `ImageNet-1K` 0.105 steps/s, 287 h for 100K steps.

Table A.3: Comparison between TIDE and CTM on five image classification datasets at equal training budget. The *Best Seed* column reports the highest-performing run with its best/final `top-1` accuracy, Acc@1 (%), and `mean` $\pm$ `std` across the multi seeded training for both architectures. CTM is retrained based on [15] for all the ablations with identical seeds. Note that † denotes the mean `top-1` accuracy drop of more than 1%, and ‡ indicates an effective batch size of 64. Moreover, the highest `top-1` accuracy for each dataset is in bold.

			TIDE				**CTM**			
			`mean` $\pm$ `std`		**Best Seed**		`mean` $\pm$ `std`		**Best Seed**	
Task	**Steps**	d_{model}	best	final	best	final	best	final	best	final
MNIST	50K	256	**99.64$_{\pm 0.03}$**	**99.61$_{\pm 0.04}$**	**99.67**	**99.63**	99.54$_{\pm 0.03}$	99.52$_{\pm 0.02}$	99.56	99.54
	100K	256	**99.54$_{\pm 0.05}$**	96.73$_{\pm 4.67}$†	**99.59**	91.33†	**99.54$_{\pm 0.05}$**	**99.49$_{\pm 0.07}$**	**99.59**	**99.57**
	200K	256	**99.60$_{\pm 0.05}$**	**99.53$_{\pm 0.07}$**	**99.64**	**99.60**	99.59$_{\pm 0.06}$	99.49$_{\pm 0.08}$	**99.64**	99.57
F-MNIST	50K	256	**94.21$_{\pm 0.21}$**	**94.15$_{\pm 0.25}$**	**94.42**	**94.42**	92.51$_{\pm 0.11}$	92.45$_{\pm 0.19}$	92.63	92.63
	200K	256	**94.12$_{\pm 0.19}$**	**93.85$_{\pm 0.10}$**	**94.23**	**93.96**	92.53$_{\pm 0.24}$	92.43$_{\pm 0.27}$	92.80	92.73
CIFAR-10	600K	512	**90.62$_{\pm 0.06}$**	**90.54$_{\pm 0.07}$**	**90.68**	**90.62**	86.23$_{\pm 0.34}$	85.98$_{\pm 0.35}$	86.59	86.31
CIFAR-100	600K	718	62.23$_{\pm 0.16}$	61.39$_{\pm 0.82}$	62.38	61.97	65.19$_{\pm 0.48}$	**65.01$_{\pm 0.39}$**	65.70	**65.44**
	600K‡	718	**65.68$_{\pm 0.18}$**	64.50$_{\pm 0.60}$†	**65.79**	65.08				
ImageNet-1K	100K	4096	**67.22$_{\pm 1.52}$**	**67.01$_{\pm 1.58}$**	**68.74**	**68.59**	52.67$_{\pm 1.79}$	51.47$_{\pm 1.62}$	54.46	53.09
	200K	4096	**68.40$_{\pm 1.91}$**	**68.18$_{\pm 2.25}$**	**69.55**	**69.50**	62.78$_{\pm 1.31}$	61.81$_{\pm 1.50}$	64.06	62.88†
	500K	4096	—	—	69.90	69.87	—	—	**71.78**	**69.88**†

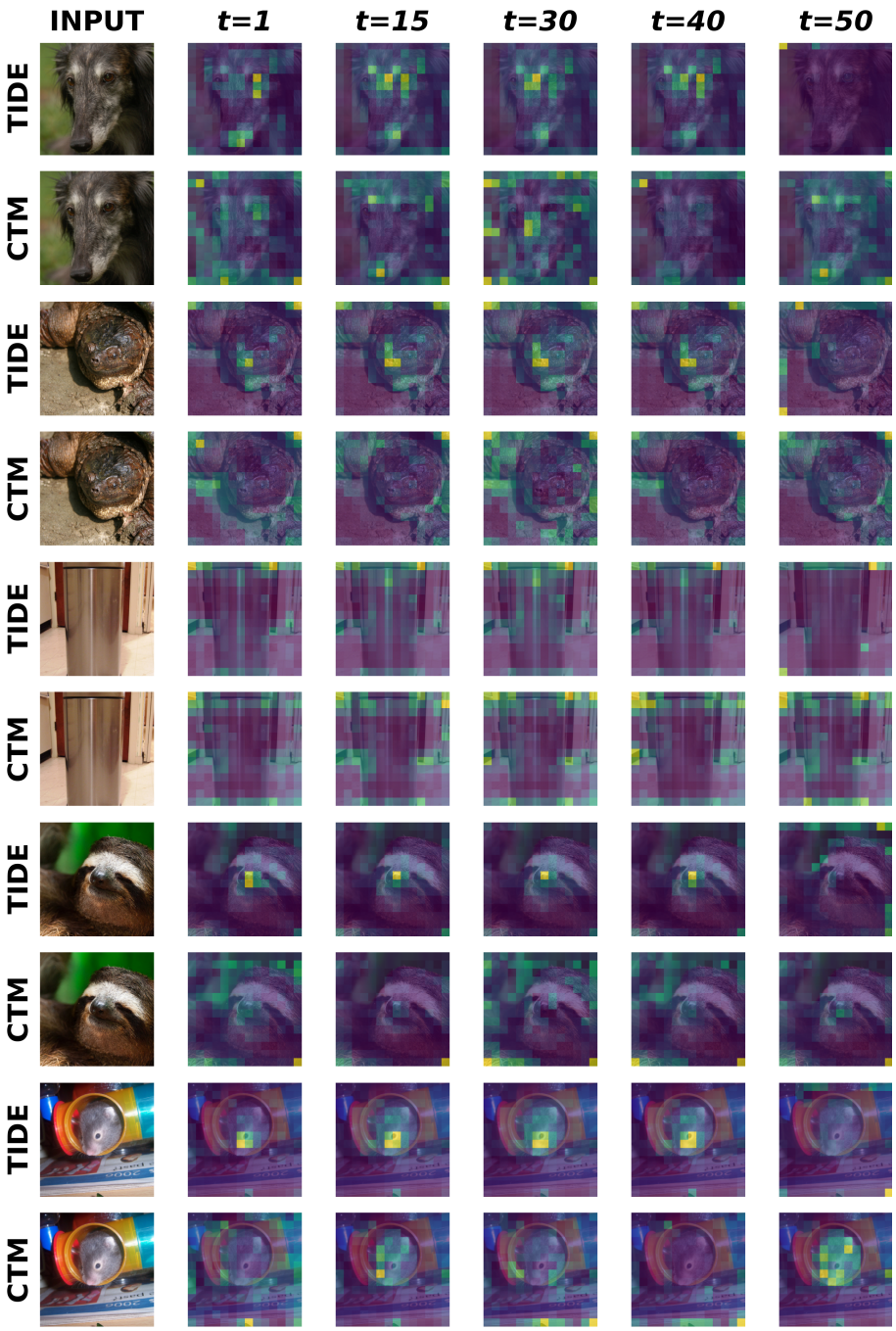

Figure A.1: Temporal evolution of mean attention as saliency per internal computation step for TIDE versus CTM for multiple cases of successful classification.

F.4 Low-variance Estimator

Let $w(\text{step}) \in [0, 1]$ be a multiplicative weight for the per-step gradient estimator of (A.26), thus, its squared deviation must be bounded component-wise by $(1 - w(\text{step}))^2 \leq 1$ as $w \equiv 1$. The cosine scheduled ramp (A.24) with $w(t) = \frac{1}{2}(1 - \cos(\pi t))$ on $t = (\text{step} - t_s)/T_w \in [0, 1]$, admits the closed form $\frac{1}{T_w} \int_{t_s}^{t_s+T_w} (1 - w(s))^2 \, \mathrm{d}s = 3/8$, where t_s is the step at which the cosine warm-up initializes, and T_w denote its length. Therefore, it can be stated that the cosine warm-up acts as a controlled, decaying perturbation of the immediate-on $w \equiv 1$ schedule, contributing on average $3/8$ squared deviation.

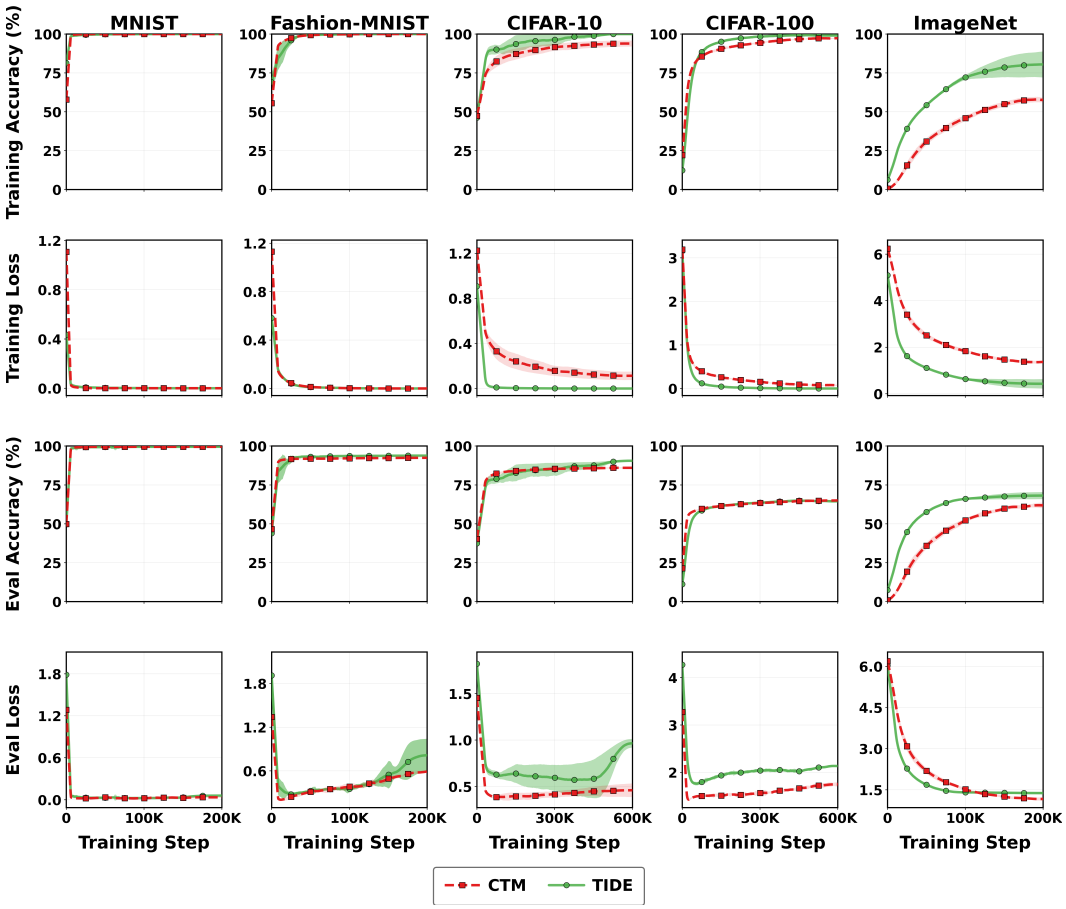

Figure A.2: Comparison between TIDE and CTM based on their training and validation curves, while providing `top-1` accuracy (%), and `mean` $\pm$ `std` total losses across all the trained seeds.

G Additional Experiments & Ablations

This section presents the main benchmark results, ablation studies, and robustness analysis. Per-dataset benchmarks with their related training and evaluation metrics are reported, while the comparative analysis of their performance against CTM as a baseline is included in Appendix G.1. Furthermore, the ablation study of hyperparameters on the `MNIST` and `Fashion-MNIST` datasets is reported across various subsections. Appendix G.2 focuses on the optimal E-I ratio, $\rho^*_{EI} = n_E/n_I$, and its effect on the resulted `top-1` accuracy. Appendix G.3 provides the ablation study of the population-specific time constants τ_E, τ_I, while Appendix G.4 is focused on the number of internal computation steps T, required for stability of training. Furthermore, the game-theoretic loss weight λ_{game} is studied in Appendix G.5, and the results for lateral inhibition iteration count K_{WTA} are provided in Appendix G.6. Thereafter, we analyze the training stability of `ImageNet-1K` via its gradient-norm curves and spectral radii during training in Appendix G.7. Further component-wise analysis of the architectures and their contribution to the performance of TIDE is reported in Appendix G.8. Additional analysis is performed to investigate the robustness of the backbone in Appendix G.9. Finally, we provide a detailed comparative analysis of the performance of CTM and TIDE under perturbations, while investigating their robustness in OOD cases in Appendix G.10.

G.1 Main Benchmark

The `top-1` accuracy of TIDE alongside CTM is reported in Table A.3. We trained both TIDE and CTM with the identical hyperparameters per dataset provided in Table A.1. To directly compare the convergence of each of the architectures, both are trained with similar training budget to analyze their sample efficiency. Note that all the provided results and metrics are based on our own re-run of the CTM under identical training. In addition to quantitative analysis, qualitative results are provided, showing the visual temporal evolution of attention for TIDE and CTM to enable direct comparison of the behavior of both architectures (cf. Figure A.1). We additionally analyzed the performance of ResNet-18 as a backbone for all datasets as part of leave-one-out

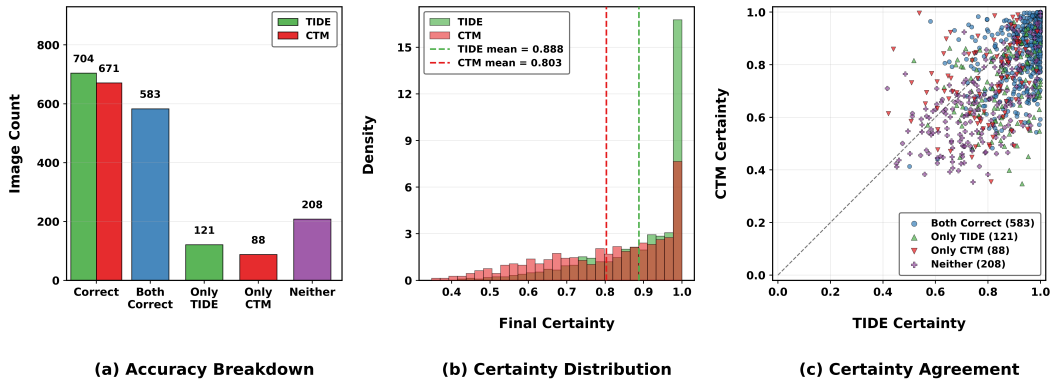

Figure A.3: TIDE versus CTM accuracy and certainty on a diverse 1000 randomly selected images subset of the `ImageNet-1K` validation set per each class. (a) Accuracy breakout and comparison across TIDE and CTM. (b) Certainty distribution for accurate classification and `mean` confidence comparison. (c) Agreement analysis based on all classifications across the randomly selected samples.

ablation study (cf. Appendix G.8), which shows significant performance degradation compared to both variants of HRF.

The remaining discussion focuses on the multi-seeded training of TIDE and the results achieved compared with CTM. Note that the backbone differs significantly between TIDE and CTM as the main aim of TIDE is to achieve bio-plausible end-to-end training by utilizing HRF and its two variants. We use an identical d_{model} for both TIDE and CTM while investigating each dataset as shown in Table A.3. Additionally, evaluation curves based on the combined results of multi-seeded training of TIDE and CTM are provided in Figure A.2.

It has been observed that TIDE can achieve higher evaluation accuracy and exceeds CTM on `MNIST` (99.67 vs. 99.64), `Fashion-MNIST` (94.42 vs. 92.80), `CIFAR-10` (90.68 vs. 86.59), `CIFAR-100` (65.79 vs. 65.70) and `ImageNet-1K`-200 K (69.55 vs. 64.06), where the CTM-200 K baseline is under-trained relative to CTM 500 K-step result of 71.78% which is achieved via our re-training and is also reported by [15]. We further investigate backbone selection by initially replacing it with ResNet-18, which overall reduces `top-1` accuracy across all datasets, as shown in Appendix G.8. We conjecture this is due to the projection enforcing Dale's principle and its poor interaction with batch-normalization statistics inherited from the ResNet stem. Thus, for the `ImageNet-1K` dataset, a revised ResNet-style deep HRF is devised to address this issue.

As shown in Figure A.3, we perform a diversity-based analysis in which we randomly select 1000 images across all classes in `ImageNet-1K` to compare TIDE and CTM. TIDE achieved 70.4% against the 500 K steps trained CTM's 67.1% `top-1` accuracy. Figure A.3 shows the decoupled of classification from the internal steps for both models, where TIDE outperforms CTM by achieving a higher accuracy rate. Furthermore, it has been observed that TIDE has higher mean certainty than CTM (0.89 vs. 0.80), indicating that E-I-based dynamics produce more decisive readouts than CTM by relying on its WTA and surprise-gated memory. Both CTM and TIDE have a prediction agreement of 79.1% for correctly classified images, indicating substantial agreement well above chance.

Table A.4: E-I population ratio ablation study. Baseline $n_E/d_{\text{model}} = 0.8$ is in bold. Final E-I activity ratio (post-training) is reported to facilitate analysis of stability, given the target has been set as $\rho^*_{EI} = 4.0$.

	`MNIST`			`Fashion-MNIST`	
n_E/d_{model}	**best**	**final**	ρ_{EI}	**best**	**final**
0.6	99.53	98.95	3.94	93.53	93.51
0.7	99.55	99.39	4.01	93.67	93.61
0.8	**99.53**	**99.51**	**4.01**	**93.53**	**86.19**
0.9	99.55	99.28	4.01	93.49	93.34

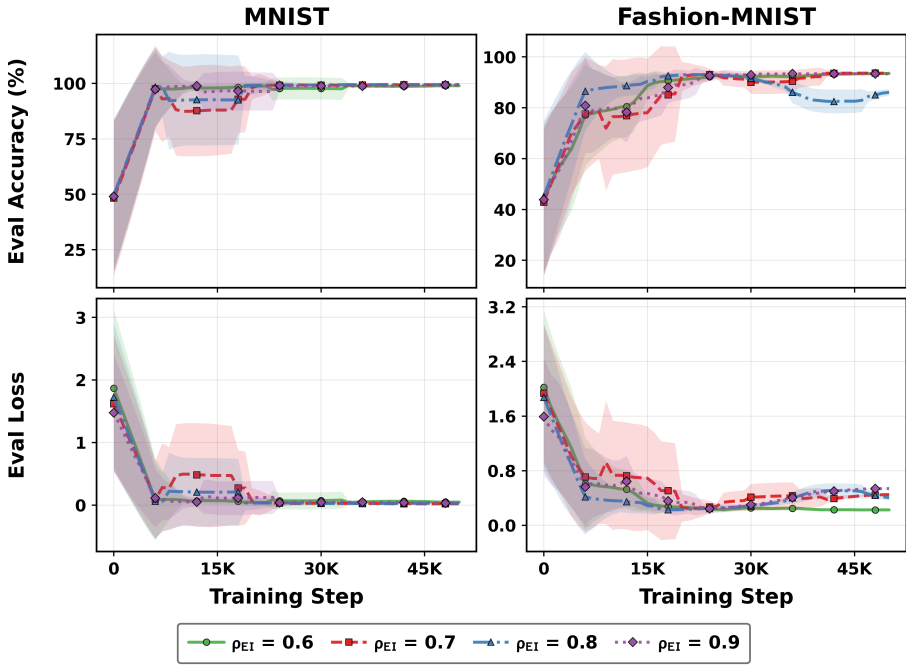

Figure A.4: Evaluation curves during ablation study to analyze E-I population ratio on `MNIST` and `Fashion-MNIST`.

G.2 E-I Ratio Analysis

Table A.4 provides the results of the E-I ratio ablation study, where we analyze four population split configurations, $\rho^*_{EI} = n_E/n_I$, on `MNIST` and `Fashion-MNIST`. Values $\{0.6, 0.7, 0.8, 0.9\}$ correspond to n_E/d_{model}, hence $\rho^*_{EI} = n_E/(d_{\text{model}} - n_E) \in \{1.5, 2.33, 4.0, 9.0\}$. The biologically-motivated default ratio 80:20 ($n_E/d_{\text{model}} = 0.8$) achieves the baseline `MNIST` result of 99.51% and is within 0.02 percentage points of the best variant ($n_E/d = 0.7$) on `MNIST`.

As illustrated in Figure A.4, the `Fashion-MNIST` run with the default E-I ratio exhibits post-peak drift (86.19%) that disappears at $n_E/d \in \{0.6, 0.7, 0.9\}$. Across all configurations, ρ_{EI} values for `MNIST` column in Table A.4 track the 80:20 desired target ratio to within 2%, indicating that the activity-ratio regularizer $\mathcal{L}_{\text{EI}}$ (A.15) is effective regardless of the population split.

G.3 Time-constant Analysis

Table A.5 provides the ablation study on the excitatory and inhibitory membrane time constants. The biological regime requires $\tau_I < \tau_E$; we verify this by performing two parallel ablation studies where either $\tau_I = 5$ ms is held fixed, or $\tau_E = 20$ ms is set as a constant, as shown in Figures A.5 and A.6.

A noticeable performance degradation is observed when $\tau_E = 10$ ms; both `MNIST` and `Fashion-MNIST` accuracies decrease by 1 and 4 percentage points, respectively. This is a direct result of $\Delta t/\tau_E = \alpha_E = 0.1$ rather than 0.05 when $\tau_E/\tau_I = 2$, which causes an overshoot in the inhibitory response, visibly destabilizing the training, as shown in Figure A.5.

Table A.5: Population time-constant ablation (ms). Baseline values $(\tau_E, \tau_I) = (20, 5)$ ms are in bold.

	`MNIST`		`Fashion-MNIST`			`MNIST`		`Fashion-MNIST`	
τ_E (ms)	**best**	**final**	**best**	**final**	τ_I (ms)	**best**	**final**	**best**	**final**
10	98.69	95.97	93.91	89.45	3	99.58	99.56	93.64	93.50
15	99.59	99.47	93.80	93.62	**5**	**99.58**	**99.52**	**93.15**	**92.56**
20	**99.62**	**99.61**	**94.23**	**93.73**	7	99.49	99.48	94.00	93.88
25	99.61	99.61	94.25	94.05	10	99.55	99.55	93.78	93.73
30	99.54	99.50	92.95	92.85					

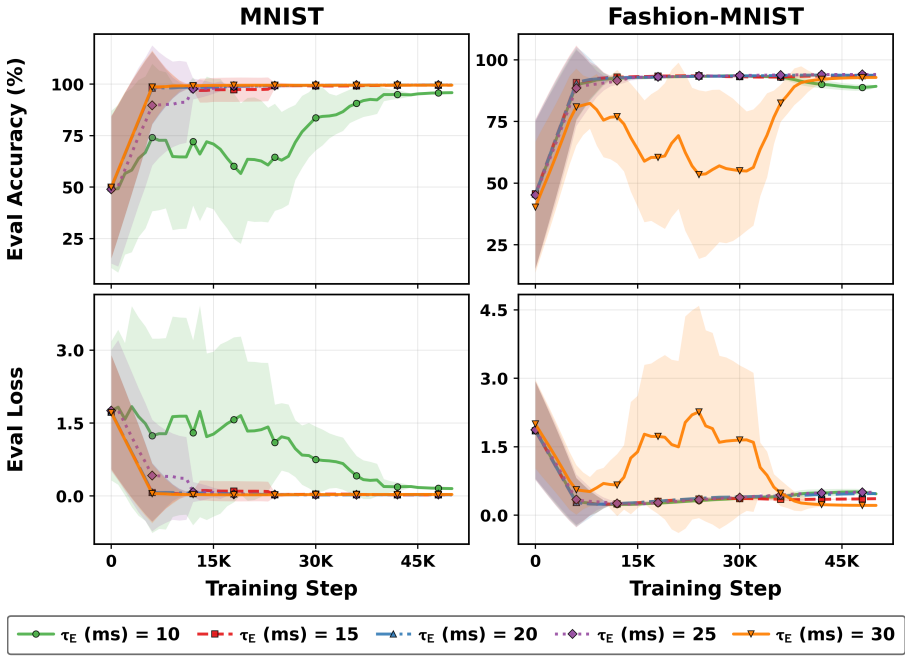

Figure A.5: Evaluation curves during ablation study to analyze τ_E impact on `MNIST` and `Fashion-MNIST`.

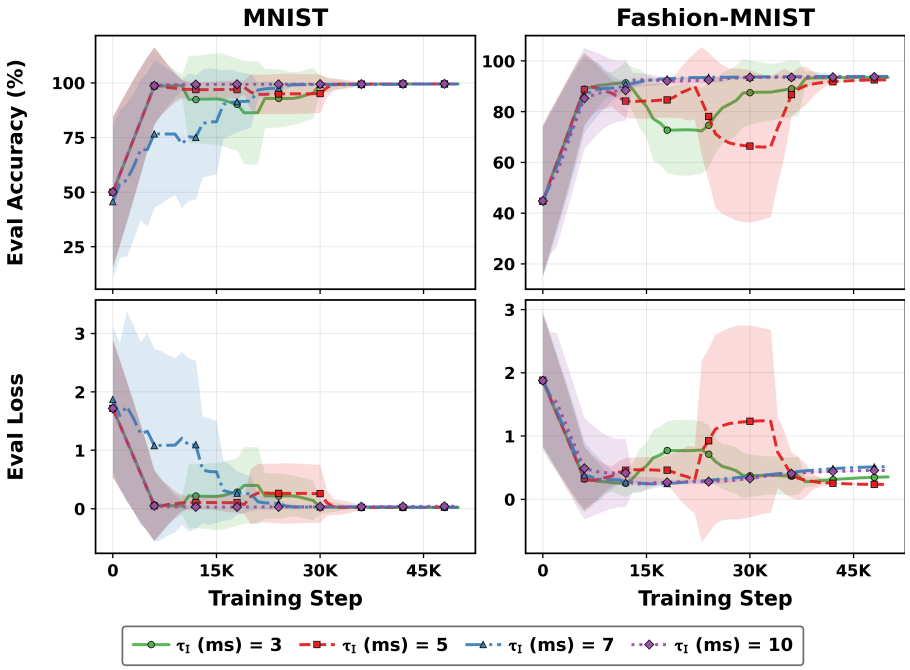

Figure A.6: Evaluation curves during ablation study to analyze τ_I impact on `MNIST` and `Fashion-MNIST`.

During the τ_I ablation, a mild preference for $\tau_I = 7$ on `Fashion-MNIST` is observed as shown in Table A.5. Although no catastrophic instability was observed during the τ_I ablation, given $\tau_I > \alpha_I \, \tau_E/2 = 0.5$ ms, which satisfies the stability condition outlined in Appendix C.2, a sudden performance degradation was observed for both `MNIST` and `Fashion-MNIST` when $\tau_I = 5$ as illustrated by Figure A.6.

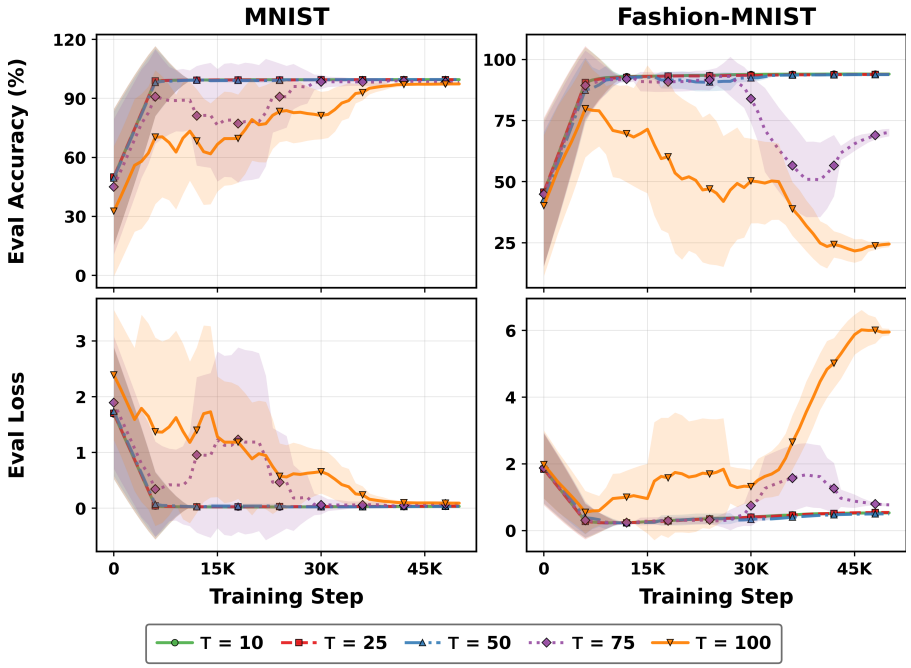

Figure A.7: Evaluation curves during ablation study to analyze computation step T and its effect on `MNIST` and `Fashion-MNIST`.

G.4 Iteration-depth Analysis

An ablation analysis of the internal computation steps T is performed while all other parameters are kept fixed (cf. Table A.6). Furthermore, we provide a wall-clock comparison for the given epoch based on the estimated time per classification, with varying duration T, which directly affects both training duration and the inference phase. Moreover, as illustrated by Figure A.7, longer internal computation steps do not guarantee higher performance, which was observed for both `MNIST` and `Fashion-MNIST`.

It has been shown that the best accuracy is unaffected when $T \in [10,\ 50]$ for both datasets, and the accuracy deviation is $< 0.31\%$ within this range. As mentioned, for $T = 100$, performance degrades substantially, especially, for `Fashion-MNIST`, where the final accuracy collapses entirely to 24.45%, as shown in Figure A.7. We attribute this instability and collapse to two major compounding effects: i) The NLM memory length limit, $M = 25$, which might not be ideal for longer T due to loss of information. ii) The BPTT horizon $K = 0$ of Algorithm A.1 for smaller image datasets such as `MNIST` and `Fashion-MNIST`, which increases the accumulated gradient variance and further destabilizes the optimization given longer T. Therefore, as with `ImageNet-1K`, either truncated-BPTT with $K \leq T/4$ or jointly scaling M is required to ensure stability .

G.5 Game-theoretic Loss Weight Analysis

The auxiliary game-theoretic weight λ_{game} of (A.20) is further studied as shown in Table A.7. Figure A.8 illustrates that the task accuracy remains performant across all four configurations `MNIST`; however, significant

Table A.6: Internal computation steps, T, ablation. Baseline $T = 50$ is in bold. The Time factor column reports training wall-clock relative to the baseline.

	`MNIST`		`Fashion-MNIST`		
T	**best**	**final**	**best**	**final**	**Time factor**
10	99.58	99.54	94.17	94.00	$\times$ 0.24
25	99.58	99.51	93.90	93.86	$\times$ 0.52
50	**99.60**	**99.57**	**93.86**	**93.78**	$\times$ 1.00
75	99.54	98.11	93.86	70.42	$\times$ 1.51
100	98.77	97.34	92.05	24.45	$\times$ 2.02

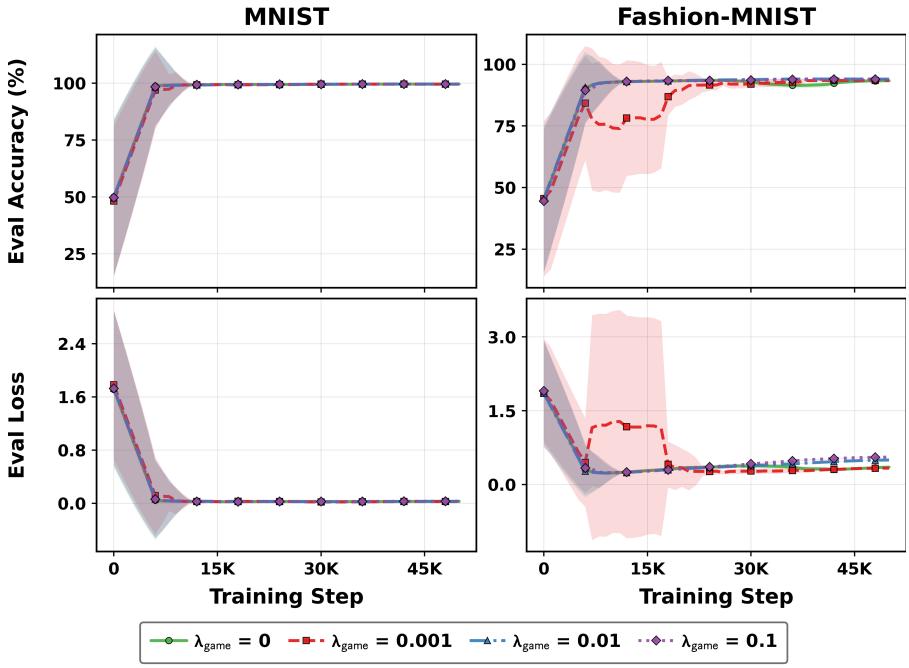

Figure A.8: Evaluation curves during ablation study to analyze the impact of λ_{game} on `MNIST` and `Fashion-MNIST`.

E-I population collapse is observed at $\lambda_{\text{game}} \in \{10^{-2}, 10^{-1}\}$ to $\rho_{EI} = 0.01$, indicating complete deactivation of inhibitory pathways. Therefore, the game-theoretic loss, $\mathcal{L}_{\text{game}}$, must be weighted comparably to the task loss, $\mathcal{L}_{\text{task}}$, to prevent the regularizer from overwhelming the task loss and fully silencing the inhibitory population.

Due to the inherent preference for $\mathcal{L}_{\text{task}}$, the best and final `top-1` accuracy is not directly affected for either `MNIST` or `Fashion-MNIST`. Moreover, as shown in Table A.7, the collapse of ρ_{EI} is not only dependent on the magnitude of $\mathcal{L}_{\text{game}}$ but also dataset dependent, as no significant changes were observed for `Fashion-MNIST`. Note that due to the neuron-inspired assumption of TIDE, λ_{game} must be $\leq 10^{-3}$ for stability of E-I circuit.

G.6 Lateral Inhibition Iterations

The lateral inhibition circuit and its internal iterations are investigated to analyze the effect of K_{WTA} on the `top-1` accuracy and stability of the TIDE, while obtaining the optimal value for K_{WTA} as shown in Table A.8.

As shown in Figure A.9, the default $K_{\text{WTA}} = 5$ does not perform well on both `MNIST` and `Fashion-MNIST`. Furthermore, the best performance is achieved when $K_{\text{WTA}} = \{3, 10\}$; thus improving the `top-1` accuracy by $+1.14\%$ and $+0.02\%$ for `Fashion-MNIST` and `MNIST`, respectively.

Given that the lateral inhibition is based on RELU without a reliance on $\mathrm{Softmax}$, while W_{EI}^{lat} is initialized to a uniform distribution bounded by $1/\sqrt{n_E}$ for the given training length of 50 K steps, the dynamics typically converge to $K_{\text{WTA}} \leq 3$. The early termination signal is typically triggered within three steps; thus, increasing K_{WTA} to higher values without decreasing the termination tolerance will only result in longer, idle iterations during inference.

Table A.7: Game-theoretic loss weight λ_{game} ablation. Baseline $\lambda_{\text{game}} = 10^{-3}$ is in bold. The final E-I activity ratio ρ_{EI} with target 4.0 is reported in italics for interpretability.

	`MNIST`			`Fashion-MNIST`		
λ_{game}	**best**	**final**	$\boldsymbol{\rho_{EI}}$	**best**	**final**	$\boldsymbol{\rho_{EI}}$
0	99.64	99.60	*4.00*	93.75	93.30	*3.99*
$\mathbf{10^{-3}}$	**99.59**	**99.53**	***3.99***	**93.61**	**93.46**	***3.98***
10^{-2}	99.64	99.59	*0.01*	94.15	93.95	*4.00*
10^{-1}	99.62	99.61	*0.01*	94.03	93.91	*3.99*

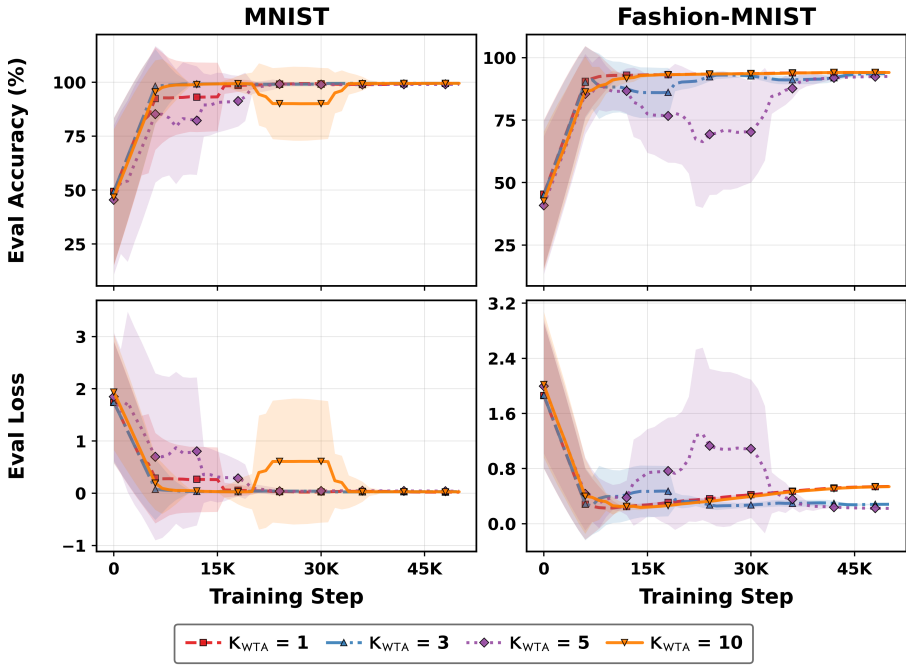

Figure A.9: Evaluation curves during ablation study to analyze the impact of K_{WTA} on `MNIST` and `Fashion-MNIST`.

The resulting failure mode given $K_{\text{WTA}}=5$ is mainly due to the location of the baseline in a local minimum; however, $K_{\text{WTA}}=5$ is retained given further experimentation (cf. Table 2) has demonstrated that this value can be used across all datasets without performance degradation. Consequently, we used $K_{\text{WTA}}=5$ across all our benchmarks (cf. Table A.2).

G.7 Training-stability Diagnostics

Given the stability issues in E-I circuits highlighted in [18], further analysis was conducted to investigate training stability, potential failure modes, and the approaches needed to achieve stable results.

As illustrated in Figures A.10 and A.11, the Perron radii converged to desired value without overshooting for both 100 K and 200 K multi-seeded `ImageNet-1K` training, where $\widehat{\lambda}_{\text{P}}(W_{EE}) = 14.69 \pm 0.27$ and $\widehat{\lambda}_{\text{P}}(W_{II}) = 6.998 \pm 0.006$, since the $\text{ReLU}(\widehat{\lambda}-\tau)^2$ results in a hinge equilibrates for the penalty gradient against the task loss gradient at the boundary, $\tau_{EE} = 15$ and $\tau_{II} = 7$ are achieved. The soft-hinge gradient remains active, providing non-vanishing pressure toward the LDS-compatible region defined in C.3. Furthermore, the regularized spectral norms grow throughout training while exceeding the Perron clamps at $\|W_{EE}\|_2 = 58.6 \pm 7.6$, $\|W_{II}\|_2 = 72.7 \pm 50.1$, $\sigma_{\max}(W_{EI}) = 63.2 \pm 22.5$, and $\sigma_{\max}(W_{IE}) = 55.8 \pm 12.7$, where $\|W\|_2 = \sigma_{\max}(W)$ denotes the largest singular value of W, i.e., the spectral norm characterizing the worst-case amplification of any input vector by W. However, this growth does not violate the LDS condition of Corollary C.5, so local asymptotic stability of the Wilson-Cowan recurrence (Theorem C.4) is preserved. Together with empirical observations from

Table A.8: Lateral-inhibition iteration count K_{WTA} ablation study. Baseline $K_{\text{WTA}} = 5$ is in bold. $K_{\text{WTA}} = 0$ (no lateral inhibition) is omitted because the corresponding runs are provided in Appendix G.8.

	`MNIST`		`Fashion-MNIST`	
K_{WTA}	**best**	**final**	**best**	**final**
1	99.53	99.46	94.10	94.10
3	99.58	99.49	93.67	93.64
5	**99.42**	**98.99**	**93.06**	**92.54**
10	99.49	99.47	94.20	94.07

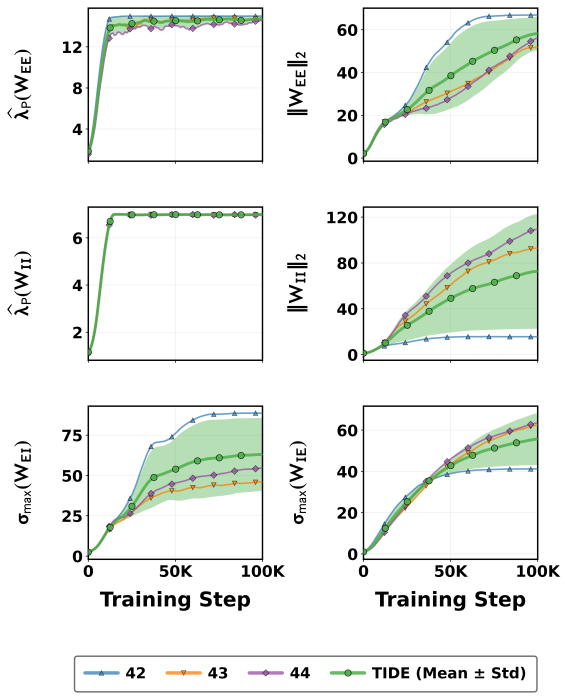

Figure A.10: Spectral values during 100 K multi-seeded `ImageNet-1K` training, Rows 1–2: Perron estimates $\widehat{\lambda}_{\mathrm{P}}$ (*left*) and spectral norm $\|\cdot\|_2$ (*right*) for W_{EE} and W_{II}. Row 3: cross-population $\sigma_{\max}(W_{EI})$ and $\sigma_{\max}(W_{IE})$.

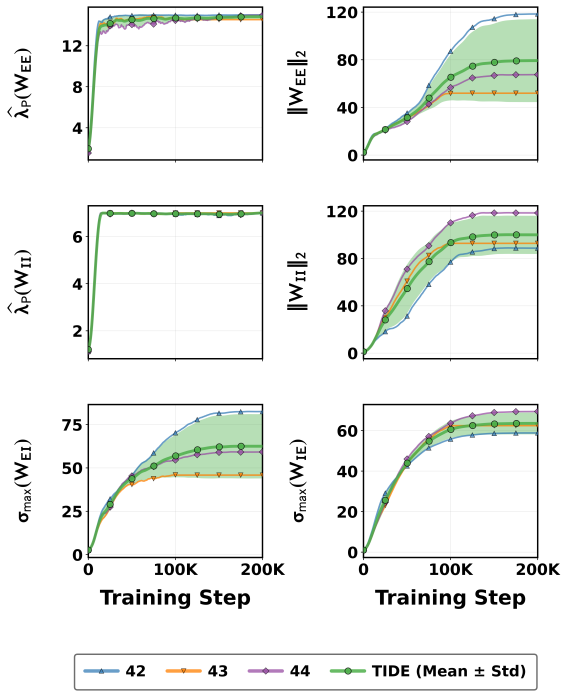

Figure A.11: Spectral values during 200 K multi-seeded `ImageNet-1K` training, Rows 1–2: Perron estimates $\widehat{\lambda}_{\mathrm{P}}$ (*left*) and spectral norm $\|\cdot\|_2$ (*right*) for W_{EE} and W_{II}. Row 3: cross-population $\sigma_{\max}(W_{EI})$ and $\sigma_{\max}(W_{IE})$.

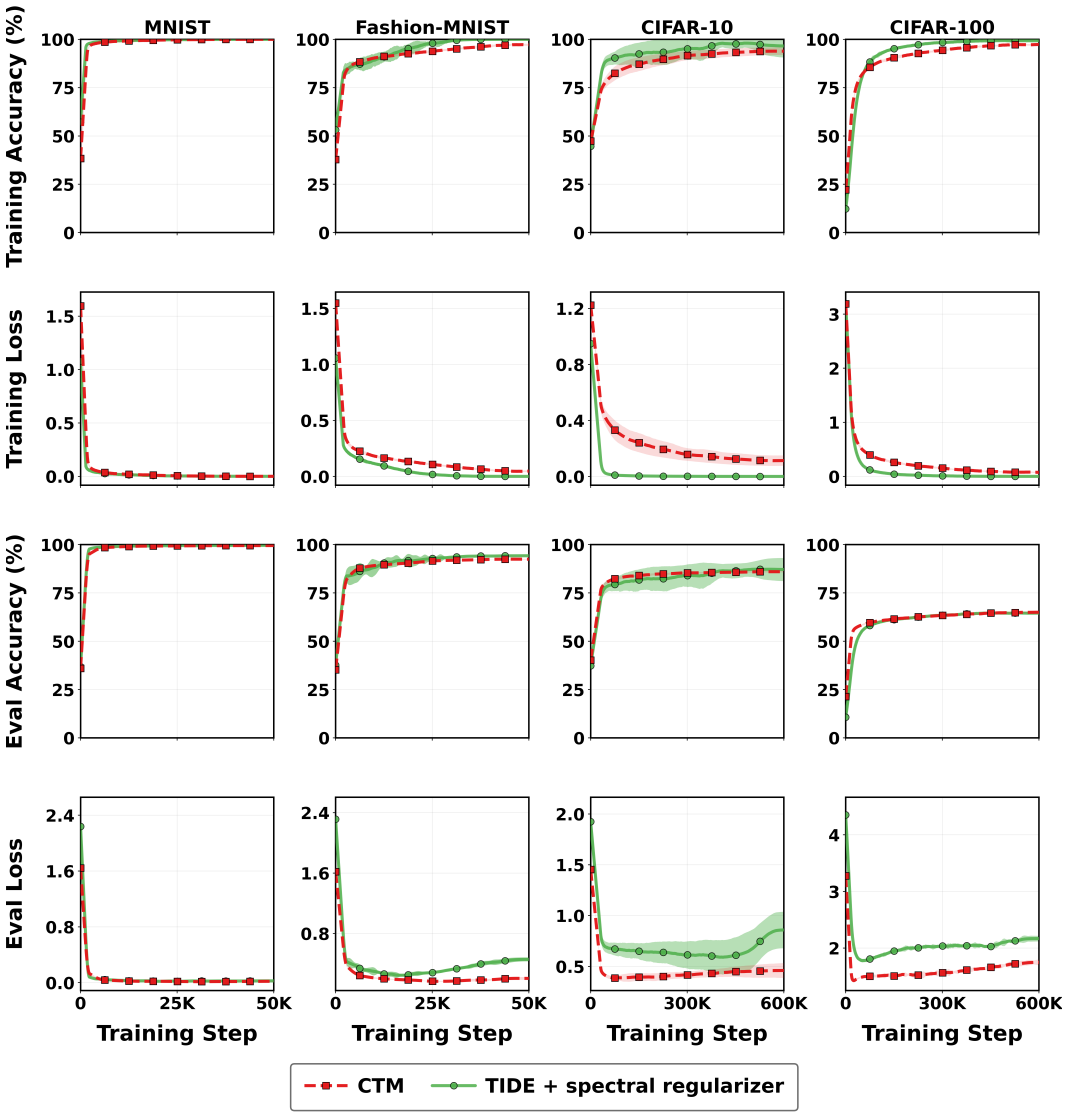

Figure A.12: Comparison between TIDE with its spectral regularization and CTM. Both `top-1` accuracy (%), and `mean` $\pm$ `std` total losses across all the trained seeds are provided.

multi-seeded training across various datasets, the stability mechanisms described in Appendix C are effective at providing stable learning in asymmetric architectures.

G.8 Architectural Component Analysis

To analyze the contribution of each component to the overall performance of TIDE, we perform a leave-one-out ablation, as shown in Table 3. This study shows the direct compounded effect of biologically-motivated constraints: removing any component lowers architecture's performance, and the highest `top-1` accuracy is achieved only when all components are included.

Setup: All architectural variants are trained and evaluated under identical training budget and hyperparameter settings as indicated in Tables A.1 and A.2. Furthermore, when investigating the spectral regularizer loss and its effect on the stability of the final `top-1` accuracy for the remaining datasets, λ_{spec} is set to 0.1 (cf. Table 3).

Spectral regularizer: With the additional stabilization from the regularizer, TIDE achieves overall better performance and its `top-1` accuracy on both `MNIST` and `Fashion-MNIST` increases as shown in Table 3. However, the stabilization observed when comparing Figure A.12 with Figure A.2 comes at the cost of lower `top-1` accuracy on both `CIFAR-10` and `CIFAR-100`.

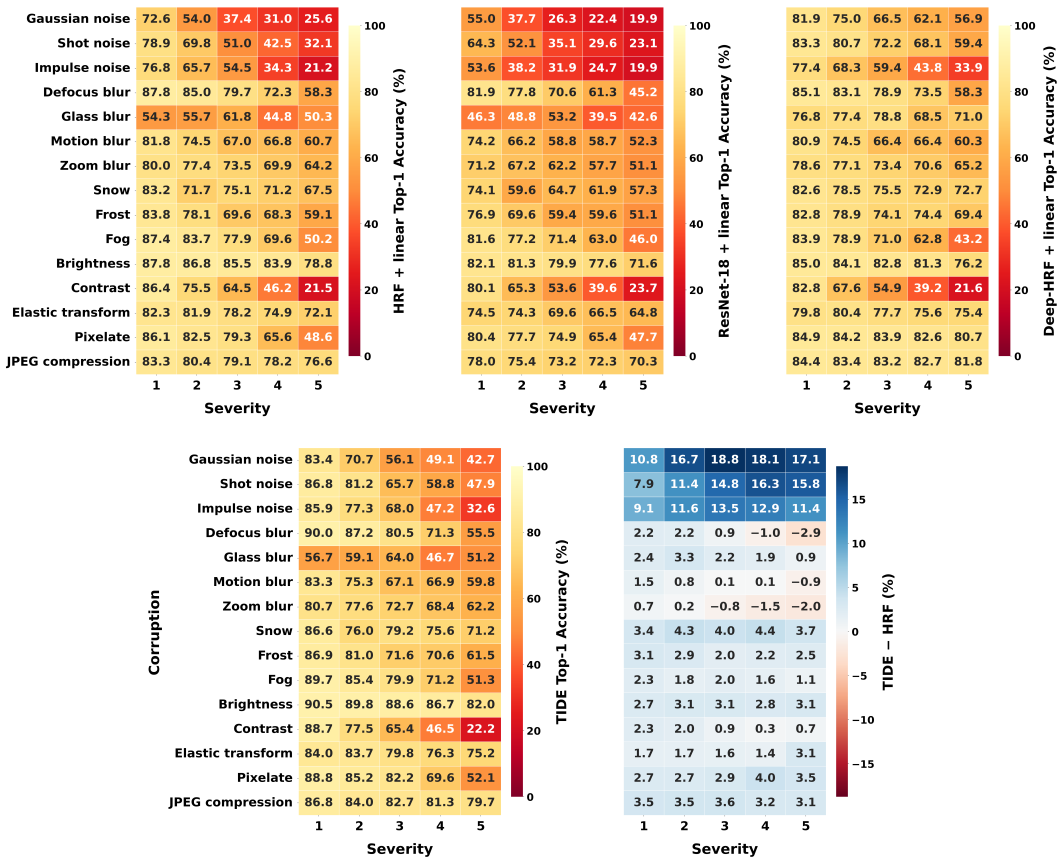

Figure A.13: Robustness analysis using CIFAR-10-C [48] for corrupted images with perturbations. Top: *Left* panel presents results for shallow HRF backbone with linear classifier, while the *center* panel reports the `top-1` accuracy under perturbations of the linear classifier with ResNet-18 backbone, and *right* panel presents the results for the deep HRF backbone. Bottom: *Left* panel presents results for TIDE, and *right* panel reports the differences between TIDE and the linear classifier with the shallow HRF backbone.

G.9 Backbone Analysis & Robustness

To separate the contribution of HRF from that of TIDE's recurrent dynamics, each backbone used in this study is trained together with a linear classifier (cf. Table A.9) to analyze their performance on unperturbed datasets. Additional robustness analysis are performed on all the variants under several perturbation severities by using `CIFAR-10-C` and `CIFAR-100-C` [48] (Figures A.13 and A.14) to further investigate the robustness gains against the conventional ResNet-18 backbone.

Setup: Each linear classifier uses the averaged 64 value tokens from its backbone and applies a single linear layer trained via cross-entropy loss for 50 K steps. To compare TIDE's recurrent dynamics against the linear classifier, we train all the models under an identical training budget. Note that all the following studies are performed on perturbed datasets, and we use only models that are solely trained on unperturbed data to investigate generalization and OOD robustness.

CIFAR-10-C & CIFAR-100-C: `CIFAR-10-C` and `CIFAR-100-C` consist of the 10 K test images of `CIFAR-10` and `CIFAR-100` [47], respectively, at 32×32 resolution with 10 and 100 classes, corrupted with the 15 standard corruption similar to `ImageNet-C` [48]. As shown in Figure A.13, TIDE outperforms shallow HRF by an average of $+4.28\%$ across all perturbations. Similarly, Figure A.14 illustrates that TIDE can outperform its counter part by an average of $+4.99\%$ across `CIFAR-100-C`.

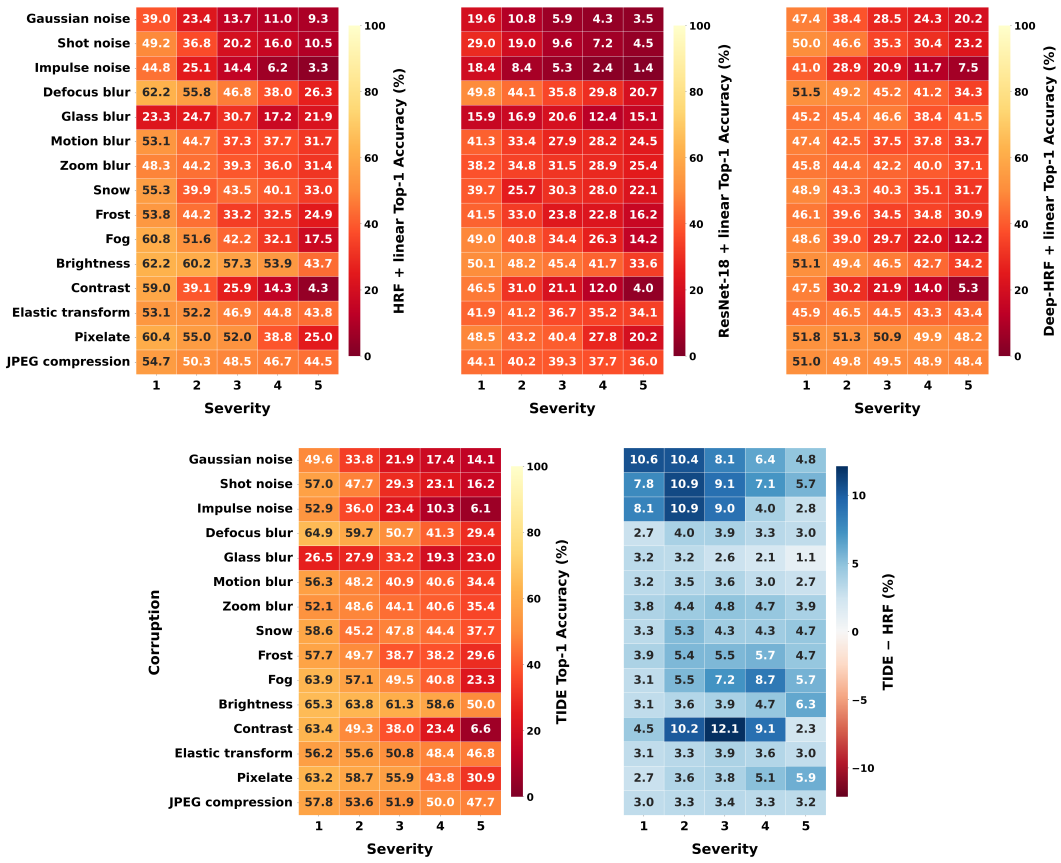

Figure A.14: Robustness analysis using CIFAR-100-C [48] for corrupted images with perturbations. Top: *Left* panel presents results for shallow HRF backbone with linear classifier, while the *center* panel reports the `top-1` accuracy under perturbations of the linear classifier with ResNet-18 backbone, and *right* panel presents the results for the deep HRF backbone. Bottom: *Left* panel presents results for TIDE, and *right* panel reports the differences between TIDE and the linear classifier with the shallow HRF backbone.

G.10 Robustness & OOD Evaluation

The previous experiments in Appendix G.1 primarily report `top-1` accuracy on unperturbed data. To further investigate if leveraging neuro-inspired motifs provides measurable benefits for the learning outcomes beyond CTM, we evaluate the robustness by investigating 100 K, 200 K and 500 K steps checkpoint for TIDE trained solely on `ImageNet-1K` against the similarly trained CTM checkpoints on three established OOD benchmarks: `ImageNet-C` [48], `ImageNet-R` [50], and `Tiny-ImageNet` [49] without retraining, fine-tuning, or test-time adaption or augmentation. Note that Section 5 reports only the 500 K comparison for `ImageNet-C`.

Table A.9: Performance analysis of different backbones with the linear classifier on unperturbed datasets. The table provides the mean best and final `top-1` accuracy, Acc@1 (%), as the `mean` $\pm$ `std` across the seeds for evaluation during training. Moreover, the highest `top-1` accuracy for each dataset is in bold.

	MNIST		F-MNIST		CIFAR-10		CIFAR-100	
Configuration	best	final	best	final	best	final	best	final
Shallow HRF	**99.59$_{\pm 0.01}$**	99.54$_{\pm 0.03}$	**94.12$_{\pm 0.06}$**	**94.06$_{\pm 0.10}$**	**86.14$_{\pm 0.15}$**	**86.09$_{\pm 0.08}$**	**52.98$_{\pm 0.24}$**	**52.96$_{\pm 0.22}$**
ResNet-18	99.52$_{\pm 0.04}$	99.48$_{\pm 0.03}$	93.52$_{\pm 0.14}$	93.42$_{\pm 0.12}$	82.45$_{\pm 0.20}$	82.39$_{\pm 0.23}$	50.60$_{\pm 0.09}$	50.49$_{\pm 0.12}$
Deep HRF	**99.59$_{\pm 0.02}$**	**99.55$_{\pm 0.04}$**	92.44$_{\pm 0.18}$	92.39$_{\pm 0.21}$	85.52$_{\pm 0.10}$	85.45$_{\pm 0.16}$	51.46$_{\pm 0.29}$	51.35$_{\pm 0.31}$

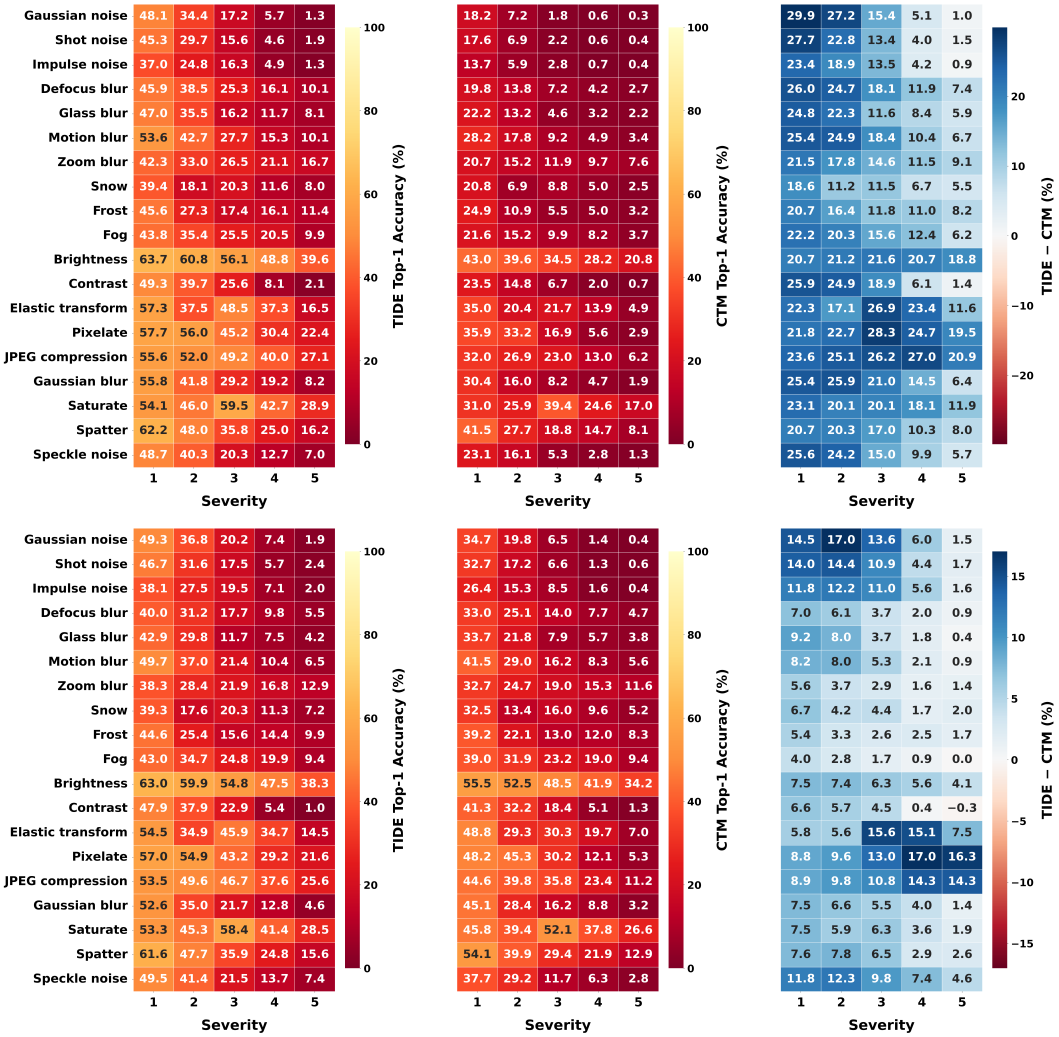

Figure A.15: Robustness analysis using ImageNet-C [48] for corrupted images with perturbations. *Left* panel presents results for TIDE, *center* panel corresponds to CTM, and *right* panel reports differences. Top: 100 K training budget. Bottom: 200 K training budget.

Setup: Both models are evaluated in inference mode with batch sizes 32 under the standard `ImageNet-1K` preprocessing conditions stated in Appendix F.1. The `top-1` best accuracy for both models across all training budget are selected for direct comparison and analysis.

ImageNet-C: `ImageNet-C` consists of the 50 K validation images of `ImageNet-1K` modified under different perturbation severity levels [48]. Following the ablation protocol from [50], we evaluate `top-1` accuracy of TIDE and CTM as shown in Figure A.15. Overall, TIDE has higher mean `top-1` accuracy than CTM under all training budgets: 16.8% at 100 K (30.6% against 13.9%), and 6.4% at 200 K (28.9% against 22.5%).

ImageNet-R: `ImageNet-R` consists of 30 K renditions with various styles and categories, including art, graphic, painting, drawn from a 200-class subset of `ImageNet-1K`. Following the ablation protocol outlined by [50], we evaluate `top-1` accuracy of TIDE and CTM as shown in Figure A.16. Overall, TIDE has higher `top-1` accuracy than CTM over all renditions under all training budgets: 11.5% at 100 K (25.75% against 14.3%), 5.6% at 200 K (26.5% against 20.9%), and 0.29% at 500 K (25.75% against 25.46%). CTM shows better robustness on tattoos and deviantart as illustrated in Figure A.16. Overall, TIDE provides more robustness against OOD and shifted distributions via rendition.

Tiny-ImageNet: `Tiny-ImageNet` consists of 10 K validation images at 64×64 resolution, evaluated against the 1000-way `ImageNet-1K` softmax via the `TinyImageNet`→`ImageNet-1K` mapping [49]. Both models are

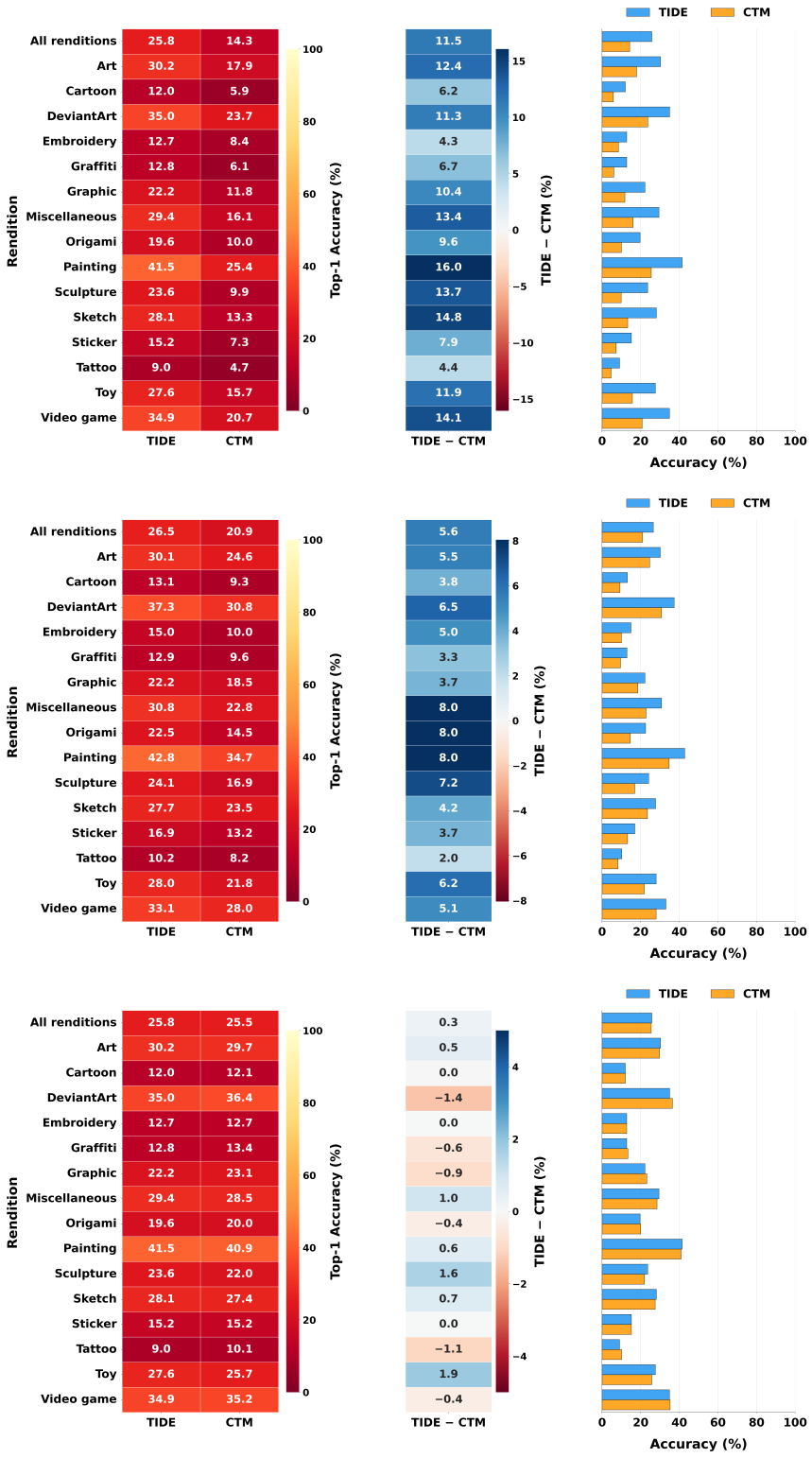

Figure A.16: Robustness analysis on `ImageNet-R` [50] in the 100 K (top), 200 K (middle) and 500 K (bottom) comparisons. *Left* panel presents `top-1` (%) for TIDE and CTM, *center* panel reports the differences and *right* panel is the accuracy bar.

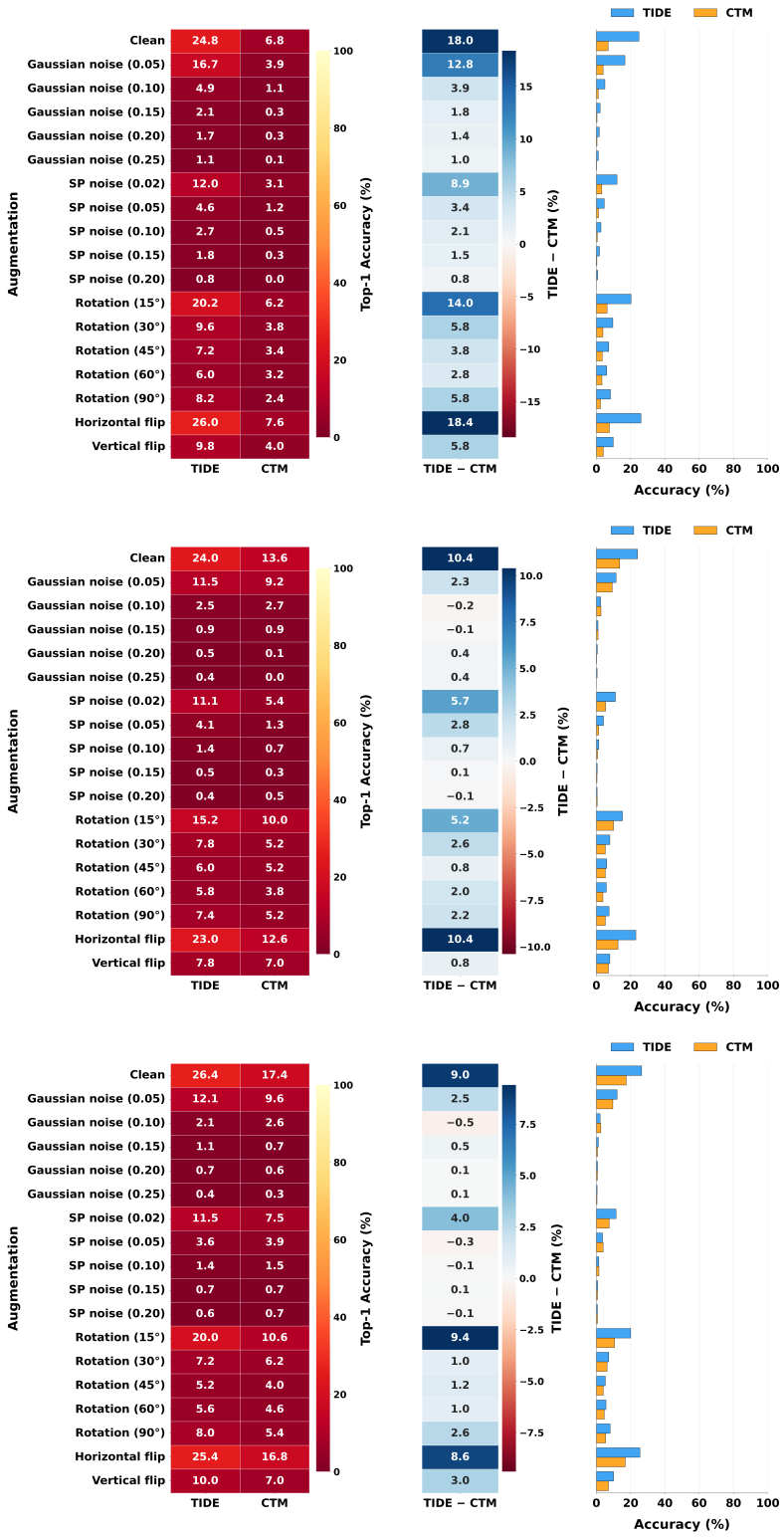

Figure A.17: OOD analysis on `Tiny-ImageNet` [49] in the 100 K (top), 200 K (middle) and 500 K (bottom) comparisons. *Left* panel present `top-1` (%) for TIDE and CTM, *center* panel reports the differences and *right* panel is the accuracy bar.

evaluated zero-shot on 17 corruption augmentations applied post-upsampling, thereby enabling a comparison of architectural robustness against the bilinear $64 \rightarrow 224$ upsampling artifact under additional augmentations. Overall, TIDE has higher mean `top-1` accuracy compared to CTM over all augmentations for all training budgets: 5.5% at 100 K (8.0% against 2.4%), 2.1% at 200 K (6.3% against 4.1%) and 1.9% at 500 K (6.8% against 4.9%). Moreover, as illustrated in Figure A.17, TIDE outperforms CTM on all augmentations, specifically horizontal flip and 15° rotation. TIDE retains 32.1%, 26.1% and 25.8% of its clean `top-1` accuracy against 35.8%, 30.3% and 28.0% for CTM, where retention is defined as the mean `top-1` accuracy over all the perturbations divided by the clean `top-1`.

H Open Questions & Limitations

In this section, we present several remaining open issues and highlight the limitations of the proposed framework. Moreover, future work must address these issues both theoretically and empirically to further investigate the credibility of asymmetric-based architectures in achieving integrated robustness under a variety of domains.

Global strict-convexity in TIDE: Given that Theorem A.3 assumes $\mathcal{L}$ is μ-strongly convex over $\mathcal{W}_{\mathrm{Dale}}$, which does not hold for the full non-convex training loss, the $\mathcal{O}(\log 1/\epsilon)$ convergence rate carries over only locally to a strict local minimum. Therefore, at most the convergence guarantee applies only to a local minimum with a positive-definite Hessian.

Exact LDS for TIDE at convergence: Corollary C.5 is sufficient for condition monitoring at runtime, and for `ImageNet-1K`, it has been observed that the soft spectral discussed in Appendix G.7 stabilizes at its target Perron radii throughout training, which is consistent with LDS being at convergence. However, we do not prove that this must hold, and the symmetrized effective matrix $\frac{1}{2}(W_{\mathrm{eff}} + W_{\mathrm{eff}}^{\top}) - I$ must be negative-definite. Thus, leaving the training-time LDS enforcement as an open issue that requires further investigation.

Balance regime: Given that the Definition A.7 is formal, and not derived directly from first principles for TIDE, it remains open whether TIDE dynamics can drive the model into the region defined by [26].

Rotation equivariance of HRF: As noted in Remark D.4, the implemented HRF backbones are not rotation equivariant, and only Stage zero of the HRF backbone under DoG initialization can be considered as rotation equivariant. To achieve rotation equivariance, a multi-scale bank for the shallow HRF and a per-stage bank for the deep HRF are required. Moreover, a rotation-equivariant deep HRF would need additional steerable filter banks in all its residual stages.

Internal computation steps: It has been observed that increasing T as in Appendix G.4 causes instability rather than improvement. A principled scaling of T with sequence length requires further investigation for a more adaptive processing of input data.

Limitations: The main limitation of both TIDE and CTM is their required sequential feedforward data processing, which prohibits shorter runtime and prevents more robust investigation. Therefore, further investigation is required to assess the feasibility of leveraging neuromorphic hardware and its dynamics, not only to enhance data preprocessing but also to implement a Look-up-Table-based approach to further reduce the training duration of the proposed framework. Given that we have already lowered the training boundary by around $50\,\%$ to achieve results comparable to CTM, indicating the soundness of our approach, further studies are still needed to assess the reachability of an even sparser representation. Additionally, due to the lack of comparable implementations using similar approaches, detailed benchmarking within the same architectural family remains limited and challenging. Although CTM uses LSTM as its baseline, it is not a direct comparison; however, their results can be directly translated to TIDE given that we use CTM as our baseline.