Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contextual Activation Functions: A Unified Framework for Context-Dependent Neural Computation in Temporal Systems

Mohamed Ben Ammar Chraiti

magneo.project@gmail.com

Research Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11271506/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Contextual Activation Functions: A Unified Framework for Context-Dependent Neural Computation in Temporal Systems

Mohamed Ben Ammar Chraiti[1*]

[1]Unaffiliated Researcher, Tunis, Tunisia.

Corresponding author(s). E-mail(s): magneo.project@gmail.com;

Standard neural networks rely on fixed elementwise activation functions, which assume that the nonlinear response is stationary and independent of context. Yet many real-world tasks are inherently regime-dependent: the appropriate computation may vary with external signals such as time, system state, or task identity, as in sequential modeling, adaptive control, and multi-task learning.

We introduce ***contextual activation functions*** **(CAF)** as a foundational **family of adaptive activation maps** in which the nonlinear response becomes explicitly driven by an exogenous context. Given a pre-activation vector $a \in \mathbb{R}^H$ and a context vector $s \in \mathbb{R}^S$, the canonical **pure CAF** operator is

$$\Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a), \qquad g(a, s) = \sigma\big(W_g[a; s] + b_g\big) \in (0, 1)^H,$$

where ρ is a scalar base activation applied componentwise and g is a learnable sigmoid gate. Rather than proposing a single model, we *establish the fundamental analytic properties of this family*: we derive the Jacobian of the operator—which decomposes into a gated diagonal term and a gate-induced coupling term—and use it to characterize local sensitivity, gradient propagation, and conditioning. Under separate smoothness, boundedness, and Lipschitz assumptions, we establish the corresponding regularity and stability properties, and we characterize gradient attenuation in saturated gating regimes. From an approximation-theoretic standpoint, on every compact product domain $K_a \times K_s$, the uniform closure of the affine-gate pure-CAF family contains the standard activation map and also contains genuinely context-dependent maps. One-hidden-layer pure CAF networks preserve universal approximation; more

substantively, universality already holds for the restricted subfamily in which the base pre-activation depends only on the primary input x and the gate depends only on the context s. We characterize standard activations as limiting cases and describe context-only, diagonal, and full gates as structured members of the CAF family, and we quantify the parametric and computational overhead, which scales as $O(H^2 + HS)$ per layer for full affine gating and linearly in the context dimension S. We finally derive an exact finite-sample risk identity for a target-profiled linearized surrogate used to compare an exogenous gate channel with a fitted self-gated member. Under explicit profiling, transportability, covariance, regularization, and Taylor-remainder assumptions, the identity yields a local selection rule. Transfer to an actually re-optimized nonlinear network additionally requires a profile-alignment condition, and prospective empirical validation is left to future work. The criterion separates the target- distribution approximation gain, ridge shrinkage bias, and estimation variance for the profiled linearized surrogate under explicit profiling and transportability assumptions. Its transfer to the nonlinear retrained network requires an additional profile-alignment bound and a controlled Taylor remainder. We provide estimable diagnostics and a sample-splitting protocol, while leaving prospective empirical validation of the criterion to future work. As a behavioral study rather than a performance race, we also evaluate recurrent CAF forecasters on Solar, Bike, EPF-NP, and one target sensor from METR-LA under history-only and known-future-context contracts, using local recurrent references, persistence-matched specialized architectures, structural ablations, and a state-preserving extension; the results show dataset- and regime-dependent competitiveness rather than universal predictive dominance.
Keywords— Contextual activation functions; Gated nonlinearities; Conditional computation; Universal approximation; Gradient flow; Regime-switching.

1 Introduction

Motivation.

Deep networks are commonly built as compositions of affine transformations and fixed elementwise nonlinearities. This design is conceptually simple and computationally efficient, but it imposes a strong stationarity assumption: the nonlinear response of each unit is identical across tasks, environments, or operating regimes. In many applications, however, the input–output relationship is inherently context-dependent. Examples include multi-task learning (where a task descriptor selects a computational regime), sequential modeling (where latent system state changes over time), and control or perception systems operating under varying conditions. These settings motivate architectures in which nonlinear computation can adapt to context in a structured and learnable way.

The CAF framework.

We propose the *contextual activation function* (CAF) framework (Definition 1), in which an activation map depends explicitly on a context vector $s \in \mathbb{R}^S$. A mapping

belongs to the CAF framework if it takes the form

$$\Phi(a, s) = g(a, s) \odot \psi(a), \qquad a \in \mathbb{R}^H, \quad s \in \mathbb{R}^S,$$

where $\psi : \mathbb{R}^H \to \mathbb{R}^H$ is a base response map and $g(a, s) \in (0, 1)^H$ is a context-dependent gate. This framework is intentionally broad: it captures the central principle of context-dependent modulation of a base response without committing to a specific parameterization of either ψ or g. Our contribution is not a single instance of this form but a characterization of the *family* it defines and of the properties shared by its members.

Within this framework, the theoretical analysis is anchored on the *pure CAF* operator (Definition 2), which constitutes the canonical member:

$$\Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a), \qquad g(a, s) = \sigma\big(G([a; s])\big),$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a scalar base activation applied componentwise, σ is the logistic sigmoid applied componentwise, and $[\cdot;\cdot]$ denotes vector concatenation. The term *pure* emphasizes that the output consists solely of the contextually modulated nonlinear response, without any residual or state-preserving pathway.

Under the affine gate parameterization $G([a; s]) = W_g^{(a)} a + W_g^{(s)} s + b_g$ (Section 2), the pure operator gives rise to a *structured variant hierarchy* (Section 2.6): all three cases share the same output rule and differ only in gate structure.

- **Context-only** ($W_g^{(a)} = 0$): the gate depends exclusively on context; the pre-activation influences the output solely through $\rho(a)$.
- **Diagonal** ($W_g^{(a)} = \text{diag}(w)$): componentwise self-modulation is added to context conditioning.
- **Full** ($W_g^{(a)}$ unrestricted): cross-unit interactions in the gate pre-activation are allowed.

For sequential settings, the pure operator extends naturally to the *temporal pure CAF* (Definition 3), in which the gate additionally depends on a compact temporal summary u_t of recent hidden states:

$$z_t = W_g^{(a)} a_t + W_g^{(u)} u_t + W_g^{(s)} s_t + b_g, \qquad y_t = \sigma(z_t) \odot \rho(a_t).$$

This temporal form makes explicit a distinction we return to throughout the paper: the gate receives an *endogenous* context (a_t, u_t), summarizing the unit's own recent dynamics, and an *exogenous* context s_t of external covariates. Beyond pure CAF and its variants, the framework also accommodates *extensions built on the CAF principle*, which modify the output rule itself while retaining an explicit context-modulated nonlinear contribution. Two representative extensions are analyzed: *residual CAF*,

$$\Phi_{\text{res}}(a, s) = a + \alpha\, g(a, s) \odot \rho(a),$$

which combines the modulated response with an identity pathway, and *state-preserving CAF*,

$$h_t = \beta_t \odot h_{t-1} + (\mathbf{1} - \beta_t) \odot \big(g_t \odot \rho(a_t)\big),$$

which interpolates the CAF response with the previous hidden state in recurrent settings. Throughout the paper, these extensions are treated separately from pure CAF and its variants, as they require a distinct theoretical treatment. Unless explicitly stated otherwise, the analytic results on operator properties, gradient flow, conditioning, expressivity, and complexity refer to the pure CAF operator; extensions such as residual and state-preserving CAF are discussed as separate architectural directions requiring their own analysis.

Related work and positioning.

Learnable activation functions can be organized by the signal that controls their adaptation. A first class learns shape parameters that become fixed after training, including adaptive piecewise-linear units, PReLU, and globally or locally adaptive slopes [Agostinelli et al.(2015)Agostinelli, Hoffman, Sadowski, and Baldi, He et al.(2015)He, Zhang, Ren, and Sun, Jagtap et al.(2020a)Jagtap, Kawaguchi, and Karniadakis, Jagtap et al.(2020b)Jagtap, Kawaguchi, and Karniadakis]. A second class adapts to the current unit value or feature representation itself, as in Swish-β, ACON-C, Meta-ACON, Dynamic ReLU, and broader multiplicative-interaction mechanisms [Ramachandran et al.(2017)Ramachandran, Zoph, and Le, Ma et al.(2021)Ma, Zhang, Liu, and Sun, Chen et al.(2020)Chen, Dai, Liu, Chen, Yuan, and Liu, Jayakumar et al.(2020)Jayakumar, Czarnecki, Menick, Schwarz, Rae, Osindero, Teh, Harley, and Pascanu]. CAF occupies a third *exogenous-context* category when its gate uses an explicitly identified descriptor s supplied separately from the pre-activation; its full and temporal forms may additionally combine this exogenous signal with endogenous pre-activation or state summaries. Thus, the defining distinction is not merely that the activation is learnable, but that an exogenous contextual variable is part of the activation map itself.

CAF also connects to several established forms of conditional computation, but differs in *where* context enters the computation and *what* it modulates. FiLM and conditional normalization apply context-dependent affine transformations to intermediate feature maps [Perez et al.(2018)Perez, Strub, de Vries, Dumoulin, and Courville, Dumoulin et al.(2017)Dumoulin, Shlens, and Kudlur, Huang and Belongie(2017)]. A context-only pure CAF can be viewed algebraically as a restricted FiLM transformation applied to $\rho(a)$ when the FiLM shift is zero and the scale is a sigmoid gate. The substantive distinction appears when the gate also depends on the pre-activation, when structural restrictions on $W_g^{(a)}$ control diagonal or cross-unit coupling, and when the activation-level Jacobian is the object of analysis. SE-type mechanisms reweight channels using input-dependent summary statistics [Hu et al.(2018)Hu, Shen, and Sun]. Gated Linear Units and related multiplicative interaction models apply elementwise gates to representations [Dauphin et al.(2017)Dauphin, Fan, Auli, and Grangier, Wu et al.(2016)Wu, Zhang, Zhang, Bengio, and Salakhutdinov]; some overlap algebraically with CAF-style products, but they do not provide the same

activation-centric framework, structured gate taxonomy, or systematic Jacobian analysis. HyperNetworks generate weights as functions of a conditioning signal [Ha et al.(2017)Ha, Dai, and Le], introducing context dependence at the parameter level rather than at the level of the nonlinear response. Mixture-of-Experts architectures perform conditional computation through routing across subnetworks [Jacobs et al.(1991)Jacobs, Jordan, Nowlan, and Hinton, Shazeer et al.(2017)Shazeer, Mirhoseini, Maziarz, Davis, Le, Hinton, and Dean], operating at a coarser granularity than the unit-level modulation of CAF. In recurrent architectures, LSTM and GRU gates modulate state transitions and memory flow across time [Hochreiter and Schmidhuber(1997), Cho et al.(2014)Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk, and Bengio]: these act at the level of recurrent state updates and are complementary to, rather than interchangeable with, CAF. A detailed comparison is provided in Section 8. In contrast, CAF is formulated explicitly at the level of the activation map $\Phi(a, s)$ itself, enabling a direct theoretical analysis of how context reshapes the local nonlinearity, its Jacobian structure, and the backward gradient signal. Specialized forecasting architectures such as DeepAR [Salinas et al.(2020)Salinas, Flunkert, Gasthaus, and Januschowski], Informer [Zhou et al.(2021)Zhou, Zhang, Peng, Zhang, Li, Xiong, and Zhang], and N-BEATS [Oreshkin et al.(2020)Oreshkin, Carpov, Chapados, and Bengio] provide relevant points of comparison for sequential modeling, but their probabilistic and complexity-oriented foundations are *orthogonal* to the activation-level question studied here. They situate CAF conceptually, and selected representatives are also used as empirical comparators under the common information and persistence contracts defined in Section 13; these comparisons assess complete forecasting systems rather than claiming that CAF must outperform them.

What is new.

CAF deliberately builds on standard primitives—elementwise nonlinearities, sigmoid gates, and multiplicative modulation—but uses them as a design principle for activation-level adaptation driven by an explicit exogenous context. To our knowledge, the distinctive contribution is to formulate this as a family of adaptive activations in which the external variable s enters the activation map itself, to separate s from endogenous pre-activation and state signals, and to organize context-only, diagonal, and fully coupled gates within a common parameterization. The analysis then characterizes the resulting activation Jacobian, gradient and saturation behavior, approximation properties, and computational overhead. Extensions that alter the canonical output rule $g(a, s) \odot \rho(a)$ are kept separate so that these claims are not conflated with broader mechanisms for conditional computation.

Contributions.

The main contributions are ordered by decreasing theoretical weight. The local selection criterion remains a theoretical contribution, whereas the final contribution reports a separate empirical evaluation of the CAF forecaster and its variants.

C1. CAF framework and canonical formulation (Section 2). We introduce the CAF framework (Definition 1) and its canonical member, the pure CAF operator (Definition 2), together with its structured variant hierarchy (context-only, diagonal, full), its temporal extension (Definition 3), and its extensions built on the CAF principle (residual CAF, state-preserving CAF). We establish the relationship to standard activations as limiting cases (Proposition 1).

C2. Analytic properties of pure CAF (Section 3). We establish C^r smoothness (Theorem 3), derive explicit Jacobian formulas decomposing the pre-activation and context sensitivity (Proposition 5), prove global Lipschitz continuity (Proposition 7), and characterize local monotonicity, clarifying how contextual gating changes local conditioning relative to standard activations.

C3. Gradient-flow analysis (Section 4). We decompose the backward signal into a gated diagonal term and a gate-induced cross-component coupling term, and derive a local Jacobian-norm bound for affine gating (Theorem 9). We formally characterize gradient suppression under gate saturation near zero (Proposition 13) and asymptotic recovery of the base-activation Jacobian under saturation near one (Proposition 14), and characterize gradient suppression at the gate-parameter level (Proposition 15).

C4. Expressivity and universal approximation (Sections 5 and 5.3). We show that, on every compact product domain $K_a \times K_s$, the uniform closure of the affine-gate pure-CAF family contains the standard activation map and that the family also contains genuinely context-dependent maps (Corollary 19). We retain the exact embedding of standard one-hidden-layer networks as a short lemma and prove the stronger result that universality already holds for the context-only affine-gate subfamily, with base pre-activations depending only on x and gates depending only on s (Theorem 21).

C5. Parametric and computational complexity (Section 9). We quantify additional parameters and computation relative to a standard layer for each structured gate variant. For the full affine gate, the overhead scales as $O(H^2+HS)$ per layer, with the context-dimension dependence being linear in S (Proposition 24). Structured restrictions (context-only, diagonal) yield $O(HS)$ and $O(H + HS)$ overheads respectively.

C6. Local profiled criterion for exogenous-channel selection (Section 12). We derive an exact finite-sample risk identity for a target-profiled linearized surrogate used to compare the $+s$ and $-s$ members. The identity separates the target-distribution approximation gain $\mathscr{I}_s$, ridge shrinkage bias $\mathscr{B}_s(\lambda)$, and estimation variance $\mathscr{V}_s(\lambda)$. We state explicit profiling, transportability, covariance, regularization, and Taylor-remainder assumptions. An additional profile-alignment assumption is required to transfer the result from the profiled surrogate to an actually re-optimized nonlinear network. We provide estimable diagnostics and a sample-splitting or cross-fitting protocol, while leaving prospective empirical validation to future work.

C7. Empirical evaluation of recurrent CAF forecasters (Section 13). We evaluate canonical CAF, structured gate restrictions, initialization and persistence variants, additive-routing controls, and a state-preserving extension

on Solar, Bike, EPF-NP, and one target sensor from METR-LA. The protocol separates history-only from known-future-context information, distinguishes local persistence-sharing references from persistence-matched specialized architectures, and reports seed-wise uncertainty with multiplicity-corrected paired comparisons. The results establish practical viability in selected dataset–regime configurations while documenting adverse and neutral cases and without claiming universal superiority or prospective validation of the profiled selection criterion.

2 Mathematical Formulation

2.1 Basic Definitions

We first formalize contextual activation mechanisms in general terms and then introduce the canonical multiplicative formulation used in the analysis below.

Definition 1 (CAF framework) Let $H, S \in \mathbb{N}$. A **context-adaptive activation** is a mapping

$$\Phi : \mathbb{R}^H \times \mathbb{R}^S \to \mathbb{R}^H$$

that transforms a pre-activation vector $a \in \mathbb{R}^H$ under a context vector $s \in \mathbb{R}^S$, producing an output

$$y = \Phi(a, s) \in \mathbb{R}^H.$$

A mapping Φ is said to belong to the **CAF framework** if there exist maps

$$\psi : \mathbb{R}^H \to \mathbb{R}^H, \qquad g : \mathbb{R}^H \times \mathbb{R}^S \to [0,1]^H,$$

such that

$$\Phi(a, s) = g(a, s) \odot \psi(a),$$

where $\odot$ denotes the Hadamard product. To exclude trivial degenerate cases, we additionally assume that ψ is not identically constant.

Definition 1 is intentionally broad. It captures the central principle of the framework—namely, the context-dependent modulation of a base response map—without committing to a specific parameterization of either ψ or the gate g. This distinction is important because the theoretical analysis developed below focuses on a canonical member of the framework, whereas other structured variants can be introduced as extensions that preserve the same underlying principle. A CAF map is called strictly contextual on a domain D if there exist $(a, s_1), (a, s_2) \in D$ and an index $i \in \{1, \ldots, H\}$ such that $s_1 \neq s_2$, $g_i(a, s_1) \neq g_i(a, s_2)$, and $\psi_i(a) \neq 0$.

In the analysis below, we restrict attention to the C^1 setting. The context vector s may encode system state, temporal information, task descriptors, control variables, or other auxiliary signals relevant to the computation. Although the present analysis is stated for continuously differentiable activations, piecewise-smooth choices such as ReLU may also be treated almost everywhere, with the usual caveats at nondifferentiable points.

2.2 Canonical Pure Formulation

We now introduce the canonical pure formulation that serves as the primary analytical case.

Definition 2 (Pure CAF) Let $\rho : \mathbb{R} \to \mathbb{R}$ be a scalar base activation function, applied componentwise to vectors, and let

$$G : \mathbb{R}^{H+S} \to \mathbb{R}^H$$

be a learnable gate pre-activation map. Define

$$z(a, s) := G([a; s]) \in \mathbb{R}^H, \tag{1}$$

$$g(a, s) := \sigma(z(a, s)) \in (0, 1)^H, \tag{2}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid applied componentwise. The **pure CAF** operator is defined by

$$\Phi_{\text{pure}}(a, s) := g(a, s) \odot \rho(a) \in \mathbb{R}^H. \tag{3}$$

The pure CAF operator is the canonical formulation used for the theoretical development of the framework. Its structure separates two roles: the base activation $\rho(a)$ provides the underlying nonlinear response, while the gate $g(a, s)$ acts as a context-dependent gain that modulates this response componentwise. When $g_i(a, s) \approx 1$, the i-th component behaves similarly to the underlying activation ρ; when $g_i(a, s) \approx 0$, the corresponding response is strongly attenuated.

The term "pure" emphasizes that the output consists only of the contextually modulated nonlinear response, without any additional residual or state-preserving pathway. This formulation therefore provides a tractable setting for the mathematical analysis of the broader CAF framework.

Notation and standing assumptions.

Throughout the paper, $H, S \in \mathbb{N}$ denote positive integers, and vectors are understood as column vectors. For $v \in \mathbb{R}^n$, $\text{diag}(v)$ denotes the $n \times n$ diagonal matrix with diagonal entries $v_1, \ldots, v_n$. For matrices A, $\|A\|$ denotes the spectral norm (operator 2-norm). For vectors v, $\|v\|$ denotes the Euclidean norm and $\|v\|_\infty = \max_i |v_i|$ the infinity norm. For a differentiable vector-valued function $f : \mathbb{R}^m \to \mathbb{R}^n$, $Df(x)$ denotes the Jacobian matrix (an $n \times m$ matrix) with entries

$$(Df(x))_{ij} = \frac{\partial f_i}{\partial x_j}(x).$$

We use the following notation:

- $a \in \mathbb{R}^H$ is the pre-activation vector, and $s \in \mathbb{R}^S$ is the context vector.
- By a standard abuse of notation, $\rho(a) \in \mathbb{R}^H$ denotes the componentwise application of the scalar activation ρ to a, and $\rho'(a) \in \mathbb{R}^H$ denotes the vector of componentwise derivatives.

- Likewise, $\sigma(z) \in (0,1)^H$ denotes the componentwise application of the logistic sigmoid to a vector $z \in \mathbb{R}^H$.
- $z(a,s) = G([a;s]) \in \mathbb{R}^H$ is the gate pre-activation, and $g(a,s) = \sigma(z(a,s)) \in (0,1)^H$ is the gate output.
- $D_a G([a;s]) \in \mathbb{R}^{H\times H}$ and $D_s G([a;s]) \in \mathbb{R}^{H\times S}$ denote the Jacobian blocks of G with respect to a and s, respectively. Here $D_a G([a;s])$ denotes the Jacobian of G with respect to the first H coordinates of $[a;s]$, whereas $D_s G([a;s])$ denotes the Jacobian with respect to the last S coordinates.
- $\mathbf{1} \in \mathbb{R}^H$ denotes the all-ones vector.

Unless otherwise stated, we assume that $\rho \in C^1(\mathbb{R})$ and $G \in C^1(\mathbb{R}^{H+S})$. Additional boundedness assumptions are introduced only when required for global estimates.

Remark 1 (Scope of regularity statements) All statements invoking C^r or classical Jacobians require the stated smoothness assumptions on ρ and G. If ρ is chosen as ReLU, the corresponding differentiability statements are understood almost everywhere, and the formulas at the nondifferentiable points are not interpreted as classical derivatives. Lipschitz and contraction results may still be applied to ReLU whenever they rely only on its global Lipschitz constant rather than on classical differentiability.

In particular, when global norm bounds are needed, we assume either:

1. ρ is globally bounded and there exist constants $B_\rho, L_\rho \geq 0$ such that

$$|\rho(x)| \leq B_\rho, \qquad |\rho'(x)| \leq L_\rho \quad \text{for all } x \in \mathbb{R},$$

   or
2. the analysis is restricted to bounded subsets of $\mathbb{R}^H$, on which corresponding local bounds hold for ρ and ρ'.

When needed, we also assume either that

$$\sup_{(a,s)\in\mathbb{R}^{H+S}} \|DG([a;s])\| \leq L_G,$$

or that the analysis is restricted to bounded subsets on which $\|DG\|$ admits a corresponding local bound.

Under the global boundedness assumption in item (1), the componentwise extension of ρ satisfies

$$\|\rho(a)\|_\infty \leq B_\rho, \qquad \|\rho'(a)\|_\infty \leq L_\rho \quad \text{for all } a \in \mathbb{R}^H.$$

For affine G, one may take $L_G = \|W_g\|$.

Unless explicitly stated otherwise, all theoretical statements in the remainder of the paper refer to the pure CAF operator (3).

2.3 Structured Gate Parameterization

For most of the theoretical analysis, we consider an affine gate parameterization of the form

$$G([a; s]) = W_g[a; s] + b_g = W_g^{(a)}a + W_g^{(s)}s + b_g, \tag{4}$$

where

$$W_g \in \mathbb{R}^{H\times(H+S)}, \qquad W_g^{(a)} \in \mathbb{R}^{H\times H}, \qquad W_g^{(s)} \in \mathbb{R}^{H\times S}, \qquad b_g \in \mathbb{R}^H.$$

This parameterization induces an affine dependence on both the pre-activation and the context, followed by sigmoid modulation. It is analytically tractable and captures several structured cases of interest.

In particular, the matrix $W_g^{(a)}$ determines how the gate depends on the pre-activation itself, which yields a natural hierarchy of gate structures:

- **context-only gate:** $W_g^{(a)} = 0$, so the modulation depends exclusively on the context s;
- **diagonal gate:** $W_g^{(a)} = \mathrm{diag}(w)$ for some $w \in \mathbb{R}^H$, so each component is self-modulated independently;
- **unconstrained gate:** $W_g^{(a)}$ is unrestricted, allowing cross-component interactions in the gate pre-activation.

These structured cases are useful both conceptually and for empirical design. They should be understood as instances of the same pure CAF formulation rather than as distinct activation principles. More expressive gate architectures, such as multilayer perceptrons, can be analyzed in a similar way, at the cost of additional nonlinear terms in the derivative calculations.

2.4 Temporal Pure CAF

For sequential settings, the context may include a summary of recent dynamics in addition to the instantaneous external signal.

Definition 3 (Temporal pure CAF) At time t, let $a_t \in \mathbb{R}^H$ be the pre-activation, let $s_t \in \mathbb{R}^S$ be an external context signal, and let $h_{t-m}, \ldots, h_{t-1} \in \mathbb{R}^H$ denote a window of recent hidden states, where $m \geq 1$. Let

$$u_t = \mathscr{T}(h_{t-m}, \ldots, h_{t-1}) \in \mathbb{R}^U,$$

where $\mathscr{T} : (\mathbb{R}^H)^m \to \mathbb{R}^U$ is a temporal summarizer and U denotes its output dimension. A temporal pure CAF is then defined by

$$z_t = W_g^{(a)}a_t + W_g^{(u)}u_t + W_g^{(s)}s_t + b_g \in \mathbb{R}^H, \tag{5}$$

$$g_t = \sigma(z_t), \tag{6}$$

$$y_t = g_t \odot \rho(a_t), \tag{7}$$

where

$$W_g^{(a)} \in \mathbb{R}^{H\times H}, \qquad W_g^{(u)} \in \mathbb{R}^{H\times U}, \qquad W_g^{(s)} \in \mathbb{R}^{H\times S}, \qquad b_g \in \mathbb{R}^H.$$

This formulation extends the pure CAF operator to sequential settings by allowing the gate to depend not only on the instantaneous context s_t but also on a compact representation u_t of recent temporal behavior. In this sense, the total context entering the gate may be viewed as the pair (u_t, s_t). The multiplicative CAF structure is preserved, with the nonlinear response $\rho(a_t)$ being modulated componentwise by a context-dependent gate.

The same structured gate taxonomy introduced above carries over directly to the temporal setting. For example, one may impose $W_g^{(a)} = 0$ to obtain a context-only temporal gate, or restrict $W_g^{(a)}$ to be diagonal to obtain componentwise self-modulation.

For parameter and operation counts, this temporal gate includes the additional block $W_g^{(u)} \in \mathbb{R}^{H \times U}$. Thus, relative to a standard layer, the full temporal gate contributes $O(H^2 + HU + HS)$ gate parameters and multiply-add operations per step, before accounting for the cost of the temporal summarizer $\mathscr{T}$. The non-temporal complexity formulas in Section 9 correspond to the special case without this u_t input, or equivalently to counting only the activation-level pure CAF operator.

2.5 Relationship to Standard Activations and Structured Cases

Standard activations arise as limiting or structured cases of the pure CAF formulation.

Proposition 1 (Limiting and structured cases) *Under the affine gate parameterization (4), the following hold:*

1. *Let $\alpha \to +\infty$. If*
$$W_g^{(a)} = 0, \qquad W_g^{(s)} = 0, \qquad b_g = \alpha \mathbf{1},$$
then, for every fixed $(a, s) \in \mathbb{R}^{H+S}$,
$$g(a, s) = \sigma(\alpha)\mathbf{1} \to \mathbf{1} \quad \text{and} \quad \Phi_{\text{pure}}(a, s) \to \rho(a).$$
Thus, the base activation ρ is recovered as a pointwise limiting case of pure CAF.
2. *If $W_g^{(a)} = 0$, then*
$$g(a, s) = \sigma(W_g^{(s)} s + b_g),$$
so the gate depends only on context and implements context-dependent scaling of the base activation.
3. *If $W_g^{(s)} = 0$, then*
$$g(a, s) = \sigma(W_g^{(a)} a + b_g),$$
so the gate depends only on the pre-activation and implements pre-activation-dependent modulation. If $W_g^{(a)}$ is diagonal, this reduces to componentwise self-modulation.
4. *If both $W_g^{(a)} = 0$ and $W_g^{(s)} = 0$, then*
$$\Phi_{\text{pure}}(a, s) = \sigma(b_g) \odot \rho(a),$$
so the operator reduces to a fixed componentwise rescaling of $\rho(a)$.
5. *Let ρ be the identity, set $W_g^{(s)} = 0$, $b_g = 0$, and $W_g^{(a)} = \beta I_H$ for a scalar β. Then*
$$\Phi_{\text{pure}}(a, s) = a \odot \sigma(\beta a),$$

which is the componentwise Swish-β activation, independent of the context s [Ramachandran et al.(2017)Ramachandran, Zoph, and Le]. Thus Swish-β is an exact diagonal, context-free member rather than merely a limiting case; the scope of the theoretical results for this embedding is specified in Proposition 2.

This exact embedding relates CAF to self-gated customized activations such as Swish and the broader ACON family [Ramachandran et al.(2017)Ramachandran, Zoph, and Le, Ma et al.(2021)Ma, Zhang, Liu, and Sun], while CAF's exogenous channel distinguishes the contextual members considered here.

Proposition 2 (Embedding of adaptive activations) *Several standard adaptive activations can be represented either as pure CAF members or as CAF extensions.*

1. *Swish-β, $a \mapsto a\sigma(\beta a)$, is the diagonal, context-free pure CAF member obtained with identity base activation, $W_g^{(s)} = 0$, $b_g = 0$, and $W_g^{(a)} = \mathrm{diag}(\beta)$. Results whose hypotheses require a bounded or non-polynomial ρ do not apply to this embedding directly; global Lipschitz continuity and universality follow instead from separate arguments.*
2. *PReLU belongs to the pure CAF framework whenever its negative-slope parameter satisfies $\alpha \in [0, 1]$: choose a piecewise base response $\rho_\alpha(a)_i = a_i$ for $a_i \geq 0$ and $\rho_\alpha(a)_i = \alpha a_i$ for $a_i < 0$, and use the constant gate $g \equiv \mathbf{1}$ as a limiting pure-CAF member.*
3. *ACON-C can be written coordinatewise as $(p_1 - p_2)a\,\sigma(\beta(p_1 - p_2)a) + p_2 a$. When $p_1, p_2 \in [0, 1]$, its multiplicative self-gated component is a diagonal, context-free CAF-type component, while the additive linear term places the complete ACON-C formula in a residual or parameterized-response CAF extension.*
4. *The $K = 1$ case of Dynamic ReLU, equivalently a squeeze-and-excitation style channel reweighting, is a context-only CAF member when the routing statistics are treated as the context s and the dynamic bias is absent or absorbed into an extension with an additive term.*
5. *Meta-ACON is a CAF-type self/adaptive gate when the meta-network output that determines β is treated as a context variable, for example s equal to pooled statistics of the input feature representation.*
6. *APL and the adaptive activations of Jagtap et al. are best viewed as instances of the broader CAF principle with a parameterized base response rather than as pure CAF members with the fixed canonical product $g(a, s) \odot \rho(a)$.*

Consequently, the operator-level theorems proved in this article—including the Jacobian, Lipschitz, gradient, conditioning, and universal-approximation results— apply directly only to the members that preserve the pure CAF form and its stated assumptions. Embeddings that require additive terms or a parameterized base response are conceptual inclusions in the broader CAF design principle and require separate verification.

Hence, pure CAF interpolates between standard elementwise activations and richer pre-activation- and context-dependent modulation schemes within a unified parameterization. This canonical form will serve as the mathematical anchor for the subsequent analysis, while more elaborate CAF variants may be viewed as structured extensions of the same underlying principle.

2.6 Variants of Pure CAF and Extensions Built on the CAF Principle

The distinction between *variants of pure CAF* and *extensions built on the CAF principle* is important both conceptually and mathematically. Variants of pure CAF preserve the canonical output rule

$$\Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a),$$

and differ only in the structural form of the gate. By contrast, extensions built on the CAF principle modify the output rule itself while retaining an explicit context-modulated nonlinear contribution.

Variants of pure CAF.

Within the affine gate parameterization

$$g(a, s) = \sigma\big(W_g^{(a)} a + W_g^{(s)} s + b_g\big),$$

the following special cases remain instances of pure CAF, since their output retains the form $\Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a)$.

- **Context-only pure CAF:**

  $$W_g^{(a)} = 0, \qquad \Phi_{\text{ctx}}(a, s) = \sigma\big(W_g^{(s)} s + b_g\big) \odot \rho(a).$$

  In this case, the gate depends only on the context, and the pre-activation influences the output solely through the base activation $\rho(a)$.
- **Diagonal pure CAF:**

  $$W_g^{(a)} = \operatorname{diag}(w), \qquad w \in \mathbb{R}^H,$$

  and

  $$\Phi_{\text{diag}}(a, s) = \sigma\big(\operatorname{diag}(w) a + W_g^{(s)} s + b_g\big) \odot \rho(a).$$

  Equivalently,

  $$\Phi_{\text{diag}}(a, s) = \sigma\big(w \odot a + W_g^{(s)} s + b_g\big) \odot \rho(a).$$

  This variant allows only componentwise self-modulation in the gate pre-activation.
- **Unconstrained pure CAF:**

  $$W_g^{(a)} \in \mathbb{R}^{H \times H} \text{ arbitrary}, \qquad \Phi_{\text{unc}}(a, s) = \sigma\big(W_g^{(a)} a + W_g^{(s)} s + b_g\big) \odot \rho(a).$$

  This is the most expressive affine-gate instance of pure CAF, allowing cross-component interactions in the gate pre-activation.

These operators should be understood as *variants of pure CAF*, not as different activation principles. They share the same multiplicative output rule and differ only in the structural constraints imposed on the gate.

Extensions built on the CAF principle.

An operator is regarded here as an extension built on the CAF principle when the output rule is modified beyond the canonical pure form, while preserving an explicit context-modulated nonlinear contribution.

A general extension may be written abstractly as

$$\Phi_{\mathrm{ext}}(a, s) = \mathscr{A}\big(a, \rho(a), g(a, s)\big),$$

where $\mathscr{A} : \mathbb{R}^H \times \mathbb{R}^H \times (0,1)^H \to \mathbb{R}^H$ combines the modulated nonlinear response with additional pathways, residual terms, or state variables, and is no longer restricted to the pure multiplicative form

$$\mathscr{A}(a, \rho(a), g) = g \odot \rho(a).$$

Two representative extensions are as follows.

- **Residual CAF:**

  $$\Phi_{\mathrm{res}}(a, s) = a + \alpha\, g(a, s) \odot \rho(a), \qquad \alpha \geq 0.$$

  Here the contextually modulated nonlinear response is combined with an identity pathway on the pre-activation. This preserves the CAF modulation mechanism while extending the pure operator with direct signal preservation.
- **State-preserving CAF:** In sequential settings, one may define

  $$h_t = \beta_t \odot h_{t-1} + (\mathbf{1} - \beta_t) \odot \big(g_t \odot \rho(a_t)\big), \qquad h_t, h_{t-1} \in \mathbb{R}^H,$$

  where

  $$g_t := g(a_t, s_t) \in (0,1)^H$$

  and

  $$\beta_t \in (0,1)^H$$

  is an additional gate, for instance produced by a separate gating map depending on the current pre-activation and the relevant context variables. In this case, the CAF response is interpolated with the previous state, yielding a memory-preserving extension of the pure operator.

The preceding distinction will be maintained throughout the paper. When the output retains the form

$$\Phi_{\mathrm{pure}}(a, s) = g(a, s) \odot \rho(a),$$

we speak of *variants of pure CAF*. When the output rule itself is altered, for example through residual addition or state-preserving interpolation, we speak of *extensions built on the CAF principle*. This separation is important both conceptually and mathematically: the former preserve the canonical analytical form, whereas the latter generally require a distinct theoretical treatment.

3 Fundamental Properties

In this section, all results refer to the pure CAF operator introduced in Section 2. For convenience, we write

$$z(a,s) := G([a;s]), \qquad g(a,s) := \sigma(z(a,s)), \qquad y(a,s) := \Phi_{\text{pure}}(a,s) = g(a,s) \odot \rho(a).$$

3.1 Smoothness and Differentiability

Theorem 3 (Differentiability) *Let $\rho : \mathbb{R} \to \mathbb{R}$ be of class C^r and let $G : \mathbb{R}^{H+S} \to \mathbb{R}^H$ be of class C^r for some $r \geq 1$. Then the pure CAF operator*

$$\Phi_{\text{pure}}(a,s) = \sigma(G([a;s])) \odot \rho(a)$$

is of class C^r as a function of $(a,s) \in \mathbb{R}^{H+S}$.

Proof Let $\bar{\rho} : \mathbb{R}^H \to \mathbb{R}^H$ and $\bar{\sigma} : \mathbb{R}^H \to \mathbb{R}^H$ denote the componentwise extensions of ρ and σ, respectively. Since $\rho \in C^r(\mathbb{R})$, the map $\bar{\rho}$ is of class C^r. Since $\sigma(x) = 1/(1+e^{-x})$ is C^∞ on $\mathbb{R}$, the map $\bar{\sigma}$ is C^∞ on $\mathbb{R}^H$.

Let $\pi_a : \mathbb{R}^{H+S} \to \mathbb{R}^H$ be the projection $\pi_a(a,s) = a$. Then $(a,s) \mapsto \bar{\rho}(\pi_a(a,s)) = \bar{\rho}(a)$ is of class C^r. Likewise, since G is C^r, the composition $(a,s) \mapsto \bar{\sigma}(G([a;s]))$ is of class C^r. Finally, the Hadamard product of two C^r vector-valued functions is again of class C^r, hence

$$\Phi_{\text{pure}}(a,s) = \bar{\sigma}(G([a;s])) \odot \bar{\rho}(a)$$

is of class C^r. □

Corollary 4 For the standard choice $\rho(x) = \tanh(x)$ and affine G, the pure CAF operator is C^∞ in both a and s.

Proof The function tanh is C^∞ on $\mathbb{R}$, hence its componentwise extension is C^∞ on $\mathbb{R}^H$. An affine map G is C^∞, and the sigmoid is C^∞. Therefore, by closure of C^∞ under composition and Hadamard products,

$$\Phi_{\text{pure}}(a,s) = \sigma(G([a;s])) \odot \tanh(a)$$

is C^∞. □

Smoothness ensures that gradients are well defined and continuous, which facilitates the theoretical analysis and supports stable gradient-based optimization.

3.2 Jacobian Structure

Proposition 5 (Jacobian formulas) *Let*

$$z(a,s) := G([a;s]), \qquad g(a,s) := \sigma(z(a,s)), \qquad y(a,s) := \Phi_{\text{pure}}(a,s) = g(a,s) \odot \rho(a).$$

Then

$$D_a y = \operatorname{diag}(\rho(a))\, D_a g + \operatorname{diag}(g)\, \operatorname{diag}(\rho'(a)), \tag{8}$$

$$D_s y = \operatorname{diag}(\rho(a))\, D_s g, \tag{9}$$

where

$$D_a g = \operatorname{diag}\big(g \odot (\mathbf{1} - g)\big)\, D_a G([a; s]), \tag{10}$$

$$D_s g = \operatorname{diag}\big(g \odot (\mathbf{1} - g)\big)\, D_s G([a; s]). \tag{11}$$

In particular, under the affine gate parameterization (4),

$$D_a y = \operatorname{diag}(\rho(a))\operatorname{diag}\big(g \odot (\mathbf{1} - g)\big) W_g^{(a)} + \operatorname{diag}(g)\operatorname{diag}(\rho'(a)), \tag{12}$$

$$D_s y = \operatorname{diag}(\rho(a))\operatorname{diag}\big(g \odot (\mathbf{1} - g)\big) W_g^{(s)}. \tag{13}$$

Proof The identities (10) and (11) follow from the chain rule and the scalar derivative formula

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

Next, for each $i \in \{1, \dots, H\}$,

$$y_i(a, s) = g_i(a, s)\rho(a_i).$$

Differentiating componentwise with respect to a gives

$$\frac{\partial y_i}{\partial a_j} = \frac{\partial g_i}{\partial a_j}(a, s)\rho(a_i) + g_i(a, s)\rho'(a_i)\delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. This yields (8). Differentiating with respect to s gives

$$\frac{\partial y_i}{\partial s_j} = \frac{\partial g_i}{\partial s_j}(a, s)\rho(a_i),$$

which yields (9). Substituting $D_a G([a; s]) = W_g^{(a)}$ and $D_s G([a; s]) = W_g^{(s)}$ in the affine case gives (12) and (13). □

The Jacobian with respect to a therefore decomposes into a direct activation term and a gate-feedback term. The latter is absent in standard elementwise activations and captures the additional sensitivity induced by context-dependent gating.

3.3 Boundedness

Proposition 6 (Output bounds) *Assume that ρ is globally bounded, i.e., there exists $B_\rho > 0$ such that*

$$|\rho(x)| \le B_\rho \qquad \text{for all } x \in \mathbb{R}.$$

Then

$$\|\Phi_{\text{pure}}(a, s)\|_\infty \le B_\rho \qquad \text{for all } (a, s) \in \mathbb{R}^{H+S}.$$

Proof Since $g(a, s) \in (0, 1)^H$ and, componentwise, $|\rho(a_i)| \le B_\rho$, each component of $y = \Phi_{\text{pure}}(a, s)$ satisfies

$$|y_i| = |g_i(a, s)\rho(a_i)| \le |\rho(a_i)| \le B_\rho.$$

Taking the maximum over i yields the claim. □

For unbounded base activations, the gate still scales amplitudes by factors in $(0, 1)$, but no uniform global output bound follows without additional assumptions. On bounded subsets of $\mathbb{R}^H$, corresponding local bounds may still be obtained.

3.4 Lipschitz Continuity

Proposition 7 (Global Lipschitz continuity) *Assume that ρ is globally Lipschitz with constant L_ρ, that ρ is globally bounded by B_ρ, and that G is globally Lipschitz on $\mathbb{R}^{H+S}$ with constant L_G. Then the pure CAF operator*

$$\Phi_{\text{pure}}(a, s) = \sigma(G([a; s])) \odot \rho(a)$$

is globally Lipschitz in (a, s) (with respect to the Euclidean norm), with

$$\|\Phi_{\text{pure}}(a, s) - \Phi_{\text{pure}}(a', s')\| \leq \left(L_\rho + \frac{B_\rho L_G}{4}\right) \|[a; s] - [a'; s']\|$$

for all $(a, s), (a', s') \in \mathbb{R}^{H+S}$.

Proof Let $\xi = [a; s]$ and $\xi' = [a'; s']$. Let $\pi_a(\xi) = a$ and define

$$r(\xi) := \rho(\pi_a(\xi)) = \rho(a), \qquad g(\xi) := \sigma(G(\xi)).$$

Then

$$\Phi_{\text{pure}}(\xi) - \Phi_{\text{pure}}(\xi') = g(\xi) \odot \big(r(\xi) - r(\xi')\big) + \big(g(\xi) - g(\xi')\big) \odot r(\xi').$$

Using $\|u \odot v\| \leq \|u\|_\infty \|v\|$, together with $\|g(\xi)\|_\infty \leq 1$ and $\|r(\xi')\|_\infty \leq B_\rho$, we obtain

$$\|\Phi_{\text{pure}}(\xi) - \Phi_{\text{pure}}(\xi')\| \leq \|r(\xi) - r(\xi')\| + B_\rho \|g(\xi) - g(\xi')\|.$$

Since the scalar activation ρ is L_ρ-Lipschitz, its componentwise extension to $\mathbb{R}^H$ is also L_ρ-Lipschitz with respect to the Euclidean norm. Therefore,

$$\|r(\xi) - r(\xi')\| = \|\rho(a) - \rho(a')\| \leq L_\rho \|a - a'\| \leq L_\rho \|\xi - \xi'\|.$$

Moreover, the scalar sigmoid is 1/4-Lipschitz, hence its componentwise extension is 1/4-Lipschitz with respect to the Euclidean norm. Therefore,

$$\|g(\xi) - g(\xi')\| \leq \frac{1}{4} \|G(\xi) - G(\xi')\| \leq \frac{L_G}{4} \|\xi - \xi'\|.$$

Combining the two estimates yields

$$\|\Phi_{\text{pure}}(\xi) - \Phi_{\text{pure}}(\xi')\| \leq \left(L_\rho + \frac{B_\rho L_G}{4}\right) \|\xi - \xi'\|,$$

as claimed. □

If ρ is only locally bounded and locally Lipschitz, and G is locally Lipschitz, then the same argument yields local Lipschitz continuity of Φ_{pure} on bounded subsets of $\mathbb{R}^{H+S}$.

3.5 Monotonicity Properties

Unlike standard elementwise activations, the pure CAF operator need not be globally monotone in a, because the gate $g(a, s)$ may itself depend on the pre-activation vector. Nevertheless, one can characterize local monotonicity componentwise.

Proposition 8 (Local componentwise monotonicity) *Under the standing C^1 assumptions on ρ and G, fix $i \in \{1, \ldots, H\}$, $s \in \mathbb{R}^S$, and all coordinates a_j with $j \neq i$. For each $x \in \mathbb{R}$, define*

$$a^{(i,x)} := (a_1, \ldots, a_{i-1}, x, a_{i+1}, \ldots, a_H) \in \mathbb{R}^H$$

and

$$f_i(x) := \big(\Phi_{\text{pure}}(a^{(i,x)}, s)\big)_i.$$

Then

$$f_i'(x) = \frac{\partial g_i}{\partial a_i}(a^{(i,x)}, s)\,\rho(x) + g_i(a^{(i,x)}, s)\,\rho'(x).$$

Consequently:

- *if $f_i'(x_0) > 0$, then there exists a neighborhood of x_0 on which f_i is strictly increasing;*
- *if $f_i'(x_0) < 0$, then there exists a neighborhood of x_0 on which f_i is strictly decreasing;*
- *if $f_i'(x) \geq 0$ throughout some neighborhood of x_0, then f_i is nondecreasing on that neighborhood.*

Moreover, on an interval $I \subset \mathbb{R}$, the function f_i is nondecreasing on I whenever $f_i'(x) \geq 0$ for all $x \in I$.

Proof By definition, $f_i(x) = g_i(a^{(i,x)}, s)\,\rho(x)$. Differentiating with respect to x yields

$$f_i'(x) = \frac{\partial g_i}{\partial a_i}(a^{(i,x)}, s)\,\rho(x) + g_i(a^{(i,x)}, s)\,\rho'(x).$$

The strict local monotonicity statements follow from the sign of $f_i'(x_0)$ and continuity of f_i', which implies that the same strict sign holds on a neighborhood of x_0. The corresponding monotonicity then follows from the one-dimensional mean value theorem. The nondecreasing claims follow from the standard one-dimensional criterion for monotonicity. □

Under the affine gate parameterization (4), one further has

$$\frac{\partial g_i}{\partial a_i}(a^{(i,x)}, s) = g_i(a^{(i,x)}, s)\big(1 - g_i(a^{(i,x)}, s)\big)(W_g^{(a)})_{ii},$$

and therefore

$$f_i'(x) = g_i(a^{(i,x)}, s)\big(1 - g_i(a^{(i,x)}, s)\big)(W_g^{(a)})_{ii}\rho(x) + g_i(a^{(i,x)}, s)\rho'(x).$$

Thus, even when ρ is monotone, the pure CAF operator may become non-monotone through the interaction between the gate sensitivity and the base activation. This additional flexibility may increase expressive power, but it also makes the response explicitly context-dependent.

4 Gradient Flow Analysis

Understanding gradient propagation through the pure CAF operator is essential for assessing trainability and identifying potential optimization difficulties, particularly in relation to the classical vanishing and exploding gradient phenomena [Bengio et al.(1994)Bengio, Simard, and Frasconi, Glorot and Bengio(2010), Pascanu et al.(2013)Pascanu, Mikolov, and Bengio].

In this section, all statements refer to the pure CAF operator introduced in Section 2. For convenience, we write

$$z(a,s) := G([a;s]), \qquad g(a,s) := \sigma(z(a,s)), \qquad y(a,s) := \Phi_{\text{pure}}(a,s) = g(a,s)\odot\rho(a).$$

4.1 Backward Signal Decomposition

Let $\mathscr{L}$ be a scalar loss function, and let

$$\delta_y := \nabla_y \mathscr{L} \in \mathbb{R}^H$$

denote the incoming gradient at the output of the activation. By Proposition 5, the gradient with respect to the pre-activation a is given by

$$\nabla_a \mathscr{L} = (D_a y(a,s))^\top \delta_y = (D_a g(a,s))^\top \operatorname{diag}(\rho(a))\,\delta_y + \operatorname{diag}(\rho'(a))\,\operatorname{diag}(g(a,s))\,\delta_y, \tag{14}$$

where

$$y(a,s) := \Phi_{\text{pure}}(a,s) = g(a,s) \odot \rho(a).$$

Using (10), this may equivalently be written as

$$\nabla_a \mathscr{L} = \big(D_a G([a;s])\big)^\top \operatorname{diag}\big(g(a,s)\odot(\mathbf{1}-g(a,s))\big)\operatorname{diag}(\rho(a))\,\delta_y + \operatorname{diag}(\rho'(a))\,\operatorname{diag}(g(a,s))\,\delta_y. \tag{15}$$

In particular, under the affine gate parameterization (4), one has

$$\nabla_a \mathscr{L} = (W_g^{(a)})^\top \operatorname{diag}\big(g(a,s)\odot(\mathbf{1}-g(a,s))\big)\operatorname{diag}(\rho(a))\,\delta_y + \operatorname{diag}(\rho'(a))\,\operatorname{diag}(g(a,s))\,\delta_y. \tag{16}$$

The gradient with respect to the context variable s is obtained analogously from (9):

$$\nabla_s \mathscr{L} = (D_s y(a,s))^\top \delta_y = (D_s g(a,s))^\top \operatorname{diag}(\rho(a))\,\delta_y. \tag{17}$$

Using (11), this becomes

$$\nabla_s \mathscr{L} = \big(D_s G([a;s])\big)^\top \operatorname{diag}\big(g(a,s) \odot (\mathbf{1} - g(a,s))\big)\operatorname{diag}(\rho(a))\,\delta_y. \tag{18}$$

In the affine case (4), this reduces to

$$\nabla_s \mathscr{L} = (W_g^{(s)})^\top \operatorname{diag}\big(g(a,s) \odot (\mathbf{1} - g(a,s))\big)\operatorname{diag}(\rho(a))\,\delta_y. \tag{19}$$

Remark 2 (Interpretation) Equation (14) decomposes the backward signal into two distinct contributions. The term

$$\operatorname{diag}(\rho'(a))\,\operatorname{diag}(g(a,s))\,\delta_y$$

is a gated version of the standard componentwise derivative and is attenuated coordinatewise by the factors $g_i(a,s) \in (0,1)$. The term

$$(D_a g(a,s))^\top \operatorname{diag}(\rho(a))\,\delta_y$$

arises from the dependence of the gate on the pre-activation. When $D_a g(a,s)$ is non-diagonal, this second contribution introduces cross-component gradient couplings that are absent from standard elementwise activations.

4.2 Local Jacobian Bounds

We begin with the standard bound on the derivative of the logistic sigmoid.

Lemma 1 (Sigmoid derivative bound)**.** *For the logistic sigmoid* $\sigma(x) = 1/(1+e^{-x})$*,*

$$0 < \sigma'(x) \le \frac{1}{4} \qquad \textit{for all } x \in \mathbb{R},$$

and the upper bound is attained at $x = 0$*.*

Proof Since

$$\sigma'(x) = \sigma(x)\big(1-\sigma(x)\big),$$

the derivative is maximized when $\sigma(x) = 1/2$, which occurs at $x = 0$. Therefore,

$$\sigma'(x) \le \frac{1}{4} \qquad \text{for all } x \in \mathbb{R}.$$

Moreover, since $\sigma(x) \in (0,1)$ for every $x \in \mathbb{R}$, one has $\sigma'(x) > 0$ for all $x \in \mathbb{R}$. □

The preceding estimate yields a simple operator-norm bound for the Jacobian with respect to the pre-activation in the affine-gate setting.

Theorem 9 (Local Jacobian norm bound for affine gating) *Assume the affine gate parameterization* (4)*. Then, for every* $(a,s) \in \mathbb{R}^{H+S}$*,*

$$\|D_a y(a,s)\| \le \|g(a,s)\|_\infty \|\rho'(a)\|_\infty + \frac{1}{4}\|\rho(a)\|_\infty \|W_g^{(a)}\|. \tag{20}$$

Proof Under the affine gate parameterization, Proposition 5 gives

$$D_a y = \operatorname{diag}(\rho(a)) \operatorname{diag}\big(g(a,s) \odot (\mathbf{1} - g(a,s))\big) W_g^{(a)} + \operatorname{diag}(g(a,s)) \operatorname{diag}(\rho'(a)).$$

Taking operator norms and using the triangle inequality yields

$$\|D_a y\| \le \Big\|\operatorname{diag}(\rho(a)) \operatorname{diag}\big(g(a,s) \odot (\mathbf{1} - g(a,s))\big) W_g^{(a)}\Big\| + \Big\|\operatorname{diag}(g(a,s)) \operatorname{diag}(\rho'(a))\Big\|.$$

For diagonal matrices, the operator norm equals the infinity norm of the diagonal vector. Hence

$$\Big\|\operatorname{diag}(g(a,s)) \operatorname{diag}(\rho'(a))\Big\| \le \|g(a,s)\|_\infty \|\rho'(a)\|_\infty.$$

Likewise,

$$\Big\|\operatorname{diag}(\rho(a)) \operatorname{diag}\big(g(a,s) \odot (\mathbf{1} - g(a,s))\big) W_g^{(a)}\Big\| \le \|\rho(a)\|_\infty \|g(a,s) \odot (\mathbf{1} - g(a,s))\|_\infty \|W_g^{(a)}\|.$$

By Lemma 1,

$$\|g(a,s) \odot (\mathbf{1} - g(a,s))\|_\infty \le \frac{1}{4}.$$

Combining the preceding estimates gives (20). □

Remark 3 (Interpretation of the Jacobian bound) The estimate (20) separates the local sensitivity of the pure CAF operator into two contributions. The first term is inherited from the base activation and is attenuated by the gate magnitude. The second term is specific to contextual gating and is controlled by the amplitude of the base activation together with the size of the pre-activation coupling matrix $W_g^{(a)}$. In particular, large values of $\|W_g^{(a)}\|$ may increase the local Jacobian norm even when the direct derivative term is strongly attenuated.

Corollary 10 (Uniform affine-gate Jacobian bound) Assume the standing global bounds

$$\|\rho(a)\|_\infty \le B_\rho, \qquad \|\rho'(a)\|_\infty \le L_\rho \qquad \text{for all } a \in \mathbb{R}^H.$$

Then, for every $(a, s) \in \mathbb{R}^{H+S}$,

$$\|D_a y(a, s)\| \le L_\rho + \frac{1}{4} B_\rho \|W_g^{(a)}\|.$$

Proof By Theorem 9,

$$\|D_a y(a, s)\| \le \|g(a, s)\|_\infty \|\rho'(a)\|_\infty + \frac{1}{4} \|\rho(a)\|_\infty \|W_g^{(a)}\|.$$

Since $\|g(a, s)\|_\infty \le 1$, the assumed bounds imply

$$\|D_a y(a, s)\| \le L_\rho + \frac{1}{4} B_\rho \|W_g^{(a)}\|,$$

as claimed. □

Remark 4 When $\|W_g^{(a)}\|$ is moderate, the local Jacobian norm remains controlled. This does not by itself rule out gradient explosion in a deep or recurrent model, but it helps mitigate excessive gradient growth at the layer level.

4.3 Conditioning and Numerical Stability

The local conditioning of the activation map is governed by the Jacobian $D_a y(a, s) \in \mathbb{R}^{H \times H}$.

Definition 4 (Jacobian condition number) If $D_a y(a, s)$ is invertible, its condition number is

$$\kappa\big(D_a y(a, s)\big) = \|D_a y(a, s)\| \, \big\|D_a y(a, s)^{-1}\big\|,$$

where $\|\cdot\|$ denotes the spectral norm.

The quantities $\|D_a y(a, s)\|$, $\sigma_{\min}(D_a y(a, s))$, and $\kappa(D_a y(a, s))$ describe different aspects of local behavior. The norm $\|D_a y(a, s)\|$ controls local forward amplification of perturbations in the pre-activation. The smallest singular value controls local loss of sensitivity and, when positive, the stability of local inversion. The condition number

$$\kappa(D_a y(a, s)) = \frac{\sigma_{\max}(D_a y(a, s))}{\sigma_{\min}(D_a y(a, s))}$$

measures anisotropy of the local transformation and sensitivity of inversion, rather than absolute gradient attenuation alone. For example, the matrix $D_a y = \varepsilon I_H$ has condition number equal to one for every $\varepsilon > 0$, although its gradients vanish as $\varepsilon \to 0$.

Remark 5 If $D_a y(a, s)$ is singular, one may instead consider a pseudoinverse-based conditioning measure such as

$$\|D_a y(a, s)\| \, \big\|(D_a y(a, s))^\dagger\big\|,$$

where $(D_a y(a, s))^\dagger$ denotes the Moore–Penrose pseudoinverse.

Proposition 11 (Perturbation bounds for the conditioning of affine-gate CAF) *Under the affine gate parameterization* (4)*, write*

$$D_a y(a, s) = D(a, s) + E(a, s), \tag{21}$$

where

$$D(a, s) := \operatorname{diag}\big(g(a, s) \odot \rho'(a)\big), \tag{22}$$

$$E(a, s) := \operatorname{diag}\big(\rho(a)\big) \operatorname{diag}\big(g(a, s) \odot (\mathbf{1} - g(a, s))\big) W_g^{(a)}. \tag{23}$$

Define

$$d_{\min}(a, s) := \min_{1 \le i \le H} \left|g_i(a, s)\rho'(a_i)\right|, \qquad d_{\max}(a, s) := \max_{1 \le i \le H} \left|g_i(a, s)\rho'(a_i)\right|.$$

Then

$$\sigma_{\min}\big(D_a y(a, s)\big) \ge \max\left\{0,\, d_{\min}(a, s) - \|E(a, s)\|\right\}, \tag{24}$$

$$\sigma_{\max}\big(D_a y(a, s)\big) \le d_{\max}(a, s) + \|E(a, s)\|. \tag{25}$$

Moreover,

$$\|E(a, s)\| \le \frac{1}{4} \|\rho(a)\|_\infty \|W_g^{(a)}\|. \tag{26}$$

Consequently, if

$$d_{\min}(a, s) > \|E(a, s)\|,$$

then $D_a y(a, s)$ is invertible and its spectral condition number satisfies

$$\kappa\big(D_a y(a, s)\big) \le \frac{d_{\max}(a, s) + \|E(a, s)\|}{d_{\min}(a, s) - \|E(a, s)\|}. \tag{27}$$

Proof For every unit vector $v \in \mathbb{R}^H$, the reverse triangle inequality gives

$$\|(D + E)v\| \ge \|Dv\| - \|Ev\| \ge \sigma_{\min}(D) - \|E\|.$$

Taking the minimum over all unit vectors yields

$$\sigma_{\min}(D + E) \ge \sigma_{\min}(D) - \|E\|.$$

Since singular values are nonnegative, this gives

$$\sigma_{\min}(D + E) \ge \max\{0, \sigma_{\min}(D) - \|E\|\}.$$

Likewise, for every unit vector v,

$$\|(D + E)v\| \le \|Dv\| + \|Ev\| \le \sigma_{\max}(D) + \|E\|,$$

and therefore

$$\sigma_{\max}(D+E) \le \sigma_{\max}(D) + \|E\|.$$

Because D is diagonal,

$$\sigma_{\min}(D) = d_{\min}(a,s), \qquad \sigma_{\max}(D) = d_{\max}(a,s),$$

which proves (24) and (25).

Furthermore, submultiplicativity of the spectral norm and the identity $\|\operatorname{diag}(v)\| = \|v\|_\infty$ give

$$\|E(a,s)\| \le \|\rho(a)\|_\infty \|g(a,s) \odot (\mathbf{1} - g(a,s))\|_\infty \|W_g^{(a)}\|.$$

Since $0 < g_i(a,s) < 1$ and

$$g_i(a,s)\big(1 - g_i(a,s)\big) \le \frac{1}{4},$$

we obtain (26).

Finally, if $d_{\min}(a,s) > \|E(a,s)\|$, then the lower bound on $\sigma_{\min}(D_a y)$ is strictly positive, so $D_a y$ is invertible. Using

$$\kappa(D_a y) = \frac{\sigma_{\max}(D_a y)}{\sigma_{\min}(D_a y)}$$

together with the preceding singular-value bounds proves (27). □

Corollary 12 (Exact conditioning of context-only pure CAF) For a context-only affine gate, $W_g^{(a)} = 0$, and therefore

$$E(a,s) = 0, \qquad D_a y(a,s) = \operatorname{diag}\big(g(a,s) \odot \rho'(a)\big).$$

If

$$\min_{1 \le i \le H} \big|g_i(a,s)\rho'(a_i)\big| > 0,$$

then the Jacobian is invertible and

$$\kappa\big(D_a y(a,s)\big) = \frac{\max_i |g_i(a,s)\rho'(a_i)|}{\min_i |g_i(a,s)\rho'(a_i)|}. \tag{28}$$

If at least one diagonal term vanishes, then $D_a y(a,s)$ is singular.

Remark 6 (Interpretation) Proposition 11 provides a sufficient condition for local invertibility and an explicit upper bound on the condition number. It does not imply that pre-activation-dependent gating systematically improves or degrades conditioning. Instead, it quantifies how the coupling term E perturbs the singular values of the gated diagonal term D. In the context-only case, the coupling term vanishes and the local conditioning is determined exactly by the spread of the nonzero diagonal factors $g_i(a,s)\rho'(a_i)$.

4.4 Local Conditioning of the Activation Map

We now formalize the fact that gate saturation near zero suppresses the local backward influence of the corresponding output component with respect to the pre-activation variables.

Proposition 13 (Gradient suppression under gate saturation) *Assume the affine gate parameterization (4). Let* $\{(a^{(n)}, s^{(n)})\}_{n\geq 1}$ *be a sequence in* $\mathbb{R}^{H+S}$*, and fix* $i \in \{1, \dots, H\}$*. If*

$$g_i(a^{(n)}, s^{(n)}) \to 0 \qquad \text{as } n \to \infty,$$

and if the sequences $\rho(a_i^{(n)})$ *and* $\rho'(a_i^{(n)})$ *remain bounded, then*

$$\frac{\partial y_i}{\partial a_j}(a^{(n)}, s^{(n)}) \to 0 \qquad \text{for all } j = 1, \dots, H.$$

Proof By (12), the (i, j)-entry of $D_a y$ is

$$\frac{\partial y_i}{\partial a_j}(a, s) = \rho(a_i)\, g_i(a, s)\big(1 - g_i(a, s)\big)(W_g^{(a)})_{ij} + g_i(a, s)\rho'(a_i)\delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. Evaluating this expression at $(a^{(n)}, s^{(n)})$ gives

$$\frac{\partial y_i}{\partial a_j}(a^{(n)}, s^{(n)}) = \rho(a_i^{(n)})\, g_i(a^{(n)}, s^{(n)})\big(1 - g_i(a^{(n)}, s^{(n)})\big)(W_g^{(a)})_{ij} + g_i(a^{(n)}, s^{(n)})\rho'(a_i^{(n)})\delta_{ij}.$$

Since $g_i(a^{(n)}, s^{(n)}) \to 0$, one also has

$$g_i(a^{(n)}, s^{(n)})\big(1 - g_i(a^{(n)}, s^{(n)})\big) \to 0.$$

Because $\rho(a_i^{(n)})$ and $\rho'(a_i^{(n)})$ remain bounded along the sequence, both terms on the right-hand side converge to zero. Therefore,

$$\frac{\partial y_i}{\partial a_j}(a^{(n)}, s^{(n)}) \to 0 \qquad \text{for all } j = 1, \dots, H.$$

□

Remark 7 (Selective gradient blocking) Proposition 13 shows that strongly suppressed gates induce a form of selective gradient blocking: output components whose gate values approach zero become weakly coupled to the backward signal through the local activation map. In this sense, the gate acts as a learned deterministic masking mechanism. This behavior is conceptually related to gating effects in recurrent architectures [Hochreiter and Schmidhuber(1997), Cho et al.(2014)Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk, and Bengio], although the present setting is purely feedforward at the level of the local activation operator.

A complementary statement holds when the gate saturates near one.

Proposition 14 (Asymptotic recovery of the base-activation Jacobian) *Assume the affine gate parameterization (4). Let* $\{(a^{(n)}, s^{(n)})\}_{n\geq 1}$ *be a sequence in* $\mathbb{R}^{H+S}$*, and fix* $i \in \{1, \dots, H\}$*. If*

$$g_i(a^{(n)}, s^{(n)}) \to 1 \qquad \text{as } n \to \infty,$$

and if the sequences $\rho(a_i^{(n)})$ *and* $\rho'(a_i^{(n)})$ *remain bounded, then*

$$\frac{\partial y_i}{\partial a_j}(a^{(n)}, s^{(n)}) - \rho'(a_i^{(n)})\delta_{ij} \to 0 \qquad \text{for all } j = 1, \dots, H.$$

Proof Again by (12),

$$\frac{\partial y_i}{\partial a_j}(a,s) = \rho(a_i)\, g_i(a,s)\big(1 - g_i(a,s)\big)(W_g^{(a)})_{ij} + g_i(a,s)\rho'(a_i)\delta_{ij}.$$

Subtracting $\rho'(a_i)\delta_{ij}$ from both sides yields

$$\frac{\partial y_i}{\partial a_j}(a,s) - \rho'(a_i)\delta_{ij} = \rho(a_i)\, g_i(a,s)\big(1 - g_i(a,s)\big)(W_g^{(a)})_{ij} + \big(g_i(a,s) - 1\big)\rho'(a_i)\delta_{ij}.$$

Evaluating along the sequence $(a^{(n)}, s^{(n)})$, the factor $g_i(a^{(n)}, s^{(n)})(1-g_i(a^{(n)}, s^{(n)}))$ tends to zero because $g_i(a^{(n)}, s^{(n)}) \to 1$, and the second term also tends to zero because $g_i(a^{(n)}, s^{(n)}) - 1 \to 0$ while $\rho'(a_i^{(n)})$ remains bounded. Since $\rho(a_i^{(n)})$ remains bounded as well, the first term converges to zero. Hence

$$\frac{\partial y_i}{\partial a_j}(a^{(n)}, s^{(n)}) - \rho'(a_i^{(n)})\delta_{ij} \to 0 \qquad \text{for all } j = 1, \dots, H.$$

□

4.5 Gate-Parameter Saturation and Initialization

The preceding results concern derivatives of the activation output with respect to its input. Saturation also affects the optimization of the gate parameters through the derivative of the sigmoid pre-activation.

Proposition 15 (Gate-gradient suppression under saturation) *Let $g_i = \sigma(z_i)$ be one coordinate of the gate, and consider a sequence along which $|z_i| \to \infty$. Assume that the upstream derivative*

$$\frac{\partial \mathscr{L}}{\partial g_i}$$

remains bounded along this sequence. Then

$$\frac{\partial \mathscr{L}}{\partial z_i} = \frac{\partial \mathscr{L}}{\partial g_i}\,\sigma'(z_i) \to 0.$$

Proof By the chain rule,

$$\frac{\partial \mathscr{L}}{\partial z_i} = \frac{\partial \mathscr{L}}{\partial g_i}\,\sigma'(z_i).$$

Since

$$\sigma'(z_i) = \sigma(z_i)\big(1 - \sigma(z_i)\big) \to 0 \qquad \text{as } |z_i| \to \infty,$$

and $\partial \mathscr{L}/\partial g_i$ remains bounded, the product converges to zero. □

Corollary 16 (Parameter-gradient suppression under affine gating) Assume the affine gate parameterization

$$z = W_g^{(a)} a + W_g^{(s)} s + b_g.$$

Fix a gate coordinate i and consider a sequence along which $|z_i| \to \infty$. Assume that $\partial \mathscr{L}/\partial g_i$ remains bounded and that the sequences a and s remain bounded. Then

$$\frac{\partial \mathscr{L}}{\partial z_i} \to 0.$$

Moreover, the corresponding parameter-gradient rows satisfy

$$\nabla_{(W_g^{(a)})_{i,:}} \mathscr{L} = \frac{\partial \mathscr{L}}{\partial z_i} a^\top \to 0,$$

$$\nabla_{(W_g^{(s)})_{i,:}} \mathscr{L} = \frac{\partial \mathscr{L}}{\partial z_i} s^\top \to 0,$$

and

$$\frac{\partial \mathscr{L}}{\partial (b_g)_i} = \frac{\partial \mathscr{L}}{\partial z_i} \to 0.$$

Thus, saturation directly suppresses the gate-preactivation gradient and, under bounded inputs, the gradients of the corresponding affine gate parameters.

Proof The first statement follows from Proposition 15. The remaining identities follow from the affine parameterization of z_i and the chain rule. Since a and s remain bounded while $\partial \mathscr{L}/\partial z_i \to 0$, all three parameter-gradient expressions converge to zero. □

Remark 8 (Initialization trade-off) Proposition 17 shows that a constant gate $g = \sigma(\alpha)\mathbf{1}$ approaches the standard activation map as $\alpha \to +\infty$. However, the same limit gives $\sigma'(\alpha) \to 0$, so initialization near $g = 1$ can suppress the initial gradient received by the gate parameters. By contrast, $\alpha \approx 0$ places the gate near $1/2$, where the sigmoid derivative is maximal, but initially rescales the base response. Thus, initialization involves a trade-off between proximity to the base activation and trainability of the gate; no universally optimal choice follows from these local bounds alone.

4.6 Comparison with Standard Activations

For a standard elementwise activation

$$y_{\text{std}} = \rho(a),$$

the Jacobian with respect to a is simply

$$D_a y_{\text{std}} = \operatorname{diag}(\rho'(a)).$$

By contrast, for the pure CAF operator, Proposition 5 gives

$$D_a y = \operatorname{diag}(\rho(a))\, D_a g + \operatorname{diag}(g) \operatorname{diag}(\rho'(a)).$$

Thus, the Jacobian of the pure CAF operator consists of two conceptually distinct parts:

1. a gated diagonal term
$$\operatorname{diag}(g) \operatorname{diag}(\rho'(a)),$$
which attenuates the base derivative componentwise since $g_i(a, s) \in (0, 1)$;

2. a gate-induced term
$$\operatorname{diag}(\rho(a))\, D_a g,$$
which may introduce cross-component gradient couplings whenever $D_a g$ is non-diagonal.

Hence, the pure CAF operator not only modulates the magnitude of local gradient flow on a componentwise basis, but may also redistribute gradients across components through the dependence of the gate on the pre-activation. This additional flexibility may increase expressive power, but it also makes the local backward dynamics more sensitive to the structure and scale of the gate map.

4.7 Gate-Restricted Local Bound

The affine-gate Jacobian bound of Theorem 9 uses only the global fact that each sigmoid gate coordinate lies in $(0, 1)$. If, on a specified domain $D \subseteq \mathbb{R}^H \times \mathbb{R}^S$, one has an independent certified bound

$$\bar{g}_D := \sup_{(a,s)\in D} \|g(a, s)\|_\infty < 1,$$

then the same proof gives the domain-restricted estimate

$$\|D_a y(a, s)\| \le \bar{g}_D L_\rho + \frac{1}{4} B_\rho \|W_g^{(a)}\|, \qquad (a, s) \in D.$$

This is not a contraction certificate for a recurrent forecaster unless the state–context trajectory is also certified to remain inside D and all additional recurrent, temporal-summary, and autoregressive feedback paths are included. In particular, evaluating the gate only at its initialization point can be useful as an initialization diagnostic, but it does not by itself establish a post-training Lipschitz or stability bound.

4.8 Illustrative Special Cases

The following cases specialize the preceding Jacobian formulas. They are included as analytic illustrations rather than as additional theorems, and do not enlarge the standing assumptions of the results above.

ReLU base activation.

Consider $\rho(x) = \max(0, x)$. Although ReLU is not of class C^1, the formulas below hold away from $a_i = 0$ and may be interpreted almost everywhere (or in a subgradient sense at the kink). The i-th pure-CAF coordinate is

$$y_i(a, s) = g_i(a, s) \max(0, a_i).$$

For $a_i > 0$, one has $y_i(a, s) = g_i(a, s)a_i$, and therefore, for every j,

$$\frac{\partial y_i}{\partial a_j} = a_i \frac{\partial g_i}{\partial a_j} + g_i(a, s)\delta_{ij}.$$

In particular,

$$\frac{\partial y_i}{\partial a_i} = g_i(a,s) + a_i \frac{\partial g_i}{\partial a_i}.$$

For $a_i < 0$, the output coordinate vanishes locally, so

$$\frac{\partial y_i}{\partial a_j} = 0 \qquad \text{for every } j.$$

Thus, on the active ReLU branch, contextual gating contributes the additional term $a_i\,\partial g_i/\partial a_i$ to the local slope. Its sign and magnitude are parameter- and context-dependent; no systematic amplification or attenuation follows without further assumptions.

Pure context dependence.

For a context-only affine gate,

$$g(a,s) = \sigma\big(W_g^{(s)} s + b_g\big), \qquad W_g^{(a)} = 0,$$

one has $D_a g(a,s) = 0$. Proposition 5 then reduces to

$$D_a y(a,s) = \operatorname{diag}\big(g(a,s)\big)\,\operatorname{diag}\big(\rho'(a)\big).$$

Hence the gate modulates the componentwise derivative without introducing cross-unit coupling through the activation Jacobian. This is consistent with the exact conditioning formula in Corollary 12.

Near-identity gate limit.

Let

$$W_g^{(a)} = 0, \qquad W_g^{(s)} = 0, \qquad b_g = \alpha \mathbf{1}.$$

Then $g(a,s) = \sigma(\alpha)\mathbf{1} \to \mathbf{1}$ as $\alpha \to +\infty$, and consequently

$$y(a,s) \to \rho(a), \qquad D_a y(a,s) \to \operatorname{diag}\big(\rho'(a)\big)$$

pointwise. The output convergence is uniform on compact sets whenever ρ is continuous. The Jacobian convergence is uniform on compact sets whenever ρ' is bounded there, in particular when ρ' is continuous. This limit clarifies recovery of the base activation, while Remark 8 explains why using a strongly saturated near-identity gate at initialization can reduce gate-parameter gradients.

5 Expressivity and Universal Approximation

In this section, we examine the expressive consequences of the pure CAF formulation introduced in Section 2. We first compare the standard and pure CAF mechanisms at the level of individual activation maps, where the relevant statement is one of uniform-closure inclusion and context-dependent separation. We then record the exact

embedding of standard one-hidden-layer networks as a short lemma. The main result is the stronger context-only universality theorem: even when each base pre-activation depends only on the primary input x and each affine gate depends only on the context s, finite sums of the resulting product units are dense on compact product domains.

5.1 Expressive Extension at the Level of Activation Maps

With the base activation ρ fixed, the standard activation mechanism corresponds to the context-independent map

$$\Phi_{\mathrm{std}}(a, s) := \rho(a), \qquad (a, s) \in \mathbb{R}^H \times \mathbb{R}^S,$$

whereas, under the affine gate parameterization (4), the pure CAF activation map is

$$\Phi_{\mathrm{pure}}(a, s) = \sigma\big(W_g^{(a)} a + W_g^{(s)} s + b_g\big) \odot \rho(a).$$

Because the logistic sigmoid takes values in $(0, 1)$, the map Φ_{std} is not realized exactly by a single pure CAF activation with finite parameters unless ρ vanishes on the domain of interest. Nevertheless, Φ_{std} belongs to the uniform closure of the affine-gate pure CAF family on compact sets.

Proposition 17 (Approximate recovery of the standard activation map) *Let $K_a \subset \mathbb{R}^H$ and $K_s \subset \mathbb{R}^S$ be compact, and assume that ρ is continuous. Then, for every $\varepsilon > 0$, there exists a pure CAF activation map with affine gate parameterization, denoted Φ_ε, such that*

$$\sup_{(a,s) \in K_a \times K_s} \|\Phi_\varepsilon(a, s) - \Phi_{\mathrm{std}}(a, s)\|_\infty < \varepsilon.$$

Proof Since ρ is continuous and K_a is compact, the componentwise map $a \mapsto \rho(a) \in \mathbb{R}^H$ is bounded on K_a. Define

$$M := \sup_{a \in K_a} \|\rho(a)\|_\infty < \infty.$$

If $M = 0$, then $\rho(a) = 0$ for all $a \in K_a$, and the claim is trivial. Assume $M > 0$. Since $\sigma(\alpha) \to 1$ as $\alpha \to +\infty$, one can choose $\alpha > 0$ such that

$$|1 - \sigma(\alpha)|\, M < \varepsilon.$$

Set

$$W_g^{(a)} = 0, \qquad W_g^{(s)} = 0, \qquad b_g = \alpha \mathbf{1}.$$

Then the gate is constant:

$$g(a, s) = \sigma(\alpha)\mathbf{1},$$

and the corresponding pure CAF activation map is

$$\Phi_\varepsilon(a, s) = \sigma(\alpha)\rho(a).$$

Therefore,

$$\|\Phi_\varepsilon(a, s) - \Phi_{\mathrm{std}}(a, s)\|_\infty = |1 - \sigma(\alpha)|\, \|\rho(a)\|_\infty \le |1 - \sigma(\alpha)|\, M < \varepsilon$$

for all $(a, s) \in K_a \times K_s$. This proves the claim. The uniform closure here is taken in the space $C(K_a \times K_s; \mathbb{R}^H)$ equipped with the uniform norm. The argument does not claim uniform convergence on the whole unbounded domain $\mathbb{R}^{H+S}$. □

Proposition 18 (Strict expressive separation at the level of activation maps) *Let $K_a \subset \mathbb{R}^H$ and $K_s \subset \mathbb{R}^S$ be compact. Assume that K_s contains at least two distinct points and that there exists $a^\star \in K_a$ such that $\rho(a^\star)$ is not the zero vector. Then there exists a pure CAF activation map with affine gate parameterization that cannot be represented by the context-independent standard activation map $(a, s) \mapsto \rho(a)$.*

Proof Choose $s_1, s_2 \in K_s$ with $s_1 \neq s_2$, and define

$$u := s_1 - s_2 \in \mathbb{R}^S, \qquad W_g^{(a)} = 0, \qquad W_g^{(s)} = \mathbf{1}u^\top, \qquad b_g = 0.$$

Then each gate coordinate satisfies

$$g_i(a, s) = \sigma(u^\top s), \qquad i = 1, \dots, H,$$

so the resulting pure CAF activation map is

$$\Phi_{\text{pure}}(a, s) = \sigma(u^\top s)\rho(a).$$

Moreover,

$$u^\top s_1 - u^\top s_2 = (s_1 - s_2)^\top (s_1 - s_2) = \|s_1 - s_2\|^2 > 0,$$

hence, since σ is strictly increasing,

$$\sigma(u^\top s_1) \neq \sigma(u^\top s_2).$$

Because $\rho(a^\star)$ is not the zero vector, it follows that

$$\Phi_{\text{pure}}(a^\star, s_1) \neq \Phi_{\text{pure}}(a^\star, s_2).$$

By contrast, any context-independent standard activation map of the form $(a, s) \mapsto \rho(a)$ takes the same value at $(a^\star, s_1)$ and $(a^\star, s_2)$. The constructed map is strictly contextual in the sense of Definition 1, and therefore cannot be represented by the standard activation mechanism. □

Corollary 19 (Uniform-closure inclusion and context-dependent separation) Under the assumptions of Proposition 18, the uniform closure of the affine-gate pure-CAF family on $K_a \times K_s$ contains the standard activation map and also contains genuinely context-dependent maps. The strict separation established here is at the level of a fixed activation map; it is not a width-efficiency result and does not establish superiority of the entire pure-CAF network class over standard networks.

Proof The inclusion in the uniform closure follows from Proposition 17, and strictness follows from Proposition 18. □

Remark 9 The preceding results are stated at the level of individual activation maps. Thus, exact recovery of the standard activation map by a single pure CAF activation is generally impossible with finite logistic gate parameters, except in degenerate cases where ρ vanishes on the domain of interest. By contrast, at the network level, exact recovery of standard one-hidden-layer models is possible through a constant gate together with a compensating rescaling of the output weights, as shown below.

5.2 Universal Approximation

We now consider one-hidden-layer feedforward networks equipped with a pure CAF activation layer. Let $x \in \mathbb{R}^d$ denote the primary input and $s \in \mathbb{R}^S$ the context. For a network of width N and output dimension m, let

$$a(x,s) = W_x x + W_s s + b_a \in \mathbb{R}^N, \tag{29}$$

$$h(x,s) = \sigma\big(W_g^{(a)} a(x,s) + W_g^{(s)} s + b_g\big) \odot \rho(a(x,s)) \in \mathbb{R}^N, \tag{30}$$

$$F_N(x,s) = V h(x,s) + c \in \mathbb{R}^m. \tag{31}$$

Here, the vector $a(x,s) \in \mathbb{R}^N$ denotes the hidden-layer pre-activation induced by the input-context pair (x,s). In the universal approximation result below, the pair (x,s) is viewed jointly as an input in $\mathbb{R}^{d+S}$. The parameters satisfy

$$W_x \in \mathbb{R}^{N\times d}, \quad W_s \in \mathbb{R}^{N\times S}, \quad b_a \in \mathbb{R}^N, \quad W_g^{(a)} \in \mathbb{R}^{N\times N}, \quad W_g^{(s)} \in \mathbb{R}^{N\times S}, \quad b_g \in \mathbb{R}^N, \quad V \in \mathbb{R}^{m\times N}, \quad c \in \mathbb{R}^m.$$

Standing assumption for this subsection.

Assume that the base activation $\rho : \mathbb{R} \to \mathbb{R}$ is continuous and non-polynomial. Under this classical condition, standard one-hidden-layer networks with activation ρ are dense in $C(K;\mathbb{R})$ for compact $K \subset \mathbb{R}^{d+S}$ [Cybenko(1989), Hornik(1991), Leshno et al.(1993)Leshno, Lin, Pinkus, and Schocken, Pinkus(1999)]; the extension to $\mathbb{R}^m$-valued targets follows by approximating each coordinate separately.

Lemma 2 (Exact embedding of a standard one-hidden-layer network)**.** *Every standard one-hidden-layer network*

$$\widetilde{F}_N(x,s) = \widetilde{V}\,\rho\big(\widetilde{W}[x;s] + \widetilde{b}\big) + \widetilde{c}$$

is realized exactly by a pure CAF network of the form (29)–(31)*.*

Proof Write $\widetilde{W} = [W_x\ W_s]$ and set $b_a = \widetilde{b}$. Choose any $c_0 \in (0,1)$, set $W_g^{(a)} = W_g^{(s)} = 0$ and $b_g = \log(c_0/(1-c_0))\mathbf{1}$, and rescale the output parameters as $V = c_0^{-1}\widetilde{V}$ and $c = \widetilde{c}$. The hidden gate is then the constant $c_0\mathbf{1}$, so the factors c_0 and c_0^{-1} cancel and $F_N = \widetilde{F}_N$ on $\mathbb{R}^{d+S}$. □

Theorem 20 (Universal approximation for pure CAF networks with affine gates) *Let $\rho : \mathbb{R} \to \mathbb{R}$ be continuous and non-polynomial. Let $K \subset \mathbb{R}^{d+S}$ be compact, let $f \in C(K;\mathbb{R}^m)$, and let $\varepsilon > 0$. Then there exist a width $N \in \mathbb{N}$ and parameters in* (29)–(31) *such that*

$$\sup_{(x,s)\in K} \|f(x,s) - F_N(x,s)\| < \varepsilon.$$

Proof By the classical universal approximation theorem for continuous non-polynomial activations [Hornik(1991), Leshno et al.(1993)Leshno, Lin, Pinkus, and Schocken,

Pinkus(1999)], there exist $N \in \mathbb{N}$ and a standard one-hidden-layer network

$$\widetilde{F}_N(x,s) = \widetilde{V}\,\rho\big(\widetilde{W}[x;s] + \widetilde{b}\big) + \widetilde{c}$$

such that

$$\sup_{(x,s)\in K} \|f(x,s) - \widetilde{F}_N(x,s)\| < \varepsilon.$$

By Lemma 2, there exists a pure CAF network F_N of the form (29)–(31) such that

$$F_N(x,s) = \widetilde{F}_N(x,s) \qquad \text{for all } (x,s) \in \mathbb{R}^{d+S}.$$

Therefore,

$$\sup_{(x,s)\in K} \|f(x,s) - F_N(x,s)\| = \sup_{(x,s)\in K} \|f(x,s) - \widetilde{F}_N(x,s)\| < \varepsilon.$$

This proves the claim. □

Remark 10 The affine-gate restriction does not obstruct universal approximation. Constant gates are already realizable within the affine-gate pure CAF family, and they suffice to recover the classical one-hidden-layer universal approximators exactly through a compensating rescaling of the output weights. This embedding result is useful but immediate; the next theorem establishes universality for a strictly constrained, genuinely context-modulated subfamily.

5.3 Universality with Context-Only Affine Gates

We now impose a separation that is absent from Theorem 20. For every hidden unit, the base pre-activation depends only on the primary input x, whereas the affine gate depends only on the context s:

$$a_k(x) = w_k^\top x + b_k, \qquad g_k(s) = \sigma(v_k^\top s + c_k).$$

Thus no context enters the base pre-activation and no primary input enters the gate. Despite this restriction, the resulting pure-CAF networks remain universal on compact product domains.

Theorem 21 (Universal approximation with context-only affine gates) *Let $K_x \subset \mathbb{R}^d$ and $K_s \subset \mathbb{R}^S$ be compact, and let $f \in C(K_x \times K_s; \mathbb{R}^m)$. Assume that the base activation $\rho : \mathbb{R} \to \mathbb{R}$ is continuous and non-polynomial. Then, for every $\varepsilon > 0$, there exist $N \in \mathbb{N}$ and parameters $u_k \in \mathbb{R}^m$, $w_k \in \mathbb{R}^d$, $v_k \in \mathbb{R}^S$, and $b_k, c_k \in \mathbb{R}$, $k = 1, \dots, N$, such that*

$$f_{\mathrm{CAF}}(x,s) = \sum_{k=1}^{N} u_k\,\sigma(v_k^\top s + c_k)\,\rho(w_k^\top x + b_k) \tag{32}$$

satisfies

$$\sup_{(x,s)\in K_x\times K_s} \|f(x,s) - f_{\mathrm{CAF}}(x,s)\| < \varepsilon.$$

Consequently, one-hidden-layer pure-CAF networks whose base pre-activations depend only on x and whose affine gates depend only on s are dense in $C(K_x \times K_s; \mathbb{R}^m)$ for the uniform norm.

Proof For scalar targets, Stone–Weierstrass implies that finite sums of separable functions $\sum_{r=1}^{R} \varphi_r(x)\psi_r(s)$ are dense in $C(K_x \times K_s)$. Because ρ is continuous and non-polynomial, finite linear combinations of $x \mapsto \rho(w^\top x+b)$ are dense in $C(K_x)$. Because the logistic sigmoid is continuous and sigmoidal, finite linear combinations of $s \mapsto \sigma(v^\top s+c)$ are dense in $C(K_s)$. Multiplying the two finite approximations expresses each approximated separable term as a finite sum of units

$$\sigma(v^\top s + c)\, \rho(w^\top x + b),$$

which have exactly the context-only pure-CAF form in (32). Choosing the two factor-approximation tolerances sufficiently small yields the prescribed uniform error. Applying the scalar argument coordinatewise proves the vector-valued case. The complete error estimate is given in Appendix A. □

Remark 11 (Scope of the result) The theorem uses neither a term $W_s s$ in the base pre-activation nor a term $W_g^{(a)} a$ in the gate. It therefore proves universality for a strictly smaller subfamily than Theorem 20. Like any universal approximation theorem, it is qualitative: it does not establish a width or depth advantage over standard networks.

5.4 Approximation Efficiency

Universal approximation is a qualitative statement and does not, by itself, address the number of hidden units required to achieve a prescribed error tolerance. Pure CAF networks may nevertheless provide more compact representations for target functions exhibiting context-dependent regime changes or context-modulated nonlinear structure.

Conjecture 1 (Potential efficiency gains for context-dependent regime structure) There exist classes of continuous functions on compact domains with context-dependent regime structure for which pure CAF networks may achieve a given approximation error with smaller hidden width than standard networks using the same base activation.

Establishing rigorous width- or depth-efficiency bounds for pure CAF networks remains an open problem.

5.5 Inductive Bias Analysis

Pure CAF activations induce an inductive bias toward solutions in which the base nonlinear response $\rho(a)$ is modulated by context. It is useful to distinguish two sources of information entering the temporal gate: the *endogenous* signals—the pre-activation a and its recent temporal summary u—and the *exogenous* covariate context s. Suppressing the exogenous channel ($W_g^{(s)} = 0$) yields a self-gated temporal configuration, whereas setting $W_g^{(a)} = 0$ yields the distinct context-only reduction. These are different axes: exogenous-channel inclusion is controlled by $W_g^{(s)}$, while the context-only/diagonal/full hierarchy describes the structure of $W_g^{(a)}$. The two channels are beneficial under different conditions:

- **Endogenous modulation** is advantageous when the target is temporally coherent and self-persistent, so that a self-gated response refines a structured, autocorrelated signal rather than fitting exogenous drivers;
- **Exogenous modulation** through s can be advantageous when its incremental predictive information is sufficiently large relative to its dimension, noise level, and the effective estimation and conditioning cost; when s is weakly informative or excessively noisy, gating on s may add estimation variance without a compensating predictive gain;
- in both channels, selective computation or soft masking is the operative mechanism.

Conversely, if the target is essentially independent of a given context channel, the corresponding gate parameters may increase statistical and optimization complexity without a corresponding gain in approximation or generalization. In particular, a multiplicative gate is not always the most effective vehicle for a strongly informative exogenous context: even when s carries genuine predictive information, injecting it additively (e.g. concatenation or FiLM-style modulation) may outperform a saturating multiplicative gate (cf. Section 8).

A precise local characterization of this decision is developed in Section 12. Its estimable quantities are $\widehat{\mathscr{I}_s}$, $\widehat{\mathscr{B}}_s(\lambda)$, $\widehat{\mathscr{V}_s}(\lambda)$, $\widehat{\kappa}_\lambda$, $\widehat{d}_{2,\lambda}$, and a held-out Taylor-error diagnostic. Prospective validation of these quantities is left to future work.

6 Gate Saturation and Context-Variation Diagnostics

The quantities considered in this section are descriptive diagnostics rather than measures of network capacity or CAF-specific approximation theorems. We distinguish three properties of a learned gate: its degree of saturation, its variation across observations, and its statistical dependence on context. These properties cannot, in general, be inferred from a single entropy functional.

Throughout this section, $A \in \mathbb{R}^H$ denotes the random pre-activation vector at a representative layer, C denotes the corresponding context variable, and

$$g(A, C) = \big(g_1(A, C), \ldots, g_H(A, C)\big) \in (0, 1)^H$$

denotes the resulting random gate vector. Unless otherwise stated, logarithms are natural.

6.1 Normalized Binary Gate Entropy

Define the binary entropy function

$$h_{\mathrm{b}}(p) := -p \log p - (1 - p) \log(1 - p), \qquad p \in [0, 1],$$

with the convention $0 \log 0 = 0$. The width-normalized binary gate entropy is defined by

$$\overline{\mathscr{H}}_{\text{gate}} := \frac{1}{H} \mathbb{E}_{(A,C)} \left[\sum_{i=1}^{H} h_{\text{b}}\big(g_i(A,C)\big) \right]. \tag{33}$$

Since

$$0 \le h_{\text{b}}(p) \le \log 2 \qquad \text{for every } p \in [0,1],$$

with equality at the upper bound if and only if $p = 1/2$, it follows that

$$0 \le \overline{\mathscr{H}}_{\text{gate}} \le \log 2.$$

The normalization by H removes the direct linear dependence on layer width and therefore facilitates comparisons across layers, although such comparisons still depend on the distributions of their pre-activations and contexts.

Equation (33) is a pointwise saturation diagnostic. Small values indicate that gate coordinates are typically close to the saturated states 0 or 1, whereas values close to $\log 2$ indicate that gate coordinates are typically close to $1/2$. A large value does not imply substantial variation across observations or contexts. For example, a gate that is identically equal to $1/2$ attains the maximum binary entropy while having zero observational variance.

This functional treats each value $g_i(A,C)$ as the parameter of an auxiliary Bernoulli variable. Unless the architecture actually samples such a Bernoulli variable, $\overline{\mathscr{H}}_{\text{gate}}$ is not the Shannon entropy of the empirical distribution of gate values. It is instead a descriptive functional measuring their average degree of non-saturation.

6.2 Variation Across Observations

Variation of the learned gates across the data distribution must be measured separately. Since every gate coordinate lies in $(0,1)$, it is bounded and square-integrable. Define

$$\mathscr{V}_{\text{gate}} := \frac{1}{H} \sum_{i=1}^{H} \operatorname{Var}\big(g_i(A,C)\big). \tag{34}$$

Unlike the entropy functional in Equation (33), $\mathscr{V}_{\text{gate}}$ measures dispersion of gate values across observations.

For example, an almost constant gate near $1/2$ has high normalized binary entropy but low observational variance. Conversely, a gate that alternates between values near 0 and 1 can have low average binary entropy but high observational variance. Thus, entropy and observational variation describe different properties of the learned gate.

For empirical reporting, Equation (34) may be estimated by the corresponding sample variances computed on held-out observations. This statistic describes overall gate variability but does not, by itself, attribute that variability specifically to the context.

6.3 Within- and Between-Context Variation

Because every $g_i(A, C)$ is square-integrable, the law of total variance gives, for any context variable C,

$$\operatorname{Var}\big(g_i(A,C)\big) = \mathbb{E}_C\left[\operatorname{Var}\big(g_i(A,C) \mid C\big)\right] + \operatorname{Var}_C\left(\mathbb{E}\big[g_i(A,C) \mid C\big]\right). \tag{35}$$

The first term describes variation within contexts, whereas the second describes variation of the conditional gate mean between contexts. Accordingly, define

$$\mathscr{V}_{\text{within}} := \frac{1}{H}\sum_{i=1}^{H} \mathbb{E}_C\left[\operatorname{Var}\big(g_i(A,C) \mid C\big)\right], \tag{36}$$

$$\mathscr{V}_{\text{between}} := \frac{1}{H}\sum_{i=1}^{H} \operatorname{Var}_C\left(\mathbb{E}\big[g_i(A,C) \mid C\big]\right). \tag{37}$$

Averaging Equation (35) over the H coordinates yields the exact decomposition

$$\mathscr{V}_{\text{gate}} = \mathscr{V}_{\text{within}} + \mathscr{V}_{\text{between}}. \tag{38}$$

These identities also hold when C is continuous. In empirical applications, one may instead use a discretized or grouped context variable. In that case, the resulting between-context term refers to the chosen grouping and does not necessarily capture all variation associated with the original continuous context.

A strictly positive value of $\mathscr{V}_{\text{between}}$ means that, for at least one gate coordinate, the conditional mean

$$\mathbb{E}\big[g_i(A,C) \mid C\big]$$

is not almost surely constant. It therefore provides descriptive evidence that average gate usage differs across observed contexts. It does not isolate a direct effect of the context input on the gate. The same pattern may arise because the conditional distribution of A changes with C, even when the gate has no explicit dependence on C. Nor does the decomposition establish a causal effect without additional intervention or identification assumptions.

6.4 Statistical Dependence on Context

A possible dependence diagnostic is

$$I\big(C; g(A,C)\big). \tag{39}$$

When finite, this quantity measures marginal statistical dependence between the context and the gate vector. It does not isolate the direct contribution of C beyond the pre-activation A. In particular, it may be positive even for a context-blind gate $g(A)$ whenever A and C are statistically dependent.

A technical difficulty also arises for deterministic mappings of continuous variables. Mutual information is an extended-valued quantity and may be infinite when the joint

distribution of C and $g(A, C)$ is singular with respect to the product of their marginal distributions. Consequently, the empirical use of Equation (39) requires an explicitly specified observation or regularization model, together with a specified estimator.

If the objective is to quantify contextual dependence beyond the information already present in A, one may instead consider a regularized conditional quantity such as

$$I\big(C; g(A, C) + \xi \mid A\big), \tag{40}$$

where ξ is an independent continuous noise variable with a specified distribution. The added noise regularizes the deterministic gate representation, while conditioning on A separates contextual dependence from dependence mediated solely through the pre-activation. Even in this case, the value depends on the selected noise model and must be estimated with a stated procedure. Mutual-information quantities are therefore treated here as optional empirical diagnostics rather than as established theoretical properties of CAF.

6.5 Relation to Conditional Prediction

Assume that the target T takes values in a finite or countable alphabet and that

$$H(T \mid X) < \infty.$$

Let $\mathscr{Q}_{XC}$ denote the class of all conditional probability mass functions given (X, C), and let $\mathscr{Q}_X$ denote the corresponding class given X. Assume that these classes include the true conditional distributions. The classical conditional-coding identity then gives

$$\begin{aligned}
\inf_{q \in \mathscr{Q}_{XC}} \mathbb{E}\big[-\log q(T \mid X, C)\big] &= H(T \mid X, C) && (41)\\
&= H(T \mid X) - I(T; C \mid X) && (42)\\
&\leq H(T \mid X) && (43)\\
&= \inf_{q \in \mathscr{Q}_X} \mathbb{E}\big[-\log q(T \mid X)\big]. && (44)
\end{aligned}$$

The first and last equalities follow from the decomposition of expected cross-entropy into conditional entropy and an expected Kullback–Leibler divergence [Cover and Thomas(2006)]. Under the stated finiteness assumption, the inequality is strict if and only if

$$I(T; C \mid X) > 0.$$

Thus, access to context reduces the unrestricted optimal expected code length precisely when the target retains conditional information about C beyond X. This is a standard information-theoretic fact and is not specific to CAF. It does not imply that a restricted or trained CAF predictor attains the optimal conditional distribution, nor does it establish that multiplicative gating is the most effective mechanism for exploiting the available information.

Remark 12 (Scope of the diagnostics) The preceding quantities answer distinct questions:

- $\overline{\mathscr{H}}_{\text{gate}}$ is a pointwise non-saturation diagnostic: smaller values indicate gate values closer to 0 or 1, whereas larger values indicate gate values closer to 1/2;
- $\mathscr{V}_{\text{gate}}$ measures variation across observations;
- $\mathscr{V}_{\text{within}}$ and $\mathscr{V}_{\text{between}}$ separate within-context variation from variation in conditional gate means;
- $I(C; g(A, C))$, when finite, measures marginal statistical dependence between context and gate values, but not necessarily dependence beyond that mediated by A;
- a regularized conditional quantity such as $I(C; g(A, C) + \xi \mid A)$ may be used to study contextual dependence beyond A, subject to the specified noise model and estimator.

None of these quantities is, without an additional theorem, a measure of effective network capacity, approximation power, causal importance, or computational savings. In particular, no entropy–capacity relation is claimed.

7 Dynamical Systems Perspectives

For recurrent architectures with pure CAF, the hidden-state update defines a context-dependent discrete-time dynamical system. When the external input or the context varies with time, the resulting dynamics are generally non-autonomous. A simple recurrent model with input and bias takes the form

$$\begin{aligned} a_t &= W_h h_t + W_x x_t + b_a, \\ h_{t+1} &= g(a_t, s_t) \odot \rho(a_t), \end{aligned} \tag{45}$$

where $h_t \in \mathbb{R}^H$ is the hidden state, $x_t \in \mathbb{R}^d$ is the input, $s_t \in \mathbb{R}^S$ is the context at time t, and

$$W_h \in \mathbb{R}^{H \times H}, \qquad W_x \in \mathbb{R}^{H \times d}, \qquad b_a \in \mathbb{R}^H.$$

Under the affine gate parameterization (4), one may write

$$g(a_t, s_t) = \sigma\big(W_g^{(a)} a_t + W_g^{(s)} s_t + b_g\big).$$

The simpler recurrent form

$$h_{t+1} = g(W h_t, s_t) \odot \rho(W h_t)$$

is recovered as the special case $W_x = 0$ and $b_a = 0$ with $W_h = W$.

7.1 Fixed Points and Context Dependence

If the input and context are constant, namely $x_t \equiv x$ and $s_t \equiv s$, then fixed points satisfy

$$\begin{aligned} h^\star &= g(a^\star, s) \odot \rho(a^\star), \\ a^\star &= W_h h^\star + W_x x + b_a. \end{aligned} \tag{46}$$

Remark 13 (Context-dependent fixed points) Equation (46) depends explicitly on the context variable s. Consequently, the location and number of fixed points may vary with the context,

so that different contexts can induce different equilibrium structures and, more generally, different dynamical regimes.

7.2 Local Linearization and Stability

For fixed x and s, define

$$\Phi(h; s, x) := g(a(h), s) \odot \rho(a(h)), \qquad a(h) := W_h h + W_x x + b_a.$$

Let $h^\star$ be a fixed point, and let

$$a^\star := a(h^\star) = W_h h^\star + W_x x + b_a.$$

A first-order linearization around $h^\star$ gives

$$h_{t+1} - h^\star \approx J(h^\star; s, x)\,(h_t - h^\star), \tag{47}$$

where the Jacobian with respect to h is

$$J(h^\star; s, x) = \big[\operatorname{diag}\big(g(a^\star, s)\big)\operatorname{diag}\big(\rho'(a^\star)\big) + \operatorname{diag}\big(\rho(a^\star)\big)\,D_a g(a^\star, s)\big]\,W_h. \tag{48}$$

This is the recurrent analogue of the layerwise Jacobian formula established in Proposition 5, composed with the recurrent affine map $h \mapsto a(h)$.

In the affine-gate case, one further has

$$D_a g(a^\star, s) = \operatorname{diag}\big(g(a^\star, s) \odot (\mathbf{1} - g(a^\star, s))\big)\,W_g^{(a)}.$$

Equivalently, if

$$g^\star := g(a^\star, s),$$

then

$$J(h^\star; s, x) = \Big[\operatorname{diag}(g^\star)\operatorname{diag}(\rho'(a^\star)) + \operatorname{diag}(\rho(a^\star))\operatorname{diag}\big(g^\star \odot (\mathbf{1} - g^\star)\big)W_g^{(a)}\Big]W_h.$$

Local asymptotic stability of the fixed point is guaranteed if the spectral radius satisfies

$$\varrho\big(J(h^\star; s, x)\big) < 1.$$

A stronger but simpler sufficient condition is

$$\|J(h^\star; s, x)\| < 1$$

for any matrix norm subordinate to a vector norm.

Remark 14 (Potential gate-mediated stabilization) Small gate values attenuate the direct contribution

$$\operatorname{diag}\big(g(a^\star, s)\big)\operatorname{diag}\big(\rho'(a^\star)\big)W_h,$$

and may therefore help stabilize recurrent dynamics in regimes where the gate-induced Jacobian term

$$\operatorname{diag}\big(\rho(a^\star)\big)\, D_a g(a^\star, s)\, W_h$$

remains moderate. Thus, contextual gating introduces an additional mechanism by which the model can reshape the local stability of recurrent dynamics.

Remark 15 (Leaky or residual recurrent update) A common variant is the leaky update

$$h_{t+1} = (1-\lambda)h_t + \lambda\, g(a_t, s_t) \odot \rho(a_t), \qquad \lambda \in (0, 1].$$

For constant x and s, this variant has the same fixed points as (46). Its local Jacobian at a fixed point is

$$J_\lambda(h^\star; s, x) = (1-\lambda) I_H + \lambda\, J(h^\star; s, x),$$

which interpolates between the identity map and the contextual recurrent Jacobian. This provides an additional degree of control over local stability.

7.3 Uniform Contraction, Context-Dependent Equilibria, and Tracking

The spectral-radius condition above is local: it characterizes the behavior in a neighborhood of a fixed point whose existence is assumed. The following result gives a global sufficient condition for existence, uniqueness, continuous dependence on context, and tracking under a time-varying context. Throughout the result, the input x is fixed and

$$\Phi_s(h) := \Phi(h; s, x).$$

Theorem 22 (Uniform contraction and contextual equilibrium tracking) *Let $\mathscr{C} \subseteq \mathbb{R}^S$ be nonempty and fix $x \in \mathbb{R}^d$. Assume that there exist constants $q \in [0, 1)$ and $L_s \geq 0$ such that, for every $h, \widetilde{h} \in \mathbb{R}^H$ and every $s, \widetilde{s} \in \mathscr{C}$,*

$$\|\Phi(h; s, x) - \Phi(\widetilde{h}; s, x)\| \leq q\|h - \widetilde{h}\|, \tag{49}$$

$$\|\Phi(h; s, x) - \Phi(h; \widetilde{s}, x)\| \leq L_s\|s - \widetilde{s}\|. \tag{50}$$

Then the following statements hold.

1. *For each $s \in \mathscr{C}$, there exists a unique fixed point $h^\star(s; x) \in \mathbb{R}^H$ satisfying*

$$h^\star(s; x) = \Phi\big(h^\star(s; x); s, x\big).$$

Moreover, for the autonomous iteration $h_{t+1} = \Phi(h_t; s, x)$,

$$\|h_t - h^\star(s; x)\| \leq q^t \|h_0 - h^\star(s; x)\|. \tag{51}$$

2. *The equilibrium map is Lipschitz on $\mathscr{C}$:*

$$\|h^\star(s; x) - h^\star(\widetilde{s}; x)\| \leq \frac{L_s}{1-q}\|s - \widetilde{s}\|. \tag{52}$$

3. *Let $\{s_t\}_{t\geq 0} \subseteq \mathscr{C}$ and consider the non-autonomous dynamics*

$$h_{t+1} = \Phi(h_t; s_t, x).$$

Defining

$$e_t := \|h_t - h^\star(s_t; x)\|,$$

one has, for every $t \geq 0$,

$$e_t \leq q^t e_0 + \frac{L_s}{1-q} \sum_{k=0}^{t-1} q^{t-1-k} \|s_{k+1} - s_k\|. \tag{53}$$

In particular, if $\sup_{k\geq 0} \|s_{k+1} - s_k\| \leq \delta$, *then*

$$e_t \leq q^t e_0 + \frac{L_s \delta}{(1-q)^2}(1-q^t), \qquad \limsup_{t\to\infty} e_t \leq \frac{L_s \delta}{(1-q)^2}. \tag{54}$$

Proof For each fixed $s \in \mathscr{C}$, the map $\Phi_s : \mathbb{R}^H \to \mathbb{R}^H$ is a contraction with constant q. Since $\mathbb{R}^H$ is complete, the Banach fixed-point theorem gives a unique fixed point $h^\star(s; x)$. Repeated application of (49) yields (51).

For $s, \widetilde{s} \in \mathscr{C}$, the fixed-point identities and the triangle inequality give

$$\begin{aligned}
\|h^\star(s;x) - h^\star(\widetilde{s};x)\| &= \|\Phi(h^\star(s;x); s, x) - \Phi(h^\star(\widetilde{s};x); \widetilde{s}, x)\| \\
&\leq \|\Phi(h^\star(s;x); s, x) - \Phi(h^\star(\widetilde{s};x); s, x)\| \\
&\quad + \|\Phi(h^\star(\widetilde{s};x); s, x) - \Phi(h^\star(\widetilde{s};x); \widetilde{s}, x)\| \\
&\leq q\|h^\star(s;x) - h^\star(\widetilde{s};x)\| + L_s\|s - \widetilde{s}\|.
\end{aligned}$$

Rearranging proves (52).

For the non-autonomous dynamics, insert the intermediate equilibrium $h^\star(s_t; x)$ to obtain

$$\begin{aligned}
e_{t+1} &= \|\Phi(h_t; s_t, x) - h^\star(s_{t+1}; x)\| \\
&\leq \|\Phi(h_t; s_t, x) - \Phi(h^\star(s_t; x); s_t, x)\| + \|h^\star(s_t; x) - h^\star(s_{t+1}; x)\| \\
&\leq q e_t + \frac{L_s}{1-q}\|s_{t+1} - s_t\|.
\end{aligned}$$

Iterating this recursion proves (53). If the context increments are bounded by δ, summing the geometric series gives (54). □

Theorem 22 replaces the informal notion of a context-modulated attractor family by a precise statement. In the uniformly contractive regime, each frozen context has the singleton global attractor $\{h^\star(s; x)\}$, these equilibria vary Lipschitz-continuously with s, and the trajectory remains close to the moving equilibrium whenever the context increments are small. The result does not assert analogous properties outside the contractive regime, where multiple equilibria or more general attractors may occur.

Corollary 23 (A sufficient contraction condition for affine-gate pure CAF) Consider the recurrent pure-CAF map

$$\Phi(h; s, x) = \sigma\Big(W_g^{(a)} a(h) + W_g^{(s)} s + b_g\Big) \odot \rho(a(h)), \qquad a(h) = W_h h + W_x x + b_a. \tag{55}$$

Assume that ρ is globally bounded and globally Lipschitz, with

$$|\rho(r)| \leq B_\rho, \qquad |\rho(r) - \rho(r')| \leq L_\rho |r - r'| \quad \text{for all } r, r' \in \mathbb{R}.$$

Define

$$q_{\mathrm{CAF}} := \|W_h\| \left(L_\rho + \frac{B_\rho}{4}\|W_g^{(a)}\|\right). \tag{56}$$

If $q_{\mathrm{CAF}} < 1$, then the assumptions of Theorem 22 hold with

$$q = q_{\mathrm{CAF}}, \qquad L_s = \frac{B_\rho}{4}\|W_g^{(s)}\|. \tag{57}$$

Consequently, for every fixed x and every context s, the recurrent map has a unique globally attracting equilibrium. Its dependence on context and its tracking error satisfy (52)–(54) with the constants in (57).

Proof Write $a = a(h)$ and $\widetilde{a} = a(\widetilde{h})$. Using $\|u \odot v\| \le \|u\|_\infty \|v\|$, the fact that every gate coordinate lies in $(0, 1)$, and the $1/4$-Lipschitz continuity of the logistic sigmoid, one has

$$\begin{aligned}
\|\Phi(h; s, x) - \Phi(\widetilde{h}; s, x)\| &\le \|\rho(a) - \rho(\widetilde{a})\| + B_\rho \|g(a, s) - g(\widetilde{a}, s)\| \\
&\le \left(L_\rho + \frac{B_\rho}{4}\|W_g^{(a)}\|\right) \|a - \widetilde{a}\| \\
&\le q_{\mathrm{CAF}} \|h - \widetilde{h}\|.
\end{aligned}$$

For fixed h, the pre-activation $a(h)$ does not depend on s, and therefore

$$\begin{aligned}
\|\Phi(h; s, x) - \Phi(h; \widetilde{s}, x)\| &\le B_\rho \|g(a(h), s) - g(a(h), \widetilde{s})\| \\
&\le \frac{B_\rho}{4}\|W_g^{(s)}\| \|s - \widetilde{s}\|.
\end{aligned}$$

Thus (49) and (50) hold with the stated constants, and the conclusion follows from Theorem 22. □

8 Relationship to Existing Conditional Modulation Mechanisms

For brevity, we refer to context-adaptive activations as pure *CAF* in this section. Pure CAF is related to several existing conditional-computation and context-modulation mechanisms, but it differs from them in both the locus and the form of contextual adaptation [Perez et al.(2018)Perez, Strub, de Vries, Dumoulin, and Courville, Ha et al.(2017)Ha, Dai, and Le, Hochreiter and Schmidhuber(1997), Cho et al.(2014)Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk, and Bengio, Jacobs et al.(1991)Jacobs, Jordan, Nowlan, and Hinton, Shazeer et al.(2017)Shazeer, Mirhoseini, Maziarz, Davis, Le, Hinton, and Dean].

The central distinction is that CAF introduces context dependence directly at the level of the activation map. In its pure form, the activation output is given by

$$\Phi_{\mathrm{pure}}(a, s) = g(a, s) \odot \rho(a),$$

where the gate $g(a, s)$ modulates the base nonlinearity as a function of both the pre-activation a and the context s. This should be contrasted with other conditional mechanisms, which typically act at the level of features, weights, recurrent state updates, or expert routing.

8.1 Comparison with Adaptive Activation Functions

Adaptive activation functions differ primarily in the source of the adaptation signal. This distinction matters mathematically because it determines whether the activation can be analyzed as a fixed trained nonlinearity, a self-gated function of the current input, or a map that depends on an explicitly external context. Table 1 summarizes the distinction.

Table 1: CAF compared with adaptive activation families.

Family	Source of adaptation	What varies	Mathematical consequence
Parameterized activations such as PReLU, APL, and adaptive slopes	Learned parameters fixed after training	Shape or slope of the scalar response	The trained activation is fixed at inference; no explicit context derivative $D_s\Phi$ is present
Self- or input-conditioned activations such as Swish-β, ACON-C, Meta-ACON, Dynamic ReLU, and SE-style gates	Current activation value or statistics of the input representation	Gate, slope, or channel weights generated from the input itself	Adaptation is endogenous to the representation; Jacobian terms are driven by the same signal being transformed
CAF	Explicit exogenous descriptor s, optionally combined with a or temporal summaries	Activation-level gate multiplying the base response	The activation map is a function $\Phi(a, s)$ with analyzable derivatives in both a and s; this separates exogenous contextual sensitivity from self-gating

This comparison motivates the terminology used here: CAF is not merely a new trained scalar nonlinearity, but a family of adaptive activations whose defining feature is the availability of an explicit exogenous context at activation level.

8.2 Comparison with FiLM

Feature-wise Linear Modulation (FiLM) applies context-dependent affine transformations to intermediate features [Perez et al.(2018)Perez, Strub, de Vries, Dumoulin, and Courville]. In a schematic

form, a FiLM layer may be written as

$$a = Wx + b, \qquad h = \rho(a), \qquad y_{\text{FiLM}} = \gamma(s) \odot h + \beta(s), \qquad \gamma(s), \beta(s) \in \mathbb{R}^H.$$

By contrast, a pure CAF layer takes the form

$$a = Wx + b, \qquad \Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a).$$

Remark 16 (CAF versus FiLM) When $\beta(s) = 0$ and $\gamma(s) = g(s)$, a context-only pure CAF is algebraically a restricted FiLM transformation applied to $\rho(a)$. The framework becomes distinct from this restricted case when the gate also depends on a, when structural constraints on $W_g^{(a)}$ determine diagonal or cross-unit coupling, or when the activation-level Jacobian is the object of analysis. The contribution is therefore not the multiplicative product itself, but an activation-centric formalization, a structured taxonomy, and a unified theoretical analysis.

8.3 Comparison with HyperNetworks

HyperNetworks generate weights or biases as functions of a conditioning signal [Ha et al.(2017)Ha, Dai, and Le]. In schematic form,

$$W(s) = \text{HyperNet}(s), \qquad y = \rho\big(W(s)x + b(s)\big).$$

A pure CAF layer instead keeps the main affine transformation fixed and introduces context dependence in the nonlinear stage:

$$a = Wx + b, \qquad \Phi_{\text{pure}}(a, s) = g(a, s) \odot \rho(a).$$

Remark 17 (CAF versus HyperNetworks) HyperNetworks introduce context dependence by modifying the parameters of the affine map itself. CAF introduces context dependence by modulating the nonlinear response associated with a fixed pre-activation. Accordingly, HyperNetworks act at the parameter level, whereas CAF acts at the level of the activation map.

8.4 Comparison with Recurrent Gating

LSTM and GRU architectures use gates to regulate information flow in recurrent state updates [Hochreiter and Schmidhuber(1997), Cho et al.(2014)Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk, and Bengio]. For example, a GRU-type update may be written schematically as

$$h_t = (\mathbf{1} - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \qquad z_t = \sigma\big(W_z[x_t; h_{t-1}] + b_z\big).$$

Table 2: Schematic comparison of CAF with related conditional modulation mechanisms.

Mechanism	Schematic form	Gate source	Object modulated	Additive term	Relation to CAF
CAF	$g(a, s) \odot \rho(a)$	a and/or s	activation response	no in pure form	activation-centric family
FiLM	$\gamma(s) \odot h + \beta(s)$	context s	feature representation	yes	context-only CAF is a restricted zero-shift case
GLU/SwiGLU	value branch times gate branch	input branch	representation branch	no	overlaps for selected branch identifications; SwiGLU is generally non-sigmoid and unbounded
SE	channel reweighting from pooled features	input statistics	channels	usually no	data-dependent channel modulation
HyperNetworks	$\rho(Vx + b(s))$	conditioning signal	parameters	implicit	context enters weights, not local activation
LSTM/GRU	recurrent state gates	state/input	memory update	yes	complementary state-transition gating
MoE	$\sum_k w_k(x) f_k(x)$	router input	experts/subnetworks	mixture sum	coarser conditional routing

By contrast, a temporal CAF mechanism modulates the activation map itself, for example through

$$y_t = g_t \odot \rho(a_t), \qquad g_t = \sigma\big(W_g^{(a)} a_t + W_g^{(u)} u_t + W_g^{(s)} s_t + b_g\big).$$

Remark 18 (CAF versus recurrent gating) Recurrent gates regulate state transitions and memory flow across time, whereas CAF gates regulate the local nonlinear response of the activation map. The two mechanisms therefore operate at different levels and should be viewed as complementary rather than interchangeable.

8.5 Comparison with Mixture-of-Experts

Mixture-of-experts (MoE) models perform conditional computation by routing inputs across specialized subnetworks [Jacobs et al.(1991)Jacobs, Jordan, Nowlan, and Hinton, Shazeer et al.(2017)Shazeer, Mirhoseini, Maziarz, Davis, Le, Hinton, and Dean]. In abstract form,

$$y = \sum_{k=1}^{K} w_k(x) f_k(x), \qquad w_k(x) = \mathrm{softmax}(r(x))_k.$$

A pure CAF layer instead performs fine-grained modulation at the unit level:

$$y_i = g_i(a, s)\rho(a_i).$$

Remark 19 (CAF versus mixture-of-experts) MoE performs conditional computation through expert selection or soft routing at the network or subnetwork level. CAF instead performs context-dependent modulation within a single layer, at the level of individual activation units. Thus, MoE and CAF differ primarily in the granularity and architectural locus at which adaptation is introduced.

Remark 20 (Positioning) The preceding comparisons indicate that CAF is best understood as a context-dependent activation mechanism. Its distinctive feature is that contextual modulation is introduced directly into the activation map itself, rather than through post-activation feature transformations, dynamic weight generation, recurrent state gating, or expert routing.

9 Parameter and Computational Complexity

We analyze the cost of a single affine-gate pure CAF layer with input $x \in \mathbb{R}^d$, context $s \in \mathbb{R}^S$, hidden dimension H, and base activation ρ. To isolate the overhead induced by contextual gating, we consider the common setting in which the main affine map depends only on the primary input, while the context enters through the gate:

$$a = W_x x + b_a, \qquad g = \sigma\big(W_g^{(a)} a + W_g^{(s)} s + b_g\big), \qquad y = g \odot \rho(a),$$

with

$$W_x \in \mathbb{R}^{H\times d}, \quad b_a \in \mathbb{R}^H, \quad W_g^{(a)} \in \mathbb{R}^{H\times H}, \quad W_g^{(s)} \in \mathbb{R}^{H\times S}, \quad b_g \in \mathbb{R}^H.$$

Remark 21 If the main affine map also depends explicitly on the context, for example

$$a = W_x x + W_x^{(s)} s + b_a,$$

then both the standard layer and the pure CAF layer acquire an additional HS parameters and an additional $O(HS)$ cost in the main affine computation. The gate-induced overhead derived below remains unchanged.

9.1 Parameter Count

A standard layer with activation ρ uses:

- main weights $W_x \in \mathbb{R}^{H \times d}$, contributing Hd parameters;
- main bias $b_a \in \mathbb{R}^H$, contributing H parameters.

Hence its total parameter count is

$$H(d+1).$$

An affine-gate pure CAF layer uses, in addition:

- gate weights $W_g^{(a)} \in \mathbb{R}^{H \times H}$, contributing H^2 parameters;
- gate weights $W_g^{(s)} \in \mathbb{R}^{H \times S}$, contributing HS parameters;
- gate bias $b_g \in \mathbb{R}^H$, contributing H parameters.

Therefore, the total parameter count becomes

$$Hd + H + H^2 + HS + H = H(d + H + S + 2).$$

The additional number of learned parameters relative to the standard layer is thus

$$H^2 + HS + H = H(H + S + 1).$$

9.2 Computational Cost

Forward pass.

The dominant operations are:

1. computing the pre-activation

$$a = W_x x + b_a,$$

which costs $O(Hd)$;

2. computing the gate pre-activation

$$W_g^{(a)} a + W_g^{(s)} s + b_g,$$

which costs $O(H^2 + HS)$;

3. applying the sigmoid and the base activation, and forming

$$y = g \odot \rho(a),$$

which costs $O(H)$.

Hence the total forward cost is

$$O(Hd + H^2 + HS).$$

For comparison, a standard layer has forward cost

$$O(Hd + H).$$

Thus, the affine-gate overhead in the forward pass is

$$O(H^2 + HS).$$

Backward pass.
Backpropagation through the pure CAF layer requires:

- gradients through the elementwise product and the base activation, costing $O(H)$;
- gradients through the sigmoid, costing $O(H)$;
- gradients through the affine gate map, whose dominant cost is $O(H^2 + HS)$.

Accordingly, the additional backward cost induced by the affine gate is of order

$$O(H^2 + HS),$$

up to constant factors and implementation details. The total backward cost for the layer is therefore

$$O(Hd + H^2 + HS),$$

of which $O(H^2 + HS)$ is gate-specific overhead. Thus, at the level of dominant terms, the backward-pass overhead matches the forward-pass overhead. In a sequence model unrolled over T steps, applying a dense affine gate at every step gives gate-specific cost $O(T(H^2 + HS))$, excluding the cost of the rest of the recurrent or sequence model.

Remark 22 (Structured gate reductions) The preceding complexity bounds correspond to the fully affine gate. For the structured cases introduced in Section 2, the gate cost may be reduced. If $W_g^{(a)} = 0$ (context-only gate), the gate-specific forward and backward overhead becomes $O(HS)$. If $W_g^{(a)} = \mathrm{diag}(w)$ (diagonal gate), the corresponding overhead becomes $O(H + HS)$. Thus, the structural restrictions discussed earlier also yield lower parameter and computational complexity.

9.3 Memory Requirements

Relative to a standard layer, one must additionally store the gate pre-activation and gate output:

$$z = W_g^{(a)} a + W_g^{(s)} s + b_g \in \mathbb{R}^H, \qquad g = \sigma(z) \in \mathbb{R}^H.$$

The additional activation storage for one example and one layer is $O(H)$ for z and g. This excludes the gate parameters $O(H^2 + HS)$, their gradients, optimizer states, and saved activations required by backpropagation. In a sequence model unrolled over T steps, storing gate activations scales as $O(TH)$ per layer, subject to checkpointing and implementation details.

9.4 Scalability with Context Dimension

Proposition 24 (Linear dependence on the context dimension) *For fixed hidden dimension H, the S-dependent part of the additional computational cost introduced by the affine gate scales linearly with the context dimension S, namely as*

$$O(HS).$$

Equivalently, the total gate-specific overhead is

$$O(H^2 + HS),$$

and its dependence on S is linear.

Proof The context-dependent part of the gate pre-activation is

$$W_g^{(s)} s,$$

where $W_g^{(s)} \in \mathbb{R}^{H \times S}$. Computing this matrix-vector product requires $O(HS)$ operations. The remaining gate-related term $W_g^{(a)} a$ contributes $O(H^2)$ operations and is independent of S. Therefore, for fixed H, the dependence of the gate overhead on the context dimension is linear in S. □

In particular, for fixed hidden dimension H, the dependence of the gate-induced overhead on the context dimension S is linear.

10 Taxonomy of Prospective CAF Extensions

The theoretical results of this paper concern the pure CAF operator and the structured affine-gate variants analyzed in Sections 2–4. The constructions listed below are extensions of the same design principle. Unless a result is explicitly cited, they are neither analyzed theoretically nor evaluated empirically in this work and should not be interpreted as established contributions. The state-preserving construction is the exception on the empirical side: a finite recurrent realization is evaluated in Section 13.7.3, while a dedicated theory and broader validation remain open.

Table 3: Taxonomy and status of prospective CAF extensions. The table records possible constructions and the additional analysis required before making performance, stability, or computational-efficiency claims.

Extension	Schematic construction	Intended role	Status and required validation
Multi-source context	$g = \sigma\Big(W_g^{(a)} a + \sum_m W_g^{(s,m)} c^{(m)} + b_g\Big)$	Combine several contextual channels in one gate.	Algebraic generalization only; requires identifiability, ablation, and complexity analysis for the individual sources.
Attention-based gate	$g_i = \sigma\Big(\sum_j \alpha_{ij} v_j\Big)$ with attention weights α_{ij}	Aggregate a set- or sequence-valued context separately for each unit.	Not evaluated here; requires comparison with standard attention and an analysis of the additional sequence-length-dependent cost.
Parameterized base response	$y = g(a, s) \odot \rho_\theta(a)$	Learn the shape of the base response jointly with contextual modulation.	Not tested in this work; the study of learned or adaptive base responses is left as a prospective direction.
Stochastic gate	$g = \sigma(G([a; s]) + \epsilon)$	Introduce randomized modulation.	Not analyzed or evaluated here; no regularization benefit is claimed without a specified noise model, objective, and expected-risk analysis.
Hierarchical context	$s^{(\ell+1)} = f_{\mathrm{ctx}}^{(\ell)}(s^{(\ell)}, h^{(\ell)})$	Update the context representation across network depth.	Architectural proposal only; requires stability, complexity, and controlled ablation studies.
Continuous-time CAF	$\dot{h}(t) = g(F_{\mathrm{dyn}}(h(t), t), s(t)) \odot \rho(F_{\mathrm{dyn}}(h(t), t))$	Modulate a continuous-time vector field by an external context path.	Well-posedness requires explicit Carathéodory-type assumptions; see the qualification below. No neural-ODE experiment is reported.
Thresholded gate	$\widetilde{g}_i = g_i \mathbf{1}_{\{g_i > \tau\}}$	Produce zero-valued inactive coordinates after thresholding.	Thresholding alone does not reduce realized FLOPs or latency. Computational savings require sparse kernels or conditional execution that skips inactive units or blocks, followed by measured speed and memory comparisons.

Qualification for the continuous-time extension.

Let

$$f(t,h) := g(F_{\text{dyn}}(h,t), s(t)) \odot \rho(F_{\text{dyn}}(h,t)) \,.$$

A standard sufficient framework for local well-posedness is the following. The context path $s : [0,T] \to \mathbb{R}^S$ is measurable and locally essentially bounded; for each fixed h, the map $t \mapsto f(t,h)$ is measurable; and, for every compact set $K \subset \mathbb{R}^H$, there exist integrable functions $m_K, L_K \in L^1([0,T])$ such that, for almost every $t \in [0,T]$,

$$\begin{aligned} \|f(t,h)\| &\le m_K(t), && h \in K, \\ \|f(t,h) - f(t,\widetilde{h})\| &\le L_K(t)\|h - \widetilde{h}\|, && h, \widetilde{h} \in K. \end{aligned}$$

Under these Carathéodory conditions, the initial-value problem $\dot{h} = f(t,h)$ has a unique local absolutely continuous solution. These conditions are not implied merely by separate local Lipschitz statements for F_{dyn}, g, and ρ when $s(t)$ is arbitrary. Global existence requires additional growth bounds preventing finite-time escape.

Scope.

The entries in Table 3 organize future research rather than enlarge the set of results proved in this article. In particular, this paper makes no claim that stochastic perturbations regularize CAF, that thresholded gates reduce actual computation, or that a learned base response outperforms a fixed activation.

11 Design hypotheses and research directions

The theoretical framework of contextual activations suggests several design principles for neural architectures.

11.1 Modularity Principle

Contextual activations provide a modular mechanism for incorporating context into neural computation. Unlike architectural modifications that tightly couple context to a specific network design, the broader CAF framework can in principle be inserted into many existing layer types, enabling flexible composition across architectures.

Design implication.

Practitioners can experiment with contextual modulation at different depths or in different submodules, identifying where context-dependent computation is most useful without redesigning the full architecture.

11.2 Gate inspectability

Gate activations provide a direct means of inspecting context-dependent behavior: visualizing $g(a,s)$ can show which units are emphasized or suppressed under different contexts. Such visualizations are descriptive diagnostics, not causal explanations by themselves.

Design implication.

In applications where interpretability matters—such as medical decision support, autonomous systems, or scientific modeling—contextual activations may facilitate post hoc inspection of how context alters internal computation, provided that the resulting hypotheses are validated by appropriate interventions or ablations.

11.3 Adaptive Computation

Gate values indicate the relative contribution of activation coordinates, but a small gate value does not reduce the operations performed by a standard dense implementation. Actual computational savings require a thresholding or routing rule together with sparse kernels or conditional execution that skips inactive units or blocks. No such implementation or measured speedup is evaluated in this work; conditional execution is therefore a future engineering direction rather than an established property of CAF.

11.4 Multi-Task and Transfer Learning

Context vectors can encode task identity, domain information, or side constraints, allowing a single shared model to modulate its behavior across tasks.

Design implication.

In multi-task settings, contextual activations may offer a parameter-efficient alternative to duplicating entire subnetworks or allocating task-specific layers, but this remains a task-dependent design hypothesis to be tested empirically.

11.5 Temporal and Sequential Modeling

For sequential data, time-varying context can encode external temporal structure, such as time-of-day effects, seasonal trends, known regimes, or control signals.

Design implication.

When relevant temporal structure is available as an auxiliary signal, contextual activations provide a possible design route for adapting to such patterns without relying solely on explicit recurrent mechanisms, provided that this structure is encoded in the context variable.

12 Application: Data-Dependent Member Selection

As an application of the preceding analysis, we study the local decision of whether to activate an exogenous gate channel around a fitted self-gated member. The gate-structure hierarchy controlled by $W_g^{(a)}$ and the channel- inclusion choices controlled by $W_g^{(u)}$ and $W_g^{(s)}$ are distinct axes. The analysis below derives an exact risk identity for the $+s$ versus $-s$ decision under explicit local assumptions and specifies the diagnostics needed for a future prospective application.

12.1 The selection problem

Write the total context entering the gate as the pair (u_t, s_t), where u_t is the *endogenous* summary of recent target dynamics and s_t is the *exogenous* covariate context (Def. 3). Selecting a member amounts to deciding which of the channels $\big(W_g^{(a)}, W_g^{(u)}, W_g^{(s)}\big)$ is active. We focus on the local decision studied below: whether to activate the exogenous channel s (the $+s$ member of the temporal CAF family) or to suppress it ($-s$, a self-gated configuration).

We frame the decision as an excess-risk comparison. Let $R(\cdot)$ denote the expected prediction risk of a trained member. Activating s changes the risk by

$$\Delta R_{+s} \;=\; \underbrace{-\,\Delta\mathrm{Bias}}_{\text{approximation gain}} \;+\; \underbrace{\Delta\mathrm{Var}}_{\text{estimation cost}} \;. \tag{58}$$

The self-gated member is preferable precisely when $\Delta R_{+s} \geq 0$, i.e. when the estimation cost of the extra channel is not repaid by an approximation gain.

12.2 A Rigorous Local Criterion for a Target-Profiled Linearized Surrogate

This subsection gives a finite-sample selection result for a target-profiled linearized surrogate associated with the exogenous gate channel. The result is exact for the profiled ridge problem defined below. It should not be interpreted as an unconditional identity for the risk of an arbitrary fully retrained nonlinear network.

The self-gated member is used to construct a reference residual and common parameter scores. The exogenous channel is then represented by its local tangent score, after removing the linear directions generated by the shared parameters. The resulting criterion compares approximation gain, ridge shrinkage bias, and estimation variance in the target-profiled linear surrogate. A transfer statement for the nonlinear or actually re-optimized network requires additional profile-alignment and Taylor-remainder assumptions.

12.2.1 Nested models and local tangent directions

Let $f_{-s}(\xi;\eta)$ denote the scalar prediction of the complete network when the exogenous gate block is suppressed. Here ξ collects the endogenous pre-activation, the temporal summary, and every other covariate shared by the two members, while $\eta \in \mathbb{R}^{d_0}$, with $d_0 := \dim(\eta)$, collects all their common trainable parameters. Thus $f_{\pm s}$ denote network-level predictions, not only the layer-level pure CAF map Φ_{pure}. Let

$$B := W_g^{(s)} \in \mathbb{R}^{H\times S}, \qquad \theta := \mathrm{vec}(B) \in \mathbb{R}^p, \qquad p := HS,$$

where $\mathrm{vec}(B)$ stacks the columns of B in order, i.e., $\mathrm{vec}(B) = (B_{:,1}^\top, \ldots, B_{:,S}^\top)^\top$. For a single exogenous gate block with hidden dimension H and context dimension S, one has $p = HS$. If several exogenous gate blocks are activated jointly, θ is formed by

concatenating the vectorized blocks and

$$p = \sum_{\ell=1}^{L_{\mathrm{blk}}} H_\ell S_\ell.$$

All subsequent formulas remain valid with this generalized value of p. The member with an active exogenous channel is denoted by $f_{+s}(\xi, s; \eta, \theta)$ and is nested with the self-gated member in the sense that

$$f_{+s}(\xi, s; \eta, 0) = f_{-s}(\xi; \eta). \tag{59}$$

Fix a reference parameter η_0, obtained by fitting the self-gated member on an auxiliary sample. On an independent selection sample $\{(\xi_i, s_i, t_i)\}_{i=1}^{n_{\mathrm{sel}}}$, define

$$e_i := t_i - f_{-s}(\xi_i; \eta_0), \tag{60}$$

$$z_i := \nabla_\eta f_{-s}(\xi_i; \eta_0) \in \mathbb{R}^{d_0}, \tag{61}$$

$$\phi_i := \nabla_\theta f_{+s}(\xi_i, s_i; \eta_0, \theta)|_{\theta=0} \in \mathbb{R}^p. \tag{62}$$

For a vector-valued forecast, each output coordinate and forecast horizon is stacked as a separate scalar row; hence n_{sel} denotes the resulting number of scalar prediction observations. It is local to the selection fold and is distinct from any global sample-size notation used elsewhere in the manuscript.

The score ϕ_i contains the total derivative through the downstream network. To relate it to the CAF gate, consider one CAF layer at the reference member. If

$$g_i^{\mathrm{ref}} = \sigma\Big(W_g^{(a)} a_i + W_g^{(u)} u_i + b_g\Big), \qquad A_i := \mathrm{diag}(\rho(a_i))\,\mathrm{diag}\big(g_i^{\mathrm{ref}} \odot (\mathbf{1} - g_i^{\mathrm{ref}})\big),$$

then, with the column-stacking convention above, the immediate derivative of the CAF output with respect to θ is

$$D_\theta\big[g_i(B) \odot \rho(a_i)\big]\big|_{B=0} = s_i^\top \otimes A_i \in \mathbb{R}^{H \times HS}. \tag{63}$$

Here $\otimes$ denotes the Kronecker product. More explicitly, if $J_i^{\mathrm{down}} \in \mathbb{R}^{1\times H}$ is the Jacobian of the scalar network prediction with respect to this CAF output, this layer contributes

$$\phi_i^{\mathrm{layer}} = \big[J_i^{\mathrm{down}}\big(s_i^\top \otimes A_i\big)\big]^\top \in \mathbb{R}^{HS}. \tag{64}$$

Summing the corresponding chain-rule contributions over every occurrence and all recurrent paths yields the total score in (62). Thus the local design accounts for both the exogenous covariates and the sensitivity of the fitted CAF predictor.

Let $e = (e_1, \ldots, e_{n_{\mathrm{sel}}})^\top$, $Z = (z_1^\top; \ldots; z_{n_{\mathrm{sel}}}^\top)$, and $\Psi = (\phi_1^\top; \ldots; \phi_{n_{\mathrm{sel}}}^\top)$. The shared parameters may be re-optimized in both members. To remove directions that can already be explained, to first order, by the shared parameters, define the orthogonal projector

$$M_Z := I_{n_{\mathrm{sel}}} - ZZ^\dagger, \tag{65}$$

where $Z^\dagger$ denotes the unique Moore–Penrose pseudoinverse of Z. Let $\nu := \operatorname{rank}(M_Z)$ and assume $\nu \geq 1$; if $\operatorname{rank}(Z) = n_{\mathrm{sel}}$, no nonzero profiled residual direction remains and the local exogenous correction is not identifiable by this criterion. Choose $Q \in \mathbb{R}^{n_{\mathrm{sel}} \times \nu}$ such that

$$Q^\top Q = I_\nu, \qquad QQ^\top = M_Z. \tag{66}$$

The profiled residual and exogenous design are

$$r := Q^\top e \in \mathbb{R}^\nu, \qquad X := Q^\top \Psi \in \mathbb{R}^{\nu \times p}. \tag{67}$$

This residualization is the local analogue of comparing the two nested members after allowing both of them to adjust the common parameters η.

Selection-fold and target-distribution profiling.

The matrix Q provides coordinates only for the nuisance-orthogonal subspace of the finite selection fold; it is not an operator that can be applied to a new observation. We therefore distinguish the empirical profiling operator used on the selection fold from its population counterpart under the target or deployment distribution.

For a scalar selection-fold vector $v \in \mathbb{R}^{n_{\mathrm{sel}}}$, define

$$\mathscr{R}_{\mathrm{sel}} v := M_Z v. \tag{68}$$

For a matrix, $\mathscr{R}_{\mathrm{sel}}$ is applied columnwise. Because $QQ^\top = M_Z$ and $Q^\top Q = I_\nu$, the reduced coordinates in Equation (67) satisfy

$$r = Q^\top \mathscr{R}_{\mathrm{sel}} e = Q^\top e, \tag{69}$$

$$X = Q^\top \mathscr{R}_{\mathrm{sel}} \Psi = Q^\top \Psi, \tag{70}$$

and, for every $\theta \in \mathbb{R}^p$,

$$\|r - X\theta\|^2 = \|M_Z(e - \Psi\theta)\|^2. \tag{71}$$

Thus Q is only an isometric coordinate representation of the empirically profiled selection-fold quantities.

To define profiling for a new observation, let (e_0, z_0, ϕ_0) denote the raw reference residual, nuisance score, and exogenous score under the target distribution:

$$e_0 := t_0 - f_{-s}(\xi_0; \eta_0), \tag{72}$$

$$z_0 := \nabla_\eta f_{-s}(\xi_0; \eta_0) \in \mathbb{R}^{d_0}, \tag{73}$$

$$\phi_0 := \nabla_\theta f_{+s}(\xi_0, s_0; \eta_0, \theta)|_{\theta=0} \in \mathbb{R}^p. \tag{74}$$

All target-distribution moments below are conditional on the fitted reference parameter η_0 and on any auxiliary data used to construct it. Define

$$\Sigma_{zz}^{\mathrm{tar}} := \mathbb{E}[z_0 z_0^\top], \tag{75}$$

and assume that the second moments required below are finite. For any scalar square-integrable target random variable v_0, define the target profiling operator

$$\mathscr{R}_{\mathrm{tar}} v_0 := v_0 - z_0^\top (\Sigma_{zz}^{\mathrm{tar}})^\dagger \mathbb{E}[z_0 v_0]. \tag{76}$$

For vector-valued variables, the operator is applied coordinatewise. The operator $\mathscr{R}_{\mathrm{tar}}$ is a no-intercept linear projection onto the span of the nuisance score z_0. If a constant nuisance direction is also to be removed, the score vector must be augmented as

$$\widetilde{z}_0 := (1, z_0^\top)^\top$$

and all profiling formulas must be written in terms of $\widetilde{z}_0$. The profiled target residual and exogenous score are therefore

$$r_0 := \mathscr{R}_{\mathrm{tar}} e_0, \tag{77}$$

$$x_0 := \mathscr{R}_{\mathrm{tar}} \phi_0 = \phi_0 - \left[(\Sigma_{zz}^{\mathrm{tar}})^\dagger \mathbb{E}[z_0 \phi_0^\top]\right]^\top z_0 \in \mathbb{R}^p. \tag{78}$$

This definition removes, under the target distribution, the linear nuisance directions generated by the common parameter score z_0. It is the out-of-sample analogue of applying M_Z on the selection fold and does not require extending the fold-specific matrix Q to a new observation. In an empirical implementation, the moments in Equations (75)– (78) must be estimated on an independent validation fold or by cross-fitting; the final test outcomes must not be used to construct the profiling operator.

For later use, define the raw nonlinear correction at the reference common parameter by

$$d_0(\theta) := f_{+s}(\xi_0, s_0; \eta_0, \theta) - f_{-s}(\xi_0; \eta_0). \tag{79}$$

Its nuisance-profiled target version is $\mathscr{R}_{\mathrm{tar}} d_0(\theta)$.

Remark 23 (Scope of target-distribution profiling) The target operator $\mathscr{R}_{\mathrm{tar}}$ is a population linear projection residual under the target distribution. It is not, in general, the same operator as the random selection-fold projector M_Z, and it is not automatically equal to the effect of re-optimizing the common parameters η on the selection fold. Therefore, the quantities r_0 and x_0 define a target-profiled linear surrogate. An exact statement about the risk of an actually re-optimized nonlinear member requires an additional profile-alignment assumption. Without Assumption 4, the selection criterion should be reported only as a criterion for the target-profiled linearized surrogate.

12.2.2 Explicit assumptions

Assumption 1 (Local profiled model) Conditionally on (η_0, X), the profiled residual satisfies

$$r = X\theta_\star + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0, \qquad \mathrm{Cov}(\varepsilon \mid X) = \sigma^2 I_\nu, \tag{80}$$

for a fixed local parameter $\theta_\star \in \mathbb{R}^p$ and $0 < \sigma^2 < \infty$.

Assumption 2 (Independent profiled target observation) Conditionally on the fitted reference parameter and on any auxiliary data used to construct the target profiling operator $\mathscr{R}_{\mathrm{tar}}$, let (r_0, x_0) be defined by Equation (78). The raw target observation (e_0, z_0, ϕ_0) is independent of the complete selection sample (X, ε) and obeys the profiled local model

$$r_0 = x_0^\top \theta_\star + \varepsilon_0, \qquad \mathbb{E}[\varepsilon_0 \mid x_0] = 0, \qquad \mathbb{E}[\varepsilon_0^2 \mid x_0] = \sigma^2. \tag{81}$$

Moreover,

$$C := \mathbb{E}[x_0 x_0^\top] \succeq 0, \qquad \mathrm{tr}(C) < \infty. \tag{82}$$

The same local coefficient $\theta_\star$ is assumed to govern the profiled selection model in Assumption 1 and the profiled target model above. This is a transportability assumption between the selection and target distributions; arbitrary concept shift is outside the scope of the criterion.

If C and the target profiling moments are replaced by estimates from an independent validation set, all subsequent risk statements are additionally conditional on those validation covariates and on the fitted profiling operator. Cross-fitting may be used to avoid reusing an observation both to estimate $\mathscr{R}_{\mathrm{tar}}$ and to evaluate a target risk or a Taylor remainder.

Assumption 3 (Local validity on the selection and target distributions) There exist $\delta > 0$ and $L < \infty$ such that the following two conditions hold for every $\theta \in \mathbb{R}^p$ with $\|\theta\| \le \delta$.

(i) **Selection-fold remainder.** Let $d_{\mathrm{sel}}(\theta) \in \mathbb{R}^{n_{\mathrm{sel}}}$ collect the raw nonlinear prediction corrections

$$[d_{\mathrm{sel}}(\theta)]_i := f_{+s}(\xi_i, s_i; \eta_0, \theta) - f_{-s}(\xi_i; \eta_0). \tag{83}$$

After applying the empirical nuisance profile and expressing it in the Q-coordinates, each reduced coordinate satisfies

$$[Q^\top \mathscr{R}_{\mathrm{sel}} d_{\mathrm{sel}}(\theta)]_j = x_j^\top \theta + q_j^{\mathrm{sel}}(\theta), \qquad |q_j^{\mathrm{sel}}(\theta)| \le \frac{L}{2}\|\theta\|^2, \qquad j = 1, \dots, \nu, \tag{84}$$

where $x_j^\top$ is row j of X.

(ii) **Target-distribution remainder.** For an independent target observation, the nonlinear correction in Equation (79), after applying the target profiling operator, satisfies

$$\mathscr{R}_{\mathrm{tar}} d_0(\theta) = x_0^\top \theta + q_0^{\mathrm{tar}}(\theta), \qquad |q_0^{\mathrm{tar}}(\theta)| \le \frac{L}{2}\|\theta\|^2 \quad \text{almost surely under the target distribution.} \tag{85}$$

The first condition controls the accuracy of the local design used to estimate the exogenous direction. The second is a separate transport condition: it requires the same first-order representation, with a quadratic remainder bound, for observations drawn from the target distribution on which prediction risk is evaluated. The linearized risk identity does not require an almost-sure bound on the ridge estimator. The nonlinear transfer result is invoked only on the local event

$$E_\delta := \left\{ \|\widehat{\theta}_\lambda\| \le \delta \right\}.$$

If a high-probability rather than deterministic local statement is desired, one may assume

$$\mathbb{P}(E_\delta \mid X) \ge 1 - \alpha$$

for a specified $\alpha \in (0, 1)$. In that case, the nonlinear conclusion is conditional on E_δ unless an additional tail bound is supplied. In practice, both $\|\widehat{\theta}_\lambda\|$ and held-out versions of the selection and target Taylor errors should be reported.

Assumption 1 is a finite-sample, fixed-design local model. Homoscedasticity is used only to obtain the closed-form trace term below. In particular, stacking several output coordinates or horizons can induce heteroscedasticity or within-forecast dependence. With a general conditional covariance Ω, the same proof replaces $\sigma^2 X^\top X$ by $X^\top \Omega X$; the resulting estimator-covariance contribution is $\nu^{-2} H_\lambda^{-1} X^\top \Omega X H_\lambda^{-1}$.

Assumption 4 (Profile alignment for actual refitting) Let $\mathscr{D}_{\mathrm{sel}}$ denote the complete selection sample, and let $\widehat{\eta}_{+s}(\theta; \mathscr{D}_{\mathrm{sel}})$ denote the common-parameter fit produced by the actual training or refitting procedure for the $+s$ member when the exogenous local parameter is fixed at θ. Define the actual target correction

$$d_0^{\mathrm{act}}(\theta) := f_{+s}(\xi_0, s_0; \widehat{\eta}_{+s}(\theta; \mathscr{D}_{\mathrm{sel}}), \theta) - f_{-s}(\xi_0; \eta_0).$$

Assume that there exists a finite function $\Gamma_{\mathrm{prof}}(\delta)$ such that, for every $\|\theta\| \le \delta$,

$$\left(\mathbb{E}\left[\left| \mathscr{R}_{\mathrm{tar}} d_0^{\mathrm{act}}(\theta) - \mathscr{R}_{\mathrm{tar}} d_0(\theta) \right|^2 \middle| \mathscr{D}_{\mathrm{sel}} \right] \right)^{1/2} \le \Gamma_{\mathrm{prof}}(\delta).$$

This condition quantifies the discrepancy between the target-profiled frozen reference model and the correction produced by the actual re-optimization of the common parameters.

12.2.3 Ridge estimator, Hessian, and approximation gain

For $\lambda > 0$, estimate the exogenous direction by

$$\widehat{\theta}_\lambda := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{\nu} \|r - X\theta\|^2 + \lambda \|\theta\|^2 \right\}. \tag{86}$$

Define

$$G := \frac{1}{\nu} X^\top X, \qquad H_\lambda := G + \lambda I_p. \tag{87}$$

The matrix $2H_\lambda$ is the Hessian of the objective in (86), and $H_\lambda \succ 0$ even when G is singular. The estimator has the closed form

$$\widehat{\theta}_\lambda = H_\lambda^{-1} \frac{X^\top r}{\nu}. \tag{88}$$

Define the conditional target risk of the self-gated local member by

$$R_{-s}^{\mathrm{lin}} := \mathbb{E}\left[r_0^2 \mid X \right]. \tag{89}$$

The population approximation gain supplied by the exogenous tangent space is defined, on the same conditional target distribution, by

$$\mathscr{I}_s := R_{-s}^{\mathrm{lin}} - \inf_{\theta \in \mathbb{R}^p} \mathbb{E}\left[(r_0 - x_0^\top \theta)^2 \mid X \right] = \theta_\star^\top C \theta_\star. \tag{90}$$

The last identity follows from Assumption 2 and the stipulated independence of the target observation from (X, ε). Therefore, $\mathscr{I}_s$ is not raw mutual information and is not an informal notion of relevance: it is the reduction in target-distribution squared

approximation error achieved by the best exogenous correction in the local model. It is zero if and only if $C^{1/2}\theta_\star = 0$.

The conditioning of the regularized estimation problem is measured by

$$\kappa_\lambda := \kappa(H_\lambda) = \frac{\mu_{\max}(G) + \lambda}{\mu_{\min}(G) + \lambda}, \tag{91}$$

where zero eigenvalues are included in $\mu_{\min}(G)$. This definition is always finite for $\lambda > 0$ and concerns the profiled exogenous Hessian, not the layerwise Jacobian $D_a y$. The CAF Jacobian influences κ_λ through the score matrix Ψ in (63), but the two condition numbers are not interchangeable.

12.2.4 Exact conditional risk identity

Theorem 25 (Exact local risk decomposition for the target-profiled linear surrogate) *Suppose Assumptions 1 and 2 hold. In addition to R^{lin}_{-s} in (89), define*

$$R^{\mathrm{lin}}_{+s,\lambda} := \mathbb{E}\Big[(r_0 - x_0^\top \widehat{\theta}_\lambda)^2 \mid X\Big]. \tag{92}$$

The expectation is over both the selection-sample noise and the independent target observation. This quantity is the conditional target risk of the target-profiled linear surrogate. Without an additional profile-alignment assumption, it is not identified with the risk of a fully retrained nonlinear $+s$ network. Then

$$R^{\mathrm{lin}}_{-s} = \sigma^2 + \mathscr{I}_s, \tag{93}$$

and

$$R^{\mathrm{lin}}_{+s,\lambda} = \sigma^2 + \underbrace{\lambda^2 \theta_\star^\top H_\lambda^{-1} C H_\lambda^{-1} \theta_\star}_{\mathscr{B}_s(\lambda)} + \underbrace{\frac{\sigma^2}{\nu} \operatorname{tr}\big(C H_\lambda^{-1} G H_\lambda^{-1}\big)}_{\mathscr{V}_s(\lambda)}. \tag{94}$$

Consequently, the exact local risk gap is

$$\boxed{\Delta R^{\mathrm{lin}}_{+s}(\lambda) := R^{\mathrm{lin}}_{+s,\lambda} - R^{\mathrm{lin}}_{-s} = -\mathscr{I}_s + \mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda).} \tag{95}$$

Thus the self-gated target-profiled linear surrogate has no larger local target risk if and only if

$$\boxed{\mathscr{I}_s \le \mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda).} \tag{96}$$

Proof By Assumption 1 and (88),

$$\widehat{\theta}_\lambda = H_\lambda^{-1}\left(G\theta_\star + \frac{X^\top \varepsilon}{\nu}\right), \tag{97}$$

$$\mathbb{E}[\widehat{\theta}_\lambda \mid X] = H_\lambda^{-1} G \theta_\star = \theta_\star - \lambda H_\lambda^{-1} \theta_\star, \tag{98}$$

$$\operatorname{Cov}(\widehat{\theta}_\lambda \mid X) = \frac{\sigma^2}{\nu} H_\lambda^{-1} G H_\lambda^{-1}. \tag{99}$$

The self-gated correction is zero. Hence, by Assumption 2,

$$R^{\mathrm{lin}}_{-s} = \mathbb{E}\Big[(x_0^\top \theta_\star + \varepsilon_0)^2 \mid X\Big] = \sigma^2 + \theta_\star^\top C \theta_\star.$$

For the ridge member, independence and centering of ε_0 give

$$R^{\mathrm{lin}}_{+s,\lambda} = \sigma^2 + \mathbb{E}\Big[(\widehat{\theta}_\lambda - \theta_\star)^\top C(\widehat{\theta}_\lambda - \theta_\star) \mid X\Big].$$

Applying the identity $\mathbb{E}[v^\top Cv] = (\mathbb{E}v)^\top C(\mathbb{E}v) + \mathrm{tr}(C\,\mathrm{Cov}(v))$ together with (98) and (99) yields (94). Subtracting (93) proves (95) and (96). □

The theorem separates three quantities with distinct meanings for the target-profiled linear surrogate: $\mathscr{I}_s$ is the local approximation gain; $\mathscr{B}_s(\lambda)$ is the shrinkage bias caused by ridge regularization; and $\mathscr{V}_s(\lambda)$ is the finite-sample estimation variance. In particular, the number p of additional gate parameters does not by itself determine the variance. The relevant quantity is the spectrum of the profiled design through $CH_\lambda^{-1}GH_\lambda^{-1}$.

12.2.5 Spectral form and explicit conditioning bounds

Corollary 26 (Matched-design spectral decomposition) Under the assumptions of Theorem 25, suppose that the target covariance equals the selection Gram matrix, $C = G$. Let $G = U\,\mathrm{diag}(\mu_1, \ldots, \mu_p)U^\top$, with $\mu_j \geq 0$, and write $b := U^\top\theta_\star$. Then

$$\mathscr{I}_s = \sum_{j=1}^{p} \mu_j b_j^2, \tag{100}$$

$$\mathscr{B}_s(\lambda) = \sum_{j=1}^{p} \frac{\lambda^2\mu_j}{(\mu_j+\lambda)^2} b_j^2, \tag{101}$$

$$\mathscr{V}_s(\lambda) = \frac{\sigma^2}{\nu} \sum_{j=1}^{p} \left(\frac{\mu_j}{\mu_j+\lambda}\right)^2. \tag{102}$$

The quantity

$$d_{2,\lambda} := \sum_{j=1}^{p} \left(\frac{\mu_j}{\mu_j+\lambda}\right)^2 \leq \mathrm{rank}(G) \leq \min(\nu, p) \tag{103}$$

is the second-order effective dimension of the exogenous tangent space, and $\mathscr{V}_s(\lambda) = \sigma^2 d_{2,\lambda}/\nu$.

Proof Because G and $H_\lambda = G + \lambda I_p$ have the same eigenvectors, substitution of their simultaneous eigendecomposition into (90) and (94) gives (100)– (102). The rank bound follows because every summand in (103) belongs to $[0, 1]$ and vanishes when $\mu_j = 0$. □

The matched-design expression shows why replacing the variance by a generic $O(p/\nu)$ term is unnecessarily crude: only identifiable exogenous directions contribute, and ridge continuously discounts weak directions.

Proposition 27 (Conditioning-sensitive finite-sample bounds) *Let* $\mu_{\max} := \mu_{\max}(G)$ *and* $\mu_{\min} := \mu_{\min}(G)$. *Under the assumptions of Theorem 25,*

$$\mathscr{B}_s(\lambda) \leq \frac{\lambda^2\|C\|\,\|\theta_\star\|^2}{(\mu_{\min}+\lambda)^2}, \tag{104}$$

$$\mathscr{V}_s(\lambda) \le \frac{\sigma^2}{\nu}\frac{\operatorname{tr}(C)\,\mu_{\max}}{(\mu_{\min}+\lambda)^2}. \tag{105}$$

Equivalently, since $(\mu_{\min}+\lambda)^{-2} = \kappa_\lambda^2(\mu_{\max}+\lambda)^{-2}$,

$$\mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda) \le \frac{\kappa_\lambda^2}{(\mu_{\max}+\lambda)^2}\left[\lambda^2\|C\|\,\|\theta_\star\|^2 + \frac{\sigma^2}{\nu}\operatorname{tr}(C)\mu_{\max}\right]. \tag{106}$$

If, additionally, $C \succeq c_{\min} I_p$ *for some* $c_{\min} > 0$*, and if* $r_G := \operatorname{rank}(G) \ge 1$ *and* $\mu_{\min}^+$ *is the smallest strictly positive eigenvalue of* G*, then*

$$\mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda) \ge \frac{c_{\min}}{(\mu_{\max}+\lambda)^2}\left[\lambda^2\|\theta_\star\|^2 + \frac{\sigma^2}{\nu} r_G \mu_{\min}^+\right]. \tag{107}$$

If $r_G = 0$*, then* $G = 0$ *and the variance term* $\mathscr{V}_s(\lambda)$ *is zero. In this degenerate case, the lower bound is interpreted using only the bias contribution, and the quantity* $\mu_{\min}^+$ *is not defined. Consequently, the following are conservative sufficient conditions:*

$$\mathscr{I}_s > \frac{\kappa_\lambda^2}{(\mu_{\max}+\lambda)^2}\left[\lambda^2\|C\|\,\|\theta_\star\|^2 + \frac{\sigma^2}{\nu}\operatorname{tr}(C)\mu_{\max}\right] \quad \Longrightarrow \quad +s \text{ is preferred,} \tag{108}$$

$$\mathscr{I}_s \le \frac{c_{\min}}{(\mu_{\max}+\lambda)^2}\left[\lambda^2\|\theta_\star\|^2 + \frac{\sigma^2}{\nu} r_G \mu_{\min}^+\right] \quad \Longrightarrow \quad -s \text{ is preferred.} \tag{109}$$

Proof Since $H_\lambda \succeq (\mu_{\min}+\lambda) I_p$, $\|H_\lambda^{-1}\| \le (\mu_{\min}+\lambda)^{-1}$, which gives (104). Moreover,

$$0 \preceq H_\lambda^{-1} G H_\lambda^{-1} \preceq \frac{\mu_{\max}}{(\mu_{\min}+\lambda)^2} I_p.$$

Taking the trace after multiplication by $C \succeq 0$ gives (105). Equation (106) follows from (91).

For the lower bound, $C \succeq c_{\min} I_p$ implies

$$\mathscr{B}_s(\lambda) \ge c_{\min}\lambda^2\|H_\lambda^{-1}\theta_\star\|^2 \ge \frac{c_{\min}\lambda^2\|\theta_\star\|^2}{(\mu_{\max}+\lambda)^2}.$$

On the range of G, each positive eigenvalue satisfies

$$\frac{\mu_j}{(\mu_j+\lambda)^2} \ge \frac{\mu_{\min}^+}{(\mu_{\max}+\lambda)^2}.$$

Therefore,

$$\mathscr{V}_s(\lambda) \ge \frac{\sigma^2 c_{\min}}{\nu}\operatorname{tr}\big(H_\lambda^{-1} G H_\lambda^{-1}\big) \ge \frac{\sigma^2 c_{\min}}{\nu}\frac{r_G \mu_{\min}^+}{(\mu_{\max}+\lambda)^2},$$

which proves (107). The two selection statements follow from the exact identity (95). □

12.2.6 Control of the nonlinear remainder

The preceding results are exact for the profiled linearized model. The next statement quantifies the scope of the approximation.

Proposition 28 (Nonlinear transfer for the target-profiled surrogate) *Suppose Assumption 3 holds. Fix a vector* $\theta \in \mathbb{R}^p$ *satisfying* $\|\theta\| \le \delta$*, and define*

$$R_{+s}^{\mathrm{lin}}(\theta) := \mathbb{E}\Big[(r_0 - x_0^\top\theta)^2 \mid X\Big], \tag{110}$$

$$R_{+s}^{\mathrm{nonlin}}(\theta) := \mathbb{E}\big[(r_0 - \mathscr{R}_{\mathrm{tar}} d_0(\theta))^2 \mid X\big]. \tag{111}$$

Then

$$\left|R_{+s}^{\mathrm{nonlin}}(\theta) - R_{+s}^{\mathrm{lin}}(\theta)\right| \le L\|\theta\|^2 \sqrt{R_{+s}^{\mathrm{lin}}(\theta)} + \frac{L^2}{4}\|\theta\|^4. \tag{112}$$

Consequently, if $\theta = \widehat{\theta}_\lambda$ and the event $E_\delta = \{\|\widehat{\theta}_\lambda\| \le \delta\}$ occurs, then the right-hand side is bounded by

$$L\delta^2 \sqrt{R_{+s}^{\mathrm{lin}}(\widehat{\theta}_\lambda)} + \frac{L^2\delta^4}{4}.$$

Therefore, the ordering of the target-profiled nonlinear surrogate and the self-gated member is determined by the linearized risk margin whenever

$$\left|R_{+s}^{\mathrm{lin}}(\theta) - R_{-s}^{\mathrm{lin}}\right| > L\|\theta\|^2 \sqrt{R_{+s}^{\mathrm{lin}}(\theta)} + \frac{L^2}{4}\|\theta\|^4.$$

If this margin condition fails, the local result does not determine the nonlinear ordering.

Proof Set

$$u := r_0 - x_0^\top \theta, \qquad q_0 := \mathscr{R}_{\mathrm{tar}} d_0(\theta) - x_0^\top \theta.$$

By Assumption 3,

$$|q_0| \le \frac{L}{2}\|\theta\|^2 \qquad \text{almost surely.}$$

The nonlinear residual is $u - q_0$, and hence

$$(u - q_0)^2 - u^2 = q_0^2 - 2uq_0.$$

Taking conditional expectations and applying the Cauchy–Schwarz inequality gives

$$\left|R_{+s}^{\mathrm{nonlin}}(\theta) - R_{+s}^{\mathrm{lin}}(\theta)\right| \le \mathbb{E}[q_0^2 \mid X] + 2\sqrt{\mathbb{E}[u^2 \mid X]\,\mathbb{E}[q_0^2 \mid X]}.$$

Using the bound on q_0 yields Equation (112). □

Under Assumption 4, fix θ with $\|\theta\| \le \delta$ and define, under the common conditioning on the complete selection sample $\mathscr{D}_{\mathrm{sel}}$,

$$R_{+s}^{\mathrm{act}}(\theta) := \mathbb{E}\Big[\big(r_0 - \mathscr{R}_{\mathrm{tar}} d_0^{\mathrm{act}}(\theta)\big)^2 \Big| \mathscr{D}_{\mathrm{sel}}\Big], \tag{113}$$

$$R_{+s}^{\mathrm{lin,sel}}(\theta) := \mathbb{E}\big[(r_0 - x_0^\top \theta)^2 \big| \mathscr{D}_{\mathrm{sel}}\big]. \tag{114}$$

Let

$$\Gamma_\delta := \Gamma_{\mathrm{prof}}(\delta) + \frac{L\delta^2}{2}.$$

The triangle inequality in conditional L^2, followed by the same risk comparison used above, gives

$$\left|R_{+s}^{\mathrm{act}}(\theta) - R_{+s}^{\mathrm{lin,sel}}(\theta)\right| \le 2\Gamma_\delta \sqrt{R_{+s}^{\mathrm{lin,sel}}(\theta)} + \Gamma_\delta^2.$$

The same statement applies to any $\mathscr{D}_{\mathrm{sel}}$-measurable fitted value of θ, including $\widehat{\theta}_\lambda$ on the event E_δ. Thus, the local selection conclusion is valid for the actually re-optimized member only when the corresponding linearized risk margin dominates this enlarged remainder bound.

12.2.7 Estimable diagnostics and decision protocol

The criterion is operational only if its inputs are estimated without reusing the evaluation outcomes. A valid implementation uses sample splitting or cross-fitting:

1. Fit the self-gated reference parameter η_0 on an auxiliary fold.
2. On a disjoint selection fold, compute e, the nuisance scores Z, and the exogenous scores Ψ by automatic differentiation; then construct Q, r, X, G, and H_λ from (67)–(87).
3. Choose λ by nested validation, not by the final evaluation fold. Report κ_λ, the effective dimension $d_{2,\lambda}$, and the observed Taylor remainder.
4. Estimate the target covariance C on an independent validation fold. Estimate σ^2 from out-of-fold profiled residuals. The local parameter $\theta_\star$ may be estimated by a cross-fitted regression of r on X; uncertainty in $\mathscr{I}_s$, $\mathscr{B}_s$, and $\mathscr{V}_s$ should be reported by a paired or cluster-aware bootstrap that respects the temporal sampling unit.
5. Apply the exact plug-in criterion corresponding to (96), and confirm the selected member on a final untouched test fold.

An equivalent held-out estimate of the gain achieved by any correction $\widetilde{\theta}$, fitted independently of a validation observation, is

$$\mathscr{I}_s(\widetilde{\theta}) := \mathbb{E}\Big[r_0^2 - (r_0 - x_0^\top \widetilde{\theta})^2\Big] = 2\widetilde{\theta}^\top \mathbb{E}[x_0 r_0] - \widetilde{\theta}^\top C \widetilde{\theta}. \tag{115}$$

At $\widetilde{\theta} = \theta_\star$ this equals $\mathscr{I}_s$. Equation (115) therefore provides a directly estimable, out-of-sample diagnostic and avoids interpreting raw correlation between s and the target as incremental exogenous information.

Remark 24 (Scope of the criterion) The criterion is exact for the target-profiled linearized surrogate under Assumptions 1 and 2. It is not a universal statement that noisy context always increases risk, nor does it claim that multiplicative gating is the optimal vehicle for informative context. Its use for the nonlinear CAF model requires a small Taylor remainder and a risk margin large enough to satisfy Proposition 28. If these conditions fail, member selection must be based on direct nested validation rather than on the local criterion.

12.3 Estimable Diagnostics and Empirical Scope

The local selection criterion of Theorem 25 is operational only when its terms are estimated independently of the final evaluation outcomes. This subsection therefore defines the quantities required for a future prospective application of the criterion by sample splitting or cross-fitting.

12.3.1 Diagnostics implied by the local risk identity

Recall the exact local risk gap

$$\Delta R^{\mathrm{lin}}_{+s}(\lambda) = -\mathscr{I}_s + \mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda), \tag{116}$$

where $\mathscr{I}_s$ is the target-distribution approximation gain, $\mathscr{B}_s(\lambda)$ is the shrinkage bias, and $\mathscr{V}_s(\lambda)$ is the finite-sample estimation variance. The local criterion selects the exogenous member when this gap is negative and the self-gated member when it is nonnegative.

Let $\widehat{\theta}_\lambda$ be fitted on a selection fold. Separately, let $\widetilde{\theta}_\star$ be a cross-fitted estimate of the local population coefficient $\theta_\star$, obtained without using the validation observations below. Let $\{(x_j^{\mathrm{val}}, r_j^{\mathrm{val}})\}_{j=1}^{m_{\mathrm{val}}}$ be an independent validation sample of profiled exogenous scores and residuals. Define

$$\widehat{C} := \frac{1}{m_{\mathrm{val}}} \sum_{j=1}^{m_{\mathrm{val}}} x_j^{\mathrm{val}} (x_j^{\mathrm{val}})^\top, \tag{117}$$

$$\widehat{\mathscr{I}_s} := \frac{1}{m_{\mathrm{val}}} \sum_{j=1}^{m_{\mathrm{val}}} \left[(r_j^{\mathrm{val}})^2 - \left(r_j^{\mathrm{val}} - (x_j^{\mathrm{val}})^\top \widetilde{\theta}_\star \right)^2 \right]. \tag{118}$$

Equation (118) is the empirical version of the held-out gain in Equation (115), evaluated at an independently estimated population direction. Unlike a raw correlation between s and the target, it measures the incremental contribution of the exogenous tangent space after profiling out the shared local directions. At $\widetilde{\theta}_\star = \theta_\star$, its conditional expectation is exactly $\mathscr{I}_s$. With an estimated $\widetilde{\theta}_\star$, its uncertainty and estimation error must be propagated through the cross-fitting or bootstrap procedure. The ridge estimator $\widehat{\theta}_\lambda$ is kept distinct from $\widetilde{\theta}_\star$: substituting the former in Equation (118) would estimate the realized gain of the regularized correction and then adding the ridge bias and variance below would double count parts of the same prediction error.

On the selection fold, let

$$\widehat{G} = \frac{1}{\nu} X^\top X, \qquad \widehat{H}_\lambda = \widehat{G} + \lambda I_p, \tag{119}$$

and estimate the profiled residual variance by an out-of-fold estimator $\widehat{\sigma}^2$. Using the cross-fitted estimate $\widetilde{\theta}_\star$ defined above, set

$$\widehat{\mathscr{B}}_s(\lambda) := \lambda^2 \widetilde{\theta}_\star^\top \widehat{H}_\lambda^{-1} \widehat{C} \widehat{H}_\lambda^{-1} \widetilde{\theta}_\star, \tag{120}$$

$$\widehat{\mathscr{V}}_s(\lambda) := \frac{\widehat{\sigma}^2}{\nu} \operatorname{tr}\left(\widehat{C} \widehat{H}_\lambda^{-1} \widehat{G} \widehat{H}_\lambda^{-1} \right). \tag{121}$$

The resulting plug-in estimate of the local risk gap is

$$\widehat{\Delta R}^{\mathrm{lin}}_{+s}(\lambda) := -\widehat{\mathscr{I}_s} + \widehat{\mathscr{B}}_s(\lambda) + \widehat{\mathscr{V}}_s(\lambda). \tag{122}$$

This decomposition should be reported term by term. Reporting only the sign of (122) would hide whether a decision is driven by weak incremental signal, regularization bias, or finite-sample variance.

Two complementary spectral diagnostics describe the estimability of the exogenous tangent space:

$$\widehat{\kappa}_\lambda := \kappa(\widehat{H}_\lambda) = \frac{\mu_{\max}(\widehat{G}) + \lambda}{\mu_{\min}(\widehat{G}) + \lambda}, \tag{123}$$

$$\widehat{d}_{2,\lambda} := \operatorname{tr}\left[\left(\widehat{G}(\widehat{G} + \lambda I_p)^{-1}\right)^2\right] = \sum_{j=1}^{p} \left(\frac{\mu_j(\widehat{G})}{\mu_j(\widehat{G}) + \lambda}\right)^2. \tag{124}$$

Here $\widehat{\kappa}_\lambda$ concerns the regularized profiled estimation problem, not the layerwise activation Jacobian $D_a y$. The CAF Jacobian affects the score matrix X through the chain rule, but $\kappa(D_a y)$ and $\widehat{\kappa}_\lambda$ are not interchangeable. Similarly, the nominal number p of exogenous gate parameters is not by itself the effective variance cost; weak or unidentifiable directions are discounted by $\widehat{d}_{2,\lambda}$.

12.3.2 Local-validity diagnostics

A decision obtained from Equation (122) is meaningful for the nonlinear CAF predictor only when the local approximation is adequate. We therefore accompany the risk diagnostics with

$$\|\widehat{\theta}_\lambda\|.$$

For each validation observation, let $\widehat{\mathscr{R}}_{\mathrm{val},-j}$ denote a cross-fitted validation profiling operator constructed without using observation j. Define the profiled nonlinear correction

$$\Delta f^{\mathrm{nonlin}}_{j,\mathrm{prof}} := \widehat{\mathscr{R}}_{\mathrm{val},-j}\left[f_{+s}(\xi_j^{\mathrm{val}}, s_j^{\mathrm{val}}; \eta_0, \widehat{\theta}_\lambda) - f_{-s}(\xi_j^{\mathrm{val}}; \eta_0)\right].$$

The held-out Taylor diagnostic is then

$$\widehat{\mathscr{E}}_{\mathrm{Taylor}} := \frac{1}{m_{\mathrm{val}}} \sum_{j=1}^{m_{\mathrm{val}}} \left|\Delta f^{\mathrm{nonlin}}_{j,\mathrm{prof}} - (x_j^{\mathrm{val}})^\top \widehat{\theta}_\lambda\right|^2. \tag{125}$$

The same profiling operator must be used in the nonlinear correction and in the score x_j^{val}. If a raw nonlinear correction is used instead, it must be compared with the corresponding raw, unprofiled score. The norm verifies whether the fitted correction remains inside the local ball of Assumption 3, while $\widehat{\mathscr{E}}_{\mathrm{Taylor}}$ measures the first-order approximation error on observations not used to fit $\widehat{\theta}_\lambda$.

The nonlinear transfer bound of Proposition 28 further requires the linearized risk margin to dominate the Taylor-remainder contribution. Therefore, a small estimated gap close to zero should not be converted into a categorical selection claim merely because its point estimate has one sign.

12.3.3 Decision protocol and uncertainty

For each candidate value of λ, the following protocol is used:

1. fit the self-gated reference member on an auxiliary training fold;
2. construct the profiled residuals and exogenous scores on a disjoint selection fold, and fit $\widehat{\theta}_\lambda$ there;
3. choose λ by nested validation rather than by final-test performance;
4. estimate $\widehat{C}$, $\widehat{\mathscr{I}_s}$, $\widehat{\sigma}^2$, $\widehat{\mathscr{B}}_s(\lambda)$, and $\widehat{\mathscr{V}}_s(\lambda)$ without using the final test fold;
5. report $\widehat{\kappa}_\lambda$, $\widehat{d}_{2,\lambda}$, $\|\widehat{\theta}_\lambda\|$, and $\widehat{\mathscr{E}}_{\mathrm{Taylor}}$ as validity diagnostics;
6. quantify uncertainty in $\widehat{\Delta R}^{\mathrm{lin}}_{+s}(\lambda)$ with a paired or cluster-aware bootstrap whose resampling unit respects the temporal dependence of the forecasting task;
7. confirm the selected member once on a final untouched test fold.

Let $[L_\Delta, U_\Delta]$ denote a confidence interval for the local risk gap. We use the conservative three-way rule

$$\begin{cases} U_\Delta < 0, & \text{select the exogenous member } (+s), \\ L_\Delta > 0, & \text{select the self-gated member } (-s), \\ 0 \in [L_\Delta, U_\Delta], & \text{declare the local criterion inconclusive.} \end{cases} \tag{126}$$

The inconclusive outcome is important: it prevents a small, uncertain local risk difference from being presented as evidence for either member. If the Taylor diagnostic is poor or the fitted correction leaves the local region, the local criterion is not applied and member selection reverts to direct nested validation, as stated in Remark 24.

A prospective empirical validation should compute all diagnostics before examining the final $+s/-s$ test contrast, apply Equation (126), and compare the selected member with the winner on an untouched test fold. The present article derives the criterion and its protocol but does not claim that such a prospective validation has already been performed.

13 Experimental Evaluation

This section evaluates the recurrent CAF forecaster on four temporal datasets under two explicit information regimes. It first defines the common forecasting, preprocessing, optimization, and statistical contracts; it then compares CAF with local and specialized references, studies gate structure, initialization, persistence, additive routing, and state preservation, and closes by relating the finite-forecaster evidence to the scope of the theoretical results. All empirical claims are conditional on the reported target, partition, model roster, adapter, and optimization budget, and none is presented as a proof of an operator-level theorem.

13.1 Experimental Protocol, Reference Models, and Ablation Studies

The protocol is designed to separate three questions: the predictive value of the canonical CAF relative to reference models, the contribution of its individual components, and the effect of extensions that modify its architecture. We therefore distinguish comparisons between architectures, ablations of the pure CAF, and complementary controls. The analytic properties of the CAF operator (Sections 3–4) motivate these comparisons without constituting a prior guarantee of predictive gain, training convergence, or stability of the complete forecaster.

Forecasting task and information contract.

For a dataset d, a window consists of a history of length L_d and a forecasting horizon P_d. At forecast origin τ, the historical inputs are

$$\mathscr{H}_\tau = \big(X_{\tau-L_d+1:\tau},\, S_{\tau-L_d+1:\tau},\, Y_{\tau-L_d+1:\tau}\big),$$

where $x_t \in \mathbb{R}^D$ denotes the observed covariates, $s_t \in \mathbb{R}^S$ the context, and $y_t \in \mathbb{R}$ the single target. Every model outputs $\widehat{Y}_{\tau+1:\tau+P_d} \in \mathbb{R}^{P_d\times 1}$. The target history is supplied explicitly to all forecasting models; it is not information reserved to the CAF.

Two regimes are distinguished. In the `history` regime, only the histories are accessible and the future-context tensor is replaced by zero in the training, validation, and test sets. In the `known_future` regime, models may additionally exploit $S_{\tau+1:\tau+P_d}$, restricted to variables genuinely known at origin τ, such as calendar descriptors or scheduled commands. No future observation of the target or of the sensors is permitted. When a loader reconstructs a calendar from an assumed origin, this provenance must be reported: it is a reconstructed temporal index, not necessarily an authenticated civil calendar. The two regimes constitute distinct experiments on the same partitions; their results are not aggregated as if they relied on the same information.

Partitions, windows, and preprocessing.

For chronological forecasting, the observations are divided into three consecutive and disjoint raw blocks devoted to training, validation, and test. Sliding windows with unit stride are constructed strictly within each block: neither the history nor the horizon of a window crosses a partition boundary. Hence a block producing $N_{d,b}$ windows, for $b \in \{\mathrm{tr}, \mathrm{val}, \mathrm{te}\}$, contains

$$B_{d,b} = N_{d,b} + L_d + P_d - 1 \tag{127}$$

observations. The numbers of windows, history lengths, horizons, and temporal bounds are specific to each dataset and recorded in its experimental configuration. This protocol does not allow absolute errors to be compared directly with published results obtained under other partitions or with histories overlapping the boundaries.

Standardization of X and Y uses exclusively the statistics of the raw training block, which are then retained for validation and test:

$$x^{\mathrm{n}}_{t,j} = \frac{x_{t,j} - \mu^{\mathrm{tr}}_{x,j}}{\sigma^{\mathrm{tr}}_{x,j} + \varepsilon}, \qquad y^{\mathrm{n}}_t = \frac{y_t - \mu^{\mathrm{tr}}_y}{\sigma^{\mathrm{tr}}_y + \varepsilon}, \qquad \varepsilon = 10^{-8}. \tag{128}$$

Any categorical vocabularies and learned context transformations are likewise estimated on the training block. In addition, the CAF and the local models center X^{n} and the history Y^{n} by their temporal mean within the observed window. The external adapters retain the internal preprocessing of their architecture: the evaluated TimeXer configuration keeps its internal normalization enabled (`use_norm=True`). The normalization behavior of CrossLinear is not asserted here because it depends on the exact pinned implementation used by the adapter. Any remaining-useful-life experiments on FD002 fall under a separate protocol, with engine-wise partitioning and a different target contract; they are not assimilated to this observed-target forecasting benchmark.

Canonical CAF: recurrent realization of the pure CAF.

Let H denote the hidden dimension, $\rho = \tanh$, and σ the logistic sigmoid. To lighten notation, the variables below are on the normalized scale and $x_t^{\mathrm{c}}, y_t^{\mathrm{c}}$ denote the centered histories. The canonical CAF, identified as **C0**, uses the pre-activations

$$a_t^{\mathrm{E}} = W_{\mathrm{E}}[x_t^{\mathrm{c}}; y_t^{\mathrm{c}}] + b_{\mathrm{E}} + W_h h_{t-1}^{\mathrm{E}}, \tag{129}$$

$$a_k^{\mathrm{D}} = w_{\mathrm{D}} r_{k-1} + b_{\mathrm{D}} + W_h h_{k-1}^{\mathrm{D}}, \qquad k = 1, \dots, P_d. \tag{130}$$

The encoder's initial state is zero and its final state directly initializes the decoder. The recurrent matrix W_h and the gate parameters are shared between the two phases, whereas their input projections are distinct. In each phase, the recurrent step is

$$u_t = \frac{1}{m_t} \sum_{j=1}^{m_t} h_{t-j}, \qquad m_t = \min(m, \text{number of available preceding states}), \tag{131}$$

$$z_t = W_g^{(a)} a_t + W_g^{(u)} u_t + W_g^{(s)} s_t + b_g, \qquad g_t = \sigma(z_t), \tag{132}$$

$$h_t = g_t \odot \tanh(a_t). \tag{133}$$

We set $u_t = 0$ when no preceding state is available. The sliding memory is carried over from the encoder to the decoder; during decoding, s_t denotes the future context permitted by the information regime.

Equation (133) is exactly the multiplicative rule of the *temporal pure CAF* (Definition 3) with augmented context (u_t, s_t) and base activation $\rho = \tanh$. The current exogenous context does not enter additively in (129)–(130): at fixed primary input and recurrent history, $D_{s_t} a_t = 0$. This separation concerns the current routing; past context may still influence a_t through earlier states. In the implementation, the gate bias is parameterized as $b_g = b_{g,0} + \tilde{b}_g$, where $b_{g,0}$ is a fixed declared constant and $\tilde{b}_g$ is the learned component initialized at zero; since biases are excluded from weight decay, this parameterization has no effect on the objective and serves traceability only.

Implementation correspondence.

In the accompanying implementation, the gate in Equations (132)–(133) is implemented by `PureCAFGate`. The C0, C0o, C0p, and C0op entries use `CanonicalCAF`, which excludes additive context from the encoder and decoder base pre-activations.

AC6 and AC8 use `TemporalCAFv2`, whose encoder and decoder base paths additionally receive the context permitted by the information regime. These class names document the executable realization of the protocol and do not define additional mathematical CAF variants.

The head computes $r_k = V\mathscr{D}(h_k^{\mathrm{D}}) + c$, where $\mathscr{D}$ denotes dropout, disabled at evaluation. For the CAF, this dropout acts only on the head, never directly on the recurrent state. The decoder operates in free-running autoregressive mode, without injection of future targets, and the gradient flows through the feedback of its own output r_k. The accompanying base architectural setting uses $H = 48$, a temporal window $m = 3$, and a dropout rate of 0.05; any architectural deviation must be fixed before the experiments and recorded, not chosen on the basis of the test set.

Compensated initialization and reference variants.

C0 uses the following recurrent compensated initialization. For the constant gate level $c_0 \in (0, 1)$,

$$W_g^{(a)} = W_g^{(u)} = W_g^{(s)} = 0, \qquad b_g = \log\frac{c_0}{1-c_0}\,\mathbf{1}, \qquad W_h = \frac{\widetilde{W}_h}{c_0}, \qquad V = \frac{\widetilde{V}}{c_0}. \tag{134}$$

The output bias c is not rescaled. With matched inputs, initial state, and decoding, the network satisfies, at initialization and in deterministic mode,

$$h_t = c_0\widetilde{h}_t, \qquad \widehat{Y}^{\mathrm{CAF}} = \widehat{Y}^{\tanh}, \tag{135}$$

where the tilde denotes its own recurrent twin with tanh activation. The identity propagates throughout a deterministic free-running rollout by induction: if the two models have matched initial states, use the same supplied inputs, feedback convention, and persistence offset, then equality of their current outputs gives the same next decoder input, and the compensated recurrent and output maps preserve Equation (135) at the following step. This is an exact forward identity at matched parameters, not an equivalence of training trajectories. Independent dropout masks need not preserve pathwise equality. Moreover, because the compensated parameterization rescales W_h and V by c_0^{-1}, an equal-coefficient quadratic penalty assigns c_0^{-2} times as much penalty to these blocks, and Euclidean gradient descent is not invariant under the reparameterization. Thus the deterministic embedding is limited to the matched initial forward map and does not imply identical regularization, gradients, stochastic trajectories, or optimized solutions.

In the accompanying configuration, `CAF_Q` sets $c_0 = 0.5$, so that the derivative of the gate sigmoid initially equals $c_0(1 - c_0) = 1/4$, the maximum of σ'. This choice follows the initialization trade-off discussed in Remark 8. The separate configuration quantity `CAF_GO` denotes the nominal bias level g_0 used under standard initialization and must not be conflated with c_0, even when the two numerical values coincide.

C0o retains the same topology but initializes $\widetilde{W}_h$ orthogonally before compensation. C0 and C0o therefore have the same number of parameters. **C0p** and **C0op** are their respective variants without the persistence shortcut. This choice modifies both

Table 4: Experimental CAF variants. Symbols refer to the protocol equations. Initialization changes are distinguished from structural changes.

Code	Name	Definition and question studied
AC1	CAF without exogenous context	$W_g^{(s)} = 0$; contribution of exogenous modulation.
AC2	Diagonal-gate CAF	$W_g^{(a)} = \mathrm{diag}(w)$; contribution of inter-unit coupling in the gate.
AC3	Exogenous-only-gate CAF	$W_g^{(a)} = W_g^{(u)} = 0$; $g_t = \sigma(W_g^{(s)} s_t + b_g)$.
AC4	CAF without temporal summary	$W_g^{(u)} = 0$; explicit contribution of the summary u_t, without removing the recurrence $W_h h_{t-1}$.
AC5	Standard-initialization CAF	Same topology as C0, non-zero random gate matrices, no $1/c_0$ compensation.
AC6	Additive-context CAF, standard	$A_{\mathrm{E}} s_t$ and $A_{\mathrm{D}} s_{\tau+k}$ added to the bases, with standard initialization.
AC7	CAF-SP, state preserving	$h_t = (\mathbf{1} - \lambda_t) \odot h_{t-1} + \lambda_t \odot (g_t \odot \tanh(a_t))$; standard initialization.
AC8	Additive-context CAF, compensated	Same additive routing as AC6, but compensation and constant c_0 of C0.

the initialization of the autoregressive feedback and the assembly of the forecast:

$$\begin{aligned} &\text{C0/C0o:} \quad r_0 = 0, \quad \widehat{y}_{\tau+k} = y_\tau + r_k, \\ &\text{C0p/C0op:} \quad r_0 = y_\tau, \; \widehat{y}_{\tau+k} = r_k. \end{aligned} \tag{136}$$

Persistence is a choice of the forecaster, not a residual term of the CAF activation operator. It is shared with the trainable local models but is not added to the external architectures.

Structural ablations, initialization, and extensions.

The `ablation` suite fixes **C0** as anchor and comprises exactly **AC1**–**AC8**. Path removals apply to both the encoder and the decoder. Unless stated otherwise, these variants retain rule (133) and the persistence shortcut of C0. The taxonomy follows the distinction between *variants of pure CAF* and *extensions built on the CAF principle* of Section 2.6.

Relation of the gate restrictions to the theory. AC1, AC2, AC3, and AC4 remain structured members of the pure CAF. At fixed context (u_t, s_t), its local Jacobian, specializing Proposition 5 to the affine gate and $\rho = \tanh$, reads

$$D_a \Phi = \mathrm{diag}\big(g \odot (\mathbf{1} - \tanh^2(a))\big) + \mathrm{diag}\big(\tanh(a) \odot g \odot (\mathbf{1} - g)\big) W_g^{(a)}. \tag{137}$$

AC2 removes the inter-unit couplings of the second term. AC3 removes that term entirely, so that the local conditioning is governed by Corollary 12, and also removes the temporal gate path: it realizes the exogenous *context-only* member studied by the approximation theory (Theorem 21), and is not a single-block ablation. This correspondence does not transfer a universal approximation theorem to a fixed-width recurrent forecaster. AC1 examines the predictive usefulness of the exogenous channel, without by itself constituting a validation of the local profiled selection criterion of Section 12.

Standard initialization. For AC5, AC6, and AC7, the bias is initialized at $b_g = \log(g_0/(1-g_0))\mathbf{1}$, with $g_0 = 0.5$ in the accompanying configuration. The active blocks $W_g^{(a)}, W_g^{(u)}, W_g^{(s)}$ are initialized randomly with standard deviation 0.02, and W_h, V are not divided by c_0. Hence, even when $g_0 = c_0$, the standard gate is not constant over inputs. C0–AC5 therefore compares two *initialization schemes* and does not allow their difference to be attributed solely to the $1/c_0$ rescaling.

Additive routing of context. AC6 and AC8 use

$$a_t^{\mathrm{E},+} = a_t^{\mathrm{E}} + A_{\mathrm{E}} s_t, \qquad a_k^{\mathrm{D},+} = a_k^{\mathrm{D}} + A_{\mathrm{D}} s_{\tau+k}, \tag{138}$$

where the base terms have the same form as (129)–(130), with model-specific parameters. Their output remains $g_t \odot \tanh(a_t)$: they are pure CAFs at the operator level but not forecasters with separated canonical routing. The primary routing contrast is **C0–AC8**, since it retains the compensated initialization scheme. It nevertheless adds projection parameters and does not guarantee identical shared weights at initialization. **C0–AC6** confounds a change of routing with a change of initialization. In the `history` regime, the future context being zero, the additive decoder path receives no varying future signal. Introducing permitted context into the base is not, in itself, an information leak.

State-preserving extension. AC7 uses a second gate,

$$\lambda_t = \sigma\big(W_\lambda [a_t; s_t] + b_\lambda\big), \tag{139}$$

$$h_t = (\mathbf{1} - \lambda_t) \odot h_{t-1} + \lambda_t \odot \Phi_{\mathrm{pure}}(a_t, (u_t, s_t)). \tag{140}$$

This formulation coincides with the state-preserving extension of Section 2.6 upon setting $\beta_t = \mathbf{1} - \lambda_t$. At initialization, $W_\lambda = 0$ and $b_\lambda = \log(\lambda_0/(1-\lambda_0))\mathbf{1}$, with $\lambda_0 = 0.9$. The base remains free of additive context, but the context acts on both g_t and λ_t. AC7 is not a pure update: its additional carry path generally invalidates the compensated identity with the tanh twin because the next state is no longer a constant rescaling of the base recurrent response. The implementation therefore enforces standard initialization; C0–AC7 does not constitute a memory contrast at equal initialization. The other theoretical extension, the residual CAF $\Phi_{\mathrm{res}}(a, s) = a + \alpha\Phi_{\mathrm{pure}}(a, s)$, does not belong to this suite and must not be presented as evaluated.

Reference models and separation of comparisons.

Internal comparison. The `main_interne` suite gathers C0/C0o and the following references:

- **T0 — Persistence:** $\widehat{y}_{\tau+k} = y_\tau$, without training.
- **T-1 — Seasonal naive forecast:** $\widehat{y}_{\tau+k} = y_{\tau+k-p_d}$ if $k \le p_d$, then $\widehat{y}_{\tau+k} = \widehat{y}_{\tau+k-p_d}$, where p_d is the seasonal period of the dataset. It is included only if the history contains a full period.
- **T3 — LSTM with context:** two-layer encoder–decoder receiving $[x_t^c; y_t^c; s_t]$ in the encoder and $[r_{k-1}; s_{\tau+k}]$ in the decoder, with shared autoregressive feedback and persistence. The LSTM gates regulate memory transitions; they are not identified with the CAF activation operator (cf. Section 8).
- **T5 — FiLM-RNN with context:** after each LSTM update, its hidden state $\bar{h}_t$ is modulated as

$$h_t = \gamma(s_t) \odot \bar{h}_t + \beta(s_t), \quad \gamma(s) = A_\gamma s + b_\gamma, \quad \beta(s) = A_\beta s + b_\beta.$$

  The modulation is initialized at the identity and does not directly modify the LSTM cell state. Unlike the pure CAF, its gain is not bounded by a sigmoid and an additive shift is permitted.

External comparison. The `main_externe` suite compares exclusively **C0p/C0op** with the available external models, without adding persistence to the latter. In the `history` regime it includes **CrossLinear** [Zhou et al.(2025)Zhou, Liu, Liang, Song, and Li] and **TimeXer** [Wang et al.(2024)Wang, Wu, Dong, Qin, Zhang, Liu, Qiu, Wang, and Long], as well as **PRformer** [Yu et al.(2025)Yu, Yu, Nie, Miao, Liu, and Li] only when its inclusion is explicitly enabled. In the `known_future` regime, these models carry the suffix `+future-ctx`: their adapters add to the historical output a declared contextual branch,

$$\widehat{y}_{\tau+k} = f_M(\mathscr{H}_\tau)_k + w_2^\top \mathscr{D}_A(\mathrm{GELU}(W_1 s_{\tau+k} + b_1)) + b_2. \tag{141}$$

$\mathscr{D}_A$ denotes the optional adapter dropout. The last layer is initialized at zero ($w_2 = 0, b_2 = 0$), which initially nullifies this branch. These adaptations are presented neither as native components of the external models nor as variants of the pure CAF.

DAG [partial-exog adapter] [Qiu et al.(2025)Qiu, Zhu, Li, Wu, Yang, and Hu] is included only in the `known_future` regime. Its future exogenous tensor is $[0_D; s_{\tau+k}]$: historical covariates without a known future receive zero slots, and their future supervision is masked in the temporal regularization term. The objective retains the architectural regularization exposed by the model,

$$\mathscr{L}_{\mathrm{DAG}} = \mathscr{L}_{\mathrm{MSE}} + \beta_{\mathrm{DAG}} \left(\mathscr{R}_{\mathrm{temp,known}} + \mathscr{R}_{\mathrm{cov}}\right). \tag{142}$$

Its `DAG+temporal-only` version is not regarded as a complete reference in the history regime. External comparisons are conditional on the actual availability of their pinned dependencies and sources; any unavailable model is declared as such rather than silently replaced. The archived benchmark metadata list the effective external rosters, exclusions, provenance, and completed runs retained for the reported runs.

Appendix comparison. The `appendix` suite gathers, together with C0, the complementary references **T4** (GRU, two layers), **T6** (Transformer), **T7** (Informer), **T8**

(local adaptation of N-BEATS with covariates), and **T9** (local Gaussian adaptation of DeepAR). They are distinguished from the external models. T8 decomposes the forecast into block contributions with successive removal of a reconstruction of the history, $e^{(b+1)} = e^{(b)} - \widehat{e}^{(b)}$ and $\widehat{Y} = \sum_b \widehat{Y}^{(b)}$, before adding persistence. T9 optimizes a Gaussian negative log-likelihood,

$$\mathscr{L}_{\text{DeepAR}} = \frac{1}{NP_d} \sum_{i,k} \left[\frac{1}{2} \log(2\pi) + \log \sigma_{i,k} + \frac{(y^*_{i,k} - \mu_{i,k})^2}{2\sigma^2_{i,k}} \right], \tag{143}$$

where y^* is the normalized target, reduced by the last observed value when persistence is enabled. Decoding uses its own predicted means, without teacher forcing; point metrics are computed on these means after restoring persistence.

To avoid confounding comparisons with and without persistence, the main tables are separated into internal and external blocks. The legacy `main` suite, which mixes these families and includes C0p but not C0op, does not replace the dedicated `main_externe` comparison. The `all` suite unites the blocks without altering their interpretation.

The term "persistence-matched" describes this architectural contract: no last-value residual is added to either the persistence-free CAF forecast or the external-model forecast. The Holm-adjusted p-values reported below are taken from the standard external comparison family, which also includes the comparison with the other CAF initialization twin; analyses using a narrower `persistence_matched` family can therefore report different adjusted numerical p-values while preserving the same significance verdicts in the checked cases.

Activation controls available but outside the active suites.

The protocol also defines A1/A2/A3 on a common recurrent scaffold. With

$$a_t = b_t^{\text{input}} + Rh_{t-1} + Cs_t + Tu_t, \qquad (\gamma_t, \beta_t) = M[s_t; u_t] + b_M,$$

their pre-dropout responses are respectively

$$\text{A1 — concatenation:} \quad v_t = \tanh(a_t), \tag{144}$$

$$\text{A2 — FiLM:} \quad v_t = (\mathbf{1} + \gamma_t) \odot \tanh(a_t) + \beta_t, \tag{145}$$

$$\text{A3 — CAF:} \quad v_t = \sigma\big(G[a_t; s_t; u_t] + b_G\big) \odot \tanh(a_t). \tag{146}$$

These controls share the scaffold layers, though not necessarily the number of modulation parameters. A3 retains a pure activation but includes the context and the temporal summary in the base; it is therefore not C0. Moreover, the gate learning-rate multiplier described below applies only to the gate parameters of the canonical CAF family and not to the modulation layer of A3. The construction functions are present in the code, but no active suite selects A1/A2/A3 automatically. They must appear among the models actually evaluated only after an explicitly configured run; their mere presence in the sources is not an experiment.

Training and validation-based selection.

For each dataset, the list of random seeds $\mathscr{S}_d$, the learning-rate grid $\mathscr{G}_{\eta,d}$, and the weight-decay grid $\mathscr{G}_{\omega,d}$ are declared in the corresponding configuration. The accompanying implementation provides a common default grid, $\mathscr{G}_\eta = \{3 \times 10^{-4}, 8 \times 10^{-4}, 2 \times 10^{-3}\}$, $\mathscr{G}_\omega = \{0, 10^{-4}\}$, and ten seeds $\{0, \ldots, 9\}$; any dataset-specific deviation must be declared before the experiments. At fixed dataset and regime, the same partitions, seeds, and base grids are used for the compared models. Each trainable model and each seed receives $|\mathscr{G}_{\eta,d}|\,|\mathscr{G}_{\omega,d}|$ trials under the dataset-specific grids. The software default would give six trials. The protocol and archived benchmark metadata consistently report the reduced grids listed in Table 5: one trial on Solar main, Solar ablation/appendix, and METR-LA, and three trials on Bike and EPF-NP. These archived metadata are treated as the authoritative account of the effective campaign configurations for this section. The naive methods require neither training nor hyperparameter search.

Each trial is reinitialized with the declared seed and uses AdamW. Biases and one-dimensional tensors are excluded from weight decay. The accompanying configuration applies a linear warm-up over two epochs followed by cosine decay, and clips the global gradient norm at 1. The CAF gate parameters use a learning-rate multiplier $\eta_{g,e} = 8\eta_e$ in this configuration; all other parameters retain η_e. This difference is explicitly declared: a common base learning-rate grid therefore does not imply a strictly identical optimization policy between the CAF and all references.

The parameters retained in common across datasets, excluding the dataset-specific training batch size, are:

Parameter	Value
Maximum number of epochs per trial	25
Early-stopping patience	6
Maximum validation and test batch size	256
Timed inference repetitions	10

The training batch size is declared per dataset in Table 5: Solar and METR-LA use 128, whereas Bike and EPF-NP use 64. The number of timed inference repetitions is passed explicitly to the benchmark script (`--inference-repeats 10`); the corresponding default of the accompanying configuration class is 50 and must not be confused with the value retained here. Training batches are shuffled by a generator dedicated to the seed, independent of the random draws used for model construction. Validation and test sets are not shuffled. Apart from DeepAR and the additional DAG regularization, the objective is the MSE on the normalized target:

$$\mathscr{L}_{\mathrm{MSE}}(\theta) = \frac{1}{NP_d} \sum_{i=1}^{N} \sum_{k=1}^{P_d} \left(\widehat{y}^{\,\mathrm{n}}_{i,k}(\theta) - y^{\mathrm{n}}_{i,k}\right)^2 . \tag{147}$$

For all trainable models, including DeepAR and DAG, selection uses the *normalized validation point MSE*, without the architectural regularization term. Stopping occurs

after six consecutive epochs without strict improvement. The best validation state of each trial is retained; the learning rate–weight decay pair is chosen among these states on validation only. The selected checkpoint is then reloaded as selected, without additional training on the union of training and validation, and evaluated once on the test set to compute the scores.

The reference between C0 and C0o is chosen according to the mean of the finite validation MSEs across repetitions, with priority to C0 in case of a tie. In `main_externe`, the same rule applies to C0p/C0op, with priority to C0p. The anchor of the ablations remains C0. No test score is used to choose the reference, the hyperparameters, or the best state. Failed repetitions and their causes are retained in the report. Equal caps on trials and epochs guarantee neither equal convergence nor equal total cost; the selection between two CAF initializations is itself a budget to be declared.

Predictive metrics, variability, and cost.

Test scores are computed after inverting the target standardization. Writing $e_{i,k} = \widehat{y}_{i,k} - y_{i,k}$ and $n = N_{d,\mathrm{te}} P_d$,

$$\mathrm{MSE} = \frac{1}{n} \sum_{i,k} e_{i,k}^2, \qquad \mathrm{MAE} = \frac{1}{n} \sum_{i,k} |e_{i,k}|, \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}. \tag{148}$$

The MSE is the primary error criterion; MAE and RMSE complement its reading. The code also computes MAPE, sMAPE, and WAPE, in percent. MAPE excludes targets of magnitude less than or equal to 10^{-3}, sMAPE excludes pairs with $|\widehat{y}| + |y| \leq 10^{-8}$, and WAPE is defined only if $\sum |y| > 10^{-8}$. The variable `MAPE_coverage` denotes the percentage, on the 0–100 scale, of admissible relative errors less than or equal to 10 %; it is not the coverage of a prediction interval. Errors are computed over window–horizon pairs, without deduplicating target dates common to several windows. Results are reported per dataset and regime, with mean and empirical standard deviation across seeds (divisor $K - 1$ when at least two repetitions are available).

For a comparison against a reference R, differences are paired on the intersection of available seeds, $\Delta_s = \mathrm{MSE}_{M,s} - \mathrm{MSE}_{R,s}$; a positive value favours R. The statistical protocol specifies a two-sided Wilcoxon signed-rank test, a Holm correction [Holm(1979)] at level 0.05 within each family of comparisons, a 95 % percentile bootstrap interval of the mean difference (5000 resamples), and a paired rank-biserial correlation. The correction is not global across all datasets or all tables. Ablations are analysed within their dedicated family, without selecting a new reference according to their test scores. These statistics describe the variability of repetitions on the fixed partitions; they do not treat overlapping temporal windows as additional independent observations.

Cost is described by the number of trainable parameters, the total training time including tuning trials, and the mean latency of one forward pass on the first test batch. This latency is measured in evaluation mode, without gradient, after three warm-up passes and over **ten timed repetitions**, with CUDA synchronization when required. The batch contains at most 256 windows; this is a per-batch latency, not a per-sequence latency, and not an end-to-end measurement including data loading.

Hardware, effective batch size, and software versions must accompany the measurements. These timing repetitions are neither new training runs nor new selections on the test set.

Scope of diagnostics and traceability.

Effective configurations, seeds, grids, partitions, best states, lists of available models, and external settings must be archived with the results. The controls by permutation of the future context and by initialization from a twin without exogenous channel are optional, disabled in the base configuration, and distinct from the ablations AC1–AC8. The local Jacobian bound of Theorem 9 does not certify the stability of the complete recurrence. A possible gate-restricted diagnostic can replace the worst-case unit gate factor by an observed or certified upper bound on the gate, but this manuscript does not report such a certified invariant-domain bound as an empirical result. The temporal summary, the autoregressive feedback, and, for CAF-SP, the derivatives of λ_t add dependency paths that are absent from the local activation bound.

The implementation-level quantity `q_caf_bound_tight`, when logged, evaluates the gate at zero pre-activation rather than computing the supremum of the gate over a certified invariant state–context domain. This substitution is exact for the constant compensated gate at initialization, but it is not generally justified after training or under standard initialization with nonzero gate matrices. It should therefore be interpreted only as an initialization or logging diagnostic, not as a reported global contraction certificate.

Finally, the benchmark does not implement the full theoretical exogenous-selection procedure of Section 12 with profiling, ridge fitting, and control of the Taylor remainder. Forecasting and ablation results must therefore be interpreted as an evaluation of the described models and mechanisms, not as an automatic validation of all theoretical statements.

13.2 Experimental Questions and Interpretation Strategy

The experiments are organized by scientific question rather than as four independent dataset case studies. This organization preserves the distinction between (i) analytical properties established for the CAF operator under their stated assumptions, (ii) empirical results obtained for finite forecasters under a fixed optimization and evaluation protocol, and (iii) mechanistic interpretations that relate an observed effect to a plausible CAF pathway. The experiments can therefore document when a CAF mechanism is useful, neutral, or unfavorable in the evaluated setting, but they do not constitute additional proofs of the theoretical results.

Experimental questions.

The evaluation addresses the following seven questions.

Q1. Predictive viability. Is the canonical CAF competitive with the local recurrent and conditional references when the information contract and the persistence mechanism are controlled?

Q2. Modes of contextual utility. Does the exogenous gate channel $W_g^{(s)} s_t$ contribute when only observed context history is available, and does its utility change when admissible future context is supplied? Here, `history` does not mean "without context": past s_t remains available during encoding, while the gate can also depend on the current pre-activation a_t and the endogenous temporal summary u_t. The `known_future` regime adds only context genuinely known at the forecast origin. Because the two regimes impose different information contracts, any change in the apparent utility of the channel is interpreted descriptively rather than as a paired treatment effect.

Q3. Gate structure. Do full, diagonal, and exogenous-only gates produce the same practical behavior? In particular, this question tests whether the cross-unit coupling permitted by an unrestricted $W_g^{(a)}$ is useful relative to the structured restrictions represented by AC2 and AC3. AC3 jointly removes the pre-activation and endogenous-summary gate terms, $W_g^{(a)} a_t$ and $W_g^{(u)} u_t$; it is therefore a composite context-only restriction rather than a single-block ablation.

Q4. Endogenous temporal memory. Does the explicit summary u_t contribute beyond the recurrence already carried by $W_h h_{t-1}$? The C0–AC4 comparison isolates this gate pathway; it does not remove recurrent state propagation itself.

Q5. Initialization and recurrent dynamics. How do compensated initialization, orthogonal initialization, and the state-preserving extension affect predictive error and variability across seeds? C0–C0o and C0–AC5 concern initialization, whereas AC7 changes the update rule and is consequently interpreted as a CAF-based extension rather than as a pure-CAF ablation.

Q6. External positioning. Does CAF remain competitive with specialized forecasting architectures when persistence is neutralized? This question is addressed only within the dedicated external suite, where C0p/C0op are compared with external models that do not receive the CAF persistence shortcut.

Q7. Practical cost. Are parameter counts and measured computation consistent with the structural overhead predicted for full, diagonal, and context-only gates, and what performance–cost trade-off is observed in practice? Asymptotic operation counts and wall-clock latency are treated as related but distinct quantities.

Primary endpoint and descriptive reporting.

All predictive errors are computed after inversion of target standardization. The denormalized MSE in Equation (148) is the primary endpoint; MAE and RMSE are secondary checks used to assess whether a conclusion depends on one representation of forecast error. Results are summarized by the mean and empirical standard deviation across seeds. MAPE, sMAPE, and WAPE are reported as supplementary scale-relative diagnostics, subject to the admissibility rules in Section 13.1; they do not replace the primary MSE analysis. In particular, `MAPE_coverage` denotes the percentage, on the 0–100 scale, of admissible relative errors not exceeding 10%, not the coverage of a predictive interval. Absolute MSE values are not compared across datasets because target scales, history lengths, horizons, and sampling processes differ. The Bike Sharing, EPF-NP, and METR-LA data sources are documented by

Table 5: Dataset-specific experimental budgets. Window counts are reported as training/validation/test. "Main" denotes the internal and external suites, except for the Bike external `known_future` run noted below.

Dataset	(L, P)	Windows	Seeds	Batch	WD	Learning-rate grid, trials, and calendar
Solar	$(144, 24)$	8064/2016/4032	15 main; 10 ablation/appendix	128	10^{-4}	8×10^{-4} for main and 3×10^{-4} for ablation/appendix; 1 trial; reconstructed calendar
Bike	$(168, 24)$	1344/336/672	10; 15 for external `known_future`	64	10^{-4}	$\{3\times10^{-4}, 8\times10^{-4}, 2\times10^{-3}\}$; 3 trials; real calendar
EPF-NP	$(168, 24)$	1344/336/672	10	64	10^{-4}	$\{3\times10^{-4}, 8\times10^{-4}, 2\times10^{-3}\}$; 3 trials; real calendar and day-ahead TSO forecasts treated as known
METR-LA	$(12, 12)$	16128/4032/8064	15	128	10^{-4}	8×10^{-4}; 1 trial; real calendar; target = first sensor (`sensor_000`, source column 0)

[Fanaee-T and Gama(2014), Lago et al.(2021)Lago, Marcjasz, Schutter, and Weron, Li et al.(2018)Li, Yu, Shahabi, and Liu], respectively. The Solar series is identified here by the public preprocessed Time-Series-Library archive used by the experiments, https://huggingface.co/datasets/thuml/Time-Series-Library, because the manuscript's preprocessed file does not provide an unambiguous original dataset citation.

Table 5 makes explicit the differences in statistical power and optimization budget that constrain cross-suite interpretation. Solar uses different seed counts and learning rates between the main and ablation/appendix families; the corresponding C0 entries are separately trained models and must not be treated as repeated evaluations of one fitted model. Bike uses 15 rather than 10 seeds only for the external `known_future` suite. EPF-NP is uniform at 10 seeds, whereas METR-LA is uniform at 15 seeds but uses a single learning rate and a larger training batch. These design differences are part of the reported protocol rather than nuisance details to be averaged away. The weight decay is fixed to 10^{-4} in the reported runs. Consequently, the table reports the effective tuning budget, not the broader software default. The same configurations are reported in the archived Solar, Bike, EPF-NP, and METR-LA benchmark metadata and are therefore retained as the effective configurations of the reported runs. For METR-LA, the effective target is the first sensor column: the loader maps source column 0

to `sensor_000` and selects it as y. This identification is independently confirmed by the deterministic persistence baseline: on the prescribed test block with $L = P = 12$, source column 0 gives MSE 146.1344, MAE 4.3705, and RMSE 12.0886, matching the published values, whereas the last source column gives MSE 202.2764, MAE 6.4099, and RMSE 14.2224. For EPF-NP, the future-context regime treats the day-ahead TSO forecasts available in the dataset as known at each rolling origin; their publication calendar was not re-audited window by window, so this assumption should be stated when interpreting operational availability.

Statistical decision rules.

For a model M and the reference R declared for the same dataset, regime, and suite, seed-wise differences are

$$\Delta_s = \mathrm{MSE}_{M,s} - \mathrm{MSE}_{R,s},$$

computed on the intersection of available seeds. Thus, $\Delta_s > 0$ favors the reference. Each inferential statement is based on a two-sided paired Wilcoxon signed-rank test, with Holm correction at level 0.05 within the declared family of comparisons against that reference. The adjusted p-value is accompanied, when available, by the mean paired difference, its 95% percentile-bootstrap interval based on 5000 resamples, and the paired rank-biserial correlation. Accordingly, any claim of a difference reports four elements together: its direction, its descriptive magnitude, its variability across seeds, and its Holm-adjusted inferential result. A non-significant result means that no difference was detected under this design; it does not establish equivalence. The seeds quantify variability of repeated training on fixed chronological partitions and do not turn overlapping forecast windows into independent observations.

Comparison boundaries and robustness analyses.

Comparisons are made only within the same dataset, information regime, suite, and paired seed set. The `history` and `known_future` regimes are distinct experiments because they impose different information contracts; their scores are not pooled, and a numerical cross-regime change is not interpreted as a paired treatment effect unless an explicit matched analysis is reported. Likewise, the internal, external, ablation, and appendix suites answer different questions. The reference is selected between C0 and C0o using validation MSE only; the ablation anchor remains C0 regardless of test outcomes. C0/C0o and C0p/C0op are not interchangeable because the latter remove the persistence construction. Full all-pairs matrices, when reported, are treated as separate robustness analyses with a larger Holm family; their adjusted p-values are not mixed with those from the primary "each model versus the reference" family. This distinction is consequential for interpretation: several Bike ablation contrasts that are significant in the prespecified "versus C0" family become non-significant in the more conservative all-pairs family, and the same family-definition sensitivity applies to the Solar AC8 result, especially in `known_future`. In Bike `known_future` external comparisons, the CAF–CrossLinear contrast is also sensitive to this broader family, whereas the CAF–TimeXer contrast remains significant. Such findings are reported as sensitive to the multiplicity family rather than as robust ablation effects.

Interpretive safeguards.

The Jacobian decomposition in Equation (137), the universality results, and the complexity analysis motivate Q2–Q7, but none entails superior test error, easier optimization, or stability of the complete recurrent forecaster. In particular, a favorable C0–AC1 contrast is a signature of useful exogenous routing in the evaluated model, not a validation of the full profiled $+s/-s$ selection criterion, which is not implemented in this benchmark. Similarly, predictive errors alone do not identify gate saturation, gradient flow, local conditioning, or the causal pathway responsible for a gain; direct gate and Jacobian diagnostics would be required for such claims. Results for AC3 cannot confirm or refute context-only universal approximation, which is an asymptotic function-class statement rather than a finite-width efficiency claim. Results for AC7 are kept outside conclusions specific to pure CAF because its state-preserving path changes the recurrent update.

Finally, external comparisons are conditional on the implemented adapters and available roster. In `known_future`, the contextual branches added to CrossLinear, TimeXer, and PRformer are protocol-specific adaptations rather than canonical components of those architectures, whereas DAG is the external model with native future-context handling in this benchmark. The roster is also not identical across datasets—notably, PRformer is disabled by configuration on METR-LA. Parameter counts may illustrate the theoretical hierarchy of gate overheads, but measured tuning time and latency additionally depend on recurrent execution, the number of trials, batch size, hardware, and software. The resulting conclusions are therefore stated as dataset-, regime-, and protocol-dependent empirical evidence, not as universal predictive or computational superiority of CAF.

Scope and experimental limitations.

The empirical scope is restricted to four datasets, with one prediction target per dataset and one fixed chronological train/validation/test partition. The study therefore does not measure robustness across alternative targets, rolling or repeated temporal splits, or a broader population of datasets. These design limits apply to every comparative and mechanistic interpretation above and preclude generalizing the observed rankings beyond the evaluated targets, partitions, regimes, and optimization budgets. The study also does not include a capacity- and optimization-matched comparison against an adaptive activation without exogenous context, such as a learned Swish-β variant; consequently, the specific contribution of exogeneity relative to generic activation adaptation remains to be isolated.

13.3 Predictive Performance Across Information Regimes

This subsection evaluates the predictive viability of CAF under two distinct information contracts. In both regimes, every model receives the observed histories of the target, the covariates, and the contextual descriptors. In the `history` regime, all future-context inputs are neutralized consistently in training, validation, and testing. In the `known_future` regime, the models additionally receive only covariates genuinely available at the forecast origin, such as calendar descriptors or planned signals; future target values and future sensor observations remain excluded. The canonical CAF

Table 6: Internal predictive comparison across information regimes. The CAF reference is the validation-selected initialization twin in the corresponding suite. Entries in the two rightmost columns summarize Holm-adjusted comparisons against the trainable local references.

	`history`		`known_future`	
Dataset	CAF MSE	Comparison with local models	CAF MSE	Comparison with local models
Solar	74.81 ± 7.42 (canonical)	Not distinguished from LSTM or FiLM-RNN	66.87 ± 3.62 (canonical)	Better than LSTM and FiLM-RNN
Bike	17439.05 ± 2727.71 (orthogonal)	Not distinguished from LSTM; better than FiLM-RNN	9125.84 ± 978.42 (canonical)	Better than LSTM and FiLM-RNN
EPF-NP	26.274 ± 1.575 (orthogonal)	Worse than LSTM; not distinguished from FiLM-RNN	22.748 ± 1.387 (canonical)	Better than LSTM and FiLM-RNN
METR-LA	125.36 ± 3.51 (canonical)	Not distinguished from LSTM or FiLM-RNN	125.85 ± 3.27 (canonical)	Not distinguished from LSTM or FiLM-RNN

preserves the pure multiplicative routing of Definition 2: context modulates the base response through the gate and does not enter the base pre-activation additively. In the recurrent setting, the evaluated operator instantiates the temporal formulation of Definition 3, with an endogenous summary u_t and an exogenous descriptor s_t entering the gate as specified in Equations (132)–(133). Forecasts are generated autoregressively, and model and checkpoint selection use validation error only. This correspondence connects the implemented forecasting operator to the theoretical CAF definition; it does not transfer every operator-level theorem automatically to the complete recurrent forecaster.

Results are interpreted within, rather than across, experimental suites. The internal comparisons use the same persistence construction for CAF and the trainable local references. The external comparisons instead use CAF variants without this construction because it is not added to the specialized external architectures. Within each suite, the CAF initialization twin used as statistical reference is selected from validation MSE before test comparison. All statements of superiority or inferiority below refer to denormalized test MSE and to the paired Wilcoxon test after Holm correction. A non-significant contrast is reported as "not distinguished", not as evidence of equivalence.

The numerical values, reported configurations, adjusted significance results, and comparison directions below were cross-checked against the four archived benchmark result records for Solar, Bike, EPF-NP, and METR-LA. Those records are treated as the authoritative numerical sources for this section; the manuscript values are concordant with them at the displayed precision.

13.3.1 History-only forecasting

The historical regime asks whether CAF remains useful without access to future covariates. This is not a context-free experiment. During encoding, CAF can use observed past context, while its gate remains responsive to the current pre-activation and to the endogenous summary of recent hidden states. The final encoder state and the recent-state summary are then carried into the decoder. This routing is consistent with the temporal pure-CAF construction in Definition 3, where (u_t, s_t) forms the total gate context. Consequently, favorable results in this regime can be compatible with historical contextual modulation, endogenous adaptation, or their interaction with recurrent state propagation; prediction errors alone do not identify which path was used or causally produced the result.

Internal comparisons.

On Solar, the canonical CAF obtains a MSE of 74.81 ± 7.42, compared with 76.70 ± 4.31 for LSTM and 70.92 ± 3.95 for FiLM-RNN. Neither contrast is significant after Holm correction ($p_{\text{Holm}} = 0.9086$ and 0.7571, respectively), and the orthogonally initialized CAF is likewise not distinguished from the canonical model ($p_{\text{Holm}} = 0.9086$). CAF nevertheless improves significantly over both deterministic naive forecasts. Thus, on Solar, historical CAF is viable and considerably more compact than the trainable local references, but the experiment does not establish a predictive advantage over them.

Bike exhibits a stronger dependence on initialization. The validation-selected orthogonal CAF reaches 17439.05 ± 2727.71, whereas the canonical initialization reaches 27214.65 ± 2584.63; the difference favors the orthogonal twin ($p_{\text{Holm}} = 0.0098$). The selected CAF is not distinguished from LSTM (15508.16 ± 2096.41, $p_{\text{Holm}} = 0.4648$) or from the seasonal naive forecast ($p_{\text{Holm}} = 0.6250$), but it outperforms FiLM-RNN (21990.85 ± 1930.70, $p_{\text{Holm}} = 0.0410$) and persistence ($p_{\text{Holm}} = 0.0098$). This result supports historical viability while also showing that the training outcome can depend materially on the recurrent initialization. It does not, by itself, establish the gradient or conditioning mechanism responsible for that dependence; those mechanisms are characterized at operator level in Sections 4 and 4.4.

On EPF-NP, the validation-selected orthogonal CAF records 26.274 ± 1.575. It is not distinguished from FiLM-RNN (26.349 ± 2.102, $p_{\text{Holm}} = 1.0000$) or from the canonical CAF ($p_{\text{Holm}} = 0.8633$), but LSTM is better (23.661 ± 1.115, $p_{\text{Holm}} = 0.0117$). The seasonal naive forecast is also better, whereas CAF improves over simple persistence. EPF-NP therefore provides an explicit counterexample to any claim that historical CAF uniformly dominates conventional recurrent or seasonal mechanisms.

For the evaluated METR-LA target sensor, the canonical CAF obtains 125.36 ± 3.51 and is not distinguished from LSTM (123.16 ± 2.82), FiLM-RNN (123.78 ± 2.55), or its orthogonal twin (127.32 ± 4.04), with $p_{\text{Holm}} = 0.6233$ for all three contrasts. It significantly improves over persistence (146.13, $p_{\text{Holm}} = 0.0002$). The trained models occupy a narrow MSE range, consistent with a short-horizon setting in which recent traffic history is highly informative. In this case, no difference from the local learned references was detected after Holm correction, and CAF uses substantially fewer trainable parameters; this non-detection must not be read as formal equivalence.

Taken together, the internal results show that CAF remains operational and often competitive without future covariates, but they do not yield a uniform ranking. No difference from both local learned references was detected after Holm correction on Solar or METR-LA; CAF is favorable to FiLM-RNN but sensitive to initialization on Bike, and inferior to LSTM on EPF-NP. These observations show that the evaluated scope of CAF extends beyond future-context forecasting while also locating clear limits to that conclusion.

Persistence-matched external comparisons.

The external suite produces a more heterogeneous picture. On Solar, the validation-selected CAF without persistence obtains 80.97 ± 8.62. It is not distinguished from CrossLinear (84.10 ± 11.30, $p_{\text{Holm}} = 0.9086$), but TimeXer (65.39 ± 2.58) and PRformer (65.15 ± 3.36) are better, both with $p_{\text{Holm}} = 0.0002$. On Bike, the CAF reference obtains 11849.52 ± 1356.75 and is outperformed by CrossLinear, TimeXer, and PRformer, whose MSE values are respectively 7996.12 ± 729.09, 7334.05 ± 193.66, and 6961.36 ± 168.70; all three contrasts have $p_{\text{Holm}} = 0.0078$. EPF-NP follows the same direction: CAF obtains 29.849 ± 3.720, whereas CrossLinear, TimeXer, and PRformer obtain 24.126 ± 2.522, 24.791 ± 0.958, and 23.426 ± 1.073, respectively, again with $p_{\text{Holm}} = 0.0078$ for each contrast.

For the evaluated METR-LA target sensor, this ordering is reversed. CAF without persistence obtains 130.03 ± 10.02, compared with 141.20 ± 1.25 for CrossLinear and 139.12 ± 1.18 for TimeXer. The paired differences favor CAF after Holm correction ($p_{\text{Holm}} = 0.0060$ and 0.0067). The larger standard deviation of persistence-free CAF relative to CrossLinear (10.02 versus 1.25) also indicates greater sensitivity to initialization or stochastic optimization on the fixed split. More broadly, the CAF standard deviation is descriptively larger without persistence than for the corresponding persistence-sharing internal reference on Solar (8.62 versus 7.42), EPF-NP (3.720 versus 1.575 for the validation-selected orthogonal reference), and this METR-LA target sensor (10.02 versus 3.51). Bike is an exception: the persistence-free historical CAF reference has a smaller standard deviation than the persistence-sharing internal reference (1356.75 versus 2727.71 for the validation-selected orthogonal reference). Because these entries come from different suites rather than a paired factorial persistence experiment, this three-out-of-four pattern motivates a stability hypothesis but does not identify persistence as its cause. The lower mean in the METR-LA sensor-level comparison should therefore be read together with its reduced repeatability. This advantage is specific to the primary MSE: in history, CAF has MSE/MAE 130.03/6.33 whereas CrossLinear and TimeXer have 141.20/4.52 and 139.12/4.45, indicating a different error distribution and cautioning against reducing the comparison to a single notion of accuracy. Overall, the external historical results identify one favorable dataset, one neutral-to-unfavorable dataset, and two datasets on which specialized alternatives are superior. They support competitiveness in selected settings, not general dominance.

13.3.2 Forecasting with known future context

The future-context regime tests whether CAF can convert admissible future information into predictive improvement. For the canonical model, this information acts through activation-level modulation rather than through an additive shortcut in the base response, consistent with Definition 2 and Equation (133). The context-sensitivity formula in Equation (13) identifies the operator-level pathway by which s may affect the CAF response. The experiments test the complete forecaster under this routing, but predictive errors alone do not establish that this differential pathway caused an observed gain.

Internal comparisons.

On Solar, canonical CAF reaches 66.87 ± 3.62, improving over LSTM (76.53 ± 10.92, $p_{\text{Holm}} = 0.0134$) and FiLM-RNN (79.50 ± 8.64, $p_{\text{Holm}} = 0.0004$). It also improves over both naive forecasts, while remaining not distinguished from its orthogonal twin ($p_{\text{Holm}} = 0.3028$). The future descriptors for Solar are reconstructed from an assumed temporal origin, however, and must be described as a synthetic calendar rather than an authenticated civil calendar.

On Bike, CAF obtains 9125.84 ± 978.42, compared with 11273.51 ± 1600.21 for LSTM and 11653.32 ± 998.09 for FiLM-RNN. Both contrasts favor CAF ($p_{\text{Holm}} = 0.0098$), as do the comparisons with the seasonal and persistence baselines. The orthogonal twin is not distinguished from the canonical model ($p_{\text{Holm}} = 0.0840$). Because Bike uses an observed calendar, this experiment shows that the complete canonical CAF forecaster achieves lower MSE than the evaluated local recurrent alternatives when authentic future temporal descriptors are available. The comparison does not isolate the gate as the cause of this advantage.

EPF-NP yields the same internal direction. CAF reaches 22.748 ± 1.387, versus 25.783 ± 1.263 for LSTM and 28.043 ± 2.672 for FiLM-RNN; both differences are significant after correction ($p_{\text{Holm}} = 0.0156$). CAF also outperforms persistence and is not distinguished from either its orthogonal twin or the seasonal naive forecast. Thus, although the historical experiment favored LSTM, the separately trained future-context CAF has a lower MSE than LSTM in the internal suite. This cross-regime reversal is descriptive because the regimes impose distinct information contracts.

The evaluated METR-LA target sensor is the principal neutral case. CAF obtains 125.85 ± 3.27, while LSTM and FiLM-RNN obtain 123.80 ± 2.33 and 124.79 ± 2.94; neither difference is detected ($p_{\text{Holm}} = 0.4543$ and 0.4586). CAF still improves over persistence, but no advantage over the learned local references is detected. The short forecasting horizon and strong local autocorrelation offer a plausible explanation, but this interpretation is not a causal finding.

The internal evidence has the same direction on three datasets: CAF is better than both local trainable references on Solar, Bike, and EPF-NP when future context is available. METR-LA does not reproduce this advantage. These results are compatible with the practical relevance of activation-level contextual routing, while showing that the informational value of future context depends on the temporal system. They do not isolate the gate from the remaining components of the complete forecasting systems.

Persistence-matched external comparisons.

On Solar, CAF without persistence obtains 71.08 ± 5.89. It outperforms CrossLinear with its future-context adaptation (85.95 ± 12.86, $p_{\text{Holm}} = 0.0214$), but no difference is detected from adapted TimeXer (71.17 ± 3.85), DAG (75.03 ± 7.35), adapted PRformer (66.55 ± 4.34), or its orthogonal twin after Holm correction. CAF is therefore better than CrossLinear in this comparison, but the remaining non-significant contrasts do not establish a shared leading group or equivalence.

Bike shows the clearest regime-dependent reversal. The selected CAF reference obtains 7148.73 ± 616.75. It outperforms adapted CrossLinear (8018.34 ± 654.84, $p_{\text{Holm}} = 0.0061$) and adapted TimeXer (8516.02 ± 461.65, $p_{\text{Holm}} = 0.0006$), while no difference is detected from DAG (8241.06 ± 1326.83, $p_{\text{Holm}} = 0.0905$) or adapted PRformer (7185.35 ± 399.26, $p_{\text{Holm}} = 0.9341$). Thus, the historically superior CrossLinear and TimeXer no longer outperform CAF under the future-information contract; the non-significant DAG and PRformer contrasts are not evidence of equivalence.

EPF-NP is the strongest adverse case for CAF in the external suite. The selected CAF reference reaches 75.051 ± 10.673, whereas adapted CrossLinear, adapted TimeXer, DAG, and adapted PRformer obtain respectively 21.686 ± 2.107, 19.040 ± 1.934, 28.146 ± 6.012, and 22.476 ± 2.766. Every comparison favors the external alternative with $p_{\text{Holm}} = 0.0098$. This result is particularly important because the same canonical CAF performs well in the internal suite when persistence is shared. The persistence-free CAF errors are also worse than the deterministic naive levels reported for this dataset, so the result should be read as a failure or instability of the persistence-free CAF construction in this regime, not merely as a case where external neural architectures are stronger. It shows that the favorable internal conclusion does not transfer to the persistence-free external setting and that architecture, forecast construction, optimization, and dataset interact strongly.

On the evaluated METR-LA target sensor, CAF without persistence obtains 126.15 ± 6.03 and outperforms all available external models: adapted CrossLinear (139.60 ± 2.11), adapted TimeXer (137.60 ± 1.64), and DAG (163.05 ± 19.22), with $p_{\text{Holm}} = 0.0002$ for each contrast. The orthogonal CAF is also worse than the canonical initialization ($p_{\text{Holm}} = 0.0054$). As in the historical comparison, the MSE advantage coexists with lower MAE for CrossLinear and TimeXer: CAF has MSE/MAE 126.15/6.02, whereas CrossLinear and TimeXer have 139.60/4.75 and 137.60/4.82. This suggests that CAF mainly reduces the contribution of larger errors rather than uniformly improving every loss functional.

These external results require a scope qualification. CrossLinear, TimeXer, and PRformer do not natively consume future covariates in the evaluated implementations; their future-context versions use a declared residual adapter added for this protocol. DAG is the external architecture with native future-context handling. The comparison therefore evaluates complete systems under a common information contract, but it does not always compare the external architectures in their original unmodified form. Moreover, PRformer is disabled by configuration in the METR-LA roster, so cross-dataset win counts are not based on an identical set of competitors. These mechanism-level distinctions complement the broader theoretical positioning in Section 8.

Table 7: Qualitative synthesis of persistence-matched external comparisons. "Better" and "worse" indicate a Holm-significant MSE difference relative to at least the listed alternatives; "not distinguished" denotes absence of a detected difference, not equivalence.

Dataset	`history`	`known_future`
Solar	Not distinguished from CrossLinear; worse than TimeXer and PRformer	Better than adapted CrossLinear; not distinguished from adapted TimeXer, DAG, or adapted PRformer
Bike	Worse than CrossLinear, TimeXer, and PRformer	Better than adapted CrossLinear and TimeXer; not distinguished from DAG or adapted PRformer
EPF-NP	Worse than CrossLinear, TimeXer, and PRformer	Worse than all four external alternatives
METR-LA	Better than CrossLinear and TimeXer	Better than adapted CrossLinear, adapted TimeXer, and DAG

13.3.3 Cross-regime and cross-dataset synthesis

The principal result is an interaction between dataset, information regime, and comparison family rather than a global model ranking. Internally, future-context CAF is favored over both LSTM and FiLM-RNN on Solar, Bike, and EPF-NP, whereas no difference is detected on METR-LA. Historically, CAF is often competitive with the local references but is not uniformly superior. The contrast suggests that future context can strengthen the relative value of CAF when the available descriptors contain predictive information at the relevant horizon; it is not a formal cross-regime causal estimate.

A descriptive within-model comparison points in the same direction but must not be mistaken for a formal cross-regime test. For the canonical CAF in the internal suite, MSE changes from 74.81 to 66.87 on Solar (-10.6%), from 27214.65 to 9125.84 on Bike (-66.5%), and from 26.349 to 22.748 on EPF-NP (-13.7%), while it changes only from 125.36 to 125.85 on METR-LA ($+0.4\%$). These models are trained separately under different information contracts, and the protocol does not supply a Holm-corrected paired test of the cross-regime change. The percentages are therefore descriptive effect sizes, not estimates of a causal treatment effect of future context.

The external comparisons sharpen the interaction. Solar moves from significant defeats against TimeXer and PRformer in the historical regime to an advantage over CrossLinear in the future regime; no difference from adapted TimeXer, DAG, or adapted PRformer is detected after Holm correction. Bike undergoes a clearer inversion: all three historical external alternatives outperform CAF, whereas future-context CAF outperforms CrossLinear and TimeXer and is not distinguished from DAG or PRformer. EPF-NP resists this pattern: external architectures outperform persistence-matched CAF in both regimes, with a particularly large gap when future

context is available. The evaluated METR-LA target sensor presents the opposite profile, with CAF outperforming every available external model in both regimes despite little internal benefit from future descriptors.

These results support three conclusions. First, CAF remains viable without future covariates. Its temporal architecture permits the use of observed historical context and endogenous recurrent information, as formalized in Definition 3, but the predictive results do not identify which pathway produced the observed performance. Second, admissible future context provides an additional and often favorable information pathway, whose operator-level sensitivity is described by Equation (13); its empirical value nevertheless depends on the dataset and forecast horizon. Third, performance depends on the complete forecasting system: the strong EPF-NP difference between internal and external suites shows that persistence and architectural context cannot be ignored when interpreting the activation mechanism.

The observations are compatible with the theoretical principle that pure CAF implements an explicit context-dependent nonlinear response (Definition 2) and with the temporal decomposition into endogenous and exogenous gate inputs (Definition 3). They do not validate the context-only universal-approximation result of Theorem 21 or identify the gate as the sole cause of a predictive gain. Universality concerns representational density on compact product domains, not finite-data generalization under a fixed width and training budget. Likewise, Proposition 5 establishes possible preactivation and context-sensitivity pathways, including the gate-induced coupling term in Equation (12); forecast MSE alone does not show that these pathways were used, improved gradient propagation, or stabilized the full recurrence. The empirically justified conclusion is therefore conditional: CAF is a viable and sometimes advantageous contextual mechanism whose benefit varies with the information regime, temporal dynamics, initialization, persistence construction, and comparator family.

13.4 When Does Exogenous Context Help?

This subsection addresses the second experimental question introduced in Section 13.2: whether the exogenous gate channel contributes from observed context history, and whether its apparent utility changes when admissible future context is supplied. Following the experimental plan, the analysis proceeds in four steps: C0 is first compared with AC1 to assess the contribution of the exogenous gate route; C0 is then compared with AC3 to assess whether exogenous conditioning is sufficient for the gate; the conclusions are contrasted descriptively across information regimes; and their variation across datasets is finally related to the scope of the theoretical results.

The first contrast compares the complete canonical CAF (C0) with AC1, for which $W_g^{(s)} = 0$. AC1 therefore removes the exogenous contribution from the gate while retaining its dependence on the current pre-activation and the endogenous temporal summary. Within the `history` regime, this contrast concerns the observed historical descriptors used during encoding. Within the `known_future` regime, it concerns the complete exogenous route available under that information contract, namely historical descriptors during encoding and admissible future descriptors during decoding. Consequently, the C0–AC1 contrast in the latter regime does not, by itself, isolate the contribution of future values from that of historical values.

Table 8: Effect of removing the exogenous gate route (AC1) and restricting the gate to exogenous context (AC3). Relative changes are computed against C0 within the same dataset and regime; positive values indicate a larger MSE than C0. Adjusted p-values correspond to the prespecified Holm-corrected family of ablation comparisons against C0.

Dataset	Regime	C0 MSE	AC1: no exogenous gate route	AC3: exogenous-only gate
Solar	`history`	83.93 ± 6.36	125.28 ± 3.95; +49.3%; $p_{\text{Holm}} = 0.0156$	108.08 ± 10.53; +28.8%; $p_{\text{Holm}} = 0.0156$
Solar	`known_future`	74.70 ± 3.20	125.28 ± 3.95; +67.7%; $p_{\text{Holm}} = 0.0156$	101.40 ± 5.68; +35.7%; $p_{\text{Holm}} = 0.0156$
Bike	`history`	27214.65 ± 2584.63	29482.33 ± 2944.91; +8.3%; $p_{\text{Holm}} = 0.1289$	37195.65 ± 5172.27; +36.7%; $p_{\text{Holm}} = 0.0156$
Bike	`known_future`	9125.84 ± 978.42	29482.33 ± 2944.91; +223.1%; $p_{\text{Holm}} = 0.0156$	24048.15 ± 1948.52; +163.5%; $p_{\text{Holm}} = 0.0156$
EPF-NP	`history`	26.349 ± 1.440	28.668 ± 0.738; +8.8%; $p_{\text{Holm}} = 0.0684$	28.818 ± 1.808; +9.4%; $p_{\text{Holm}} = 0.1172$
EPF-NP	`known_future`	22.748 ± 1.387	28.668 ± 0.738; +26.0%; $p_{\text{Holm}} = 0.0156$	34.076 ± 2.819; +49.8%; $p_{\text{Holm}} = 0.0156$
METR-LA	`history`	125.36 ± 3.51	127.68 ± 4.25; +1.9%; $p_{\text{Holm}} = 0.4280$	122.68 ± 5.22; −2.1%; $p_{\text{Holm}} = 0.3186$
METR-LA	`known_future`	125.85 ± 3.27	127.68 ± 4.25; +1.5%; $p_{\text{Holm}} = 0.5840$	126.37 ± 2.41; +0.4%; $p_{\text{Holm}} = 1.0000$

The second contrast compares C0 with AC3, for which $W_g^{(a)} = W_g^{(u)} = 0$. The AC3 gate therefore depends only on exogenous context, $g_t = \sigma(W_g^{(s)} s_t + b_g)$. This restriction applies to the gate inputs: the forecaster still has endogenous base, recurrent, and autoregressive pathways. The contrast thus asks whether exogenous conditioning is sufficient for the gate, or whether gate conditioning also benefits from the current pre-activation and recent hidden-state summary. All comparisons are conducted within the dedicated ablation suite, with C0 fixed as the reference and denormalized MSE as the primary endpoint. Cross-regime changes are used only descriptively because the two regimes correspond to separately trained models under different information contracts. The optional future-context permutation control is not included because it was disabled in the reported benchmark; no claim below relies on an unexecuted diagnostic.

Contribution of the exogenous gate route.

Solar provides the clearest evidence that exogenous modulation is useful even without future descriptors. In the `history` regime, removing the exogenous gate route increases MSE from 83.93 ± 6.36 to 125.28 ± 3.95, a relative increase of 49.3% ($p_{\text{Holm}} = 0.0156$). Because future context is neutralized in this regime, this within-regime contrast detects a contribution from the observed historical descriptors used during encoding. In the `known_future` regime, C0 records 74.70 ± 3.20, whereas AC1 remains at 125.28 ± 3.95; the 67.7% gap is also significant ($p_{\text{Holm}} = 0.0156$). Together with the descriptively lower C0 MSE in the future-information regime and the unchanged AC1 result, this pattern is consistent with additional utility from the

admitted future descriptors. It is not a formal interaction test. Moreover, the Solar descriptors are based on a reconstructed temporal index and must not be presented as an authenticated civil calendar.

Bike shows a different pattern. In `history`, AC1 is numerically worse than C0 (29482.33 ± 2944.91 versus 27214.65 ± 2584.63), but the 8.3% difference is not detected after correction ($p_{\text{Holm}} = 0.1289$). In `known_future`, C0 records 9125.84 ± 978.42, while AC1 records 29482.33 ± 2944.91; removing the exogenous gate route is associated with a 223.1% larger MSE ($p_{\text{Holm}} = 0.0156$). The change from a weak historical contrast to a large future-regime contrast is consistent with the available future calendar descriptors being an important source of exogenous utility on Bike. Because no paired cross-regime interaction was tested, this interpretation remains descriptive.

EPF-NP follows the same qualitative transition, but less dramatically. In `history`, removing the exogenous route increases MSE from 26.349 ± 1.440 to 28.668 ± 0.738 (+8.8%), without reaching the corrected significance threshold ($p_{\text{Holm}} = 0.0684$). In `known_future`, C0 records 22.748 ± 1.387, while AC1 remains at 28.668 ± 0.738; the 26.0% degradation is significant ($p_{\text{Holm}} = 0.0156$). This pattern is consistent with useful future calendar information for the complete CAF on EPF-NP, while providing only inconclusive evidence for an additional contribution from historical exogenous descriptors.

METR-LA is the neutral case. AC1 differs from C0 by only 1.9% in `history` and 1.5% in `known_future`, with $p_{\text{Holm}} = 0.4280$ and 0.5840, respectively. No contribution of the exogenous gate route is detected under either information contract. This agrees with the weak descriptive cross-regime change observed for C0. Recent traffic history may already contain much of the useful signal at the short evaluated horizon, but this explanation is plausible rather than causally established.

AC1 has the same reported MSE in the two regimes for every dataset. This is an expected protocol consistency check: after $W_g^{(s)}$ is removed, the canonical base path does not consume exogenous context, so the information distinguishing the two regimes is unavailable to AC1. This invariance supports the intended routing implementation; it does not by itself establish that the corresponding future information is predictive for C0.

Is exogenous conditioning sufficient for the gate?

AC3 retains exogenous modulation but removes the gate's dependence on the preactivation and endogenous temporal summary. Its comparison with C0 therefore evaluates the sufficiency of exogenous conditioning for the gate, not the sufficiency of exogenous information for the entire forecaster. On Solar, AC3 is worse than C0 in both regimes: its MSE is 108.08 ± 10.53 in `history` and 101.40 ± 5.68 in `known_future`, corresponding to increases of 28.8% and 35.7%; both contrasts have $p_{\text{Holm}} = 0.0156$. Exogenous information is useful on Solar, but a gate conditioned on that information alone does not recover the performance of the full gate.

On Bike, AC3 is also inferior in both regimes. It increases MSE by 36.7% in `history` and by 163.5% in `known_future`, with $p_{\text{Holm}} = 0.0156$ in both cases. The future-regime result is especially informative: although the exogenous route is useful, removing a_t and u_t from the gate is detrimental. The experiment therefore supports

combining exogenous descriptors with endogenous gate inputs in this finite forecaster; it does not imply that the rest of AC3 lacks endogenous dynamics.

For EPF-NP, AC3 is not distinguished from C0 in `history` (28.818 ± 1.808 versus 26.349 ± 1.440, $p_{\text{Holm}} = 0.1172$), but is substantially worse in `known_future` (34.076 ± 2.819 versus 22.748 ± 1.387, $p_{\text{Holm}} = 0.0156$). AC3's descriptive cross-regime change is unfavorable even though the full CAF benefits from the exogenous route in the future-information regime. This divergence is compatible with a benefit from joint exogenous and endogenous gate conditioning, but it is not a formal test of their interaction.

For the evaluated METR-LA target sensor, AC3 is not distinguished from C0 in either regime. It is numerically better in `history` (122.68 ± 5.22 versus 125.36 ± 3.51) and nearly identical in `known_future` (126.37 ± 2.41 versus 125.85 ± 3.27), with adjusted p-values of 0.3186 and 1.0000. These results neither establish a need for the full gate nor show equivalence between the restricted and complete variants. They indicate only that the experiment detects no MSE difference between them for this target sensor.

Cross-regime and cross-dataset synthesis.

The C0–AC1 contrasts identify three descriptive profiles. First, Solar shows a detected contribution from the exogenous route in both regimes, with a larger relative gap when future descriptors are admitted. Second, Bike and EPF-NP show no detected historical effect after correction but a clear difference in `known_future`. Third, METR-LA shows no detected dependence on the exogenous route under either regime. These profiles suggest that contextual utility depends jointly on the information content of the descriptors, the forecast horizon, and the dataset dynamics. They do not constitute a formal regime-by-dataset interaction analysis.

The C0–AC3 comparison adds a second dimension. In every setting where C0–AC1 detects a contribution from the exogenous route, AC3 is also inferior to C0: Solar in both regimes, and Bike and EPF-NP in `known_future`. AC3 is additionally inferior on Bike in `history`. Within the evaluated finite-width models, these results favor combining exogenous context with the pre-activation and temporal-summary inputs of the full gate. METR-LA remains the exception, with no detected separation among these alternatives.

These conclusions refer to the prespecified family of comparisons against C0. The broader all-pairs correction is more conservative. On Bike, in particular, the ablation contrasts reported as significant against C0 are no longer significant when every pairwise comparison is included in one Holm family. The primary contrasts remain the prespecified tests for the scientific questions considered here, but this sensitivity limits the strength of mechanistic claims. The same caution is important for interpretation of the strongest positive ablation effects. On Bike in the `history` regime, the C0 ablation anchor is much weaker than its orthogonal twin and than the local LSTM reference; gains of AC2, AC7, or AC8 against C0 may therefore partly reflect a fragile anchor rather than a clean structural benefit. On Solar, the ablation and appendix families use a different learning rate from the main families, so improvements of AC6 or AC8 against the ablation C0 should be interpreted within that suite rather than transferred to the main C0 run.

Moreover, all cross-regime changes remain descriptive because the models are trained separately and the protocol does not perform a paired treatment or interaction test across regimes.

Relation to the theoretical analysis.

The ablations provide empirical evidence about finite recurrent forecasters, not proofs of operator-level theory. The C0–AC1 contrast is compatible with the context-sensitivity pathway in Equation (13), but it does not validate the profiled theoretical criterion for exogenous-channel selection in Section 12, which is not implemented in this benchmark. Likewise, the weaker performance of AC3 does not conflict with context-only universal approximation (Theorem 21): universality is an asymptotic statement about representational capacity on compact domains, whereas these experiments concern fixed-width models, finite data, and bounded training budgets. The empirically supported conclusion is narrower: in the evaluated forecasters, the usefulness of exogenous context is dataset- and regime-dependent, and the settings with a detected exogenous contribution also favor combining it with endogenous gate inputs.

13.5 Gate Structure, Cross-Unit Coupling, and Temporal Memory

The preceding analysis established whether exogenous information contributes to forecasting, but it did not determine which internal gate structure is needed to use that information effectively. We therefore examine three restrictions of the canonical CAF. AC2 replaces the unrestricted dependence on the current pre-activation with a diagonal one, removing gate-induced cross-unit coupling through $W_g^{(a)} a_t$ while retaining componentwise self-modulation, exogenous conditioning, and the temporal-summary pathway. AC3 restricts the gate to exogenous context alone, jointly removing dependence on the current pre-activation and on the temporal summary. AC4 removes only the temporal summary while preserving the full instantaneous dependence on the current pre-activation and the ordinary recurrent state transition. Consequently, C0–AC2 isolates the empirical value of cross-unit coupling induced by the current pre-activation within the gate, C0–AC4 isolates the additional temporal-summary pathway, and C0–AC3 evaluates the more restrictive exogenous-only endpoint rather than a single-component ablation.

All contrasts in this subsection belong to the dedicated ablation suite and use C0 as the fixed reference. They must therefore be interpreted independently of the validation-selected reference used in the main predictive comparison. The primary endpoint is denormalized test MSE, and inferential statements use paired Wilcoxon comparisons against C0 after Holm correction within the prespecified ablation family. The present synthesis reports the mean and empirical standard deviation across seeds, the relative MSE change, and the Holm-adjusted p-value. The dataset-specific statistical outputs additionally provide the paired mean differences, their 95% percentile-bootstrap intervals based on 5000 resamples, and paired rank-biserial correlations. These detailed statistics are reserved for the supplementary tables to keep the main table compact.

Table 9: Gate-structure and temporal-memory ablations. Each cell reports MSE (mean ± standard deviation); ablation cells additionally report the Holm-adjusted p-value against C0. Lower MSE is better.

Dataset	Regime	C0: full gate	AC2: diagonal gate	AC3: context-only gate	AC4: no temporal summary
Solar	`history`	83.93 ± 6.36	84.54 ± 7.10; $p = 1.0000$	108.08 ± 10.53; $p = 0.0156$	87.95 ± 6.27; $p = 0.6406$
Solar	`known_future`	74.70 ± 3.20	77.74 ± 5.27; $p = 0.2617$	101.40 ± 5.68; $p = 0.0156$	77.73 ± 2.93; $p = 0.0820$
Bike	`history`	27214.65 ± 2584.63	21630.61 ± 2479.61; $p = 0.0156$	37195.65 ± 5172.27; $p = 0.0156$	32675.31 ± 4938.37; $p = 0.0586$
Bike	`known_future`	9125.84 ± 978.42	10325.53 ± 1510.87; $p = 0.0156$	24048.15 ± 1948.52; $p = 0.0156$	12287.35 ± 2156.15; $p = 0.0156$
EPF-NP	`history`	26.349 ± 1.440	25.747 ± 1.531; $p = 0.9668$	28.818 ± 1.808; $p = 0.1172$	26.922 ± 1.369; $p = 0.9668$
EPF-NP	`known_future`	22.748 ± 1.387	23.806 ± 0.926; $p = 0.0391$	34.076 ± 2.819; $p = 0.0156$	22.914 ± 1.353; $p = 0.5566$
METR-LA	`history`	125.36 ± 3.51	123.71 ± 4.29; $p = 0.7183$	122.68 ± 5.22; $p = 0.3186$	126.07 ± 5.17; $p = 0.7615$
METR-LA	`known_future`	125.85 ± 3.27	124.20 ± 2.62; $p = 0.7233$	126.37 ± 2.41; $p = 1.0000$	125.58 ± 3.12; $p = 1.0000$

13.5.1 Full versus diagonal and context-only gates

The full and diagonal gates differ in whether modulation of one hidden unit may depend directly on the current pre-activation of other units. The theoretical Jacobian reflects this distinction: its direct component is diagonal, whereas the component induced by $W_g^{(a)}$ can introduce off-diagonal interactions when the gate is full. AC2 removes these interactions from the instantaneous pre-activation pathway but preserves componentwise self-modulation, exogenous conditioning, and the temporal-summary pathway. Its empirical behavior therefore indicates whether unrestricted coupling through the current pre-activation acts as a useful inductive bias under the observed task and information regime. It does not remove every possible cross-unit dependency from the complete recurrent forecaster.

On Solar, no difference between the two structures is detected. Relative to C0, the diagonal gate changes MSE from 83.93 ± 6.36 to 84.54 ± 7.10 in the `history` regime and from 74.70 ± 3.20 to 77.74 ± 5.27 in the `known_future` regime. The relative degradations are 0.7% and 4.1%, with $p_{\text{Holm}} = 1.0000$ and 0.2617, respectively. Thus, the experiment provides no evidence that unrestricted coupling through $W_g^{(a)}$ improves MSE on Solar. The full gate may still represent richer instantaneous interactions, but this representational difference does not translate into a detected advantage under the present budget. These non-significant contrasts do not establish equivalence between the two gates.

Bike exhibits a regime-dependent reversal. In `history`, the diagonal gate improves MSE from 27214.65 ± 2584.63 to 21630.61 ± 2479.61, a 20.5% reduction relative to C0 ($p_{\text{Holm}} = 0.0156$). Under `known_future`, however, it increases MSE from 9125.84 ± 978.42 to 10325.53 ± 1510.87, a 13.1% degradation ($p_{\text{Holm}} = 0.0156$). One plausible interpretation is that restricting the instantaneous pre-activation coupling regularizes the historical model, whereas the unrestricted pathway is beneficial when future calendar information is available and can be integrated with the hidden response. The observed reversal is an empirical interaction pattern, not a direct measurement of learned off-diagonal dependencies or their causal effect.

EPF-NP shows a weaker version of the same pattern. In the historical regime, AC2 is numerically better than C0 (25.747 ± 1.531 versus 26.349 ± 1.440), but no difference is detected ($p_{\text{Holm}} = 0.9668$). With known future context, the diagonal gate is worse (23.806 ± 0.926 versus 22.748 ± 1.387), corresponding to a 4.7% increase in MSE and a significant contrast ($p_{\text{Holm}} = 0.0391$). This result suggests that unrestricted instantaneous coupling can help exploit future information on EPF-NP under the evaluated protocol, although the effect is much smaller than the penalties associated with the exogenous-only restriction.

For the evaluated METR-LA target sensor, AC2 is numerically better than C0 in both regimes, but neither difference is detected. Its MSE is 123.71 ± 4.29 in `history` and 124.20 ± 2.62 in `known_future`, compared with 125.36 ± 3.51 and 125.85 ± 3.27 for C0. The adjusted p-values are 0.7183 and 0.7233. The short-horizon traffic task therefore supplies no evidence that unrestricted instantaneous coupling through $W_g^{(a)}$ improves predictive MSE; no difference between the diagonal and full gates was detected at the resolution of this experiment.

The context-only restriction is more severe because it removes both pre-activation-dependent modulation and the temporal-summary pathway. On Solar, it is worse than C0 in both regimes, increasing MSE by 28.8% in `history` and 35.7% in `known_future`; both comparisons have $p_{\text{Holm}} = 0.0156$. Bike displays still larger penalties of 36.7% and 163.5%, again significant in both regimes. On EPF-NP, the historical increase of 9.4% is not detected ($p_{\text{Holm}} = 0.1172$), whereas the future-context increase of 49.8% is significant ($p_{\text{Holm}} = 0.0156$). METR-LA remains neutral: no difference between the context-only and full gates is detected under either information contract.

Taken together, AC2 and AC3 show that gate structure cannot be ordered universally. A diagonal dependency is not distinguished from the full gate on Solar, EPF-NP in `history`, and METR-LA, and it improves Bike in `history`. The full structure is favored on Bike and EPF-NP when future context is known. By contrast, removing all endogenous gate dependence is frequently harmful, particularly where future descriptors are informative. These results suggest that performance often depends more on retaining both exogenous and endogenous gate inputs than on unrestricted instantaneous cross-unit coupling alone, although these ablations do not identify the responsible mechanism.

The context-only results do not contradict its theoretical expressivity. The relevant approximation result concerns the closure of a function class and does not guarantee that a fixed-width recurrent forecaster will be statistically or computationally efficient under finite data and a fixed optimization budget. Likewise, the empirical advantage

of a full gate on selected tasks does not verify that its theoretical off-diagonal Jacobian term caused the improvement. Establishing that mechanism would require direct diagnostics of learned coupling or targeted perturbation experiments.

13.5.2 Contribution of the temporal summary

AC4 removes the compact summary of recent hidden states from the gate while leaving the recurrent state transition intact. The comparison therefore does not test "memory versus no memory." It tests whether a dedicated gate-level summary adds detectable predictive information beyond the history already propagated by the recurrent state and the autoregressive decoder.

On Solar, removing the summary increases MSE from 83.93 ± 6.36 to 87.95 ± 6.27 in `history` and from 74.70 ± 3.20 to 77.73 ± 2.93 in `known_future`. These increases, 4.8% and 4.1%, are not significant after Holm correction ($p_{\text{Holm}} = 0.6406$ and 0.0820). The future-context result is directionally favorable to the complete model, but it remains insufficient for a claim that the temporal summary is necessary or improves MSE under this design.

Bike is the only dataset showing a statistically detected contribution from the temporal summary in one of the two regimes. In `history`, AC4 increases MSE from 27214.65 ± 2584.63 to 32675.31 ± 4938.37, a 20.1% degradation, but the adjusted test remains above the conventional threshold ($p_{\text{Holm}} = 0.0586$). With known future context, removing the summary raises MSE from 9125.84 ± 978.42 to 12287.35 ± 2156.15, a 34.6% increase that is significant ($p_{\text{Holm}} = 0.0156$). Under the primary comparison family, the temporal-summary pathway is therefore associated with a lower MSE when Bike's future calendar context is available. This result does not by itself identify how the summary interacts with the exogenous descriptors.

On EPF-NP, AC4 remains close to C0. Historical MSE changes from 26.349 ± 1.440 to 26.922 ± 1.369 (+2.2%, $p_{\text{Holm}} = 0.9668$), and future-context MSE changes from 22.748 ± 1.387 to 22.914 ± 1.353 (+0.7%, $p_{\text{Holm}} = 0.5566$). No additional contribution of the gate-level temporal summary beyond the ordinary recurrence was detected on EPF-NP at this resolution; this absence of detection is not evidence that the two variants are equivalent.

METR-LA is similarly insensitive. AC4 changes MSE by only +0.6% in `history` and −0.2% in `known_future`; the corresponding adjusted p-values are 0.7615 and 1.0000. At the short evaluated horizon, no improvement from the additional summary was detected beyond the information already conveyed by the recurrent state.

Across the eight dataset–regime combinations, a contribution of the dedicated temporal summary is detected only for Bike with known future context in the prespecified comparison family. The remaining contrasts range from small numerical differences to a near-threshold historical trend on Bike. This heterogeneity is consistent with the summary acting as a task-dependent inductive bias rather than a universally necessary form of memory. One plausible hypothesis is that it is most useful when forecasts depend jointly on recent latent dynamics and a future exogenous schedule; direct analyses of the learned gates and temporal representations would be needed to test that hypothesis.

The joint reading of AC2, AC3, and AC4 also prevents an over-interpretation of the context-only comparison. AC3 removes two endogenous gate pathways at once, yet AC4 shows that the temporal component alone is rarely associated with a detected difference. Hence, where AC3 is substantially worse but AC4 is not, the deficit cannot be attributed solely to loss of the temporal summary. Dependence on the current hidden response, its interaction with exogenous context, or another optimization consequence of the composite restriction remains a plausible explanation. Bike with known future context is the case in which both the temporal summary and the unrestricted instantaneous coupling are associated with detected differences under the prespecified comparisons against C0; this conclusion is sensitive to the broader all-pairs multiplicity family discussed below.

These conclusions remain tied to the prespecified comparisons against C0. A broader all-pairs family is more conservative and, on Bike, removes the significance of the ablation contrasts. The reported findings should therefore be read as evidence under the primary question-oriented correction family, with Bike's structural conclusions treated as sensitive to multiplicity scope. More generally, prediction-level ablations identify useful architectural biases but do not establish the detailed gradient-flow, conditioning, or stability mechanisms described by the analytical results.

13.6 Additive Context Routing Controls

The AC6 and AC8 controls test whether admitting exogenous context additively into the recurrent base improves over the canonical multiplicative-only routing. AC6 combines this additive route with standard initialization, so C0–AC6 changes routing and initialization jointly. AC8 is the closest available routing control because it preserves the compensated scheme of C0, but it still adds projection parameters and is not strictly matched in capacity or in every initial weight realization. These contrasts therefore evaluate complete variants rather than a single isolated causal factor.

The additive route is therefore highly dataset- and regime-dependent. AC8 is favorable on Solar and Bike in the historical regime under the primary family, although the Solar result is sensitive to the larger all-pairs Holm family. In contrast, AC8 is detrimental on Solar, Bike, and EPF-NP when future context is supplied, and it is not distinguished on METR-LA. AC6 follows the same adverse future-context direction on Bike and EPF-NP, but because it also changes initialization, it is best read as a complementary stress test rather than the primary routing contrast. These outcomes argue against treating additive context injection as a uniformly beneficial extension of CAF; its effect depends on the information regime, dataset dynamics, capacity change, and multiplicity family.

13.7 Initialization, Persistence, and State Preservation

This subsection separates three design choices that act at different levels of the forecasting system. Compensated and orthogonal initialization alter the starting geometry while preserving the canonical topology. Persistence modifies the forecast construction by expressing the prediction relative to the last observed target. State preservation

Table 10: Additive-context routing controls. Each AC cell reports MSE (mean ± standard deviation) and the Holm-adjusted p-value against C0 in the primary ablation family. Lower MSE is better.

Dataset	Regime	C0	AC6: additive, standard	AC8: additive, compensated
Solar	`history`	83.93 ± 6.36	74.56 ± 6.36; $p = 0.1367$	74.41 ± 5.83; $p = 0.0234$
Solar	`known_future`	74.70 ± 3.20	84.79 ± 10.30; $p = 0.0684$	96.41 ± 18.02; $p = 0.0352$
Bike	`history`	27214.65 ± 2584.63	25480.96 ± 1988.35; $p = 0.1602$	17426.03 ± 1319.06; $p = 0.0156$
Bike	`known_future`	9125.84 ± 978.42	19539.26 ± 4767.25; $p = 0.0156$	18828.23 ± 11086.37; $p = 0.0156$
EPF-NP	`history`	26.349 ± 1.440	28.171 ± 2.829; $p = 0.5273$	28.062 ± 3.132; $p = 0.6406$
EPF-NP	`known_future`	22.748 ± 1.387	31.998 ± 1.723; $p = 0.0156$	32.598 ± 2.423; $p = 0.0156$
METR-LA	`history`	125.36 ± 3.51	122.38 ± 3.29; $p = 0.2120$	124.22 ± 3.87; $p = 0.5629$
METR-LA	`known_future`	125.85 ± 3.27	127.76 ± 3.68; $p = 0.7233$	128.87 ± 4.33; $p = 0.5840$

changes the recurrent update itself by introducing an adaptive path that carries the previous state forward. These choices must not be pooled into a single architectural comparison: initialization can be evaluated between topology-identical models within a suite, persistence must be matched to the comparator family, and state preservation is an extension of pure CAF rather than another initialization of it.

13.7.1 Compensated and orthogonal initialization

The compensated initialization starts from a constant gate and rescales the recurrent transition and prediction head so that, under matched deterministic conditions, the initial free-running forecast coincides with that of a corresponding hyperbolic-tangent recurrent model. This identity is an initialization control, not a prediction about the optimized solution. The reparameterization changes the Euclidean optimization geometry and the effective quadratic penalty on the rescaled blocks; stochastic training can therefore lead to different trajectories immediately after initialization. The orthogonal twin retains the same topology, parameter count, persistence convention, and compensation, but changes the initialization of the recurrent transition.

We denote the compensated canonical anchor by C0 and its topology-identical orthogonal twin by C0o. Their persistence-free counterparts are denoted C0p and C0op, respectively. Within the ablation suite, AC5 denotes the canonical topology under standard initialization, whereas AC7 denotes the complete state-preserving extension.

The internal comparison identifies one clear initialization effect. On Bike in the `history` regime, orthogonal initialization reduces MSE from 27214.65 ± 2584.63 to 17439.05 ± 2727.71, a descriptive reduction of 35.9% ($p_{\text{Holm}} = 0.0098$). No corresponding difference is detected with known future context. On Solar, EPF-NP, and METR-LA, the topology-matched twins are not distinguished in either

Table 11: Topology-matched compensated versus orthogonal CAF in the internal suite. The last column gives the Holm-adjusted paired comparison against the validation-selected reference of that suite.

Dataset	Regime	Compensated CAF	Orthogonal CAF	Within-suite inference
Solar	`history`	74.81 ± 7.42	75.98 ± 4.43	Not distinguished; $p_{\text{Holm}} = 0.9086$
Solar	`known_future`	66.87 ± 3.62	68.74 ± 4.06	Not distinguished; $p_{\text{Holm}} = 0.3028$
Bike	`history`	27214.65 ± 2584.63	17439.05 ± 2727.71	Orthogonal better; $p_{\text{Holm}} = 0.0098$
Bike	`known_future`	9125.84 ± 978.42	9855.39 ± 822.78	Not distinguished; $p_{\text{Holm}} = 0.0840$
EPF-NP	`history`	26.349 ± 1.440	26.274 ± 1.575	Not distinguished; $p_{\text{Holm}} = 0.8633$
EPF-NP	`known_future`	22.748 ± 1.387	23.535 ± 2.307	Not distinguished; $p_{\text{Holm}} = 0.2754$
METR-LA	`history`	125.36 ± 3.51	127.32 ± 4.04	Not distinguished; $p_{\text{Holm}} = 0.6233$
METR-LA	`known_future`	125.85 ± 3.27	127.55 ± 4.47	Not distinguished; $p_{\text{Holm}} = 0.4586$

regime. Orthogonalization therefore has a material empirical effect in one internal configuration rather than constituting a uniformly superior default.

The persistence-free external suite reinforces this contingency. On EPF-NP in `history`, the compensated variant obtains 29.849 ± 3.720, whereas its orthogonal twin obtains 42.446 ± 8.608; the compensated model is better in the primary reference-based family ($p_{\text{Holm}} = 0.0078$). For the evaluated METR-LA target sensor, the same direction appears in both regimes: compensated CAF obtains 130.03 ± 10.02 versus 141.08 ± 19.00 in `history` ($p_{\text{Holm}} = 0.0479$), and 126.15 ± 6.03 versus 133.86 ± 9.14 in `known_future` ($p_{\text{Holm}} = 0.0054$). The historical contrast for this METR-LA target sensor is sensitive to multiplicity scope: its adjusted p-value becomes 0.1437 in the complete all-pairs family. By contrast, the persistence-free twins are not distinguished on Solar or Bike. Hence, the observed initialization ranking varies across datasets, regimes, and experimental suites. Because forecast construction and other protocol factors change jointly across suites, these results do not identify which factor causes that variation.

A second comparison replaces compensation with a standard initialization while keeping the ablation-suite topology fixed. Within the primary reference-based Holm family, standard initialization is worse on Bike in both regimes: MSE rises by 31.8% in `history` and 68.1% in `known_future`, with $p_{\text{Holm}} = 0.0156$ in both cases. These Bike contrasts are sensitive to multiplicity scope: they are no longer significant when all pairwise comparisons are included in a single Holm family. For the evaluated METR-LA target sensor in `history`, standard initialization improves MSE from 125.36 ± 3.51 to 122.68 ± 2.04 ($p_{\text{Holm}} = 0.0093$ in the primary reference-based family). No difference is detected on Solar or EPF-NP, or for the evaluated METR-LA target sensor with future context. Compensation thus provides an interpretable and controlled starting

point, but its exact initial embedding does not imply an optimization advantage after training.

13.7.2 Persistence-matched forecasting

Persistence is a forecasting convention, not part of the CAF activation operator. In the internal suite, CAF and the trainable local recurrent models all predict a residual relative to the last observed target, and the last value is restored when the forecast is assembled. In the external suite, no such shortcut is added to the specialized external architectures; CAF is therefore evaluated without persistence as well. This separation prevents an architectural comparison from being confounded by a last-value residual available to only one family.

The persistence-free CAF variants are not a simple ablation of the internal CAF. Removing persistence changes both the initialization of autoregressive feedback and the assembly of the prediction. Moreover, internal and external suites can differ in roster, random repetitions, optimization settings, and selected reference. Consequently, the numerical gap between an internal CAF score and an external persistence-free CAF score must not be interpreted as the causal effect of persistence. No paired, within-suite persistence ablation is available for that claim.

Within the matched external suite, the appropriate conclusion is comparative. In `history`, persistence-free CAF is not distinguished from CrossLinear on Solar, is outperformed by all available external models on Bike and EPF-NP, and outperforms CrossLinear and TimeXer for the evaluated METR-LA target sensor in terms of MSE. In `known_future`, it outperforms adapted CrossLinear on Solar; on Bike it outperforms adapted CrossLinear and TimeXer while remaining indistinguishable from DAG and adapted PRformer; on EPF-NP it is outperformed by every available external model; and for the evaluated METR-LA target sensor it outperforms all available external models. These heterogeneous rankings show that CAF can remain competitive without a persistence shortcut, but do not establish that persistence is beneficial or harmful in isolation.

The distinction is especially important on EPF-NP. CAF performs favorably in the persistence-shared internal future-context comparison but poorly in the persistence-free external future-context suite. Since several factors change between these suites, the contrast cannot be assigned uniquely to persistence. It instead demonstrates that conclusions about the activation mechanism depend on the complete forecasting system and on a comparator-matched prediction contract.

13.7.3 State-preserving extension

The state-preserving extension supplements the pure context-gated nonlinear response with an adaptive carry path from the previous hidden state. It can therefore retain information even when the newly computed response is strongly attenuated. This changes the recurrent update and introduces an additional gate; it is not a structured member of the pure CAF hierarchy. It also uses standard rather than compensated initialization because the carry path breaks the exact initial identity with the hyperbolic-tangent twin. The comparison with C0 thus evaluates a complete extension and does not isolate state preservation at equal initialization.

Table 12: Standard initialization and the state-preserving extension in the ablation suite. Relative change is measured against C0; negative values indicate lower MSE. Reported p-values are Holm-adjusted within the primary reference-based comparison family.

Dataset	Regime	C0 MSE	Standard initialization	State-preserving extension
Solar	`history`	83.93 ± 6.36	81.85 ± 4.76; -2.5%; $p_{\text{Holm}} = 1.0000$	84.33 ± 5.62; $+0.5\%$; $p_{\text{Holm}} = 1.0000$
Solar	`known_future`	74.70 ± 3.20	78.58 ± 1.68; $+5.2\%$; $p_{\text{Holm}} = 0.0684$	77.09 ± 1.59; $+3.2\%$; $p_{\text{Holm}} = 0.2617$
Bike	`history`	27214.65 ± 2584.63	35866.08 ± 4404.01; $+31.8\%$; $p_{\text{Holm}} = 0.0156$	20981.72 ± 1555.81; -22.9%; $p_{\text{Holm}} = 0.0156$
Bike	`known_future`	9125.84 ± 978.42	15341.63 ± 2539.96; $+68.1\%$; $p_{\text{Holm}} = 0.0156$	8865.78 ± 897.76; -2.8%; $p_{\text{Holm}} = 0.6953$
EPF-NP	`history`	26.349 ± 1.440	26.764 ± 1.239; $+1.6\%$; $p_{\text{Holm}} = 0.9668$	24.076 ± 0.832; -8.6%; $p_{\text{Holm}} = 0.0312$
EPF-NP	`known_future`	22.748 ± 1.387	22.536 ± 1.229; -0.9%; $p_{\text{Holm}} = 0.5508$	21.023 ± 1.067; -7.6%; $p_{\text{Holm}} = 0.0820$
METR-LA	`history`	125.36 ± 3.51	122.68 ± 2.04; -2.1%; $p_{\text{Holm}} = 0.0093$	123.06 ± 2.68; -1.8%; $p_{\text{Holm}} = 0.0718$
METR-LA	`known_future`	125.85 ± 3.27	124.37 ± 3.12; -1.2%; $p_{\text{Holm}} = 0.7571$	123.71 ± 2.97; -1.7%; $p_{\text{Holm}} = 0.7233$

The dataset-specific statistical outputs also contain the paired mean differences, their 95% percentile-bootstrap intervals based on 5000 resamples, and paired rank-biserial correlations. These detailed statistics should be retained in the supplementary statistical tables so that the main table can remain compact.

Under the primary reference-based Holm family, the complete state-preserving extension improves over C0 in two historical settings. On Bike, it reduces MSE from 27214.65 ± 2584.63 to 20981.72 ± 1555.81, a 22.9% reduction ($p_{\text{Holm}} = 0.0156$). On EPF-NP, it reduces MSE from 26.349 ± 1.440 to 24.076 ± 0.832, an 8.6% reduction ($p_{\text{Holm}} = 0.0312$). These results show that the complete extension can improve prediction in these configurations, but they do not identify retained latent state as the causal source of the gain. Moreover, the Bike contrast is no longer significant in the complete all-pairs Holm matrix and should therefore be treated as sensitive to the multiplicity family.

The effect is not uniform. On Solar, the extension is not distinguished from C0 in either regime. For the evaluated METR-LA target sensor, it is numerically better in both regimes but no difference is detected after Holm correction. With known future context, no difference is detected on Bike or EPF-NP, despite numerical reductions of 2.8% and 7.6%. The reference-based evidence for the complete extension is therefore concentrated in the historical experiments, with different robustness to multiplicity scope across datasets; the future-context experiments do not establish an additional gain.

The results motivate the state-preserving formulation as a separate extension, not as a replacement for the canonical operator. Its favorable historical results cannot

validate the theoretical claims for pure CAF because the update rule, initialization, parameter count, and dependency paths all change. In particular, local bounds derived for the activation operator do not certify the full state-preserving recurrence: the carry gate, temporal summary, and autoregressive feedback add derivative paths absent from the pure analysis. Direct stability diagnostics and an initialization-matched memory control would be required to attribute the observed gains specifically to improved state retention.

Overall, the experiments support a conditional rather than universal role for these design choices. Compensation supplies an analytically controlled starting point but does not guarantee the best trained solution; orthogonalization is associated with substantial changes in selected configurations but has no uniform direction; persistence is essential to comparator fairness but is not causally isolated by the current suites; and the complete state-preserving extension is favorable in selected historical comparisons while remaining a distinct recurrent construction beyond pure CAF.

13.8 Comparison with Alternative Conditional and Forecasting Mechanisms

CAF introduces context at the level of the nonlinear response, but the benchmark also includes mechanisms that condition or propagate information at other levels. The recurrent baselines regulate the state transition, feature-wise modulation applies a context-dependent affine transformation to intermediate representations, and the external forecasting architectures impose specialized cross-series, patch-based, or dependency-aware inductive biases. The objective of this comparison is therefore not to claim that CAF replaces these mechanisms, but to determine when activation-level modulation constitutes a competitive alternative under matched information and forecast-construction contracts.

The comparison is divided into three families. The local suite shares the persistence construction across trainable models and provides the most direct comparison with recurrent and conditional baselines. The external suite removes persistence from CAF because it is not added to the published forecasting architectures. The appendix suite broadens the empirical context with recurrent, attention-based, basis-expansion, and probabilistic references. Statistical conclusions are always drawn within a suite and regime; scores are not compared inferentially across suites.

13.8.1 Local recurrent and conditional baselines

The local LSTM baseline controls information flow through recurrent transition gates, whereas FiLM-RNN applies a learned affine modulation to hidden representations. CAF acts at a different level: its bounded gate modulates the base nonlinear response directly. All three mechanisms receive the same admissible information in a given regime, and all use the same persistence construction. This makes their comparison informative about alternative modes of context integration, while differences in topology and parameterization prevent a strict one-component causal interpretation.

In the historical regime, CAF is generally competitive but not uniformly superior. On Solar, its MSE is 74.81 ± 7.42, compared with 76.70 ± 4.31 for LSTM and $70.92 \pm$

Table 13: CAF relative to the local recurrent and conditional baselines under a shared persistence convention. "Better" and "worse" denote Holm-significant MSE contrasts; "not distinguished" denotes that no difference was detected.

Dataset	`history`	`known_future`
Solar	Not distinguished from LSTM or FiLM-RNN	Better than LSTM and FiLM-RNN
Bike	Not distinguished from LSTM; better than FiLM-RNN	Better than LSTM and FiLM-RNN
EPF-NP	Worse than LSTM; not distinguished from FiLM-RNN	Better than LSTM and FiLM-RNN
METR-LA	Not distinguished from LSTM or FiLM-RNN	Not distinguished from LSTM or FiLM-RNN

3.95 for FiLM-RNN; neither contrast is detected after Holm correction. On Bike, the validation-selected orthogonal CAF obtains 17439.05 ± 2727.71. It is not distinguished from LSTM (15508.16 ± 2096.41, $p_{\text{Holm}} = 0.4648$), but improves over FiLM-RNN (21990.85 ± 1930.70, $p_{\text{Holm}} = 0.0410$). EPF-NP is the adverse local case: CAF obtains 26.274 ± 1.575 and is outperformed by LSTM (23.661 ± 1.115, $p_{\text{Holm}} = 0.0117$), while remaining indistinguishable from FiLM-RNN. For the evaluated METR-LA target sensor, CAF, LSTM, and FiLM-RNN occupy a narrow MSE range and no pairwise difference against CAF is detected. Thus, historical CAF offers a viable alternative to local recurrent conditioning, but the ranking depends on the dataset and, on Bike, on initialization.

The known-future regime is more favorable to CAF. On Solar, CAF reaches 66.87 ± 3.62, compared with 76.53 ± 10.92 for LSTM ($p_{\text{Holm}} = 0.0134$) and 79.50 ± 8.64 for FiLM-RNN ($p_{\text{Holm}} = 0.0004$). On Bike, its MSE of 9125.84 ± 978.42 is lower than the LSTM and FiLM-RNN values of 11273.51 ± 1600.21 and 11653.32 ± 998.09, with $p_{\text{Holm}} = 0.0098$ in both comparisons. On EPF-NP, CAF obtains 22.748 ± 1.387, versus 25.783 ± 1.263 for LSTM and 28.043 ± 2.672 for FiLM-RNN; both contrasts favor CAF after correction. METR-LA again remains neutral: CAF is not distinguished from either local baseline.

Across these comparisons, the evaluated CAF forecaster is particularly competitive when admissible future context is informative. This does not imply that CAF subsumes recurrent gating or feature-wise affine conditioning. Rather, the results show that a distinct, bounded modulation of the nonlinear response can be preferable to the evaluated local alternatives on Solar, Bike, and EPF-NP under the future-context contract, while METR-LA and the historical experiments rule out a claim of universal dominance. Because the comparisons involve complete forecasters, they do not isolate activation-level modulation as the sole cause of the observed differences.

13.8.2 Persistence-matched external architectures

The external suite evaluates CAF without the last-value residual used in the local suite. CrossLinear, TimeXer, and PRformer are specialized forecasting architectures

Table 14: Persistence-matched external comparison. The statements summarize Holm-corrected MSE contrasts against the validation-selected CAF reference in each suite.

Dataset	`history`	`known_future`
Solar	Not distinguished from CrossLinear; worse than TimeXer and PRformer	Better than adapted CrossLinear; not distinguished from adapted TimeXer, DAG, or adapted PRformer
Bike	Worse than CrossLinear, TimeXer, and PRformer	Better than adapted CrossLinear and TimeXer; not distinguished from DAG or adapted PRformer
EPF-NP	Worse than CrossLinear, TimeXer, and PRformer	Worse than adapted CrossLinear, adapted TimeXer, DAG, and adapted PRformer
METR-LA	Better than CrossLinear and TimeXer	Better than adapted CrossLinear, adapted TimeXer, and DAG

that are historical by construction in their evaluated upstream forms. In the known-future regime, they receive a declared residual context adapter added for this protocol and must not be described as canonical reproductions of the original architectures. DAG natively accepts future exogenous information and is included only in that regime; its incomplete history-only reduction is excluded. PRformer is disabled by configuration in the METR-LA roster. These asymmetries mean that the experiment compares complete systems under a common information contract, but not an identical set of native architectures on every dataset.

On Solar in `history`, persistence-free CAF obtains 80.97 ± 8.62. It is not distinguished from CrossLinear (84.10 ± 11.30), but TimeXer (65.39 ± 2.58) and PRformer (65.15 ± 3.36) are better, both with $p_{\text{Holm}} = 0.0002$. With future context, CAF reaches 71.08 ± 5.89 and outperforms adapted CrossLinear (85.95 ± 12.86, $p_{\text{Holm}} = 0.0214$), while no difference is detected against adapted TimeXer, DAG, or adapted PRformer. CAF consequently moves from a mixed historical position to outperforming adapted CrossLinear when future descriptors are available, while no difference is detected from adapted TimeXer, DAG, or adapted PRformer.

Bike shows a clearer reversal. In `history`, CAF obtains 11849.52 ± 1356.75 and is outperformed by CrossLinear (7996.12 ± 729.09), TimeXer (7334.05 ± 193.66), and PRformer (6961.36 ± 168.70), with $p_{\text{Holm}} = 0.0078$ for each contrast. In `known_future`, CAF obtains 7148.73 ± 616.75 and outperforms adapted CrossLinear (8018.34 ± 654.84, $p_{\text{Holm}} = 0.0061$) and adapted TimeXer (8516.02 ± 461.65, $p_{\text{Holm}} = 0.0006$). It is not distinguished from DAG or adapted PRformer. The change illustrates that architectural rankings are conditional on the available information, although the history and future external suites use different numbers of repetitions on Bike and are not a paired cross-regime experiment.

EPF-NP is consistently unfavorable to persistence-free CAF. In the historical regime, CAF records 29.849 ± 3.720, whereas CrossLinear, TimeXer, and PRformer obtain 24.126 ± 2.522, 24.791 ± 0.958, and 23.426 ± 1.073; all three comparisons have $p_{\text{Holm}} = 0.0078$. With future context, the selected CAF reference records 75.051 ± 10.673, compared with 21.686 ± 2.107 for adapted CrossLinear, 19.040 ± 1.934 for adapted TimeXer, 28.146 ± 6.012 for DAG, and 22.476 ± 2.766 for adapted PRformer. Each external model is better with $p_{\text{Holm}} = 0.0098$. This result is a central counterexample to any claim that contextual gating alone guarantees competitiveness with specialized forecasting systems.

The evaluated METR-LA target sensor exhibits the opposite profile. In `history`, CAF reaches 130.03 ± 10.02, improving on CrossLinear (141.20 ± 1.25, $p_{\text{Holm}} = 0.0060$) and TimeXer (139.12 ± 1.18, $p_{\text{Holm}} = 0.0067$). Under known future context, CAF reaches 126.15 ± 6.03, compared with 139.60 ± 2.11 for adapted CrossLinear, 137.60 ± 1.64 for adapted TimeXer, and 163.05 ± 19.22 for DAG; all three contrasts favor CAF with $p_{\text{Holm}} = 0.0002$. The MSE advantage does not extend uniformly to MAE, for which CrossLinear and TimeXer are lower. This pattern is consistent with an advantage concentrated in reducing larger errors rather than every aspect of the error distribution.

The external evidence therefore rejects a global ordering. Specialized architectures outperform CAF on EPF-NP and on historical Bike, whereas CAF outperforms the available alternatives in MSE for the evaluated METR-LA target sensor. On Solar and future-context Bike, CAF outperforms some alternatives, while no difference is detected from others; these results do not identify a unique winner. These findings position CAF as a competitive mechanism in selected temporal systems, not as a general replacement for architectures designed around different forecasting biases.

13.8.3 Complementary appendix baselines

The appendix suite adds a context-aware GRU, Transformer, Informer, a local context adaptation of N-BEATS, and a local Gaussian adaptation of DeepAR. These models broaden the comparison but do not isolate a property of the CAF operator. The context interfaces are protocol-level adaptations, and the DeepAR comparison uses point forecasts derived from its predictive means. Detailed scores remain in the appendix; Table 15 retains only the direction of the Holm-corrected comparisons against the canonical CAF reference.

On Solar, CAF is not distinguished from GRU, Transformer, Informer, or N-BEATS in either regime and outperforms DeepAR in both. No difference is detected between CAF and GRU, Transformer, Informer, or N-BEATS, so the experiment does not identify a unique best architecture among these models. On Bike, historical CAF is outperformed by GRU, Informer, and N-BEATS, is not distinguished from Transformer, and outperforms DeepAR. With future context, this ordering reverses: CAF outperforms Transformer, Informer, N-BEATS, and DeepAR, while the difference from GRU is not detected.

EPF-NP yields a more selective result. In `history`, GRU outperforms CAF, whereas CAF outperforms Transformer, Informer, N-BEATS, and DeepAR. With future context, GRU remains better, Informer is not distinguished from CAF, and

Table 15: Summary of the five complementary comparisons per dataset and regime. A CAF win means that the canonical CAF reference has significantly lower MSE; a loss means that the alternative has significantly lower MSE.

Dataset	Regime	CAF wins	Not distinguished	CAF losses
Solar	`history`	1	4	0
Solar	`known_future`	1	4	0
Bike	`history`	1	1	3
Bike	`known_future`	4	1	0
EPF-NP	`history`	4	0	1
EPF-NP	`known_future`	3	1	1
METR-LA	`history`	2	3	0
METR-LA	`known_future`	2	3	0

CAF outperforms Transformer, N-BEATS, and DeepAR. For the evaluated METR-LA target sensor, the pattern is stable across regimes: CAF is not distinguished from GRU, Informer, or N-BEATS, and it outperforms Transformer and DeepAR.

These appendix results refine, but do not overturn, the main comparison. CAF is frequently competitive across diverse mechanism classes, particularly under known future context on Bike and against several attention-based or probabilistic references. At the same time, GRU is stronger on EPF-NP in both regimes and several alternatives outperform historical CAF on Bike. The absence of a detected difference in many Solar and METR-LA comparisons must not be interpreted as formal equivalence.

Overall, the mechanism-level evidence supports a qualified positioning. CAF provides an explicit activation-level route for contextual modulation and can match or exceed recurrent, affine-conditional, and specialized forecasting alternatives in several configurations. Its strongest local pattern appears when future context is genuinely informative; its external performance remains highly dataset-dependent. The results demonstrate practical viability and complementary inductive bias, not universal superiority or empirical proof of the theoretical properties of the CAF operator.

13.9 Model Complexity and Practical Computational Cost

The theoretical analysis in Section 9 separates the structural overhead of the CAF gate from the cost of the underlying layer. For an affine full non-temporal gate, the additional parameter count is $H^2+HS+H = H(H+S+1)$, and the gate-specific forward and backward costs are $O(H^2+HS)$. Restricting only the pre-activation block to a diagonal structure reduces that block to $O(H)$, while a context-only non-temporal gate requires $O(HS)$. In the evaluated temporal realization, however, the recent-state summary is part of the augmented gate context. AC2 still retains the dense temporal-summary block $W_g^{(u)} \in \mathbb{R}^{H\times U}$, so its gate cost is $O(H + HU + HS)$, which remains quadratic when $U = H$. Thus AC2 removes $H^2 - H$ parameters from $W_g^{(a)}$ but does not make the entire temporal gate linear in the hidden width. Unrolling over T recurrent steps therefore multiplies the corresponding gate-specific operation count by T.

Table 16: Trainable parameter counts for the canonical CAF, two structured gate restrictions, and the local learned references. The diagonal restriction removes cross-unit dependence from the current-pre-activation gate block. The exogenous-only restriction additionally removes the temporal-summary gate block; it is therefore more restrictive than a one-block ablation.

Dataset	Canonical CAF	Diagonal gate	Exogenous-only gate	LSTM	FiLM-RNN
Solar	14,065	11,809	9,457	86,065	49,201
Bike	8,113	5,857	3,505	62,641	25,969
EPF-NP	7,633	5,377	3,025	60,721	24,049
METR-LA	17,425	15,169	12,817	99,505	62,641

The additional saved gate activations scale as $O(TH)$, excluding parameters, gradients, optimizer states, and the remaining recurrent computation. These expressions characterize structural growth, not realized hardware latency.

Parameter complexity.

Table 16 reports the parameter counts of the canonical CAF, its two principal structural restrictions, and the local learned references. The canonical CAF contains between 7,633 and 17,425 trainable parameters across the four datasets. It uses 82.5%–87.4% fewer parameters than the corresponding LSTM and 68.3%–72.2% fewer than FiLM-RNN. Thus, the evaluated CAF forecaster is parametrically compact relative to these local references, even though its gate introduces an overhead relative to an otherwise matched ungated recurrent layer.

The structural ablations reproduce the ordering predicted by the gate parameterization. Relative to the full gate, the diagonal variant removes $2{,}256 = 48^2 - 48$ parameters on every dataset, corresponding to the replacement of a dense 48×48 pre-activation block by a diagonal vector. The exogenous-only variant contains $4{,}608 = 2 \times 48^2$ fewer parameters than the canonical model because it removes both dense endogenous gate blocks. This agreement is an implementation-level illustration of the theoretical complexity hierarchy. It does not imply a performance ordering: as shown in Section 13.5, the diagonal and exogenous-only restrictions can improve, preserve, or degrade MSE depending on the dataset and information regime.

Measured tuning cost.

The practical timing measure is the cumulative training time of all hyperparameter trials for one seed, averaged across seeds. It is therefore a cost of the complete selection procedure rather than the duration of a single epoch or of the final selected fit. Table 17 summarizes the reported means for the canonical CAF with and without the persistence construction and for the principal external references. Each cell reports `history/known_future`; intervals for CAF reflect the corresponding canonical configurations within the reported suites.

Parameter count and tuning time lead to different rankings. On Bike, the canonical CAF uses only 8,113 parameters, yet its reported tuning cost is approximately 450–530

Table 17: Mean cumulative tuning time per seed, in seconds. Values are reported as `history/known_future`. A dash denotes a model not included in the corresponding regime. Times are descriptive and should be compared only within a dataset and protocol configuration.

Dataset	Canonical CAF	CrossLinear	TimeXer	Other external model
Solar	∼ 283–374 / ∼ 382–422	∼ 12 / ∼ 17	∼ 24 / ∼ 32	PRformer: ∼ 495 / ∼ 624
Bike	∼ 450–523 / ∼ 450–530	∼ 9.5 / ∼ 12.4	∼ 20.6 / ∼ 24.2	PRformer: ∼ 93.2 / ∼ 96.6
EPF-NP	251.1–452.0 / 194.8–419.8	12.2 / 9.8	18.3 / 15.7	PRformer: 70.7 / 60.9
METR-LA	∼ 52–59 / ∼ 53–63	∼ 11 / ∼ 12	∼ 39 / ∼ 37	DAG: – / ∼ 181

seconds per seed, compared with about 9.5–12.4 seconds for CrossLinear and 20.6–24.2 seconds for TimeXer. EPF-NP shows the same separation: canonical CAF configurations require about 195–452 seconds across the two regimes, whereas CrossLinear and TimeXer require about 10–18 seconds. On Solar, CrossLinear and TimeXer are again markedly faster, but PRformer is slower than CAF, requiring approximately 495–624 seconds. For the evaluated METR-LA target sensor, where a single learning rate is evaluated, CAF requires about 52–63 seconds; CrossLinear remains faster, TimeXer is closer, and DAG is slower in the future-context regime.

These differences cannot be attributed to gate parameter count alone. Bike and EPF-NP evaluate three learning-rate candidates per model and seed, whereas Solar and METR-LA use one candidate in the principal suites. The realized cost also depends on early stopping, the number of executed epochs, sequential recurrent computation, forecast horizon, dataset size, batching, and the implementation of each architecture. Equal caps on trials and epochs therefore do not imply equal computation. Likewise, the gate learning-rate policy changes optimization dynamics but does not, by itself, add arithmetic operations to a forward or backward pass.

Inference-time scope.

The protocol measures the mean latency of a forward pass on the first test batch in evaluation mode, without gradients, after three warm-up passes and ten timed repetitions, with synchronization on accelerators when required. The batch contains at most 256 windows, so this quantity is a first-batch latency, not a per-sequence latency and not an end-to-end measure including data loading. However, the consolidated results available for this comparison do not provide a complete latency table accompanied by the effective batch size, hardware, and software versions for every model and dataset. We therefore do not infer an inference-speed ranking from the reported tuning times. Such a ranking would require the missing environment-qualified latency measurements and, ideally, peak-memory measurements under a common execution stack.

Performance–cost interpretation.

The empirical trade-off is dataset- and regime-dependent. CAF combines a small parameter footprint with competitive or favorable MSE in several settings, including the internal future-context comparisons on Solar, Bike, and EPF-NP and the external

comparisons on METR-LA. Nevertheless, lightweight external models can reach lower tuning times and, in some historical or EPF-NP external comparisons, lower MSE. Conversely, a larger parameter count does not guarantee a higher practical cost or better accuracy: feed-forward references can train quickly despite having substantially more parameters, while PRformer is slower than CAF on Solar but faster on Bike and EPF-NP. The experiments therefore support a qualified conclusion: CAF is parametrically compact, but its recurrent training and model-selection procedure can be computationally expensive. This is compatible with the theoretical gate-overhead hierarchy, yet it does not establish computational superiority. Realized efficiency must be assessed from joint performance, tuning time, latency, memory, and hardware-qualified measurements rather than from parameter count alone.

13.10 Theory-Guided Empirical Synthesis

The theoretical analysis and the experimental evaluation address different levels of the CAF framework. The analytical results characterize the pure CAF operator and its structured variants under explicit assumptions; the experiments evaluate finite recurrent forecasters trained with fixed datasets, optimization budgets, and information contracts. Their role is therefore not to prove the theoretical statements a second time, but to determine whether the mechanisms permitted by the theory produce identifiable empirical signatures, where those signatures disappear, and which qualifications are required when moving from an operator-level result to a complete forecasting system. Table 18 summarizes this correspondence. Unless stated otherwise, "better", "worse", and "favored" refer to denormalized test MSE, with inferential decisions taken from the prespecified primary Holm family. A non-significant result means that no difference was detected under this design; it does not establish equivalence. The underlying means, standard deviations, and Holm-adjusted p-values are reported in Tables 6–12, 10, 7, and 16–17.

Multiple modes of contextual utility.

The experiments support a broader interpretation of temporal CAF than one restricted to future covariates. In the `history` regime, the decoder receives no future context, but the encoder can still use observed historical descriptors and the gate remains conditioned on the current pre-activation and the endogenous summary. The favorable historical C0–AC1 contrast on Solar and the persistence-matched advantage over CrossLinear and TimeXer for the evaluated METR-LA target sensor show that a temporal CAF can be useful without future covariates in the evaluated configurations (Tables 8 and 7). These findings are compatible with the temporal decomposition in Definition 3; they do not identify whether historical exogenous descriptors, self-gating, the temporal summary, or their interaction is the exclusive source of a predictive gain.

Known future context provides an additional mode of utility rather than a necessary condition for CAF operation. In the internal suite, CAF outperforms both LSTM and FiLM-RNN on Solar, Bike, and EPF-NP under the `known_future` contract, while no difference is detected on METR-LA (Table 6). The exogenous-channel ablations sharpen this pattern: removing $W_g^{(s)} s_t$ is strongly detrimental on Bike and EPF-NP only when future context is admitted, is detrimental on Solar in both regimes, and has

Table 18: Theory-guided synthesis of the experimental evidence. The last column states the strongest conclusion supported by the reported experiments; it also marks the boundary beyond which a direct diagnostic or a different experimental design would be required.

Theoretical principle or mechanism	Experimental observation	Supported level of conclusion
Explicit exogenous sensitivity through $W_g^{(s)} s_t$ (Definition 3 and Proposition 5)	Removing the exogenous gate route degrades C0 on Solar in both regimes and on Bike and EPF-NP under `known_future`; no difference is detected on METR-LA. In `history`, the Solar contrast isolates a contribution from observed past descriptors because future context is neutralized (Table 8).	Empirical support for useful exogenous routing in specific dataset–regime configurations; neither universal contextual utility nor a causal attribution to the local derivative pathway is established.
One-hidden-layer networks built from context-only pure-CAF units remain universal on compact product domains under the assumptions of Theorem 21	The exogenous-only gate is often worse than the full gate, notably on Solar and Bike and on EPF-NP under `known_future`, while no difference is detected on METR-LA (Tables 8 and 9).	No contradiction arises: universal approximation concerns asymptotic function-class density, not finite-width statistical, optimization, or parameter efficiency.
A full $W_g^{(a)}$ permits gate-induced cross-unit coupling (Proposition 5)	The diagonal gate improves historical Bike, is worse on Bike and EPF-NP with known future context, and is not distinguished from the full gate in the other reported settings after the primary correction (Table 9).	Cross-unit coupling behaves as a task- and regime-dependent inductive bias, not as a uniformly beneficial structural property. Forecast errors do not directly measure the learned off-diagonal interactions.
The endogenous summary u_t provides an additional temporal gate pathway (Definition 3)	Removing u_t produces a detected degradation only on Bike under `known_future` in the prespecified comparison family; no uniform effect is detected elsewhere (Table 9).	The dedicated summary can be useful beyond ordinary recurrent state propagation, but its contribution is conditional and is not equivalent to a general "memory versus no memory" result.
Gate saturation creates a trade-off between proximity to the base activation and gate trainability (Remark 8)	Compensated, orthogonal, and standard initializations change ranking across Bike, EPF-NP, and METR-LA, while several other contrasts are not detected (Tables 11 and 12).	Initialization is empirically consequential in selected configurations, but prediction scores alone do not establish saturation, improved gradient flow, or better conditioning as the cause.
A state-preserving path changes the recurrent dynamics and lies outside pure CAF	The complete state-preserving extension improves historical MSE on Bike and EPF-NP in the primary reference-based family, but no uniform advantage is observed; the Bike result is sensitive to the multiplicity family (Table 12).	State preservation is a promising separate extension. Its results neither validate nor invalidate claims established specifically for the pure CAF operator.
A full affine gate adds $O(H^2 + HS)$ computation, with lower costs for diagonal and context-only structures (Section 9)	Parameter counts follow the predicted structural ordering, and CAF is compact relative to the local LSTM and FiLM-RNN references. Nevertheless, recurrent CAF is often slower	Parameter growth is consistent with the theoretical hierarchy, whereas wall-clock time remains

no detected effect on METR-LA (Table 8). For Solar, the future descriptors are derived from a reconstructed temporal index and should not be interpreted as an authenticated civil calendar. The empirical picture is therefore an interaction between contextual information content, horizon, temporal dynamics, and model structure rather than a global benefit of supplying more covariates. Since the two regimes contain separately trained models under different information contracts, these cross-regime patterns are descriptive and not paired causal effects.

Structure is an inductive bias, not an empirical ordering.

The ablation results do not induce a universal hierarchy from exogenous-only to diagonal to full gates. Although the full gate has the richest instantaneous coupling structure, the diagonal restriction improves historical Bike, becomes worse on Bike and EPF-NP with future context, and remains statistically undistinguished from the full gate in several other settings. Similarly, the temporal summary yields a detected benefit only in one primary comparison (Table 9). All detected Bike ablation effects reported here refer to the prespecified comparisons against C0; they are not retained under the broader all-pairs Holm family. These outcomes are consistent with the theoretical analysis: the Jacobian establishes which sensitivity and coupling terms are available, but it does not claim that every task requires them. Their practical role is that of selectable inductive biases whose approximation benefit must compensate for additional parameters, estimation variance, and optimization burden.

The exogenous-only restriction gives an especially important boundary case. Its frequent finite-sample degradation does not weaken the context-only universal approximation result. The theorem states that sufficiently large networks from that restricted family are dense on compact product domains; it does not provide a finite-width efficiency theorem, an optimization guarantee, or a sample-size bound for the recurrent forecasting architecture used here. Conversely, a favorable result for the full gate does not show that the gate-induced Jacobian term improved gradient propagation. Such a claim would require direct measurements of gate values, singular values, gradient norms, or targeted perturbations.

Initialization and extensions delimit the scope of pure CAF.

The compensated construction supplies an analytically controlled starting point, but the trained-model results confirm that this identity does not determine the optimization outcome. Orthogonal initialization is favorable on historical Bike but unfavorable in selected persistence-free EPF-NP and METR-LA comparisons; standard initialization improves historical METR-LA while degrading selected Bike configurations (Tables 11 and 12). This heterogeneity is compatible with the theoretical initialization trade-off, but it cannot be attributed causally to saturation or conditioning from forecast errors alone.

The state-preserving extension likewise illustrates the importance of keeping operator classes separate. Its favorable historical results on Bike and EPF-NP motivate further study, yet the added carry gate changes the update rule, parameter count, initialization, and derivative paths (Table 12). It is therefore evidence for a CAF-based recurrent extension, not stronger evidence for theorems proved for pure CAF.

An initialization-matched control and direct recurrent-stability diagnostics would be needed to isolate the value of state preservation.

Predictive position and computational trade-off.

The comparisons with complete forecasting systems reject both universal superiority and practical irrelevance. CAF is competitive with local learned references in several historical settings and is favored internally on three of four datasets when future context is known (Table 6). In the persistence-matched external suite, however, specialized models outperform CAF on historical Bike and on EPF-NP in both regimes, whereas CAF is favored on METR-LA and in selected Solar and future-context Bike comparisons (Table 7). Future-context external comparisons concern protocol-adapted versions of CrossLinear, TimeXer, and PRformer; DAG is the native future-context reference, and the external roster is not identical across datasets, notably because PRformer is disabled by configuration on METR-LA. These countervailing results show that the activation mechanism cannot be ranked independently of persistence, architecture, information contract, dataset dynamics, and the implemented external adapters.

The efficiency results lead to the same qualified conclusion. The canonical CAF uses substantially fewer parameters than the local LSTM and FiLM-RNN references, and its structural ablations reproduce the expected full–diagonal–context-only parameter hierarchy (Table 16). Yet a small parameter count does not imply a short tuning time: recurrent execution, the number of hyperparameter trials, early stopping, and implementation details make CAF markedly slower than CrossLinear or TimeXer on several datasets (Table 17). The asymptotic analysis and the measured costs are therefore consistent only at the level of structural growth. Hardware-qualified latency and memory measurements are still required for claims about realized inference efficiency.

Overall empirical conclusion.

Taken together, the experiments identify CAF as an explicit contextual modulation mechanism with several empirically viable modes of operation. Its utility is not confined to future covariates: historical exogenous descriptors, current pre-activations, and endogenous temporal information can all feed the gate. The clearest favorable signatures include the historical exogenous-channel effect on Solar, the future-context advantages in the local comparisons on Solar, Bike, and EPF-NP, and the external MSE advantages for the evaluated METR-LA target sensor. The Solar future-context evidence remains conditional on descriptors derived from a reconstructed temporal index rather than an authenticated civil calendar. The adverse external results on historical Solar and Bike and on EPF-NP demonstrate, in parallel, that these mechanisms do not guarantee predictive dominance.

The resulting conclusion is conditional but substantive: CAF provides a parametrically compact and theoretically interpretable route to context-dependent nonlinear computation whose practical value depends on the information carried by the context, the temporal system, gate structure, initialization, persistence convention, optimization budget, and comparison adapter. The experiments document where the mechanisms are favorable, neutral, or unfavorable in the reported finite forecasters.

They do not prove universal approximation, identify the Jacobian pathway responsible for a gain, certify recurrent stability, or validate the profiled exogenous-selection criterion. Those claims remain within their respective analytical assumptions or require dedicated prospective diagnostics.

13.11 Experimental Limitations

The experiments provide a controlled evaluation of CAF across distinct temporal systems, but their scope remains narrower than the theoretical scope of the framework. The following limitations delimit the empirical conclusions.

Dataset, target, and partition scope.

The evaluation covers four datasets, one target per dataset, and one principal chronological train–validation–test partition. The datasets represent different temporal systems, but they do not constitute a population from which universal predictive conclusions can be drawn. Likewise, repeated random seeds quantify variability due to initialization and stochastic optimization on fixed partitions; they do not measure variability across alternative temporal splits, targets, sites, years, or data-generating systems. Because forecast windows overlap, the window–horizon pairs are not treated as independent experimental replicates.

The forecasting contracts are also heterogeneous. Solar uses a history of 144 steps and a horizon of 24 steps; Bike and EPF-NP use histories of 168 steps and horizons of 24 steps; METR-LA uses 12 steps for both history and horizon. These settings are appropriate to the respective datasets but confound the effects of domain, sampling frequency, history length, and horizon in any cross-dataset interpretation. Absolute error levels are therefore not directly comparable across datasets, and the apparent value of future context cannot be attributed to horizon alone.

Context provenance and information regimes.

The protocol and the archived benchmark metadata report context derived from observed timestamps for Bike, EPF-NP, and METR-LA, whereas Solar uses a reconstructed temporal index based on an assumed origin. For METR-LA, the archived metadata specify `calendar_source=‘‘real timestamps from source’’`; this is therefore the provenance retained for the reported campaign. The loader also supports inputs without timestamps by constructing a synthetic fallback calendar, but that fallback is not the source recorded for the archived METR-LA results. The Solar calendar must be interpreted as reconstructed temporal structure, not as an authenticated civil calendar. This limits claims about operationally available future information on that dataset. More generally, the `history` and `known_future` regimes are separate experiments with different information contracts and separately trained models. Their numerical differences describe regime-dependent behavior but are not paired causal estimates of the value of future context.

The reported experiments examine contexts that are fully supplied according to the declared regime. They do not systematically test noisy, delayed, shifted, partially missing, or incorrectly specified context. Consequently, the results do not characterize robustness to the context failures that may occur in a deployed forecasting system.

Optimization-budget heterogeneity.

The number of random seeds, training batch size, and learning-rate grid are not uniform across all datasets and suites. Solar uses different seed counts and learning rates for its main and ablation/appendix families. Bike uses more seeds for the external `known_future` suite than for its other suites. Bike and EPF-NP evaluate three learning-rate candidates, whereas the principal Solar and METR-LA suites use one. Entries trained in different suites are therefore not interchangeable evaluations of one fitted model, even when they share an architectural label. Equal maximum epoch and patience settings also do not imply equal realized optimization, convergence, or total computation.

The CAF gate uses a declared optimization policy that differs from the base learning-rate policy. This is part of the evaluated training procedure, but it prevents the comparison from isolating architecture alone under strictly identical optimization dynamics. The experiments establish performance for the complete model–optimizer combinations rather than an optimizer-independent ranking of activation mechanisms.

Statistical resolution and multiplicity.

Primary inference is based on paired seed-wise differences within a fixed dataset, regime, suite, and prespecified reference-based family. The sample sizes of ten or fifteen seeds provide useful information about training variability but limited power to distinguish small effects or establish equivalence. A non-significant contrast therefore means that no difference was detected under the design; it does not show that two models are practically or statistically equivalent.

Some ablation conclusions are sensitive to the definition of the multiplicity family. In particular, several Bike contrasts that are significant in the prespecified comparisons against C0 are no longer significant when every model pair is included in a larger Holm family; the Solar additive-routing result is similarly sensitive. The primary question-oriented tests remain valid for their declared family, but these changes reduce the robustness of fine-grained mechanistic claims. Moreover, correction is performed within each declared family rather than globally across all datasets, regimes, and tables.

Comparator scope and adaptation.

The external roster is not identical across datasets: notably, PRformer is not included for METR-LA, and DAG is evaluated only when future exogenous information is available. In the `known_future` regime, CrossLinear, TimeXer, and PRformer receive protocol-specific future-context branches rather than native components of their original architectures; DAG is the external reference with native future-context handling in this benchmark. The comparison therefore concerns complete adapted systems under the stated information contract, not always the original architectures in canonical form. Differences in roster and adaptation prevent simple cross-dataset win counts from defining a universal ranking. For METR-LA, all reported claims concern only the selected `sensor_000` target among the dataset's 207 sensors; they therefore apply to this sensor, not to METR-LA as a whole.

Persistence is controlled by separating local and external suites, but it is not isolated in a common factorial experiment. The persistence-free CAF variants change both the autoregressive initialization and forecast assembly, and their runs belong to a different comparison suite. Numerical differences between CAF with and without persistence must therefore not be interpreted as the causal effect of persistence.

Finite architectural coverage.

The study evaluates one principal hidden width, one temporal-summary window, and a fixed set of pure variants and extensions. It does not map scaling behavior as a function of width, depth, context dimension, summary length, or forecast horizon. The context-only universal-approximation theorem consequently cannot be assessed through these finite-width experiments, and the observed performance of one structured gate does not imply a general width-efficiency result.

Several comparisons also modify more than one factor. The exogenous-only gate removes both the pre-activation and temporal-summary gate pathways. The state-preserving extension changes the update rule, initialization, parameter count, and derivative paths. Additive routing introduces additional projection parameters. These comparisons are informative about complete variants, but they do not isolate every component under parameter- and initialization-matched conditions.

Absence of direct mechanism diagnostics.

The evaluation reports predictive errors and structural ablations but not the learned distribution of gate values, gate entropy, saturation frequency, context-conditioned variability, gradient norms, or singular values of the activation Jacobian. Predictive improvement alone therefore cannot establish that CAF avoided saturation, improved gradient propagation, changed local conditioning, or used a particular coupling pathway. Likewise, an ablation can show that a route is useful to the evaluated system without identifying the causal mechanism through which the route contributes.

The optional future-context permutation control and the initialization-matched twin control with the exogenous gate channel removed were not part of the base reported experiments. Their absence limits the ability to distinguish genuine use of contextual content from sensitivity to architecture, optimization, or context-correlated nuisance structure.

Self-contained recurrent qualifications.

The recurrent compensation, deterministic-rollout induction, limits of the compensated embedding, and gate-aware diagnostic qualifications required by this experimental section are stated locally above rather than delegated to labels absent from the accompanying theoretical manuscript. The compensated identity assumes matched initial states; the zero initial state used in the benchmark satisfies that requirement. It remains an initialization identity under matched deterministic execution, not a statement of equal stochastic training dynamics or global recurrent stability.

Local diagnostics versus recurrent stability.

The theoretical Jacobian and contraction results concern local operators under their stated assumptions. They do not certify the complete recurrent forecaster, which additionally contains the temporal summary, encoder–decoder state transfer, autoregressive feedback, and, for the state-preserving extension, an additional carry gate. No invariant state–context domain or global bound for all these dependency paths was established empirically. The reported experiments therefore provide no proof of global recurrent stability, robustness to long rollouts, or controlled error accumulation.

Unvalidated exogenous-selection criterion.

The benchmark does not implement the full profiled $+s/-s$ selection procedure of Section 12. In particular, it does not estimate the approximation gain, shrinkage bias, estimation variance, effective dimension, conditioning term, Taylor remainder, or profile-alignment error required for transfer from the local linearized surrogate to an actually re-optimized nonlinear model. The heterogeneous C0–AC1 results motivate the selection problem, but they do not validate the theoretical decision rule.

Computational measurements.

Parameter counts and cumulative tuning times are available, but they are not interchangeable measures of efficiency. Tuning time depends on the number of trials, early stopping, recurrent execution, batching, implementation, hardware, and software. The available consolidated results do not provide a complete hardware-qualified table of first-batch latency and peak memory for every model and dataset. Accordingly, the study supports statements about parameter compactness and protocol-specific tuning cost, not universal claims of inference speed, energy efficiency, or computational superiority.

Validation of the reported numerical results.

The archived benchmark result records, with the METR-LA target correction incorporated in their final consolidated version, are the authoritative numerical sources for this section. The principal means, standard deviations, reported p-values, comparison directions, parameter counts, and dataset-specific configurations in this protocol were checked against those records and are concordant at the displayed precision. These records also document their checks against the campaign summaries, seed-level results, Wilcoxon test results, run manifests, and completion records. For METR-LA, an additional deterministic persistence reconstruction identifies source column 0 (`sensor_000`) as the target used by the campaign. Accordingly, the numerical statements in this section are treated as validated results of the reported runs.

Taken together, these limitations do not invalidate the reported comparisons; they define their proper level of generality. The experiments characterize the behavior of the evaluated forecasters under fixed data, information, and optimization contracts. Broader claims require prospective designs that vary these contracts and directly measure the mechanisms invoked by the theory.

13.12 Research Perspectives

The preceding limitations suggest a research program organized around seven priorities: prospective validation of member selection, direct observation of the gate, causal architectural controls, broader temporal robustness, practical efficiency, a separate theory of state-preserving extensions, and carefully scoped extensions beyond the present benchmark.

Prospective validation of exogenous-channel selection.

The first priority is to evaluate the local $+s/-s$ criterion as a decision rule rather than retrospectively relate it to ablation outcomes. On training-only or cross-fitted folds, future work should estimate $\widehat{\mathscr{I}_s}$, $\widehat{\mathscr{B}}_s(\lambda)$, $\widehat{\mathscr{V}}_s(\lambda)$, the conditioning diagnostic $\widehat{\kappa}_\lambda$, the effective dimension $\widehat{d}_{2,\lambda}$, and the held-out Taylor error. It should also quantify the profile-alignment error required to connect the target-profiled linearized surrogate to the actually re-optimized nonlinear network. The resulting choice between the $+s$ and $-s$ members must then be frozen and evaluated on untouched temporal folds. Success should be measured not only by prediction error but also by selection accuracy, excess risk relative to an oracle choice, and calibration of the predicted risk margin across datasets and horizons.

Direct characterization of learned gates.

A second priority is to connect predictive behavior with direct mechanism measurements. For every dataset, regime, recurrent phase, and forecast horizon, future experiments should report the distribution of g_t, the proportion of coordinates near zero or one, normalized gate entropy, variation within and between contexts, and sensitivity to controlled perturbations of s_t. These measurements should be complemented by norms and singular values of the local Jacobian, gradient norms for the gate and base pathways, and temporal profiles of these quantities during training and autoregressive rollout.

Context-permutation, time-shift, masking, and replacement controls would help determine whether a model uses the content and temporal alignment of context rather than merely responding to the presence of an additional input channel. Such diagnostics would not by themselves prove causal importance, but, combined with targeted interventions and ablations, they would provide a substantially stronger link between the theoretical sensitivity formulas and the observed forecasting behavior.

Factorial and parameter-matched architectural studies.

The effects of initialization, persistence, routing, and state preservation should be separated in a common factorial design. Compensated and standard initializations could be compared from more tightly matched shared weights; orthogonalization could be crossed with both schemes; persistence could be switched within the same suite, seeds, and hyperparameter grid; and the state-preserving extension could be compared with a pure CAF control under a matched initialization and parameter budget. Additive, multiplicative, FiLM-style, and concatenation-based context injection should likewise be tested under approximately matched capacity and optimization budgets.

For gate structure, experiments should vary full, diagonal, context-only, grouped, and low-rank forms over several widths and context dimensions. This would determine when the $O(H^2)$ coupling block provides enough predictive benefit to justify its additional estimation and computational cost. Composite restrictions should be decomposed into single-factor interventions wherever possible, particularly by separating removal of pre-activation dependence from removal of the temporal summary.

Broader temporal and contextual robustness.

Generalization should be evaluated over multiple rolling-origin splits, calendar periods, targets, sites, and datasets rather than only repeated optimization on one partition. Each dataset should be studied at several history lengths and forecast horizons, with multiple hidden widths and temporal-summary windows. Multivariate and multi-target forecasting would test whether CAF's unit-level modulation remains useful when cross-series dependence becomes central.

The information contract should also be stress-tested. Relevant conditions include noisy context, missing entries, delayed or shifted descriptors, partially known future trajectories, covariate-distribution shift, and context whose reliability changes over time. Comparisons between authentic scheduled signals and reconstructed calendar proxies would clarify how context provenance affects utility. These experiments could identify not only when exogenous context helps, but when the model should attenuate, ignore, or distrust it.

Uniform performance–cost evaluation and implementation efficiency.

The theoretical costs in Section 9 motivate structured and systems-level optimization. Diagonal, grouped, low-rank, and context-only gates should be evaluated as explicit performance–cost alternatives. Parallelizable temporal summaries, cached context projections, and fusion of the gate pre-activation, sigmoid, base activation, and elementwise product may reduce kernel-launch and memory overhead without changing the operator definition.

Future benchmarks should report, under a common hardware and software stack, training time per epoch, total selection cost, inference latency per batch and per sequence, throughput, peak memory, and, where possible, energy consumption. Measurements should state the effective batch size, precision, warm-up protocol, repetition count, and synchronization policy. Performance–cost frontiers across models and gate structures would be more informative than rankings by parameter count or tuning time alone.

A separate theory and evaluation for state-preserving CAF.

The favorable historical results of the state-preserving extension on Bike and EPF-NP justify dedicated analysis, but this extension should remain distinct from pure CAF. Its recurrent Jacobian must include derivatives of both the CAF response and the carry gate. Future theory should establish sufficient conditions for contraction or

bounded sensitivity, characterize the trade-off between state retention and responsiveness, and determine how the additional gate interacts with the temporal summary and autoregressive feedback.

Empirically, long-horizon perturbation tests, memory-retention tasks, controlled regime changes, and initialization-matched comparisons could determine whether the extension improves state preservation rather than merely altering optimization. These experiments should include direct measurements of carry-gate values and rollout sensitivity, not only final forecast error.

Extending the CAF framework beyond the present benchmark.

Further directions include parameterized base responses, hierarchical or attention-based contexts, stochastic gates, sparse conditional execution, and continuous-time formulations. Each extension changes the relevant function class, derivative paths, or realized computation and therefore requires its own theoretical assumptions and controlled evaluation. In particular, thresholding a gate does not by itself reduce realized operations: efficiency claims for sparse CAF require execution kernels that actually skip inactive units or blocks.

The immediate empirical priority is not to enlarge the architecture catalog, but to test the existing theoretical claims with mechanism-level diagnostics and prospective decisions. A successful next stage would show, on untouched temporal folds, when exogenous routing should be selected, how the learned gate uses context, and whether structured reductions preserve predictive performance while improving measured efficiency. This progression would turn the current conditional evidence into a more complete account of when contextual activation is statistically useful, computationally justified, and dynamically reliable.

14 Conclusion, Future Work, and Limitations

14.1 Conclusion

We have introduced *contextual activation functions* (CAF) not as a single model but as a *family* of activation maps $\Phi(a, s) = g(a, s) \odot \rho(a)$ that place context-dependent modulation at the level of the activation map itself, rather than in feature space, at the parameter level, or across routed subnetworks. Within this framework we established a corpus of analytic invariants for the canonical pure CAF operator: C^r smoothness, boundedness, and global Lipschitz continuity; an explicit Jacobian decomposition into a gated diagonal term and a gate-induced coupling term, with a local Jacobian-norm bound under affine gating; a characterization of gradient attenuation under gate saturation; and universal approximation for one-hidden-layer networks, holding already for the context-only subfamily in which the base pre-activation depends only on x and the gate only on s. At activation-map level, the precise inclusion statement is that, on every compact product domain $K_a \times K_s$, the uniform closure of the affine-gate pure-CAF family contains the standard activation map and also contains genuinely context-dependent maps; this does not imply a width advantage or universal network-class superiority. These results relate standard activations and structured modulation

through limiting and special cases (Proposition 1), and they quantify the parametric and computational overhead of each structured gate (Section 9).

The experimental evaluation in Section 13 complements these operator-level results with finite recurrent forecasts on four datasets. Under the declared information, persistence, and optimization contracts, CAF is competitive in selected settings and unfavorable or statistically undistinguished in others; the state-preserving extension is also evaluated as a separate recurrent architecture. These findings support the practical viability of the CAF forecaster without establishing universal predictive superiority or validating the analytic mechanisms causally.

As a theoretical application, we derived an exact local risk identity for a target-profiled linearized surrogate comparing the $+s$ and $-s$ members. Under the stated assumptions, its risk gap is

$$-\mathscr{I}_s + \mathscr{B}_s(\lambda) + \mathscr{V}_s(\lambda).$$

The identity is exact for the profiled linear surrogate. Transfer to an actually re-optimized nonlinear network requires the additional profile-alignment assumption and the local Taylor-remainder control described above. The criterion is operational only when its fold-specific diagnostics are estimated independently, the fitted correction remains in the local region, the profile-alignment error is controlled when actual refitting is considered, and the risk margin dominates the combined approximation and profiling remainder. It should therefore be viewed as a principled, testable selection rule rather than as a universal preference for any CAF member. The CAF forecaster itself has been evaluated empirically in Section 13; what remains prospectively unvalidated is this profiled selection criterion and its transfer to a re-optimized nonlinear network, not the predictive forecaster as such.

14.2 Future Work

Several directions follow naturally from the framework.

Reducing implementation overhead.

The asymptotic analysis in Section 9 identifies the additional gate cost but does not determine wall-clock latency on any particular hardware or software stack. Future work should study parallelizable temporal-gate formulations, kernel fusion of the gate pre-activation, sigmoid, and base activation, structured gates with lower asymptotic cost, and lighter temporal summaries u_t. Such engineering work should report hardware, batch size, warm-up, repetition counts, and memory measurements before making latency or training-time claims.

Prospective validation of the selection criterion.

Theorem 25 gives an exact identity for the target-profiled linearized surrogate. Applying the resulting rule to an actually re-optimized nonlinear network requires estimating and validating the profile-alignment error in Assumption 4, in addition to the Taylor remainder and risk-margin diagnostics. Future work should estimate $\widehat{\mathscr{I}_s}$, $\widehat{\mathscr{B}}_s(\lambda)$,

$\widehat{\mathscr{V}}_s(\lambda)$, $\widehat{\kappa}_\lambda$, $\widehat{d}_{2,\lambda}$, and the held-out Taylor error by sample splitting or cross-fitting, then test the decision prospectively on untouched folds.

Vehicle of the exogenous context.

The present evaluation already compares multiplicative CAF routing with additive controls, a FiLM-RNN reference, and recurrent references that concatenate context at their inputs. These finite comparisons are informative but differ in capacity, initialization, recurrence, or optimization and therefore do not isolate the routing mechanism. Future work should extend them to systematic, capacity- and initialization-matched comparisons of multiplicative, additive, FiLM-style, and concatenation-based injection across strongly context-driven signals.

Prospective extensions.

Parameterized base responses, attention-based gates, hierarchical contexts, stochastic gates, sparse execution, and continuous-time formulations remain prospective directions summarized in Table 3. Each requires dedicated theoretical analysis and controlled experiments before performance, stability, or efficiency claims can be made.

14.3 Limitations

The limitations specific to the datasets, targets, temporal partitions, optimization budgets, comparator adaptations, statistical resolution, and absent mechanism diagnostics are detailed separately in Section 13.11 ("Experimental Limitations"). We state the remaining cross-cutting theoretical scope below.

- **Local criterion scope.** The selection identity is exact for the target-profiled linearized surrogate under the stated profiling, transportability, covariance, and regularization assumptions. Its transfer to the nonlinear or actually re-optimized network additionally requires the Taylor-remainder, local-ball, and profile-alignment conditions. It is not a universal law about exogenous context.
- **Profiled surrogate versus actual refitting.** The exact risk identity concerns a target-profiled linearized surrogate. A claim about the risk of an actually re-optimized nonlinear network additionally requires the profile-alignment assumption Assumption 4.
- **Local-ball condition.** The nonlinear transfer result is local and applies when the fitted exogenous correction remains within the radius δ. If this condition is verified only with high probability, the corresponding conclusion is conditional or high-probability rather than unconditional.
- **Asymptotic cost versus realized latency.** The complexity analysis quantifies parameter and operation growth, but it does not predict realized hardware latency or training time without implementation-specific measurements.
- **No width-efficiency or performance theorem.** The universality, closure, Jacobian, and conditioning results do not establish a width advantage, predictive superiority, or computational superiority over specialized architectures.

- **Prospective validation of the local criterion.** The quantities required by the local $+s$ versus $-s$ selection rule still require independent prospective estimation and validation.

15 Declarations

1. **Competing interests:** The sole author declares that he has no competing interests relevant to this manuscript.
2. **Funding: Not** applicable. **No** external funding was received for this work.
3. **Author contribution: Mohamed Ben Ammar Chraiti**: Conceptualization, Methodology, Formal analysis, Algorithm development, Writing—original draft, Writing—review & editing. The sole author approved the final version of the manuscript.
4. **Data Availability Statement:** No new dataset was generated. The experiments analyze four public datasets: Solar, available through the preprocessed Time-Series-Library archive (https://huggingface.co/datasets/thuml/Time-Series-Library); Bike Sharing, available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset) [Fanaee-T and Gama(2014)]; EPF-NP, distributed with the official TimeXer repository (https://github.com/thuml/TimeXer) and documented by its open benchmark [Lago et al.(2021)Lago, Marcjasz, Schutter, and Weron]; and METR-LA, available through the DCRNN repository's documented data links (https://github.com/liyaguang/DCRNN) [Li et al.(2018)Li, Yu, Shahabi, and Liu]. The exact chronological partitions, target definitions, and preprocessing used here are reported in Section 13.1.
5. **Research involving human and/or animal participants: Not** applicable.
6. **Informed consent: Not** applicable.

A Detailed Proof of Context-Only Universality

This appendix supplies the complete Stone–Weierstrass and error-control argument for Theorem 21. The theorem itself, including the defining separation $a_k(x) = w_k^\top x + b_k$ and $g_k(s) = \sigma(v_k^\top s + c_k)$, is stated in the main expressivity section.

Detailed proof of Theorem 21 We first prove the scalar case $m = 1$.

1. Density of finite sums of separable functions. Consider the set

$$\mathscr{A} := \left\{ \sum_{r=1}^{R} \varphi_r(x)\psi_r(s) : \ R \in \mathbb{N},\ \varphi_r \in C(K_x),\ \psi_r \in C(K_s) \right\} \subset C(K_x \times K_s).$$

Then $\mathscr{A}$ is a subalgebra of $C(K_x \times K_s)$, it contains the constant functions, and it separates points. Indeed, if $(x, s) \neq (x', s')$, then either $x \neq x'$ and some function in $C(K_x)$ separates x and x', or $s \neq s'$ and some function in $C(K_s)$ separates s and s'. Therefore, by the Stone–Weierstrass theorem, $\mathscr{A}$ is dense in $C(K_x \times K_s)$ for the uniform norm.

Hence, given $\varepsilon > 0$, there exist $R \in \mathbb{N}$ and functions

$$\varphi_1, \dots, \varphi_R \in C(K_x), \qquad \psi_1, \dots, \psi_R \in C(K_s),$$

such that

$$\sup_{(x,s)\in K_x\times K_s}\left|f(x,s)-\sum_{r=1}^{R}\varphi_r(x)\psi_r(s)\right|<\frac{\varepsilon}{2}. \tag{149}$$

2. Approximation of the x-dependent factors. Because ρ is continuous and non-polynomial, the classical universal approximation theorem implies that finite linear combinations of the functions $x\mapsto\rho(w^\top x+b)$ are dense in $C(K_x)$ [Hornik(1991), Leshno et al.(1993)Leshno, Lin, Pinkus, and Schocken, Pinkus(1999)]. Hence, for each $r\in\{1,\dots,R\}$ and each tolerance $\eta>0$, there exist $I_r\in\mathbb{N}$, coefficients $\alpha_{r,i}\in\mathbb{R}$, vectors $w_{r,i}\in\mathbb{R}^d$, and biases $b_{r,i}\in\mathbb{R}$ such that

$$\sup_{x\in K_x}\left|\varphi_r(x)-\sum_{i=1}^{I_r}\alpha_{r,i}\rho(w_{r,i}^\top x+b_{r,i})\right|<\eta. \tag{150}$$

Define

$$\widetilde{\varphi}_r(x):=\sum_{i=1}^{I_r}\alpha_{r,i}\rho(w_{r,i}^\top x+b_{r,i}).$$

3. Approximation of the s-dependent factors. Since σ is continuous and sigmoidal, finite linear combinations of the functions $s\mapsto\sigma(v^\top s+c)$ are dense in $C(K_s)$ [Cybenko(1989), Hornik(1991), Pinkus(1999)]. Hence, for each $r\in\{1,\dots,R\}$ and each tolerance $\eta>0$, there exist $J_r\in\mathbb{N}$, coefficients $\beta_{r,j}\in\mathbb{R}$, vectors $v_{r,j}\in\mathbb{R}^S$, and biases $c_{r,j}\in\mathbb{R}$ such that

$$\sup_{s\in K_s}\left|\psi_r(s)-\sum_{j=1}^{J_r}\beta_{r,j}\sigma(v_{r,j}^\top s+c_{r,j})\right|<\eta. \tag{151}$$

Define

$$\widetilde{\psi}_r(s):=\sum_{j=1}^{J_r}\beta_{r,j}\sigma(v_{r,j}^\top s+c_{r,j}).$$

Output-bias convention.

The representation in Equation (32) does not display a separate output bias. This causes no loss of generality. Since ρ is non-polynomial, it is not identically zero, so there exists $b_\rho\in\mathbb{R}$ such that $\rho(b_\rho)\neq 0$. Setting $w=0$ gives the constant function $x\mapsto\rho(b_\rho)$, and an arbitrary scalar multiple of this unit realizes any constant function of x. Similarly, the sigmoid branch contains the constant function $\sigma(0)=1/2$ by taking $v=0$ and $c=0$. Consequently, constant terms in the factor approximations can be represented within the stated CAF unit class.

4. Approximation of each separable term by CAF units. For each r, consider the product

$$\widetilde{\varphi}_r(x)\widetilde{\psi}_r(s)=\sum_{i=1}^{I_r}\sum_{j=1}^{J_r}\alpha_{r,i}\beta_{r,j}\,\rho(w_{r,i}^\top x+b_{r,i})\,\sigma(v_{r,j}^\top s+c_{r,j}).$$

Each term in this double sum is exactly of the CAF form

$$\sigma(v^\top s+c)\,\rho(w^\top x+b),$$

that is, a gate depending only on s multiplied by a base activation depending only on x. Therefore, finite linear combinations of units of the form (32) can realize the approximation

$$\sum_{r=1}^{R}\widetilde{\varphi}_r(x)\widetilde{\psi}_r(s).$$

5. Control of the approximation error. Define

$$F_\eta(x,s) := \sum_{r=1}^{R} \widetilde{\varphi}_r(x)\widetilde{\psi}_r(s).$$

Then

$$f(x,s) - F_\eta(x,s) = \underbrace{f(x,s) - \sum_{r=1}^{R} \varphi_r(x)\psi_r(s)}_{E_1(x,s)} + \underbrace{\sum_{r=1}^{R} \big(\varphi_r(x)\psi_r(s) - \widetilde{\varphi}_r(x)\widetilde{\psi}_r(s)\big)}_{E_2(x,s)}.$$

By (149),

$$\|E_1\|_\infty < \frac{\varepsilon}{2}.$$

For the second term, fix r. Then

$$\varphi_r\psi_r - \widetilde{\varphi}_r\widetilde{\psi}_r = \psi_r(\varphi_r - \widetilde{\varphi}_r) + \widetilde{\varphi}_r(\psi_r - \widetilde{\psi}_r),$$

hence

$$\|\varphi_r\psi_r - \widetilde{\varphi}_r\widetilde{\psi}_r\|_\infty \le \|\psi_r\|_\infty\|\varphi_r - \widetilde{\varphi}_r\|_\infty + \|\widetilde{\varphi}_r\|_\infty\|\psi_r - \widetilde{\psi}_r\|_\infty.$$

By (150),

$$\|\varphi_r - \widetilde{\varphi}_r\|_\infty < \eta, \qquad \|\widetilde{\varphi}_r\|_\infty \le \|\varphi_r\|_\infty + \eta.$$

By (151),

$$\|\psi_r - \widetilde{\psi}_r\|_\infty < \eta.$$

Therefore,

$$\|\varphi_r\psi_r - \widetilde{\varphi}_r\widetilde{\psi}_r\|_\infty \le \|\psi_r\|_\infty\eta + (\|\varphi_r\|_\infty + \eta)\eta.$$

Assume $\eta \le 1$. Then

$$\|\varphi_r\psi_r - \widetilde{\varphi}_r\widetilde{\psi}_r\|_\infty \le (\|\psi_r\|_\infty + \|\varphi_r\|_\infty + 1)\eta.$$

Hence

$$\|E_2\|_\infty \le \left(\sum_{r=1}^{R}(\|\psi_r\|_\infty + \|\varphi_r\|_\infty + 1)\right)\eta.$$

Since R is fixed and all φ_r, ψ_r are continuous on compact sets, the coefficient in parentheses is finite. Choosing $\eta > 0$ sufficiently small so that

$$\left(\sum_{r=1}^{R}(\|\psi_r\|_\infty + \|\varphi_r\|_\infty + 1)\right)\eta < \frac{\varepsilon}{2},$$

we obtain

$$\|E_2\|_\infty < \frac{\varepsilon}{2}.$$

Hence

$$\|f - F_\eta\|_\infty \le \|E_1\|_\infty + \|E_2\|_\infty < \varepsilon.$$

This proves the scalar case.

6. Vector-valued case. If $f = (f_1, \dots, f_m)$ takes values in $\mathbb{R}^m$, apply the scalar argument to each coordinate f_ℓ with tolerance $\varepsilon/\sqrt{m}$. Concatenating the resulting approximating subnetworks yields a CAF network f_{CAF} such that

$$\sup_{(x,s)\in K_x\times K_s} \|f(x,s) - f_{\mathrm{CAF}}(x,s)\| < \varepsilon.$$

This completes the proof. □

B Architecture-Specific Directions

CAF-style modulation can be instantiated in convolutional, Transformer, and graph neural architectures. The formulations below show how the CAF principle can be embedded in these settings and provide first-order cost expressions where the gate structure is explicit. They are prospective design templates rather than architectures evaluated or theoretically characterized in this article. In particular, no performance, stability, or realized-efficiency claim follows from these formulations alone.

B.1 Convolutional Networks

For a convolutional layer with C channels, spatial dimensions $H_{\text{img}} \times W_{\text{img}}$, and context dimension S, two gate sharing patterns illustrate the principal cost trade-off.

- **Spatially shared gate.** A single gate vector is computed per channel and shared across spatial locations. Under a dense channelwise affine gate, the gate computation has order
$$O\big(C(C+S)\big)$$
per layer, in addition to the underlying convolution.
- **Spatially varying gate.** A separate gate is evaluated at each spatial location. Under the same dense parameterization, the gate computation has order
$$O\big(H_{\text{img}} W_{\text{img}}\, C(C+S)\big).$$

These expressions count the dense affine gate operations and do not by themselves predict wall-clock latency. Pooled-context, grouped, diagonal, or low-rank gates provide intermediate designs whose realized cost depends on the implementation.

B.2 Transformer Architectures

In Transformer architectures [Vaswani et al.(2017)Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, an the most direct use places CAF in a positionwise feed-forward block, where the context may encode position, global sequence statistics, task identity, or an external control signal. More speculative variants may modulate quantities adjacent to the attention computation
$$\operatorname{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right),$$
for example by gating the post-attention values or the output of an attention sub-layer. Directly modifying attention scores defines a different operator and is not covered by the pure-CAF results proved for $g(a,s) \odot \rho(a)$. The overhead depends on sequence length and on whether the gate is global, token-specific, or head-specific; a dedicated analysis is therefore required for each construction.

B.3 Graph Neural Networks

For a graph neural network, let $h_u^{(\ell)} \in \mathbb{R}^H$ be the state of node u at layer ℓ, let $W^{(\ell)} \in \mathbb{R}^{H\times H}$, and define an aggregated pre-activation at node v by

$$a_v^{(\ell)} := \sum_{u\in\mathcal{N}(v)} W^{(\ell)} h_u^{(\ell)} \in \mathbb{R}^H.$$

Given graph-level context s_G, a nodewise pure-CAF update can be written as

$$h_v^{(\ell+1)} = g_v^{(\ell)}\big(a_v^{(\ell)}, s_G\big) \odot \rho\big(a_v^{(\ell)}\big), \qquad g_v^{(\ell)}\big(a_v^{(\ell)}, s_G\big) \in (0,1)^H.$$

This is an instance of the pure multiplicative rule at each node, provided that the aggregation defining $a_v^{(\ell)}$ is treated as the preceding affine or message-passing operation. Context may instead be node- or edge-specific, but those choices change the sharing structure and computational cost. Their stability, expressivity, and empirical behavior remain to be established.

References

[Agostinelli et al.(2015)Agostinelli, Hoffman, Sadowski, and Baldi] Agostinelli F, Hoffman M, Sadowski P, et al (2015) Learning activation functions to improve deep neural networks. Workshop contribution, ICLR 2015, arXiv:1412.6830

[Bengio et al.(1994)Bengio, Simard, and Frasconi] Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2):157–166. https://doi.org/10.1109/72.279181

[Chen et al.(2020)Chen, Dai, Liu, Chen, Yuan, and Liu] Chen Y, Dai X, Liu M, et al (2020) Dynamic ReLU. In: Computer Vision – ECCV 2020, pp 351–367, https://doi.org/10.1007/978-3-030-58529-7_21

[Cho et al.(2014)Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk, and Bengio] Cho K, van Merriënboer B, Gülçehre Ç, et al (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 1724–1734, https://doi.org/10.3115/v1/D14-1179

[Cover and Thomas(2006)] Cover TM, Thomas JA (2006) Elements of Information Theory, 2nd edn. Wiley-Interscience, Hoboken, NJ, https://doi.org/10.1002/047174882X

[Cybenko(1989)] Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2(4):303–314. https://doi.org/10.1007/BF02551274

[Dauphin et al.(2017)Dauphin, Fan, Auli, and Grangier] Dauphin YN, Fan A, Auli M, et al (2017) Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol 70. PMLR, pp 933–941

[Dumoulin et al.(2017)Dumoulin, Shlens, and Kudlur] Dumoulin V, Shlens J, Kudlur M (2017) A learned representation for artistic style. In: International Conference on Learning Representations (ICLR)

[Fanaee-T and Gama(2014)] Fanaee-T H, Gama J (2014) Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence 2(2–3):113–127. https://doi.org/10.1007/s13748-013-0040-3

[Glorot and Bengio(2010)] Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research, vol 9. PMLR, pp 249–256

[Ha et al.(2017)Ha, Dai, and Le] Ha D, Dai AM, Le QV (2017) HyperNetworks. In: International Conference on Learning Representations (ICLR)

[He et al.(2015)He, Zhang, Ren, and Sun] He K, Zhang X, Ren S, et al (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 1026–1034, https://doi.org/10.1109/ICCV.2015.123

[Hochreiter and Schmidhuber(1997)] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[Holm(1979)] Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2):65–70

[Hornik(1991)] Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Networks 4(2):251–257. https://doi.org/10.1016/0893-6080(91)90009-T

[Hu et al.(2018)Hu, Shen, and Sun] Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 7132–7141

[Huang and Belongie(2017)] Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision (ICCV). IEEE, pp 1501–1510

[Jacobs et al.(1991)Jacobs, Jordan, Nowlan, and Hinton] Jacobs RA, Jordan MI, Nowlan SJ, et al (1991) Adaptive mixtures of local experts. Neural Computation 3(1):79–87. https://doi.org/10.1162/neco.1991.3.1.79

[Jagtap et al.(2020a)Jagtap, Kawaguchi, and Karniadakis] Jagtap AD, Kawaguchi K, Karniadakis GE (2020a) Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. Journal of Computational Physics 404:109136. https://doi.org/10.1016/j.jcp.2019.109136

[Jagtap et al.(2020b)Jagtap, Kawaguchi, and Karniadakis] Jagtap AD, Kawaguchi K, Karniadakis GE (2020b) Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. Proceedings of the Royal Society A 476(2239):20200334. https://doi.org/10.1098/rspa.2020.0334

[Jayakumar et al.(2020)Jayakumar, Czarnecki, Menick, Schwarz, Rae, Osindero, Teh, Harley, and Pascanu] Jayakumar SM, Czarnecki WM, Menick J, et al (2020) Multiplicative interactions and where to find them. In: International Conference on Learning Representations (ICLR)

[Lago et al.(2021)Lago, Marcjasz, Schutter, and Weron] Lago J, Marcjasz G, Schutter BD, et al (2021) Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Applied Energy 293:116983. https://doi.org/10.1016/j.apenergy.2021.116983

[Leshno et al.(1993)Leshno, Lin, Pinkus, and Schocken] Leshno M, Lin VY, Pinkus A, et al (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6(6):861–867. https://doi.org/10.1016/S0893-6080(05)80131-5

[Li et al.(2018)Li, Yu, Shahabi, and Liu] Li Y, Yu R, Shahabi C, et al (2018) Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: International Conference on Learning Representations (ICLR), arXiv:1707.01926

[Ma et al.(2021)Ma, Zhang, Liu, and Sun] Ma N, Zhang X, Liu M, et al (2021) Activate or not: Learning customized activation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8032–8042, https://doi.org/10.1109/CVPR46437.2021.00794

[Oreshkin et al.(2020)Oreshkin, Carpov, Chapados, and Bengio] Oreshkin BN, Carpov D, Chapados N, et al (2020) N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In: International Conference on Learning Representations

[Pascanu et al.(2013)Pascanu, Mikolov, and Bengio] Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol 28. PMLR, pp 1310–1318

[Perez et al.(2018)Perez, Strub, de Vries, Dumoulin, and Courville] Perez E, Strub F, de Vries H, et al (2018) FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI). AAAI Press, pp 3942–3951

[Pinkus(1999)] Pinkus A (1999) Approximation theory of the MLP model in neural networks. Acta Numerica 8:143–195. https://doi.org/10.1017/S0962492900002919

[Qiu et al.(2025)Qiu, Zhu, Li, Wu, Yang, and Hu] Qiu X, Zhu Y, Li Z, et al (2025) DAG: A dual correlation network for time series forecasting with exogenous variables. Current arXiv title; align with the pinned implementation version, arXiv:2509.14933

[Ramachandran et al.(2017)Ramachandran, Zoph, and Le] Ramachandran P, Zoph B, Le QV (2017) Searching for activation functions. arXiv:1710.05941

[Salinas et al.(2020)Salinas, Flunkert, Gasthaus, and Januschowski] Salinas D, Flunkert V, Gasthaus J, et al (2020) DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting 36(3):1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

[Shazeer et al.(2017)Shazeer, Mirhoseini, Maziarz, Davis, Le, Hinton, and Dean] Shazeer N, Mirhoseini A, Maziarz K, et al (2017) Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR)

[Vaswani et al.(2017)Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin]

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS). Curran Associates, Inc., pp 5998–6008

[Wang et al.(2024)Wang, Wu, Dong, Qin, Zhang, Liu, Qiu, Wang, and Long] Wang Y, Wu H, Dong J, et al (2024) TimeXer: Empowering transformers for time series forecasting with exogenous variables. In: Advances in Neural Information Processing Systems

[Wu et al.(2016)Wu, Zhang, Zhang, Bengio, and Salakhutdinov] Wu Y, Zhang S, Zhang Y, et al (2016) On multiplicative integration with recurrent neural networks. In: Advances in Neural Information Processing Systems 29 (NIPS). Curran Associates, Inc., pp 2856–2864

[Yu et al.(2025)Yu, Yu, Nie, Miao, Liu, and Li] Yu Y, Yu W, Nie F, et al (2025) PRformer: Pyramidal recurrent transformer for multivariate time series forecasting. Neural Networks 191:107769. https://doi.org/10.1016/j.neunet.2025.107769

[Zhou et al.(2021)Zhou, Zhang, Peng, Zhang, Li, Xiong, and Zhang] Zhou H, Zhang S, Peng J, et al (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11106–11115, https://doi.org/10.1609/AAAI.V35I12.17325

[Zhou et al.(2025)Zhou, Liu, Liang, Song, and Li] Zhou P, Liu Y, Liang J, et al (2025) CrossLinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp 4120–4131, https://doi.org/10.1145/3711896.3736899, arXiv:2505.23116