SAME LOSS, DIFFERENT GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentiable learning typically assumes that the scalar objective evaluated in the forward pass and the gradient supplied to the optimizer in the backward pass describe the same mathematical object. We show that this correspondence can fail when probabilistic objectives rely on finite special-function recurrences, custom backward rules, and numerical clipping. In high-dimensional von Mises–Fisher learning, real numerical implementations can produce identical forward scores and losses at the same learning state while supplying different gradients and following different optimization trajectories. We characterize the structure of this mismatch in finite-start Bessel recurrence and show that classwise radial mismatch can compose through probabilities into a locally nonconservative update field. Evaluating the accuracy of special-function values and derivatives separately is therefore insufficient to characterize the realized learning objective. Motivated by this observation, we introduce AR/FR, a fixed-depth analytic realization that constructs a potential and its derivative jointly, ensuring forward–backward coherence by construction. We establish a uniform cubic-order error bound relative to the exact Bessel ratio over the entire nonnegative concentration axis and propagate this guarantee to learning scores and objectives. As representation dimension increases, the original finite recurrence becomes sequentially deeper, whereas the worst-case AR/FR error guarantee tightens cubically, jointly providing coherence, certified fidelity, and fixed-depth computation. These results suggest that a differentiable numerical primitive is defined by both the values it realizes and the derivatives it actually supplies to the optimizer; together, they constitute the numerical realization of the learning algorithm.

1 INTRODUCTION

In differentiable machine learning, a scalar objective serves two roles: the forward pass evaluates its value, and the backward pass supplies its derivative to the optimizer to determine the next update. We usually assume that these processes describe the same mathematical object. If the program realizes a numerical objective $\widetilde{L}$, its backward should supply $\nabla\widetilde{L}$. In ordinary automatic differentiation, this correspondence is typically guaranteed by the computation graph (Baydin et al., 2018). When the objective contains high-order special functions, finite recurrences, clipping, or a custom backward, however, that guarantee can fail. Two implementations may then return identical forward scores and losses while supplying different gradients and ultimately following different training trajectories. **The same forward loss need not imply the same learning process.**[1]

This issue is particularly relevant to probabilistic learning, where objectives often depend on normalization constants and special functions that are difficult to evaluate efficiently. High-dimensional von Mises–Fisher (vMF) learning is a representative example (Scott et al., 2021; Sra, 2012): its scalar potential contains a high-order modified Bessel function, and its analytic derivative is the adjacent-order Bessel ratio

$$R_\nu(x) = \frac{I_{\nu+1}(x)}{I_\nu(x)}.$$

The Bessel order grows with representation dimension, motivating finite recurrences, ratio approximations, and custom backward rules in practical learning systems (Amos, 1974; Du et al., 2024; He

[1] Code: https://anonymous.4open.science/r/potential-consistent-vmf/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025). This raises a concrete question of numerical coherence: after approximation, do the forward potential and supplied derivative retain their original derivative relation?

We observe this decoupling in real vMF learning implementations. In PATT (He et al., 2025), the original implementation and automatic differentiation of the same finite forward produce elementwise-identical scores and losses from identical model states and inputs, yet supply different feature gradients. As the branches continue to update independently, their parameter and representation trajectories separate. An additional audit of ProCo (Du et al., 2024) detects the same type of mismatch in a second implementation, with a magnitude that depends on the numerical states visited during learning. **These observations move the question beyond local approximation error to the numerical realization of the learning objective: does the value reported by the program remain coherent with the derivative received by the optimizer?**

To explain this incoherence, we analyze the internal structure of finite-start Bessel recurrence. We prove that the sign of the ratio error is determined exactly by the parity relation between recurrence depth and target order, and characterize its high-concentration asymptotics. Clipping bounds the supplied ratio but does not restore its derivative relation with the forward potential. Consequently, deeper recurrence can improve fidelity to the exact Bessel ratio without monotonically improving coherence. After Softmax composition, forward probabilities couple with mismatched classwise radial rules, potentially producing an asymmetric Jacobian and a locally nonconservative update field. This establishes a mechanism linking special-function mismatch to the geometry of learning updates.

The mismatch also accompanies a dimension-dependent computational structure. Sequential recurrence depth grows with Bessel order, while high orders are precisely where the analytic approximation becomes tighter. Thus, as the original recurrence becomes deeper in high-dimensional vMF learning, a fixed-depth analytic replacement gains increasingly favorable approximation conditions. This motivates a paired analytic realization that avoids recurrence over the order.

We describe such a realization through two complementary requirements. *Coherence* asks whether the supplied backward differentiates the currently realized scalar objective. *Fidelity* asks whether this numerical surrogate remains sufficiently close to the intended analytic vMF objective. These properties are independent: an accurate forward approximation can be paired with a backward rule that does not belong to it, while an internally coherent numerical object can approximate the analytic target poorly. A reliable differentiable numerical primitive must address both questions.

We introduce **AR/FR**, which directly constructs a potential–derivative pair instead of designing the forward potential and backward ratio independently, so that

$$F_R'(x) = A_R(x)$$

holds in exact arithmetic. Coherence is therefore a property of the construction itself. We further prove a uniform cubic-order fidelity bound for AR relative to the exact Bessel ratio over the entire nonnegative concentration axis. Integration transfers this guarantee to the potential endpoint differences used in vMF scores and then to the fixed-state learning objective. AR/FR also has fixed computational depth, eliminating sequential recurrence over Bessel order. Since feature dimension is linearly related to the order, increasing dimension lengthens the original recurrence chain while tightening the worst-case AR/FR error guarantee as $O(p^{-3})$. **The high-dimensional regime in which recurrence becomes deeper is also the regime in which the analytic replacement receives a stronger theoretical guarantee.**

Real learning states, high-precision numerical comparisons, and complete training experiments validate this mechanism and analytic realization. They show that a paired realization can eliminate forward–backward incoherence while maintaining competitive task performance.

Our main contributions are:

1. **Numerical incoherence and its mechanism.** We identify identical forward values with different gradients in real vMF learning implementations, characterize finite-start parity, high-concentration behavior, and clipping, and show that classwise mismatch can compose through probabilities into a locally nonconservative update field.
2. **A coherent and certified paired vMF realization.** AR/FR guarantees forward–backward coherence through $F_R' = A_R$ and provides a uniform cubic-order fidelity bound over the nonnegative concentration axis, together with its propagation to learning scores and objectives.

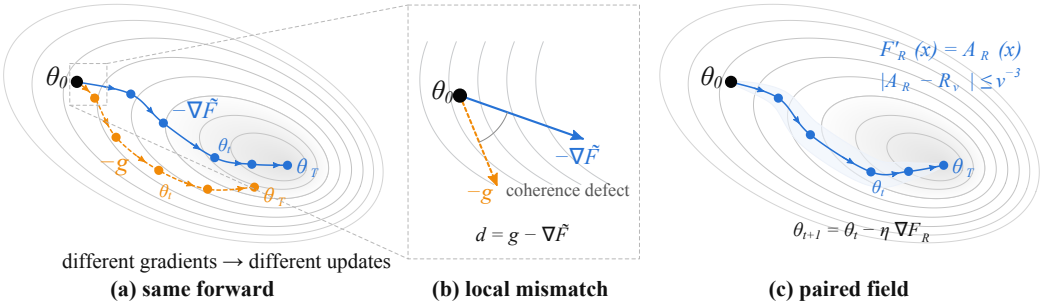


Figure 1: **Same loss, different gradients: diagnosis and paired realization.** (a) A common forward objective can supply different updates when backward rules differ. (b) The coherence defect compares the supplied backward with the derivative of that forward. (c) AR/FR constructs the potential and derivative together, combining coherence with a uniform fidelity bound.

3. **Fixed-depth high-dimensional computation.** AR/FR replaces order-dependent recurrence with fixed-depth analytic computation, revealing a complexity–accuracy reversal as dimension increases: recurrence becomes deeper while the approximation guarantee tightens. This realization substantially reduces the cost of the practical vMF computation path.

2 Coherence and Fidelity of Numerical Realizations

2.1 Coherence, Fidelity, and Realized Objectives

Let F denote an analytic scalar objective, $\widetilde{F}$ the forward function evaluated by a numerical program, and g the supplied backward rule. Its *coherence defect* is

$$d(x) = g(x) - \widetilde{F}'(x). \tag{1}$$

The realization is coherent when $d = 0$: its supplied backward differentiates its realized forward. In contrast, *fidelity* measures proximity to the intended target, through derivative errors $|g - F'|$ and forward potential-difference errors. These properties are logically independent: a realization can approximate the target accurately without being internally coherent, or be coherent while approximating the target poorly.

Two implementations can evaluate the same $\widetilde{F}$ at a shared state but supply different backward rules. They then have identical forward values and losses yet induce different local updates. Throughout this paper, "same loss" denotes equality at a shared state; after separate updates, trajectories and subsequent losses need not remain identical. For potentials, we compare endpoint differences $F(b) - F(a)$, since additive constants do not affect learning.

2.2 The vMF Value–Derivative Relation

Let the feature dimension be p and $\nu = p/2 - 1$. For $x > 0$, the vMF potential and its derivative satisfy

$$\Phi_\nu(x) = \log I_\nu(x) - \nu \log x, \qquad \Phi'_\nu(x) = R_\nu(x) = \frac{I_{\nu+1}(x)}{I_\nu(x)}, \tag{2}$$

where I_ν is the modified Bessel function of the first kind and Φ_ν is extended continuously to $x = 0$.

Fix a class state under the original stop-gradient semantics. Class j has unit direction μ_j and concentration $\kappa_j \ge 0$; f is a query feature and $\tau > 0$ is the temperature. Its scalar radius and score are

$$r_j(f) = \left\| \kappa_j \mu_j + \frac{f}{\tau} \right\|, \qquad q_j(f) = \Phi_\nu(r_j(f)) - \Phi_\nu(\kappa_j). \tag{3}$$

The score requires a forward potential difference and its radial derivative. Finite numerical evaluation can split this analytic pair into distinct computational paths.

3 WHEN FINITE RECURRENCE BREAKS COHERENCE

3.1 THREE NUMERICAL OBJECTS FROM ONE RECURRENCE

For integer orders $M > \nu \geq 1$ and $x > 0$, finite Miller backward recurrence uses

$$b_M = 1, \quad b_{M+1} = 0, \qquad b_{i-1} = \frac{2i}{x} b_i + b_{i+1}.$$

It induces a finite forward potential $\widetilde{\Phi}_{\nu,M}$ and an adjacent-order ratio. The supplied radial backward may additionally clip that ratio:

$$\widetilde{R}_{\nu,M}(x) = \frac{b_{\nu+1}}{b_\nu}, \qquad g_{\nu,M}(x) = \min\{\widetilde{R}_{\nu,M}(x), 1\}.$$

Thus, the potential, raw ratio, and supplied backward are three numerical objects. Coherence requires $g_{\nu,M} = \widetilde{\Phi}'_{\nu,M}$, which neither finite recurrence nor clipping guarantees. Appendix D.3 specifies the finite forward and its normalization.

3.2 FINITE-START PARITY AND CLIPPING

Each adjacent-ratio recurrence step is a strictly decreasing map on the positive axis and reverses the ordering relative to the exact tail. Consequently, we obtain the following all-axis result.

Theorem 1 (Finite-start parity)**.** *For integer $M > \nu \geq 1$,*

$$\operatorname{sign}\big(\widetilde{R}_{\nu,M}(x) - R_\nu(x)\big) = (-1)^{M-\nu+1}, \qquad x > 0. \tag{4}$$

This all-axis parity law reduces to $(-1)^{\nu+1}$ under $M = 2\nu$ (proof in Appendix D.4). It characterizes ratio fidelity; coherence additionally depends on the finite-forward derivative.

At high concentration with $M = 2\nu$, the raw ratio grows linearly for odd ν and decays inversely for even ν, while the exact ratio approaches one (Appendix D.5). Clipping bounds the supplied value but does not restore the potential relation. Increasing depth can improve raw ratio fidelity without monotonically improving the clipped coherence defect, which also depends on the finite-forward derivative and clipping regime. Section 5.3 examines these conditions numerically.

3.3 COMPOSITION CAN CHANGE UPDATE-FIELD GEOMETRY

A scalar radial rule $g(r)$ admits a one-dimensional primitive, so radial mismatch alone does not rule out an underlying objective. The obstruction appears after classwise composition. Let $q_j = F(r_j) - F(\kappa_j) + b_j$, with fixed offsets b_j, and let $p_j = \exp(q_j)/\sum_k \exp(q_k)$. Write $v_j = \nabla_f q_j$ and $w_j = g(r_j)\nabla_f r_j$. Cross-entropy supplies the feature field

$$G(f) = \sum_j (p_j - \mathbf{1}[j = y]) w_j.$$

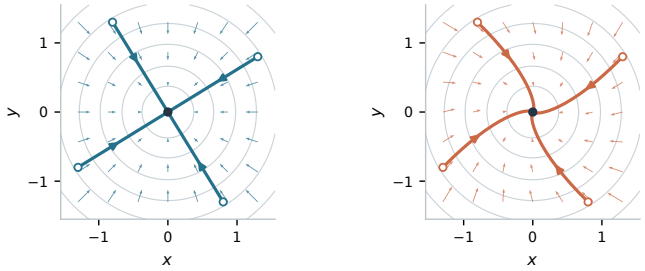
On a smooth neighborhood with fixed class state and positive radii, its antisymmetric Jacobian is

$$J_G - J_G^\top = \bar{v}\bar{w}^\top - \bar{w}\bar{v}^\top, \qquad \bar{v} = \sum_j p_j v_j, \quad \bar{w} = \sum_j p_j w_j. \tag{5}$$

Forward probabilities depend on F, while feature updates use g. Softmax couples these radial realizations, and the antisymmetric Jacobian determines whether the composed field admits a local C^2 scalar potential.

When $g = F'$, we have $w_j = v_j$ and the expression vanishes, as required for a C^2 scalar-loss gradient field.

A valid fixed-state vMF construction yields a nonzero antisymmetric Jacobian component under the supplied backward, whereas differentiating the same forward gives zero. The full construction is given in Appendix D.2, with additional numerical controls in Appendix E.5. This provides a local existence result: classwise mismatch can compose into a nonconservative feature-update field.



(a) Conservative field
Symmetric Jacobian

(b) Mismatched field
Asymmetric Jacobian

Figure 2: **Local geometry of coherent and mismatched update fields.** Both panels share the same scalar forward surface. A matched backward yields a conservative gradient field with symmetric Jacobian, whereas a mismatched backward can introduce a local rotational component and asymmetric cross-derivatives. The figure is an analytic illustration.

4 A PAIRED ANALYTIC VMF REALIZATION

4.1 PAIRED ANALYTIC CONSTRUCTION

Analytic ratio approximations and their integrated forms have previously been used in vMF learning (Scott et al., 2021). Our goal is to make the numerical realization itself satisfy both forward–backward coherence and provable fidelity to the intended objective. We construct the potential and its derivative jointly. For $\nu > 0$, $x \geq 0$, and $s(x) = \sqrt{\nu^2 + x^2}$, define

$$A_{R,\nu}(x) = \frac{x}{\nu + s(x)} - \frac{x}{2s(x)^2} + \frac{x(4\nu^2 - x^2)}{8s(x)^5}. \tag{6}$$

Let $F_{R,\nu}$ be its analytic primitive. The identity $F'_{R,\nu} = A_{R,\nu}$ makes the coherence defect vanish by construction. The forward potential and supplied derivative are therefore no longer approximated independently; both are generated from the same analytic object. The paired score is

$$\widehat{q}_j(f) = F_{R,\nu}(r_j(f)) - F_{R,\nu}(\kappa_j), \tag{7}$$

with the same fixed class state. Equation (85) gives the closed form of F_R, and Appendix F.2 describes stable endpoint evaluation.

4.2 UNIFORM FIDELITY CERTIFICATE

Coherence alone does not establish fidelity to the analytic vMF target. The exact radial derivative R_ν satisfies a Riccati differential equation, which provides a direct way to analyze the error of the analytic approximation $A_{R,\nu}$.

Theorem 2 (Uniform ratio error bound)**.** *For every $\nu \geq 10/9$ and $x \geq 0$,*

$$|R_\nu(x) - A_{R,\nu}(x)| \leq \nu^{-3}. \tag{8}$$

The proof rescales $x = \nu z$ and substitutes $A_{R,\nu}$ into the Riccati equation for the exact Bessel ratio. The resulting residual is uniformly $O(\nu^{-3})$. Constant upper and lower barriers at $\pm\nu^{-3}$ are then shown to be inward-pointing for $\nu \geq 10/9$, while the small-x expansion initializes the error strictly between them. This prevents the approximation error from crossing either barrier and yields the bound over the entire nonnegative axis. The complete argument is given in Appendix D.7.

Unlike a grid-based numerical validation, this guarantee is independent of the concentration regime visited during training. It therefore complements the state-dependent coherence analysis in Section 5.3: coherence defects can vary strongly across visited states, whereas the AR/FR fidelity certificate holds uniformly for every $x \geq 0$.

4.3 LEARNING-LEVEL FIDELITY AND FIXED-DEPTH COMPUTATION

Integrating the ratio certificate gives, for any $r, \kappa \geq 0$,

$$|[\Phi_\nu(r) - \Phi_\nu(\kappa)] - [F_{R,\nu}(r) - F_{R,\nu}(\kappa)]| \leq \nu^{-3}|r - \kappa|. \tag{9}$$

Since $|r_j(f) - \kappa_j| \leq \|f\|/\tau$, the learning score obeys

$$|q_j(f) - \widehat{q}_j(f)| \leq \frac{\|f\|}{\tau\nu^3}.$$

At fixed class states, this guarantee also propagates through the two-view learning objective without a multiplicative dependence on the number of classes or the long-tail priors. Let $\delta_i^{(v)} = \|f_i^{(v)}\|/(\tau\nu^3)$ and $\overline{\delta}^{(v)} = B^{-1}\sum_i \delta_i^{(v)}$. Replacing q_j by $\widehat{q}_j$ then gives

$$|\widehat{L}_{\mathrm{PATT}} - L_{\mathrm{PATT}}| \leq \overline{\delta}^{(2)} + \overline{\delta}^{(3)}, \tag{10}$$

for arbitrary additive class biases. Corresponding probability, margin, feature-gradient, and class-group bounds accompany the certificate in Appendix D.1 and are proved in Appendix D.8.

Finite recurrence follows an order-dependent chain, whereas AR/FR contains a fixed number of elementary operations. With $p = 2\nu + 2$, the certificate becomes $\nu^{-3} = 8/(p-2)^3$. This yields a useful high-dimensional reversal: increasing feature dimension makes the finite recurrence sequentially deeper, while the worst-case AR/FR approximation guarantee becomes tighter. Thus the regime in which recurrence becomes increasingly expensive is also the regime in which the paired analytic approximation is most strongly certified. AR/FR therefore combines coherence, certified fidelity, and fixed-depth computation in the high-dimensional setting that motivates the method.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP AND NUMERICAL REALIZATIONS

We compare three numerical realizations: *original recurrence* uses the finite forward and clipped backward; *consistent recurrence* differentiates that same forward; and *AR/FR* replaces the potential and derivative jointly. The first comparison isolates the backward rule, while AR/FR tests a paired replacement. Architecture, class-state estimation, data, optimizer, and surrounding objective are matched; implementation and training details appear in Appendices B.1 and F.1.

5.2 SAME FORWARD, DIFFERENT GRADIENTS AND TRAJECTORIES

Changing only the backward rule alters PATT optimization from both random initialization and an intermediate checkpoint. Within each fork, original and consistent recurrence share initial parameters, class and optimizer states, and augmented inputs. Shared-state comparisons yield bitwise-identical scores and losses but different feature gradients. Over 1000 updates, these local differences accumulate into persistent parameter and representation separation (Figure 3), linking numerical incoherence to the realized learning dynamics.

5.3 WHEN DOES MISMATCH OCCUR?

Coherence defects depend jointly on concentration and recurrence depth (Figure 4). At high concentration, deeper recurrence does not uniformly reduce the clipped defect: improving the raw ratio and matching the finite-forward derivative are distinct objectives. Captured PATT queries occupy this high-concentration regime. ProCo exhibits the same conditional phenomenon in a second vMF implementation (Appendix C).

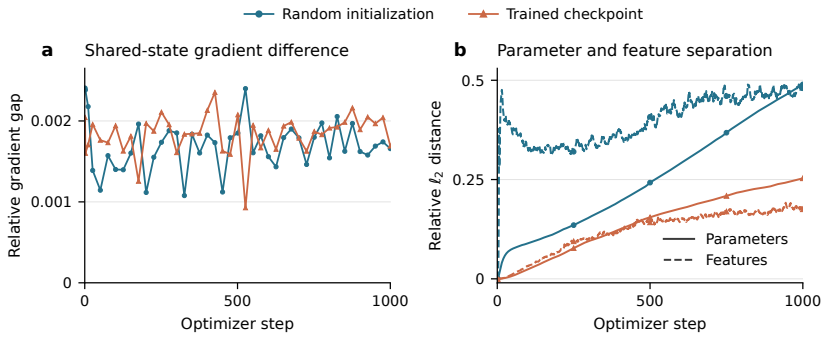


Figure 3: **Same-state gradient mismatch becomes trajectory separation.** Original and consistent recurrence share random or intermediate-training starting states and augmented inputs. Left: local gradient gaps with bitwise-identical forward values. Right: parameter and feature distances after separate updates.

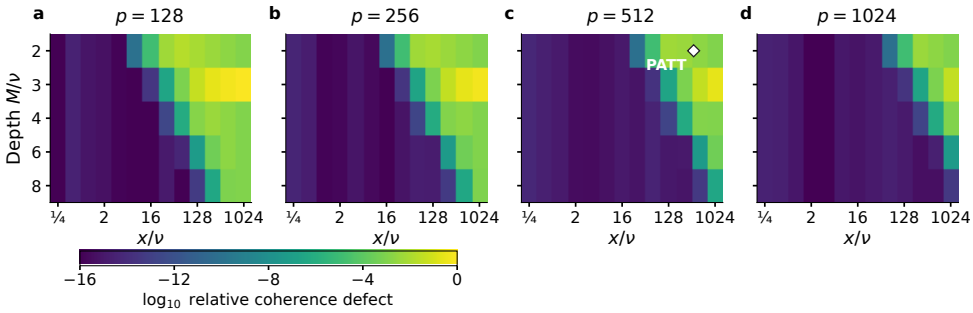


Figure 4: **Concentration and depth jointly determine coherence.** Colour shows $\log_{10}(|g_{\text{supplied}} - \widetilde{\Phi}'|/|\widetilde{\Phi}'|)$, with a 10^{-16} display floor. The diamond locates captured PATT queries at $p = 512$ and $M/\nu = 2$.

5.4 Fidelity of AR/FR

Across feature dimensions from 64 to 4096, the maximum sampled AR/FR ratio error decays with fitted log–log slope -3.02 and remains below the uniform certificate (Figure 5). This cubic trend supports the high-dimensional behavior established in Section 4.3: fidelity improves as the recurrence being replaced grows deeper. Captured-state score errors, potential-difference errors, and evaluator comparisons appear in Appendix E.2.

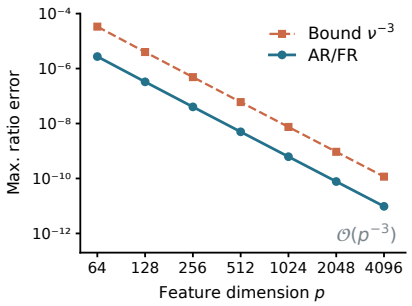


Figure 5: **Cubic error decay.** Blue: sampled maximum error. Orange: the all-axis certificate ν^{-3}.

5.5 Complete task performance

We evaluate classification and SC-OOD detection on CIFAR-10-LT and CIFAR-100-LT at imbalance factors 10, 50, and 100, together with ImageNet-LT. Table 1 reports accuracy and AUROC; Appendix G.1 includes FPR95, and Appendix B.1 gives the evaluation protocol.

The paired replacement preserves broadly comparable task performance, with changes that depend on the dataset, imbalance, and metric. On CIFAR, accuracy and OOD

detection can move in different directions, and the most imbalanced CIFAR-100 setting shows modest decreases. ImageNet-LT shows improvements across all three metrics. These results characterize the task behavior of numerical replacement across settings.

Table 1: **Classification and OOD detection under different numerical realizations.** CIFAR results report mean ± standard deviation over three runs. Metrics are percentages; IF denotes the imbalance factor. Full FPR95 results appear in Appendix G.1. Bold and underline indicate the best and second-best values within each setting.

Dataset	IF	Metric	Original recurrence	Consistent recurrence	AR/FR
CIFAR-10-LT	10	Acc. ↑	92.04 ± 0.17	$\underline{92.12} \pm 0.18$	$\mathbf{92.16} \pm 0.19$
		AUROC ↑	$\underline{93.07} \pm 0.13$	92.30 ± 0.27	$\mathbf{93.43} \pm 0.46$
CIFAR-10-LT	50	Acc. ↑	87.30 ± 0.29	$\mathbf{87.58} \pm 0.35$	$\underline{87.42} \pm 0.15$
		AUROC ↑	90.18 ± 2.46	$\underline{90.95} \pm 1.57$	$\mathbf{91.02} \pm 1.13$
CIFAR-10-LT	100	Acc. ↑	83.82 ± 0.21	$\underline{84.67} \pm 0.41$	$\mathbf{84.73} \pm 0.31$
		AUROC ↑	$\mathbf{90.66} \pm 0.13$	$\underline{89.97} \pm 0.39$	89.95 ± 0.34
CIFAR-100-LT	10	Acc. ↑	$\underline{62.37} \pm 2.43$	62.12 ± 1.84	$\mathbf{62.56} \pm 2.03$
		AUROC ↑	$\underline{78.18} \pm 2.09$	78.08 ± 1.23	$\mathbf{78.44} \pm 1.53$
CIFAR-100-LT	50	Acc. ↑	$\underline{53.26} \pm 0.19$	52.87 ± 0.31	$\mathbf{53.74} \pm 0.31$
		AUROC ↑	75.37 ± 0.68	$\underline{75.54} \pm 0.60$	$\mathbf{76.14} \pm 0.24$
CIFAR-100-LT	100	Acc. ↑	$\mathbf{50.33} \pm 0.16$	49.48 ± 0.26	$\underline{49.66} \pm 0.16$
		AUROC ↑	$\mathbf{75.78} \pm 0.33$	$\underline{75.40} \pm 0.54$	75.21 ± 0.65
ImageNet-LT	—	Acc. ↑	53.88	$\underline{54.11}$	**54.34**
		AUROC ↑	76.19	$\underline{76.87}$	**77.50**

5.6 Efficient realization and structure-aware execution

We evaluate the complete local vMF path under a unified execution protocol, covering class-state updates, score geometry, cross-entropy, and feature backward while excluding the backbone and data pipeline.

Figure 6 compares the numerical realizations and the structure-aware implementation. Consistent recurrence restores coherence by differentiating the finite forward but increases both runtime and memory. Log-domain Miller retains recurrence in a different numerical representation. AR/FR alone reduces runtime by $54.4\times$ with a similar dense-layout memory footprint. Adding direct class-statistic accumulation, factorized geometry, and compiled GPU execution yields a further $23.8\times$ runtime reduction. The overall $1293\times$ speedup and $15.8\times$ memory reduction characterize this combined realization; the memory savings arise from structure-aware execution. Runtime and incremental peak memory are measured over the same local execution boundary for all realizations; the detailed workload, memory definition, and measurement protocol are provided in Appendix F.3.

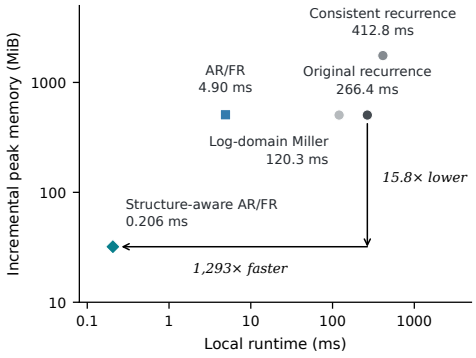


Figure 6: **Local runtime–memory trade-off.**

6 Discussion

The controlled recurrence intervention shows that the supplied derivative is part of the learning algorithm: an unchanged forward objective can lead to different optimization trajectories. AR/FR

pairs a coherent derivative with a certified forward approximation, preserving broadly comparable task performance while enabling fixed-depth execution. The Jacobian construction establishes a local mechanism rather than its prevalence across training states.

More broadly, the vMF case separates two questions that are often collapsed in differentiable numerical software. Approximation analysis asks whether a numerical value or derivative is close to an analytic reference; coherence asks whether the value and derivative supplied by the program belong to each other. The latter matters even when both individual errors are small, because optimization acts on their composition. Pairing the realized scalar with its derivative therefore provides a useful design and diagnostic principle for learning objectives built from nontrivial numerical primitives.

7 Conclusion

Finite numerical evaluation can preserve the forward objective while changing gradients and optimization trajectories. We characterize this decoupling in vMF learning and its effect on local update-field geometry. AR/FR pairs the potential and derivative to combine coherence, certified fidelity, and fixed-depth execution. Differentiable numerical primitives should be evaluated through both their realized values and supplied derivatives.

AI Use Statement. Generative AI assisted with writing and editing, feedback on theoretical derivations, experiment design and result interpretation, drafts of mathematical checks, code development, experiment organization, and figure preparation. The authors independently verified all AI-assisted mathematical claims, code, analyses, and written content against derivations, executable code, primary references, and raw experimental records, and take full responsibility for the final content.

Reproducibility Statement. Appendix B.1 describes the training and evaluation protocols, and Appendix B.2 specifies the controlled trajectory interventions. Theoretical proofs and numerical validation are provided in Appendices D and E, with the local runtime and memory protocol in Appendix F.3. The anonymous code release provides the AR/FR numerical core, a minimal coherence check, and the reported task and efficiency result tables.

References

D. E. Amos. Computation of modified bessel functions and their ratios. *Mathematics of Computation*, 28(125):239–251, 1974. doi: 10.1090/S0025-5718-1974-0333287-7.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46): 1345–1382, 2005.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10809–10819, 2020.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021. doi: 10.1007/s10107-020-01501-5.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 856–865, 2018.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5890–5904, 2024. doi: 10.1109/TPAMI.2024.3369102.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.

Yina He, Lei Peng, Yongcun Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. Long-tailed out-of-distribution detection: Prioritizing attention to tail. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):3446–3454, 2025. doi: 10.1609/aaai.v39i3.32357.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Kurt Hornik and Bettina Grün. Amos-type bounds for modified bessel function ratios. *Journal of Mathematical Analysis and Applications*, 408(1):91–101, 2013. doi: 10.1016/j.jmaa.2013.05.070.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Wenjun Miao, Guansong Pang, Xiao Bai, Tianqi Li, and Jin Zheng. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4216–4224, March 2024a. ISSN 2159-5399. doi: 10.1609/aaai.v38i5.28217.

Wenjun Miao, Guansong Pang, Jin Zheng, and Xiao Bai. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 132106–132132. Curran Associates, Inc., 2024b. doi: 10.52202/079017-4199.

F. W. J. Olver. The asymptotic expansion of bessel functions of large order. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 247(930):328–368, 1954. doi: 10.1098/rsta.1954.0021.

Andreas Plesner, Hans Henrik Brandenborg Sørensen, and Søren Hauberg. Accurate computation of the logarithm of modified bessel functions on gpus. In *Proceedings of the 38th ACM International Conference on Supercomputing*, ICS 2024, pp. 213–224. ACM, May 2024. doi: 10.1145/3650200.3656624.

Diego Ruiz-Antolín and Javier Segura. A new type of sharp bounds for ratios of modified bessel functions. *Journal of Mathematical Analysis and Applications*, 443(2):1232–1246, 2016. doi: 10.1016/j.jmaa.2016.06.011.

Tyler R. Scott, Andrew C. Gallagher, and Michael C. Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10612–10622, 2021.

Javier Segura. Simple bounds with best possible accuracy for ratios of modified bessel functions. *Journal of Mathematical Analysis and Applications*, 526(1):127211, 2023. doi: 10.1016/j.jmaa.2023.127211.

Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: And a fast implementation of $I_s(x)$. *Computational Statistics*, 27(1):177–190, 2012. doi: 10.1007/s00180-011-0232-x.

Nimeshika Udayangani, Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Exploiting inter-sample information for long-tailed out-of-distribution detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8546–8555. IEEE, February 2025. doi: 10.1109/WACV61041.2025.00828.

Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23446–23458. PMLR, 17–23 Jul 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Tong Wei, Bo-Lin Wang, and Min-Ling Zhang. EAT: Towards long-tailed out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15787–15795, 2024. doi: 10.1609/aaai.v38i14.29508.

Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 8281–8289, 2021. doi: 10.1109/ICCV48922.2021.00819.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019.

SAME LOSS, DIFFERENT GRADIENTS

Supplementary Material

Contents

A Related work **13**

B Experimental protocols and reproducibility **13**

B.1 Training and evaluation protocols 13

B.2 Controlled intervention protocols 14

B.3 Source implementation 14

C ProCo: state-dependent coherence defects **14**

D Theoretical proofs and auxiliary identities **15**

D.1 Learning-level fidelity guarantees 15

D.2 Proof of the Softmax integrability obstruction 15

D.3 PATT objective and finite-start identities 16

D.4 Proof of the finite-start sign theorem 18

D.5 High-concentration asymptotics 18

D.6 Finite-forward endpoint and defect identities 19

D.7 Proof of the AR uniform certificate 20

D.8 Proof of the PATT-specific certificate 21

E Numerical validation and mechanism analysis **22**

E.1 Frozen-state mechanism validation 22

E.2 Fidelity benchmarks and numerical references 23

E.3 Learning-level stress tests 24

E.4 Measured PATT update fields 25

E.5 Classwise geometry and full-update behavior 26

F Structure-aware execution and efficiency **27**

F.1 Paired evaluation in the host learning system 27

F.2 Stable potential differences and factorized execution 28

F.3 Unified H100 local runtime and memory protocol 28

G Extended task results **30**

G.1 Complete task metrics 30

A Related work

Supplied gradients and numerical differentiation. The possibility of using a backward rule different from the forward derivative is established, including intentional straight-through estimators for stochastic or nonsmooth units (Bengio et al., 2013; Yin et al., 2019). Automatic differentiation evaluates derivatives of numerical programs (Baydin et al., 2018); nonsmooth models and conservative set-valued fields clarify the mathematical interpretation of the resulting oracles (Bolte & Pauwels, 2020; 2021). These nonsmooth frameworks provide context for distinguishing a supplied update from a classical gradient. Our setting concerns smooth special-function objectives whose numerical forward and backward are intended to represent an analytic value–derivative relation. The contribution is a characterization of a concrete recurrence mechanism, its composition and training consequences, and a certified paired replacement.

vMF probabilistic learning. Directional clustering uses vMF mixtures to model normalized observations (Banerjee et al., 2005), while hyperspherical VAEs use vMF latent distributions (Davidson et al., 2018). Differentiating through probability distributions also motivates implicit reparameterization (Figurnov et al., 2018); that work concerns sampling gradients rather than the finite normalizer recurrence examined here. ProCo derives scores from class-conditional vMF distributions (Du et al., 2024), and PATT uses the score in long-tailed OOD learning (He et al., 2025). They provide related implementations with different visited numerical states. The supervised vMF loss of Scott et al. (2021) provides a complementary construction: it averages analytic ratio bounds and integrates that approximation to obtain a normalizer surrogate. The ratio and potential are therefore paired already at the mathematical level. Such a coherent reference separates the value of pairing from the particular approximation and certificate developed here. SC-OOD evaluates novelty after accounting for semantic overlap (Yang et al., 2021).

Contrastive representations. SimCLR learns representations from augmented positive pairs (Chen et al., 2020), and supervised contrastive learning incorporates class labels into the positive-pair construction (Khosla et al., 2020). Alignment and uniformity characterize complementary properties of contrastive representations on the hypersphere (Wang & Isola, 2020). ProCo moves from sampled pairs to class distributions; our study addresses the numerical realization of the resulting probabilistic score rather than changing the pair construction or representation objective.

Bessel computation and approximation. Classical recurrence and scaling methods (Amos, 1974; Sra, 2012), analytic ratio bounds (Hornik & Grün, 2013; Ruiz-Antolín & Segura, 2016; Segura, 2023), and uniform large-order expansions (Olver, 1954) underpin high-order special-function evaluation. CUSF supplies accurate GPU log-Bessel computation (Plesner et al., 2024) and is a relevant numerical and systems comparator. Our AR/FR pair builds on the classical asymptotic structure; its contribution comprises the paired realization, uniform certificate, and connection to the recurrence mechanisms studied here. The numerical evaluation reports regimes favoring different evaluators.

Long-tailed OOD context. Long-tailed recognition can be addressed through class-dependent margins (Cao et al., 2019), decoupled representation and classifier learning (Kang et al., 2020), or label-frequency-based logit adjustment (Menon et al., 2021). In OOD detection, maximum Softmax probability (Hendrycks & Gimpel, 2017) and energy scores (Liu et al., 2020) provide scoring rules, while outlier exposure uses auxiliary outliers during training (Hendrycks et al., 2019). PATT combines vMF-based implicit semantic augmentation with tail-focused calibration (He et al., 2025). Existing LT-OOD approaches alter training objectives, representation separation, or outlier modeling (Wei et al., 2024; Wang et al., 2022; Miao et al., 2024a); further approaches use outlier-distribution adaptation (Miao et al., 2024b) or inter-sample graph information (Udayangani et al., 2025). Our comparison holds the host learning and calibration pipeline fixed to study its numerical realization.

B Experimental protocols and reproducibility

B.1 Training and evaluation protocols

CIFAR training and readout. The CIFAR datasets originate from Krizhevsky (2009). The complete three-realization comparison uses exponential long-tail subsampling with factors 10, 50, and

100. CIFAR results report means and standard deviations over three runs, with matched initialization across realizations within each run. The shared training shell uses ResNet18 (He et al., 2016), ID batch 128 with three views, OE batch 256 from the 300k Tiny Images source, Adam, and the original augmentation pipeline. FP32 backbone and expanded geometry and FP64 scalar evaluation are shared across branches; autocast is off. The final epoch-99 weights are evaluated, without selecting a best checkpoint. All branches use matched training configurations and common initialization; the causal forks additionally share the exact augmented tensors at every update.

Evaluation applies PATT feature calibration fitted using only members of the corresponding long-tailed training split. The test set is reserved for evaluation. SC-OOD averages source-level metrics after semantic-ID handling within each of six sources.

ImageNet-LT training and evaluation. The ImageNet-LT realizations use ResNet50, 100 epochs, SGD with initial learning rate 0.1, momentum 0.9 and weight decay 5×10^{-5}, ID batch 32 and OE batch 64. Final checkpoints have zero-based epoch 99. Calibration uses deterministic class-balanced sampling from the training ID data and training-OOD data, with 30,000 calibration samples per source; held-out labels are used only for evaluation. Evaluations share manifests, transforms, seed and batch size. Classification uses the 20,000-image ImageNet-LT validation manifest and OOD detection uses the 50,000-image test-OOD manifest.

B.2 Controlled intervention protocols

PATT causal forks. Random-initialization and trained-checkpoint forks each run 1000 updates from a shared state within each fork. The second fork starts from the original branch after its first 1000 updates. Both branches have identical initial parameters, class banks, optimizer settings, and actual augmented tensors. Local probes evaluate different backward rules at one shared state; evolved-state measurements evaluate each branch after its own updates. Each fork contains 43 local probes with bitwise-equal original and consistent-recurrence forward values. Across the 43 probes, relative gradient differences remain on the order of 10^{-3} for both starting states. Relative evolved parameter distance is normalized by the original parameter norm. Both forks develop substantial parameter separation over 1000 updates.

B.3 Source implementation

The original-recurrence baseline follows the public PATT implementation used throughout our experiments, with Miller start $M = 2\nu$ and a unit-clipped custom backward. Exact revision information is provided with the submission artifacts.

C ProCo: state-dependent coherence defects

We use ProCo as a second real implementation of finite-recurrence vMF learning. The audit compares its supplied clipped radial rule with FP64 automatic differentiation of the same finite forward at $M = 2\nu$, testing coherence at fixed numerical states.

Observed states. The CIFAR100-LT pure-ProCo captures use ResNet32, $p = 128$, $\tau = 0.1$, and imbalance factor 100. Each first-batch capture contains 25,600 query radii; the end-of-epoch-1 replay contains 51,200. The ImageNet joint-training replay uses an epoch-90 checkpoint with $p = 1024$ and a new augmented batch, yielding 64,000 radii. The iNaturalist audit covers 8,142 stored class-bank concentrations at each of epochs 1, 10, 50, and 90. These bank values are reported separately from query radii.

State dependence. In the first CIFAR batch, 23% of query radii have absolute coherence defect above 10^{-6}. By epochs 100 and 200, the corresponding captures are near floating-point agreement; the ImageNet replay and iNaturalist bank audits likewise have small defects (Figure 7). Early CIFAR states include capped concentration estimates from singleton classes. The observed defect thus depends on the regime reached by the class estimator and query geometry, rather than being uniformly large throughout training.

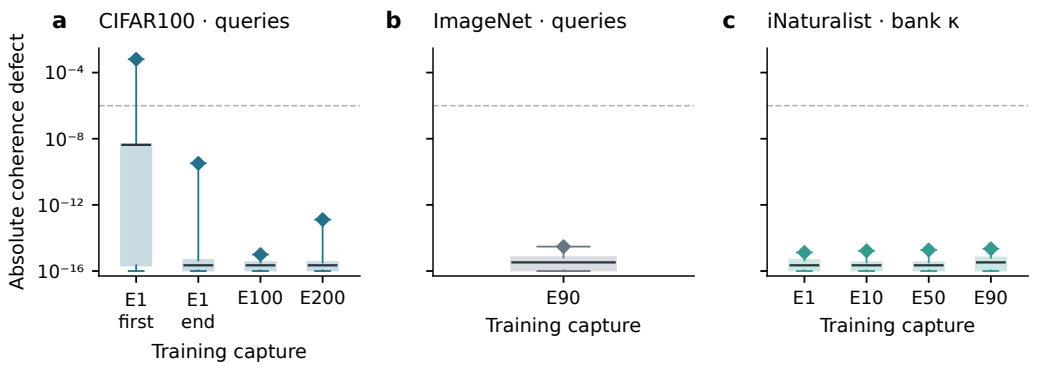


Figure 7: **ProCo coherence defects depend on the visited state.** Absolute differences compare the supplied radial rule with FP64 autodiff at $M = 2\nu$. Boxes show median and interquartile range; whiskers span extrema. (a) CIFAR queries, (b) ImageNet queries, and (c) iNaturalist class-bank concentrations. The dashed diagnostic scale is 10^{-6}; values below 10^{-16} share a display floor.

D Theoretical proofs and auxiliary identities

D.1 Learning-level fidelity guarantees

The central question is whether classes or long-tail priors amplify FR error. Class priors enter as fixed additive biases $b_j = \log \pi_j$. For view $v \in \{2, 3\}$, set $\delta_i^{(v)} = \left\| f_i^{(v)} \right\| / (\tau \nu^3)$ and $\overline{\delta}^{(v)} = B^{-1} \sum_i \delta_i^{(v)}$.

Theorem 3 (Fixed-state PATT evaluator certificate)**.** *Assume $\nu \geq 10/9$ and $\tau > 0$. Fix each view's online class state, let $K < \infty$ and additive biases b_j be arbitrary, and replace only q_j by* (7)*. Then*

$$|q_j(f_i^{(v)}) - \widehat{q}_j(f_i^{(v)})| \leq \delta_i^{(v)}, \tag{11}$$

$$|\widehat{L}_{\mathrm{PATT}} - L_{\mathrm{PATT}}| \leq \overline{\delta}^{(2)} + \overline{\delta}^{(3)}. \tag{12}$$

The bounds have no multiplicative factor in the class count, priors, or imbalance ratio. Appendix D.8 gives the proof and the corresponding Softmax-TV, margin, feature-gradient, and nonempty-group bounds.

The class-count independence follows from log-sum-exp monotonicity under a uniform coordinate perturbation, using the maximum coordinate error. It accommodates PATT's $b_j = \log \pi_j$ even under severe imbalance.

The same argument controls adjusted probabilities and feature gradients at fixed class states. The learning-level guarantee therefore connects the special-function approximation to the quantities consumed by optimization.

D.2 Proof of the Softmax integrability obstruction

Since $w_j = \nabla_f H(r_j)$, its Jacobian is symmetric. Also $\nabla_f p_j = p_j(v_j - \bar{v})$. Differentiating G and subtracting its transpose leaves

$$\sum_j p_j(w_j v_j^\top - v_j w_j^\top) - \bar{w}\bar{v}^\top + \bar{v}\bar{w}^\top.$$

For every class, w_j and v_j are collinear with $\nabla_f r_j$, so the sum is zero. Symmetry of the Hessian is necessary for a C^2 potential, proving the obstruction. When $g = F'$, the ordinary chain rule gives the final assertion. Hence the condition provides a sufficient obstruction; incoherence alone does not imply nonconservativity. □

A reproducible two-class construction. Take $p = 128$, $\nu = 63$, $M = 126$, $\tau = 1$, $f = 0$, zero class biases, $\mu_1 = e_1$, $\mu_2 = e_2$, and $(\kappa_1, \kappa_2) = (4032, 8064)$. Here $F = \widetilde{\Phi}_{63,126}$ and $g = \min\{\widetilde{R}_{63,126}, 1\}$. Both scores are zero, so both probabilities are $1/2$. The raw ratios are approximately 1.08748 and 1.56080, placing both classes strictly inside a smooth clipping region. The finite-forward derivatives are approximately 0.984317 and 0.992073. With the convention $J_{ij} = \partial G_i / \partial f_j$, direct substitution yields

$$J_{12} - J_{21} = \frac{F'(\kappa_1)g(\kappa_2) - g(\kappa_1)F'(\kappa_2)}{4\tau^2} = -1.93915 \times 10^{-3}.$$

The true derivative of the same forward produces zero antisymmetry. Class states are fixed; the local smoothness assertion excludes clipping transition points and zero query radii.

Degeneracy and controlled mismatch. If $g = cF'$ with the same constant c throughout a neighborhood, then $G = c\nabla_f L$ is conservative even when $c \neq 1$. Nonzero coherence defect alone therefore does not imply nonconservativity. At the origin of the orthogonal two-class construction, equal concentrations make the displayed antisymmetry zero; this pointwise statement does not assert integrability throughout a neighborhood. For a constant intervention strength λ, define $g_\lambda = F' + \lambda(g - F')$. Since the forward probabilities are unchanged,

$$G_\lambda - G_0 = \lambda(G_1 - G_0), \qquad J_{G_\lambda} - J_{G_\lambda}^\top = \lambda(J_{G_1} - J_{G_1}^\top).$$

The linear relation holds for the fixed-state field on smooth regions.

D.3 PATT OBJECTIVE AND FINITE-START IDENTITIES

D.3.1 THE VMF TERM INSIDE PATT

Let $p = 2\nu + 2$ be the feature dimension. For class j, let μ_j be a unit direction, $\kappa_j \geq 0$ a concentration, $f \in \mathbb{R}^p$ a query feature, and $\tau > 0$ the temperature. Define

$$r_j(f) = \|\kappa_j \mu_j + f/\tau\|, \qquad \Phi_\nu(x) = \log I_\nu(x) - \nu \log x. \tag{13}$$

At $x = 0$, Φ_ν denotes its continuous extension; finite-start sign and parity statements concern $x > 0$. The exact class score is

$$q_j(f) = \Phi_\nu(r_j(f)) - \Phi_\nu(\kappa_j). \tag{14}$$

PATT adds a fixed long-tail log prior $b_j = \log \pi_j$ before cross-entropy. For views $f_i^{(2)}, f_i^{(3)}$ with label y_i, write

$$\ell_i^{(v)} = -q_{y_i}(f_i^{(v)}) - b_{y_i} + \log \sum_{j=1}^{K} \exp(q_j(f_i^{(v)}) + b_j). \tag{15}$$

For a batch of size B, the implicit semantic augmentation term is

$$L_{\mathrm{ISAC}} = \frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{(2)} + \ell_i^{(3)} \right). \tag{16}$$

The complete original objective has the form

$$L_{\mathrm{PATT}} = L_{\mathrm{ISAC}} + \lambda_{\mathrm{TLA}} L_{\mathrm{TLA}} + \lambda_{\mathrm{OE}} L_{\mathrm{OE}}. \tag{17}$$

Only L_{ISAC} contains the Bessel score. The prior bias, TLA term, OE term, encoder, data, and feature-calibration evaluator do not depend on the choice between Miller and FR.

D.3.2 FINITE-SAMPLE CLASS STATE BEFORE BESSEL EVALUATION

For class count $n_j > 0$ and detached-feature sum S_j, PATT's cumulative-mean estimator uses

$$a_j = \frac{S_j}{n_j}, \qquad R_j = \|a_j\|, \qquad \mu_j = \frac{a_j}{R_j}, \qquad \kappa_j = h_p(R_j) = \frac{pR_j}{1 - R_j^2}. \tag{18}$$

When $R_j = 0$, we use the continuous score-geometry convention $\kappa_j = 0$ and $\kappa_j \mu_j = 0$ (equivalently, set $\mu_j = 0$); no result depends on an otherwise undefined direction. The source initializer handles $n_j = 0$ separately. Appendix Lemma 8 characterizes the finite-sample resultant. The source keeps an initializer at $n_j = 0$ and caps the raw concentration at 10^5. For batch sum s and count m, Appendix Proposition 9 gives the exact transition $a^+ = (Na + s)/(N + m)$, preserving both estimator objects and the sequential, possibly correlated view schedule without a dense one-hot expansion.

D.3.3 THE ORIGINAL FORWARD–BACKWARD OBJECTS

The exact exponential-family geometry links value and derivative:

$$\Phi'_\nu(x) = R_\nu(x), \qquad R_\nu(x) = \frac{I_{\nu+1}(x)}{I_\nu(x)}. \tag{19}$$

The original evaluator starts Miller's order recurrence at $M = 2\nu$. For a general integer $M > \nu \geq 1$, define

$$b_M = 1, \quad b_{M+1} = 0, \quad b_{i-1} = \frac{2i}{x} b_i + b_{i+1} \quad (i = M, \ldots, 1). \tag{20}$$

It is useful first to isolate the exact finite-recurrence object

$$\widetilde{\Phi}_{\nu,M}(x) = \log I_0(x) + \log \frac{b_\nu}{b_0} - \nu \log x, \tag{21}$$

$$\widetilde{R}_{\nu,M}(x) = \frac{b_{\nu+1}}{b_\nu}, \qquad \overline{R}_{\nu,M}(x) = \min\{\widetilde{R}_{\nu,M}(x), 1\}. \tag{22}$$

The source uses $\log(x + \epsilon)$ with $\epsilon = 10^{-20}$, saves $\widetilde{R}_{\nu,M}$, and supplies $\overline{R}_{\nu,M}$ in backward; Appendix D.6 retains the literal correction.

Potential-consistent evaluator. A reported forward Ψ and supplied radial backward G are *potential-consistent* on I when $\Psi' = G$ throughout I. The corresponding signed residual is $D_g^\epsilon = (\widetilde{\Phi}^\epsilon_{\nu,M})' - g = -d^\epsilon$, the negative of the coherence defect in Section 2.

D.3.4 FINITE-START PARITY IS AN ALL-AXIS EFFECT

Set

$$\widetilde{u}_i = \frac{b_{i+1}}{x b_i}, \qquad u_i = \frac{I_{i+1}(x)}{x I_i(x)}, \qquad T_i(u) = \frac{1}{2i + x^2 u}. \tag{23}$$

Then $\widetilde{u}_M = 0$ and both sequences obey $\widetilde{u}_{i-1} = T_i(\widetilde{u}_i)$ and $u_{i-1} = T_i(u_i)$.

Theorem 4 (Finite-start sign and product)**.** *For every integer $M > \nu \geq 1$ and $x > 0$,*

$$\begin{aligned} \widetilde{u}_\nu - u_\nu = (-1)^{M-\nu}(\widetilde{u}_M - u_M) \\ \times \prod_{i=\nu+1}^{M} \frac{x^2}{(2i + x^2\widetilde{u}_i)(2i + x^2 u_i)}. \end{aligned} \tag{24}$$

Consequently,

$$\operatorname{sign}(\widetilde{R}_{\nu,M} - R_\nu) = (-1)^{M-\nu+1}. \tag{25}$$

At the original start $M = 2\nu$, the sign is $(-1)^{\nu+1}$.

All factors in (24) are positive, so the theorem is an exact-arithmetic statement over the entire positive axis. Intuitively, every backward map T_i is strictly decreasing on the positive axis: the terminal approximation starts below the exact tail, and each recurrence step flips the inequality. Only the number of steps from M to ν determines which side the final ratio occupies. PATT uses $\nu = 63$ at $p = 128$ and $\nu = 255$ at $p = 512$; both are odd. Hence their raw finite-start ratios strictly exceed the exact Bessel ratio for every $x > 0$.

Theorem 5 (High-concentration parity phase)**.** *For $M = 2\nu$ and $x \to \infty$,*

$$\widetilde{R}_{\nu,2\nu}(x) = \frac{2x}{(\nu+1)(3\nu+1)} + O(x^{-1}) \to \infty \qquad (\nu \textit{ odd}), \tag{26}$$

$$\widetilde{R}_{\nu,2\nu}(x) = \frac{\nu(3\nu+2)}{2x} + O(x^{-3}) \to 0 \qquad (\nu \textit{ even}), \tag{27}$$

$$\widetilde{\Phi}_{\nu,2\nu}(x) = x - (\nu + 3/2)\log x + C_\nu + O(x^{-1}), \tag{28}$$

$$\widetilde{\Phi}'_{\nu,2\nu}(x) = 1 - \frac{\nu + 3/2}{x} + O(x^{-2}). \tag{29}$$

where the last two identities hold for odd ν.

The raw ratio therefore diverges for the odd orders used by PATT, while both the finite-forward derivative and $R_\nu = 1 - (\nu + 1/2)/x + O(x^{-2})$ approach one. By continuity, the raw ratio crosses one at least once. The concentration scan in Section 5 evaluates this crossing behavior.

Since $p = 2\nu + 2$, every $p \equiv 0 \pmod 4$ gives odd ν and hence all-axis overshoot plus high-concentration divergence; this includes $p = 128, 512$.

D.3.5 Clipping is not a potential repair

Define the finite-potential error and the two idealized $\epsilon = 0$ signed residuals by

$$\begin{aligned} E_{\nu,M} &= \widetilde{\Phi}_{\nu,M} - \Phi_\nu, \\ D_{\mathrm{raw}} &= \widetilde{\Phi}'_{\nu,M} - \widetilde{R}_{\nu,M}, \\ D_{\mathrm{clip}} &= \widetilde{\Phi}'_{\nu,M} - \overline{R}_{\nu,M}. \end{aligned} \tag{30}$$

Since $\Phi'_\nu = R_\nu$,

$$D_{\mathrm{raw}} = E'_{\nu,M} - (\widetilde{R}_{\nu,M} - R_\nu), \quad D_{\mathrm{clip}} = E'_{\nu,M} - (\overline{R}_{\nu,M} - R_\nu). \tag{31}$$

For the ideal finite-forward score $\widetilde{q}_j = \widetilde{\Phi}_{\nu,M}(r_j) - \widetilde{\Phi}_{\nu,M}(\kappa_j)$, the exact identity is

$$\boxed{\widetilde{q}_j - q_j = E_{\nu,M}(r_j) - E_{\nu,M}(\kappa_j).} \tag{32}$$

Appendix D.6 gives the literal ϵ correction, raw/clipped integral identities, and the logarithmic high-concentration finite-forward score error. Table 4 contrasts captured-state score fidelity with backward coherence.

D.4 Proof of the finite-start sign theorem

For $T_i(u) = (2i + x^2 u)^{-1}$, direct subtraction gives the one-step Möbius identity

$$T_i(a) - T_i(b) = -\frac{x^2(a-b)}{(2i + x^2 a)(2i + x^2 b)}. \tag{33}$$

The denominators are strictly positive for the recurrence sequences in (23). Apply (33) successively for $i = M, M-1, \ldots, \nu+1$. Each step contributes one negative sign and one positive factor, yielding exactly (24). Because the original terminal is $\widetilde{u}_M = 0$ while $u_M = I_{M+1}(x)/(xI_M(x)) > 0$, multiplication by $x > 0$ gives (25). Substituting $M = 2\nu$ completes the proof of Theorem 4.

D.5 High-concentration asymptotics

Write $y = x^2$. Starting from the even index $M = 2\nu$, the finite recurrence alternates between

$$\widetilde{u}_j = \frac{c_j}{y} + O(y^{-2}) \quad \text{at even } j, \qquad \widetilde{u}_j = a_j + O(y^{-1}) \quad \text{at odd } j, \tag{34}$$

where $a_j, c_j > 0$ except for the terminal coefficient $c_M = 0$. Indeed, if i is even, substitution in $\widetilde{u}_{i-1} = (2i + y\widetilde{u}_i)^{-1}$ gives $a_{i-1} = (2i + c_i)^{-1}$; if i is odd, it gives $c_{i-1} = a_i^{-1}$. Hence

$$c_{i-2} = c_i + 2i. \tag{35}$$

For odd ν,

$$a_\nu^{-1} = 2 \sum_{i=\nu+1,\nu+3,\ldots,2\nu} i = \frac{(\nu+1)(3\nu+1)}{2}, \tag{36}$$

whereas for even ν,

$$c_\nu = 2 \sum_{i=\nu+2,\nu+4,\ldots,2\nu} i = \frac{\nu(3\nu+2)}{2}. \tag{37}$$

Multiplication by x proves (26) and (27). Because the recurrence has finite depth, the rational expansions preserve the displayed remainder orders.

To obtain the forward asymptotic, telescope

$$\prod_{j=0}^{\nu-1} \widetilde{u}_j = \frac{b_\nu}{x^\nu b_0}, \qquad \widetilde{\Phi}_{\nu,2\nu} = \log I_0 + \sum_{j=0}^{\nu-1} \log \widetilde{u}_j. \tag{38}$$

For odd ν, the range $j = 0, \ldots, \nu-1$ contains $(\nu+1)/2$ even indices. Their terms in (34) each contribute $-2\log x$ plus a constant; the remaining terms approach positive constants. Combining this with

$$\log I_0(x) = x - \frac{1}{2}\log x + C_0 + O(x^{-1}) \tag{39}$$

gives (28). The finite rational expansions and the standard Bessel expansion may be differentiated term by term, proving (29).

For comparison,

$$\Phi_\nu(x) = x - (\nu + 1/2)\log x + C_\nu^\star + O(x^{-1}), \tag{40}$$

and therefore

$$E_{\nu,2\nu}(x) = -\log x + C_\nu^E + O(x^{-1}), \qquad E'_{\nu,2\nu}(x) = -x^{-1} + O(x^{-2}). \tag{41}$$

For all sufficiently large x, (26) implies $\overline{R}_{\nu,2\nu} = 1$. Thus

$$D_{\text{clip}} = -\frac{\nu + 3/2}{x} + O(x^{-2}), \tag{42}$$

$$\overline{R}_{\nu,2\nu} - R_\nu = \frac{\nu + 1/2}{x} + O(x^{-2}), \tag{43}$$

$$\widetilde{\Phi}'_{\nu,2\nu} - R_\nu = -\frac{1}{x} + O(x^{-2}). \tag{44}$$

This completes the proof of Theorem 5.

D.6 Finite-forward endpoint and defect identities

By definition,

$$\widetilde{q}_j - q_j = [\widetilde{\Phi}_{\nu,M}(r_j) - \Phi_\nu(r_j)] - [\widetilde{\Phi}_{\nu,M}(\kappa_j) - \Phi_\nu(\kappa_j)], \tag{45}$$

which is (32). The fundamental theorem of calculus and (31) give

$$\widetilde{q}_j - q_j = \int_{\kappa_j}^{r_j} (\widetilde{R}_{\nu,M} - R_\nu)\,\mathrm{d}x + \int_{\kappa_j}^{r_j} D_{\text{raw}}\,\mathrm{d}x, \tag{46}$$

$$= \int_{\kappa_j}^{r_j} (\overline{R}_{\nu,M} - R_\nu)\,\mathrm{d}x + \int_{\kappa_j}^{r_j} D_{\text{clip}}\,\mathrm{d}x. \tag{47}$$

For odd ν with both endpoints in the high-concentration regime,

$$\widetilde{q}_j - q_j = \log\frac{\kappa_j}{r_j} + O(\kappa_j^{-1} + r_j^{-1}). \tag{48}$$

Substitution of (41) at both endpoints gives (48).

For completeness, the literal source-level stabilizer obeys

$$-\nu\log(x+\epsilon) = -\nu\log x - \nu\log(1+\epsilon/x). \tag{49}$$

Thus

$$\widetilde{\Phi}^\epsilon_{\nu,M}(x) = \widetilde{\Phi}_{\nu,M}(x) - \nu\log(1+\epsilon/x), \tag{50}$$

and the literal endpoint error is

$$\widetilde{q}^\epsilon_j - q_j = E_{\nu,M}(r_j) - E_{\nu,M}(\kappa_j) - \nu[\log(1+\epsilon/r_j) - \log(1+\epsilon/\kappa_j)]\,. \tag{51}$$

Taking an endpoint difference proves (51); differentiation gives the additional $\nu\epsilon/[x(x+\epsilon)]$ term. The extra term quantifies the source stabilizer separately from finite-start error.

D.7 Proof of the AR uniform certificate

We prove Theorem 2 using a Riccati residual and constant barriers.

Set $\varepsilon = \nu^{-1}$, $x = \nu z$, and

$$H_\varepsilon(z) = \frac{I_{\nu+1}(\nu z)}{I_\nu(\nu z)}. \tag{52}$$

The adjacent-order Bessel ratio satisfies

$$\varepsilon H_\varepsilon'(z) = 1 - \frac{2+\varepsilon}{z} H_\varepsilon(z) - H_\varepsilon(z)^2. \tag{53}$$

Let $S = \sqrt{1+z^2}$ and define

$$\begin{aligned} r_0(z) &= \frac{z}{1+S}, \\ r_1(z) &= -\frac{z}{2S^2}, \\ r_2(z) &= \frac{z(4-z^2)}{8S^5}, \\ a_\varepsilon(z) &= r_0(z) + \varepsilon r_1(z) + \varepsilon^2 r_2(z). \end{aligned} \tag{54}$$

Substituting $x = \nu z$ shows that $a_\varepsilon(z) = A_{R,\nu}(x)$.

Define the residual obtained by inserting a_ε into (53):

$$\mathcal{E} = 1 - \frac{2+\varepsilon}{z} a_\varepsilon - a_\varepsilon^2 - \varepsilon a_\varepsilon'. \tag{55}$$

Direct simplification gives the exact identity

$$\begin{aligned} \mathcal{E} = \; & -\varepsilon^3 \frac{z^4 - 10z^2 + 4}{4(1+z^2)^{7/2}} \\ & -\varepsilon^4 \frac{z^2(z^2-4)^2}{64(1+z^2)^5}. \end{aligned} \tag{56}$$

For $u = z^2 \geq 0$,

$$|u^2 - 10u + 4| \leq 4(1+u)^2, \qquad u(u-4)^2 \leq 16(1+u)^3. \tag{57}$$

Consequently,

$$|\mathcal{E}(z,\varepsilon)| \leq \varepsilon^3(1 + \varepsilon/4). \tag{58}$$

Let $e = H_\varepsilon - a_\varepsilon$. Subtracting the two Riccati equations yields

$$\varepsilon e' = \mathcal{E} - Ke - e^2, \tag{59}$$

where

$$\begin{aligned} K &= \frac{2+\varepsilon}{z} + 2a_\varepsilon \\ &= \frac{2S}{z} + \frac{\varepsilon}{zS^2} + 2\varepsilon^2 r_2(z). \end{aligned} \tag{60}$$

When $0 < z \leq 2$, $r_2(z) \geq 0$. For $z \geq 2$,

$$-r_2(z) = \frac{z(z^2-4)}{8S^5} \leq \frac{1}{40}. \tag{61}$$

Therefore, for $0 < \varepsilon \leq 9/10$,

$$K \geq 2 - \frac{\varepsilon^2}{20} \geq \frac{3919}{2000}. \tag{62}$$

Choose the constant barrier $b = \varepsilon^3$. At $e = b$, using (58) and (62), the right-hand side of (59) is strictly negative. At $e = -b$, it is bounded below by

$$\varepsilon^3 \left(-\frac{49}{40} + \frac{3919}{2000} - \frac{729}{1000} \right) = \frac{11}{2000}\varepsilon^3 > 0. \tag{63}$$

Thus the vector field points into the interval $[-b, b]$ on both boundaries.

It remains to initialize the barrier argument at the singular endpoint. The small-z expansions are

$$H_\varepsilon(z) = \frac{z}{2(1+\varepsilon)} + O(z^3) \tag{64}$$

and

$$a_\varepsilon(z) = \frac{z}{2}(1 - \varepsilon + \varepsilon^2) + O(z^3). \tag{65}$$

Hence

$$e(z) = -\frac{\varepsilon^3}{2(1+\varepsilon)} z + O(z^3), \tag{66}$$

so e starts strictly between the two barriers. If a first contact with the upper barrier existed, it would require $e' \geq 0$, contradicting the inward direction above; a first contact with the lower barrier similarly requires $e' \leq 0$ and gives a contradiction. Therefore $|e(z)| < \varepsilon^3$ for all $z > 0$. Both the exact ratio and AR extend continuously to zero with value zero, yielding the non-strict result at $z = 0$. Since $\varepsilon \leq 9/10$ is equivalent to $\nu \geq 10/9$, Theorem 2 follows. □

Corollary 6 (Complexity–accuracy reversal)**.** *For integer feature dimension $p = 2\nu + 2$, the original start $M = 2\nu = p - 2$ executes exactly $p - 2$ dependent recurrence updates. Under Theorem 2, our recurrence-free ratio approximation satisfies*

$$\sup_{x \geq 0} |R_\nu(x) - A_{R,\nu}(x)| \leq \nu^{-3} = \frac{8}{(p-2)^3}. \tag{67}$$

Thus the original evaluator becomes sequentially deeper as p grows, while the FR/AR exact-arithmetic certificate tightens. Both statements concern the evaluator as a function of feature dimension.

D.8 Proof of the PATT-specific certificate

The two-view augmentation loss satisfies

$$|\widehat{L}_{\mathrm{ISAC}} - L_{\mathrm{ISAC}}| \leq \overline{\delta}^{(2)} + \overline{\delta}^{(3)}. \tag{68}$$

Under Theorem 3, put $\eta = (\tau\nu^3)^{-1}$. For each sample and view, the adjusted Softmax distributions satisfy

$$\mathrm{TV}(P_i^{(v)}, \widehat{P}_i^{(v)}) \leq \tanh(\delta_i^{(v)}/2), \tag{69}$$

an exact adjusted top-two margin greater than $2\delta_i^{(v)}$ preserves the per-view score prediction, and

$$\left\| \nabla_{f_i^{(v)}} \widehat{L}_{\mathrm{ISAC}} - \nabla_{f_i^{(v)}} L_{\mathrm{ISAC}} \right\| \leq \frac{1}{2B} \left[\frac{2}{\tau} \tanh\left(\frac{\delta_i^{(v)}}{2} \right) + 2\eta \right]. \tag{70}$$

Corollary 7 (Head, medium, and tail groups)**.** *For a nonempty sample subset G, let $L_G^{(v)} = |G|^{-1} \sum_{i \in G} \ell_i^{(v)}$ and $L_G = (L_G^{(2)} + L_G^{(3)})/2$. Then*

$$|\widehat{L}_G^{(v)} - L_G^{(v)}| \leq \frac{2}{|G|} \sum_{i \in G} \delta_i^{(v)}, \tag{71}$$

$$|\widehat{L}_G - L_G| \leq \frac{1}{|G|} \sum_{i \in G} (\delta_i^{(2)} + \delta_i^{(3)}). \tag{72}$$

For $G_i^{(v)} = 2\tau^{-1} \tanh(\delta_i^{(v)}/2) + 2\eta$,

$$\left\| \nabla_{f_i^{(v)}} \widehat{L}_G - \nabla_{f_i^{(v)}} L_G \right\| \leq \frac{G_i^{(v)}}{2|G|}, \qquad i \in G. \tag{73}$$

Since $\Phi_\nu' = R_\nu$ and $F_{R,\nu}' = A_{R,\nu}$,

$$q_j(f) - \widehat{q}_j(f) = \int_{\kappa_j}^{r_j(f)} [R_\nu(x) - A_{R,\nu}(x)] \, \mathrm{d}x. \tag{74}$$

Theorem 2 and the reverse triangle inequality yield

$$|q_j(f) - \widehat{q}_j(f)| \le \frac{|r_j(f) - \kappa_j|}{\nu^3} \le \frac{\|f\|}{\tau\nu^3}. \tag{75}$$

For one sample and view, let $s_j = q_j + b_j$ and $\widehat{s}_j = s_j + e_j$, where $|e_j| \le \delta$. Monotonicity of log-sum-exp under coordinatewise bounds gives

$$\min_j e_j \le \log \sum_j e^{s_j + e_j} - \log \sum_j e^{s_j} \le \max_j e_j. \tag{76}$$

After subtracting the target-score perturbation,

$$|\widehat{\ell}_y - \ell_y| \le \max_j e_j - \min_j e_j \le 2\delta. \tag{77}$$

The biases b_j are arbitrary and the argument introduces no sum over classes. Averaging (77) over the batch and the two views proves (68). Because TLA and OE are unchanged, their difference is zero and (12) follows.

Let $P = \mathrm{softmax}(s)$ and $\widehat{P} = \mathrm{softmax}(s + e)$. Reweighting P_j by factors in $[e^{-\delta}, e^{\delta}]$ gives $\mathrm{TV}(P, \widehat{P}) \le \tanh(\delta/2)$. A pairwise adjusted-score margin changes by at most 2δ, which proves the prediction statement.

For $v_j = \kappa_j \mu_j + f/\tau$ and $r_j = \|v_j\| > 0$,

$$\nabla_f q_j = \frac{1}{\tau} R_\nu(r_j) \frac{v_j}{r_j}, \qquad \nabla_f \widehat{q}_j = \frac{1}{\tau} A_{R,\nu}(r_j) \frac{v_j}{r_j}. \tag{78}$$

Theorem 2 therefore gives $\|\nabla_f q_j - \nabla_f \widehat{q}_j\| \le \eta$; both gradients extend continuously at $r_j = 0$. Writing a one-sample cross-entropy gradient as the Softmax-weighted score-gradient average minus the target gradient, adding and subtracting the exact gradients under $\widehat{P}$, and using $\left\|\widehat{P} - P\right\|_1 \le 2\tanh(\delta/2)$ yields

$$\left\|\nabla_f \widehat{\ell}_y - \nabla_f \ell_y\right\| \le \frac{2}{\tau}\tanh(\delta/2) + 2\eta. \tag{79}$$

The batch mean contributes $1/B$ and the two-view average contributes $1/2$, proving (70). For a nonempty group G, (77) averaged within one view gives (71); averaging its two view-specific bounds gives (72). The single-sample, single-view gradient bound is $G_i^{(v)}$. In L_G, the feature $f_i^{(v)}$ appears with coefficient $1/(2|G|)$, which proves (73).

E Numerical validation and mechanism analysis

E.1 Frozen-state mechanism validation

Frozen-state diagnostics hold the PATT features, augmented views, and class state fixed, isolating the backward rule from state updates and geometry changes.

Concentration and depth scans. We vary feature dimension, scaled concentration, and recurrence depth as in Figure 4, and test adjacent odd and even orders. Independent high-precision and finite-difference checks reproduce the predicted parity and finite-forward derivative. Increasing recurrence depth can improve raw-ratio fidelity while worsening clipped coherence, consistent with their distinct definitions. The scan covers $x/\nu \in \{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ and $M/\nu \in \{2, 3, 4, 6, 8\}$ at $p \in \{128, 256, 512, 1024\}$. The marker locates the median query radius from one captured PATT state ($p = 512$, $M/\nu = 2$); its query range is $r/\nu \in [392.35, 393.11]$.

Frozen-state backward comparison. Original recurrence uses the saved raw ratio and unit-clipped backward; consistent recurrence differentiates the identical finite-start forward. Table 2 shows exact score/loss agreement but a nonzero radial defect and aggregated feature-gradient difference in both views. Thus changing backward alone exposes the local field mismatch without a forward confound.

Table 2: **Changing only the backward rule reveals a PATT coherence defect.** Original and consistent recurrence use the same frozen CIFAR100-LT state and two augmented views (seed 3407). Score and loss columns report absolute differences; gradient columns compare the complete adjusted cross-entropy feature gradients. $D_{\text{clip}} = \widetilde{\Phi}' - g = -d$ is the signed radial residual.

View	Max score $\lvert\Delta\rvert$	Loss $\lvert\Delta\rvert$	Max $\lvert D_{\text{clip}}\rvert$	Gradient relative ℓ_2	Gradient max $\lvert\Delta\rvert$
2	0	0	2.55×10^{-3}	2.56×10^{-3}	1.40×10^{-5}
3	0	0	2.55×10^{-3}	2.56×10^{-3}	1.35×10^{-5}

Table 3: **Finite-recurrence asymptotics follow the predicted parity.** At $x = 10^9$ and $M = 2\nu$, odd and even orders are compared with their analytic limits using 100-digit arithmetic. The scaled quantity is $\widetilde{R}/x$ for odd ν and $x\widetilde{R}$ for even ν; sign refers to $\widetilde{R} - R_\nu$.

ν	p	Raw $\widetilde{R}$	Sign	Phase	Scaled	Limit	Rel. error
63	128	1.645×10^{5}	$+$	linear	1.645×10^{-4}	1.645×10^{-4}	1.2×10^{-11}
64	130	6.208×10^{-6}	$-$	inverse	6.208×10^{3}	6.208×10^{3}	1.3×10^{-11}
255	512	1.020×10^{4}	$+$	linear	1.020×10^{-5}	1.020×10^{-5}	3.2×10^{-9}
256	514	9.856×10^{-5}	$-$	inverse	9.856×10^{4}	9.856×10^{4}	3.3×10^{-9}

Odd/even dimensional behavior. Table 3 confirms linear growth for odd orders and inverse decay for even orders, in agreement with Equations (26)–(27).

Robustness across precision and geometry. We cross FP32/FP64 geometry with expanded/factorized evaluation on the same frozen view. Within every cell, original and consistent recurrence retain identical scores and losses, while their feature gradients differ. The mismatch therefore persists across the precision and geometry choices.

E.2 Fidelity benchmarks and numerical references

Table 4 separates numerical score and derivative discrepancies at captured learning states.

Table 4: **Score fidelity and backward coherence at captured PATT states.** Maximum absolute discrepancies over six CIFAR100-LT views. Score and radial fidelity use the exact Bessel reference; coherence compares the supplied backward with the derivative of its own finite forward.

Comparison	Reference	Maximum discrepancy
Finite-potential score	Exact score	3.32×10^{-3}
AR/FR same-state score	Exact score	1.17×10^{-9}
Clipped backward radial factor	Exact Bessel ratio	2.55×10^{-3}
Forward–backward coherence defect	Finite-forward derivative	2.56×10^{-3}

The three tables address complementary levels of fidelity. Table 4 evaluates captured learning states, Table 5 isolates the adjacent-order ratio over a controlled concentration grid, and Table 6 evaluates the endpoint potential differences that enter the learning score. Together they separate state-dependent implementation effects from the intrinsic approximation error of AR/FR.

Numerical references and implementation. The dimension study uses seven dimensions and 61 common scaled concentrations per dimension. Reference Bessel values are evaluated with mpmath at 60 decimal digits. Endpoint scores use $\Phi_\nu(x+1) - \Phi_\nu(x)$, so the score certificate is ν^{-3}. The production AR/FR implementation uses cancellation-resistant endpoint differences in FP64. Stable log-Miller uses the original start 2ν and normalization at order zero; logarithmic arithmetic isolates the finite-start effect from intermediate overflow. Its ratio column uses the clipped adjacent-order ratio. For CUSF, adjacent-order values define the ratio. CUSF and log-Miller errors use FP64 outputs

Table 5: **Ratio fidelity across feature dimensions.** Maximum absolute errors use 61 log-spaced concentrations $x/\nu \in [10^{-3}, 10^{3}]$ per dimension and a 60-digit Bessel reference. CUSF forms a ratio from adjacent log-Bessel values in CPU FP64; AR/FR is evaluated in 60-digit arithmetic. The certificate ν^{-3} covers the full nonnegative axis.

Dimension	CUSF	AR/FR	AR/FR certificate
64	3.37×10^{-12}	2.74×10^{-6}	3.36×10^{-5}
128	2.38×10^{-11}	3.29×10^{-7}	4.00×10^{-6}
256	2.29×10^{-11}	4.03×10^{-8}	4.88×10^{-7}
512	1.85×10^{-11}	4.99×10^{-9}	6.03×10^{-8}
1024	2.45×10^{-11}	6.21×10^{-10}	7.49×10^{-9}
2048	8.87×10^{-11}	7.74×10^{-11}	9.34×10^{-10}
4096	5.76×10^{-10}	9.66×10^{-12}	1.17×10^{-10}

Table 6: **Fidelity of unit-length potential differences.** Maximum absolute errors in $\Phi_\nu(x+1) - \Phi_\nu(x)$ use the common 61-point concentration grid at each dimension. Log-Miller retains the finite start; CUSF subtracts separately evaluated potentials; AR/FR uses its paired analytic potential. All errors use the high-precision Bessel reference.

Dimension	Log-Miller	CUSF	AR/FR
64	3.36×10^{-4}	1.89×10^{-12}	2.76×10^{-6}
128	8.29×10^{-5}	8.47×10^{-12}	3.30×10^{-7}
256	2.06×10^{-5}	2.31×10^{-11}	4.04×10^{-8}
512	5.13×10^{-6}	9.72×10^{-12}	5.00×10^{-9}
1024	1.28×10^{-6}	3.21×10^{-11}	6.21×10^{-10}
2048	2.11×10^{-7}	9.67×10^{-11}	7.74×10^{-11}
4096	7.25×10^{-9}	1.73×10^{-10}	9.67×10^{-12}

compared with the common reference. The AR/FR approximation errors are evaluated with 60-digit arithmetic.

E.3 Learning-level stress tests

Class count and prior imbalance. We generate eight normalized 128-dimensional queries and a pool of 2000 class directions with seed 3407. Concentrations span 10 to 10,000; smaller class sets are prefixes of the same randomly ordered pool. Each class set uses log-linear priors with largest-to-smallest ratios 1, 100, and 1000. All 15 configurations satisfy the score bound 4.00×10^{-5} and the mean cross-entropy bound 8.00×10^{-5}.

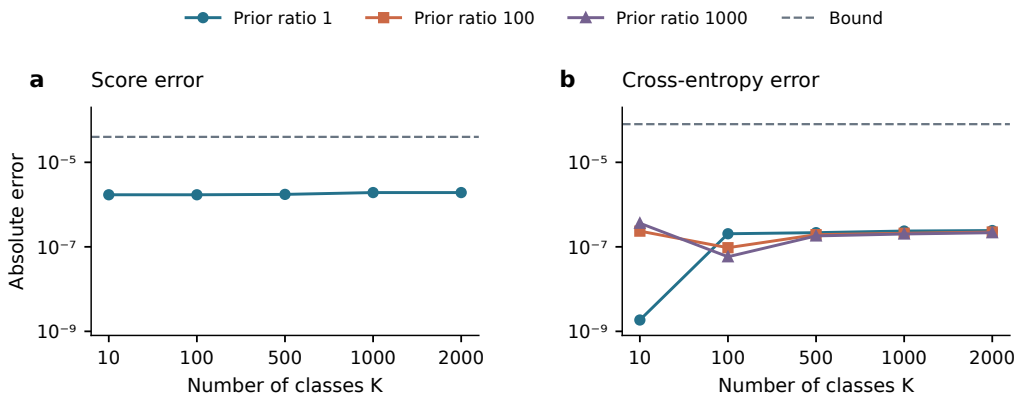


Figure 8: **Score and loss errors remain below their certificates as class count varies.** The 15 fixed-state configurations combine five class counts with prior ratios 1, 100, and 1000. (a) The three maximum score-error curves coincide and are drawn once. (b) Mean cross-entropy error is shown separately for each prior ratio. Dashed lines denote the respective theoretical bounds. Class counts are equally spaced discrete settings.

E.4 Measured PATT update fields

Figure 9 separates the supplied update from the derivative of the same finite forward on a real feature slice.

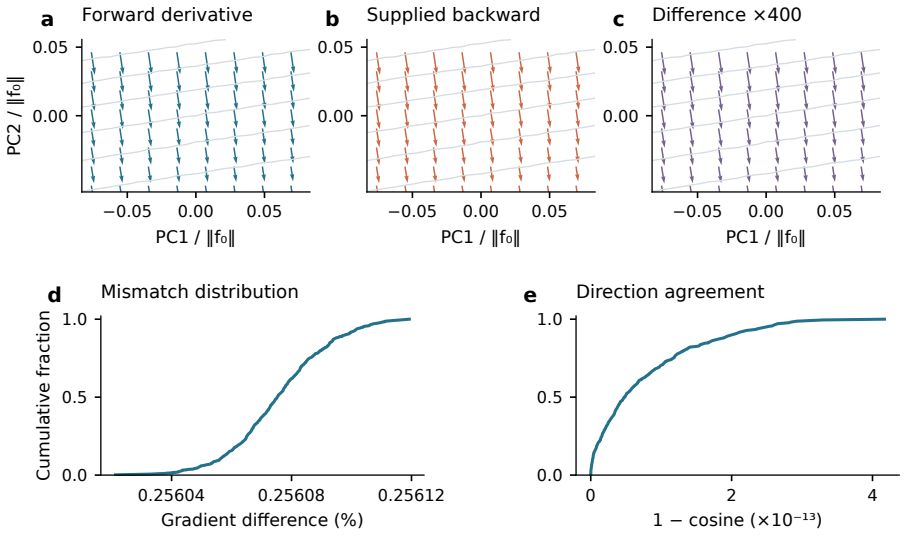


Figure 9: **A frozen PATT state exhibits mainly gradient-magnitude distortion.** The first CIFAR-10 query is varied over a 25×25 slice along two principal feature directions, with class statistics fixed. Scores and losses agree bitwise at every location. (a–b) Negative gradients from the finite forward and supplied backward, using common contours and arrow scales. (c) Their difference magnified $400\times$; each field displays 64 interior arrows. (d–e) Empirical cumulative distributions over all 625 projected-gradient pairs. Relative mismatch has median 0.2561%; $1 - \cos$ resolves the small directional difference.

E.5 CLASSWISE GEOMETRY AND FULL-UPDATE BEHAVIOR

We vary the classwise concentration ratio in the two-class construction of Appendix D.2. The resulting Jacobian antisymmetry follows the analytic prediction and vanishes at equal concentrations. A complementary fixed-forward intervention varies the mismatch strength and yields the predicted linear response in both gradient gap and antisymmetry (Figure 10).

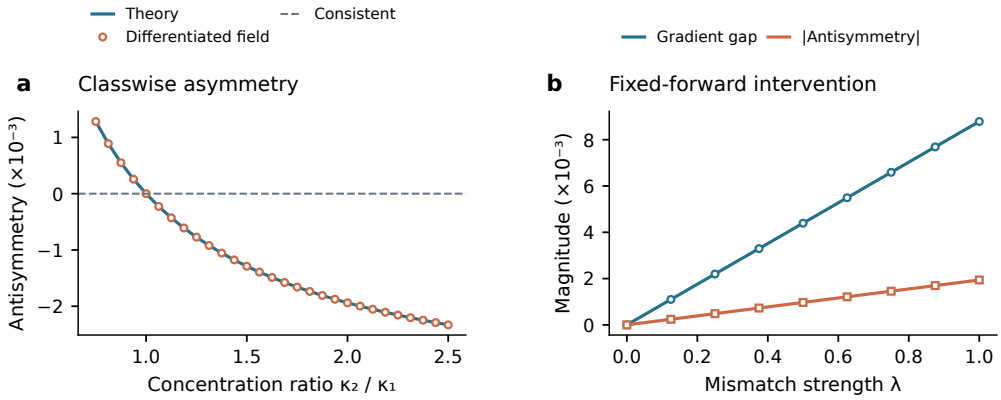


Figure 10: **Classwise mismatch controls the local Jacobian obstruction.** (a) Signed antisymmetry versus concentration ratio at 29 constructed states: lines show the analytic prediction and circles the differentiated field; the consistent field remains zero. (b) Gradient-gap norm and absolute antisymmetry across nine mismatch strengths, with predictions shown as lines and computed values as markers. All interventions retain forward loss $\log 2$.

Clipping and the high-concentration boundary. For each $\nu \in \{63, 255\}$ we evaluate concentrations $x/\nu \in [0.25, 10^5]$ at $M = 2\nu$. The clipped defect eventually decreases as $(\nu + 3/2)/x$, despite divergence of the raw ratio. This distinguishes raw ratio fidelity from clipped coherence. The saved PATT query interval $x/\nu \in [392.35, 393.11]$ lies inside the expanded range.

Full feature gradients and removal of overall scaling. We reconstruct all 256 saved PATT queries in FP64 at the same fixed class bank, using the complete 512-dimensional features, original labels, priors, and saved normalizer offsets. For consistent gradient a_i and supplied gradient b_i, set $c_i^* = a_i^\top b_i / \|a_i\|^2$ and measure $\|b_i - c_i^* a_i\| / \|b_i\|$. At this frozen state, the discrepancy is dominated by gradient magnitude: optimal per-query scalar alignment leaves a much smaller residual (Figure 11).

Full-parameter update. A one-step replay from a shared model, optimizer, and class state preserves identical forward scores and losses but yields different total parameter gradients and optimizer updates. A global rescaling of the consistent gradient does not eliminate the residual, showing that the effect is not explained by a single multiplicative factor.

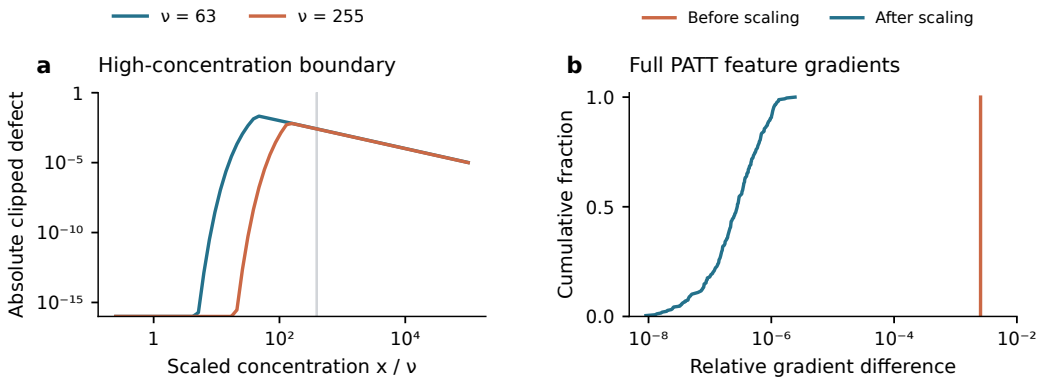


Figure 11: **Concentration and gradient scaling delimit the coherence defect.** (a) Absolute clipped defect over 130 conditions; dashed segments show the asymptote $(\nu + 3/2)/x$, and the grey interval marks saved PATT query radii. Values below 10^{-16} share a display floor. (b) Cumulative distributions for 256 frozen PATT queries in the full feature space with FP64 geometry. Relative differences use the consistent gradient norm before scaling and the supplied gradient norm after optimal per-query scaling.

F Structure-aware execution and efficiency

F.1 Paired evaluation in the host learning system

AR/FR replaces the repeated vMF score and its radial derivative. The host class-state estimand, detached-statistics semantics, view order, priors, auxiliary losses, encoder, and optimizer retain their original definitions.

Matched training realizations. The three-arm CIFAR experiment shares FP32 expanded geometry and backbone, FP64 scalar evaluation, and disabled autocast. Original and consistent recurrence use the same finite forward. AR/FR replaces the score pair; all branches retain the common class-bank update.

F.1.1 Finite-sample state identities

Lemma 8 (Finite-sample resultant decomposition)**.** *For realized $z_1, \dots, z_n \in \mathbb{R}^p$, $n \geq 1$, write $R_n = \|n^{-1} \sum_a z_a\|$. Then*

$$R_n^2 = \frac{1}{n^2} \sum_a \|z_a\|^2 + \frac{1}{n^2} \sum_{a \neq b} z_a^\top z_b. \tag{80}$$

If the vectors are i.i.d. with mean m and covariance Σ, then

$$\mathbb{E} R_n^2 = \|m\|^2 + \frac{\operatorname{tr} \Sigma}{n}; \tag{81}$$

for unit-norm features this is $n^{-1} + (1 - n^{-1}) \|m\|^2$.

Proposition 9 (Exact sum–count state reduction)**.** *For current count $N \geq 0$ and mean a, and batch count $m \geq 0$ and sum s, the original update satisfies*

$$a^+ = \frac{N}{N+m} a + \frac{s}{N+m} = \frac{Na + s}{N+m}, \qquad N^+ = N + m, \tag{82}$$

whenever $N + m > 0$; if $m = 0$, the state is unchanged.

Expanding the squared norm gives

$$\left\| \frac{1}{n} \sum_{a=1}^n z_a \right\|^2 = \frac{1}{n^2} \sum_{a=1}^n \sum_{b=1}^n z_a^\top z_b. \tag{83}$$

Separating diagonal and off-diagonal terms proves (80). If $z_1, \dots, z_n$ are i.i.d. with mean m and covariance Σ, then the sample mean has mean m and covariance Σ/n. The identity $\mathbb{E}\|X\|^2 = \|\mathbb{E}X\|^2 + \operatorname{tr}\operatorname{Cov}(X)$ therefore proves (81). For unit-norm vectors, $\operatorname{tr}\Sigma = \mathbb{E}\|z\|^2 - \|m\|^2 = 1 - \|m\|^2$. No independence statement is used for the realized decomposition.

For Proposition 9, the original update first forms the batch average s/m when $m > 0$ and uses weight $m/(N+m)$. Direct substitution gives

$$\left(1 - \frac{m}{N+m}\right) a + \frac{m}{N+m}\frac{s}{m} = \frac{Na+s}{N+m}. \tag{84}$$

When $m = 0$, the source sets the weight to zero and leaves the state unchanged. When $N = 0 < m$, the formula reduces to s/m. When $N = m = 0$, the data-derived mean is undefined and the source keeps its initializer. Thus class sums and counts are sufficient to reproduce this deterministic state transition.

F.2 Stable potential differences and factorized execution

The original per-view state-update order and stop-gradient semantics are preserved.

Direct sufficient-statistic accumulation. Class sums and counts are accumulated directly rather than through a dense $B \times K \times p$ assignment expansion. This preserves the sufficient statistics in Proposition 9 while reducing intermediate storage.

Factorized vMF geometry. Feature norms and a $B \times K$ projection matrix determine the score geometry without materializing classwise resultant vectors. Stable endpoint differences avoid cancellation between large potentials; the expressions below are evaluated in double precision.

Fixed-depth compiled execution. The paired evaluator has a fixed operation structure suitable for compilation and GPU graph replay. This reduces launch overhead in addition to the storage savings from statistics and geometry; compilation is excluded from steady-state timing.

The analytic potential is

$$\begin{aligned} F_{R,\nu}(x) = s(x) - \nu \log(\nu + s(x)) - \frac{1}{2}\log s(x) \\ + \frac{3x^2 - 2\nu^2}{24 s(x)^3}, \qquad s(x) = \sqrt{\nu^2 + x^2}, \end{aligned} \tag{85}$$

The stable endpoint increment is

$$d_j = r_j^2 - \kappa_j^2 = \frac{2\kappa_j \mu_j^\top f}{\tau} + \frac{\|f\|^2}{\tau^2}, \quad \Delta_j = \frac{d_j}{s(r_j) + s(\kappa_j)}, \tag{86}$$

With $s_1 = s(r_j)$ and $s_0 = s(\kappa_j)$, the cancellation-resistant score is

$$\begin{aligned} \widehat{q}_j = \Delta_j - \nu \log\left(1 + \frac{\Delta_j}{\nu + s_0}\right) - \frac{1}{2}\log\left(1 + \frac{\Delta_j}{s_0}\right) \\ - \frac{\Delta_j}{8 s_1 s_0} + \frac{5\nu^2 \Delta_j (s_1^2 + s_1 s_0 + s_0^2)}{24 s_1^3 s_0^3}. \end{aligned} \tag{87}$$

F.3 Unified H100 local runtime and memory protocol

Figure 6 uses a fixed state and synthetic inputs with $B = 32$, $K = 1000$, $p = 1024$, and seed 3407. The local boundary includes two-view class-state updates, score geometry, cross-entropy, and backward to features; it excludes the backbone and data loading. Each method uses seven timed repetitions after three warmups, with state restoration outside measurement. Compilation is excluded from steady-state timing. Structure-aware AR/FR uses compiled statistics and score geometry with explicit CUDA Graph replay.

Memory is the peak allocated increment within the local execution boundary above prepared inputs and warmed modules, including persistent allocations created by explicit graph capture and the replay peak. This metric measures additional local storage rather than full-process VRAM. The five forward–backward measurements are summarized in Table 7.

Table 7: **Local H100 forward–backward costs.** Median runtime and incremental peak allocation under the protocol in Appendix F.3.

Numerical realization	Runtime (ms)	Peak increment (MiB)
Original recurrence	266.40	504.89
Consistent recurrence	412.80	1755.42
Log-domain Miller	120.28	504.89
AR/FR	4.90	508.12
Structure-aware AR/FR	0.206	32.00

Attribution of the combined gain. The dense-layout comparison isolates evaluator replacement: runtime falls from 266.40 to 4.90 ms ($54.4\times$), while the peak increment remains comparable (504.89 versus 508.12 MiB). Structure-aware execution then reduces runtime from 4.90 to 0.206 ms ($23.8\times$) and the peak increment to 32.00 MiB. Relative to original recurrence, the combined reductions are $1293\times$ in runtime and $15.8\times$ in memory. The second comparison measures statistics, geometry, and compiled execution jointly; it does not isolate their individual contributions.

G EXTENDED TASK RESULTS

G.1 COMPLETE TASK METRICS

Table 8 includes all three task metrics, complementing the compact main-text table.

Table 8: **Complete classification and OOD detection results.** Metrics are percentages; CIFAR results report mean ± standard deviation over three runs. Bold and underline indicate the best and second-best values within each setting.

Dataset	IF	Metric	Original recurrence	Consistent recurrence	AR/FR
CIFAR-10-LT	10	Acc. ↑	92.04 ± 0.17	$\underline{92.12} \pm 0.18$	$\mathbf{92.16} \pm 0.19$
		AUROC ↑	$\underline{93.07} \pm 0.13$	92.30 ± 0.27	$\mathbf{93.43} \pm 0.46$
		FPR95 ↓	$\underline{26.20} \pm 1.19$	29.52 ± 1.06	$\mathbf{25.52} \pm 1.01$
CIFAR-10-LT	50	Acc. ↑	87.30 ± 0.29	$\mathbf{87.58} \pm 0.35$	$\underline{87.42} \pm 0.15$
		AUROC ↑	90.18 ± 2.46	$\underline{90.95} \pm 1.57$	$\mathbf{91.02} \pm 1.13$
		FPR95 ↓	34.51 ± 6.96	$\underline{32.84} \pm 1.94$	$\mathbf{32.23} \pm 2.89$
CIFAR-10-LT	100	Acc. ↑	83.82 ± 0.21	$\underline{84.67} \pm 0.41$	$\mathbf{84.73} \pm 0.31$
		AUROC ↑	$\mathbf{90.66} \pm 0.13$	$\underline{89.97} \pm 0.39$	89.95 ± 0.34
		FPR95 ↓	$\mathbf{33.35} \pm 0.50$	$\underline{34.33} \pm 0.71$	35.15 ± 0.79
CIFAR-100-LT	10	Acc. ↑	$\underline{62.37} \pm 2.43$	62.12 ± 1.84	$\mathbf{62.56} \pm 2.03$
		AUROC ↑	$\underline{78.18} \pm 2.09$	78.08 ± 1.23	$\mathbf{78.44} \pm 1.53$
		FPR95 ↓	57.74 ± 3.43	$\underline{57.70} \pm 1.46$	$\mathbf{57.14} \pm 2.35$
CIFAR-100-LT	50	Acc. ↑	$\underline{53.26} \pm 0.19$	52.87 ± 0.31	$\mathbf{53.74} \pm 0.31$
		AUROC ↑	75.37 ± 0.68	$\underline{75.54} \pm 0.60$	$\mathbf{76.14} \pm 0.24$
		FPR95 ↓	$\underline{62.77} \pm 1.00$	64.12 ± 0.96	$\mathbf{62.59} \pm 0.96$
CIFAR-100-LT	100	Acc. ↑	$\mathbf{50.33} \pm 0.16$	49.48 ± 0.26	$\underline{49.66} \pm 0.16$
		AUROC ↑	$\mathbf{75.78} \pm 0.33$	$\underline{75.40} \pm 0.54$	75.21 ± 0.65
		FPR95 ↓	$\mathbf{63.56} \pm 0.25$	$\underline{64.58} \pm 0.62$	64.91 ± 0.65
ImageNet-LT	—	Acc. ↑	53.88	<u>54.11</u>	**54.34**
		AUROC ↑	76.19	<u>76.87</u>	**77.50**
		FPR95 ↓	79.66	<u>78.74</u>	**76.68**