SHARPNESS CORRECTIONS NEED NOT ACT IMMEDIATELY IN MOMENTUM-BASED SAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-Aware Minimization (SAM) evaluates gradients at perturbed parameters, but under momentum the resulting sharpness correction acts through an immediate zero-lag route and a delayed route carried by the momentum state. The central question is whether a generated correction must act immediately to remain useful. Temporal Feedback SAM (TF-SAM) isolates this choice by removing the zero-lag route while preserving correction generation and its momentum-mediated influence. An exact original-iterate identity shows that this intervention replaces the fresh correction at the target update with its momentum-weighted previous counterpart. A target-matched first-order criterion then shows that the preferred route depends on the corrections' relative alignment with the current descent direction. At regular training states, the normalized-SAM correction has an explicit Jacobian, and the two routes enter different coefficients of the local closed-loop characteristic matrix polynomial. Across the evaluated image and text settings, removing zero-lag feedback does not produce a consistent loss in predictive accuracy or aggregate corruption robustness; target-matched measurements show that the relative first-order value of the routes varies across regimes. Generating sharpness information and applying it immediately are therefore separable design decisions in momentum-based SAM, while holding correction generation fixed.

1 INTRODUCTION

Background: Sharpness-Aware Minimization (SAM) improves neural-network generalization by evaluating the gradient at locally perturbed parameters rather than at the current iterate (Foret et al., 2021). Subsequent methods modify perturbation geometry (Kwon et al., 2021; Kim et al., 2022), change how the sharpness-aware signal enters the update (Yue et al., 2023; Ye et al., 2026), or reduce and schedule the additional computation used to generate that signal (Liu et al., 2022; Du et al., 2022; Zhou et al., 2025). These methods primarily vary correction construction, weighting, or generation frequency. **Observation:** *We observe that momentum introduces a distinct temporal effect after the correction has already been generated.* In standard SAM with SGD momentum (w.r.t. M-SAM), the same correction acts immediately on the current update and also enters the momentum state, thereby influencing later updates without another sharpness computation. M-SAM therefore couples a fresh zero-lag route with a momentum-mediated delayed route.

Question: This **reveals** a question that is separate from perturbation construction (Becker et al., 2025) or correction-generation schedules (Liu et al., 2022; Du et al., 2022; Zhou et al., 2025): *must an already-generated sharpness correction act immediately, or can its temporal placement itself be controlled?*

The delay matters because the two routes act on different optimization states. A correction is generated from a particular parameter state and minibatch, whereas its momentum-mediated counterpart acts after the state has evolved and is additionally weighted by β. A fresh correction may therefore contribute favorably to the current descent direction, but it may also interact adversely with it; delaying the same generated information can attenuate such an adverse interaction while potentially weakening a favorable fresh contribution. **Motivation:** We therefore infer that immediate and delayed application need not be uniformly ordered, consistent with mechanistic views of SAM that emphasize optimization dynamics and stability (Andriushchenko & Flammarion, 2022;

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.

Bartlett et al., 2023; Agarwala & Dauphin, 2023; Long & Bartlett, 2024). **Analysis:** To analyze this effect, Temporal Feedback SAM (TF-SAM) removes only the zero-lag route while preserving correction generation and its momentum-mediated path.

Theoretical results. Our analysis establishes three properties of temporal routing:**1)** Exact original-iterate dynamics show that *TF-SAM replaces M-SAM's fresh correction* at the target update with its momentum-weighted previous counterpart without discarding the generated correction; **2)** A target-matched first-order criterion shows that *the preference for immediate or delayed routing is governed by* $A_{t,0} - \beta A_{t,1}$[1] and need not have a universal ordering; **3)** The normalized-SAM correction admits *an explicit local Jacobian* at regular training states. This yields *a general closed-loop matrix polynomial*, whose coefficients capture the distinct dynamical roles of immediate and delayed routing. This structural result requires neither positive correction response nor shared eigenvectors; an interpretable scalar stability specialization is deferred to the appendix.

Experimental findings. The experiments support these predictions from both performance and mechanism perspectives: **1)** Delayed-only routing largely preserves predictive accuracy at the *M-SAM level* across the evaluated image and text tasks; **2)** Aggregate robustness is likewise *largely preserved* after removing the zero-lag route; **3)** Target-matched measurements show that an adverse fresh correction can become *less adverse* when delayed, whereas delaying a favorable fresh correction can reduce its immediate benefit. These results show that sharpness corrections need not act at zero lag to remain useful, identify post-generation temporal placement as a separable design dimension of momentum-based SAM, and motivate adaptive routing based on a target-matched score.

Structured insight: Comprehensively, we conclude that post-generation temporal placement is a separable design dimension of momentum-based SAM. Rather than fixing the zero-lag route globally, this perspective motivates adaptive temporal routing in which a target-matched score could determine how much of a newly generated correction should act immediately versus through its momentum-mediated delayed path, with the potential to retain useful sharpness feedback while limiting unfavorable immediate interactions.

Our contributions are:

- **Why delay corrections?** Momentum gives each generated sharpness correction both immediate and delayed routes, making temporal placement a separable design choice.
- **How to delay corrections?** TF-SAM removes each fresh correction from the current update but retains it in momentum, so it first affects the parameter update one step later with momentum weight β.
- **What does delay bring?** Delayed routing largely preserves M-SAM level performance while reducing adverse fresh-correction effects in unfavorable regimes, motivating target-matched adaptive routing.

2 Background and Setup

The analysis considers SAM in the standard stochastic-optimization setting underlying modern neural-network training (Robbins & Monro, 1951). This setting is especially relevant for overparameterized networks, where optimization geometry, parameterization, and stability can materially affect training dynamics (Arora et al., 2018; Mulayoff et al., 2021; Cohen et al., 2021; Ahn et al., 2022; Arora et al., 2022).

Notation. Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^n$ be the training set, where x_i and y_i are the input and target of example i, respectively, and n is the number of training examples. Let $f(x; w)$ be a model with parameter vector $w \in \mathbb{R}^p$, where p is the number of parameters. Denote the loss of example i by $\ell_i(w) = \ell(y_i; f(x_i; w))$. At iteration t, let $\mathcal{B}_t \subset [n]$ be the sampled minibatch, where $[n] := \{1, \dots, n\}$ and $B := |\mathcal{B}_t|$ is the minibatch size, and define the minibatch

[1]Here, $A_{t,0}$ and $A_{t,1}$ are the expected inner products of the current population gradient with the current and one-step-delayed corrections, conditioned on the history $\mathcal{F}_t$; β is the momentum coefficient.

loss $\ell_t(w) = B^{-1} \sum_{i \in \mathcal{B}_t} \ell_i(w)$. Thus, t indexes the training iterations, w is the variable with respect to which gradients are taken, and ℓ_t is the loss used at iteration t. For the stochastic analysis, $\ell(w; \xi)$ denotes the loss induced by $f(\cdot; w)$ on a sampled example or minibatch ξ. In particular, $\ell(w; (x_i, y_i)) = \ell_i(w)$ and, when $\xi_t = \mathcal{B}_t$, $\ell(w; \xi_t) = \ell_t(w)$.

SAM correction. SAM is motivated by minimizing the worst-case minibatch loss in a neighborhood of the current parameters (Foret et al., 2021). Using this notation, the local SAM objective at step t is

$$\ell_t^{\mathrm{SAM}}(w; \rho) = \max_{\|\epsilon\|_2 \le \rho} \ell_t(w + \epsilon), \tag{1}$$

where $\epsilon \in \mathbb{R}^p$ is a parameter perturbation, $\rho > 0$ is the neighborhood radius, and $\|\epsilon\|_2 = \|\epsilon\|_2$ is its Euclidean norm. Solving this inner maximization exactly is generally expensive, so practical SAM uses the first-order maximizer of the linearized minibatch loss. Let w_t denote the parameter vector at the beginning of iteration t, and let ∇ denote the gradient with respect to w. The clean minibatch gradient is $g_t = \nabla \ell_t(w_t) \in \mathbb{R}^p$. For $g_t \neq 0$, the first-order adversarial perturbation is $\epsilon_t = \rho g_t / \|g_t\|_2$; when $g_t = 0$, $\epsilon_t = 0$. The resulting perturbed minibatch gradient is $\hat{g}_t = \nabla \ell_t(w_t + \epsilon_t) \in \mathbb{R}^p$. The clean and perturbed gradients are related by the decomposition in Eq. 2:

$$\hat{g}_t = g_t + c_t. \tag{2}$$

The vector $c_t := \hat{g}_t - g_t \in \mathbb{R}^p$ is the perturbation-induced sharpness correction. Thus g_t is the ordinary clean descent signal, while c_t is the local change in the minibatch gradient caused by ϵ_t. This paper studies how c_t enters the optimization dynamics.

Correction temporal routing. When SAM is combined with momentum, the generated correction naturally has both an immediate and a delayed route. The momentum state evolves as

$$m_t = \beta m_{t-1} + g_t + c_t, \tag{3}$$

and introduce a zero-lag sharpness gain α in the parameter update:

$$w_{t+1} = w_t - \eta_t \left[\beta m_{t-1} + g_t + \alpha c_t \right]. \tag{4}$$

Here $m_t \in \mathbb{R}^p$ is the momentum state, m_{t-1} is its value at the previous iteration, $\beta \in [0, 1)$ is the momentum coefficient, and $\eta_t > 0$ is the learning rate at iteration t. The next parameter vector is w_{t+1}, and $\alpha \in [0, 1]$ controls only how much of the freshly generated correction acts on the current update. The initial momentum state is typically $m_{-1} = 0$. Equation 3 always stores the full correction in momentum, whereas Eq. 4 routes an α fraction of that correction through the current, zero-lag update. Standard M-SAM corresponds to $\alpha = 1$. TF-SAM corresponds to $\alpha = 0$: c_t still enters m_t, but it is absent from the current parameter update. Intermediate values $0 < \alpha < 1$ interpolate between these two structural choices. Thus TF-SAM removes the zero-lag route without changing correction generation or the delayed momentum-mediated effect.

The response-kernel analysis assumes a constant learning rate $\eta_t = \eta$. SAM-style optimizers are compared through the correction-response kernel. For a sharpness correction c_t, define the update increment by $\Delta w_j := w_{j+1} - w_j$. The notation $\Delta w_{t+k}^{(c_t)}$ denotes the component of this increment attributable to the single correction c_t, with all other gradient contributions held fixed. Here $k \in \{0, 1, 2, \ldots\}$ is the lag measured from the iteration at which c_t is generated, and $K(k)$ is the coefficient by which c_t contributes at that lag: $\Delta w_{t+k}^{(c_t)} = -\eta K(k) c_t$. Thus $K(k)$ is the coefficient in the update direction, while the parameter increment also carries $-\eta$. Under the zero-lag family above, the response to one generated correction is the sequence $K_{\alpha,\beta} = (\alpha, \beta, \beta^2, \ldots)$, where the ellipsis denotes the subsequent powers of β. Hence $K_{\text{M-SAM}} = (1, \beta, \beta^2, \ldots)$, whereas $K_{\text{TF}} = (0, \beta, \beta^2, \ldots)$. The kernel isolates temporal routing from correction generation and magnitude: M-SAM applies the correction immediately and carries it forward through momentum, whereas

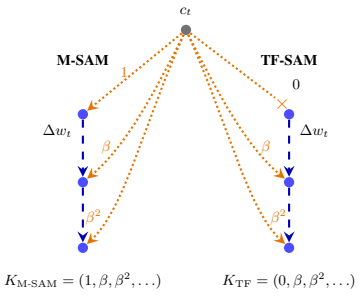


Figure 1: Response kernels for routing c_t. M-SAM applies the correction at lag 0 and carries it forward through momentum; TF-SAM removes only this zero-lag path and keeps the delayed tail.

Table 1: Response-kernel view of the routing family in Eqs. 3 and 4. The kernel column describes the temporal placement of an already generated correction c_t.

Method	Correction kernel K	Zero-lag	Main design axis
SAM	$(1, 0, 0, \ldots)$	full	perturbed gradient
M-SAM	$(1, \beta, \beta^2, \ldots)$	full	immediate plus momentum tail
TF-SAM	$(0, \beta, \beta^2, \ldots)$	none	delayed momentum correction

TF-SAM preserves only the delayed tail. The superscript (c_t) marks a contribution induced by c_t and is not an exponent. Fig. 1 illustrates this intervention at the level of a single correction impulse.

Table 1 summarizes this distinction within the update family above. In the table, "SAM" denotes the no-momentum baseline, so its correction kernel has no delayed tail, while "M-SAM" denotes the momentum-based variant.

3 Temporal Routing of Sharpness Corrections

Our analysis asks three questions: 1) ***what changes when an already-generated correction is delayed***; 2) ***when immediate or delayed application is first-order favorable at the same target update***; and 3) ***why temporal routing changes the closed-loop dynamics rather than simply rescaling the correction***. These questions are addressed by 1) an exact original-iterate identity (Proposition 1), 2) a target-matched routing criterion (Proposition 2), and 3) a local matrix-polynomial analysis (Proposition 3), respectively.

3.1 Exact temporal routing

Let $\Delta_t := w_t - w_{t-1}$ denote the parameter increment. Eliminating the momentum state from Eqs. 3 and 4 gives an exact representation of the original iterate.

Proposition 1 (Exact temporal-forcing representation)**.** *For $t \geq 1$, the unified routing family satisfies*

$$\Delta_{t+1} = \beta \frac{\eta_t}{\eta_{t-1}} \Delta_t - \eta_t [g_t + r_t^{(\alpha)}], \qquad r_t^{(\alpha)} := \alpha c_t + \beta(1-\alpha) c_{t-1}. \tag{5}$$

Thus M-SAM has $r_t^{\mathrm{M}} = c_t$, whereas TF-SAM has $r_t^{\mathrm{TF}} = \beta c_{t-1}$.

This identity isolates the routing change and makes the fresh-versus-delayed correction comparison explicit at the same target update.

The proof is in Appendix B. The identity is exact even with a time-varying step size: η_t changes the inertial prefactor and update scale, but not the selected correction lag. TF-SAM still generates and stores c_t.

3.2 Target-matched first-order routing utility

Let $F(w) = \mathbb{E}_\xi[\ell(w; \xi)]$ be the objective under the sampling distribution, $G_t := \nabla F(w_t)$, and $\mathcal{F}_t$ the history before the iteration-t minibatch is sampled. For $t \geq 1$, define the current and one-step-delayed population alignments

$$A_{t,k} := \mathbb{E}[\langle G_t, c_{t-k} \rangle \mid \mathcal{F}_t], \quad k \in \{0, 1\}, \qquad \mathcal{A}_t(\alpha) := \alpha A_{t,0} + \beta(1-\alpha) A_{t,1}. \tag{6}$$

The delayed correction c_{t-1} is already determined by $\mathcal{F}_t$; the conditional expectation keeps the notation common to both routes.

Proposition 2 (Target-matched first-order routing criterion)**.** *Under Assumption 1 and constant step size $\eta > 0$, fix the target state and history at step $t \geq 1$. The correction-attributable first-order term in the conditional one-step descent bound is $-\eta \mathcal{A}_t(\alpha)$, with*

$$\frac{\partial \mathcal{A}_t(\alpha)}{\partial \alpha} = A_{t,0} - \beta A_{t,1}. \tag{7}$$

Among $\alpha \in [0,1]$, the immediate route maximizes first-order correction alignment if $A_{t,0} > \beta A_{t,1}$, and the delayed route does so if $A_{t,0} < \beta A_{t,1}$. At equality, all routes have the same first-order correction alignment.

This criterion identifies the first-order preferred route at a fixed target update.

The full bound, including momentum and smoothness terms, is proved in Appendix B.2. The criterion does not compare full updates or training trajectories.

The population comparison motivates the realized target-matched diagnostic used in Section 4:

$$q_t^{\mathrm{imm}} := \frac{\langle c_t, g_t \rangle}{\|g_t\|^2}, \qquad q_t^{\mathrm{delay}} := \frac{\langle c_{t-1}, g_t \rangle}{\|g_t\|^2}, \qquad D_t := \beta q_t^{\mathrm{delay}} - q_t^{\mathrm{imm}}, \quad t \geq 1,\ g_t \neq 0. \tag{8}$$

Both correction terms are evaluated against the same target minibatch gradient g_t; only their age differs. Positive D_t favors the delayed route at the local first-order level, while negative D_t favors the immediate route. These realized minibatch measurements are descriptive diagnostics, not unbiased estimators of the population alignments. The measurement protocol and phase-wise results appear in Appendix D.7.

3.3 Temporal routing changes the closed-loop dynamics

The response matrix in the local analysis comes from the actual fixed-radius, normalized-SAM correction. Fix a minibatch gradient map g and a regular point $\bar{w}$ with $\bar{g} = g(\bar{w}) \neq 0$. Write $\bar{u} = \bar{g}/\|\bar{g}\|$, $\bar{\epsilon} = \rho \bar{u}$, and $c(w) = g(w + \rho g(w)/\|g(w)\|) - g(w)$ where $g(w) \neq 0$. Let D denote a Jacobian and I_p the $p \times p$ identity.

Lemma 1 (Normalized-SAM response at a regular point)**.** *Suppose g is continuously differentiable near $\bar{w}$ and $\bar{w} + \bar{\epsilon}$. With $H_0 = Dg(\bar{w})$ and $H_+ = Dg(\bar{w} + \bar{\epsilon})$, the correction has Jacobian*

$$C_{\bar{w}} := Dc(\bar{w}) = H_+ \left[I_p + \frac{\rho}{\|\bar{g}\|}(I_p - \bar{u}\bar{u}^\top) H_0 \right] - H_0. \tag{9}$$

Hence $c(\bar{w} + \delta) = c(\bar{w}) + C_{\bar{w}}\delta + o(\|\delta\|)$.

The response $C_{\bar{w}}$ need not be positive semidefinite or commute with H_0. Appendix C derives the Jacobian, its exact Hessian-drift decomposition, and the constant-curvature special case. Fixed-radius normalization is singular at $g = 0$, so the local analysis applies at regular trajectory points.

Compare neighboring routed trajectories driven by the same minibatch sequence. At a regular reference state $(\bar{w}_t, \bar{m}_{t-1})$, set $H_t = Dg_t(\bar{w}_t)$ and $C_t = Dc_t(\bar{w}_t)$, with the minibatch fixed in each map. For $\delta s_t = (\delta w_t^\top, \delta m_{t-1}^\top)^\top$, the first-order difference obeys

$$\delta s_{t+1} = \mathcal{J}_{\alpha,t} \delta s_t + o(\|\delta s_t\|), \qquad \mathcal{J}_{\alpha,t} = \begin{pmatrix} I_p - \eta_t(H_t + \alpha C_t) & -\eta_t \beta I_p \\ H_t + C_t & \beta I_p \end{pmatrix}. \tag{10}$$

To expose the routing structure, fix η and freeze $H_t = H$, $C_t = C$ at a regular operating point; the reference trajectory need not be stationary.

Proposition 3 (General closed-loop routing polynomial)**.** *For arbitrary real $p \times p$ frozen response matrices H, C, the eigenvalues of $\mathcal{J}_\alpha$ are exactly the roots of $\det Q_\alpha(r) = 0$, where*

$$Q_\alpha(r) = r^2 I_p - r[(1+\beta)I_p - \eta(H + \alpha C)] + \beta[I_p + \eta(1-\alpha)C]. \tag{11}$$

No commutativity of H and C is required.

Its coefficients distinguish the immediate and delayed routes in the local dynamics.

Specifically, immediate response enters the coefficient of r through αC, while delayed response enters the constant coefficient through $\beta(1-\alpha)C$. This is a property of the frozen local variational system, whose roots need not separate into scalar modes. Appendix C gives the proof and the general Schur criterion.

Stability interpretation. For a real common invariant direction with $Hv = \lambda v$ and $Cv = \kappa v$, the scalar characteristic polynomial has coefficients

$$\tau_\alpha = 1 + \beta - \eta(\lambda + \alpha\kappa), \qquad \delta_\alpha = \beta[1 + \eta(1-\alpha)\kappa]. \tag{12}$$

Immediate and delayed responses therefore affect distinct coefficients. Depending on the signed local response κ, shifting correction mass between the routes can move stability boundaries in different directions. The full aligned-mode Schur conditions are in Appendix C. Temporal routing cannot in general be represented by scalar rescaling of the current correction: it changes both the lag acting at the target update and where the local correction response enters the closed loop.

4 Experiments

The experiments test whether the immediate sharpness route is needed in practice by comparing stochastic gradient descent (SGD), M-SAM, and TF-SAM across image and text models, using benchmark comparisons, correction-geometry diagnostics, controlled temporal-routing ablations, and corruption robustness. The image models are Residual Networks (ResNets); the text models are TinyBERT and Generative Pre-trained Transformer 2 (GPT-2) for natural language processing (NLP). The datasets are CIFAR-10, CIFAR-100, the two-class Stanford Sentiment Treebank (SST-2), and its five-class version (SST-5). Code and experiment artifacts are available at `https://anonymous.4open.science/r/iclr2027-tfsam-75FF`; detailed training settings and reproduction commands are provided in Appendix D.2. Appendix D.5 provides shared-direction loss slices as a qualitative view of the local checkpoint geometry, following standard loss landscape visualization practice (Li et al., 2018).

4.1 Main comparison and regime dependence

Table 2: Final test accuracy (mean $\pm$ standard deviation) and paired differences in percentage points. Positive gaps favor TF-SAM. The reported 95% confidence intervals (CIs) are not equivalence tests.

Dataset	Model	SGD	M-SAM	TF-SAM	TF-SAM–M-SAM gap [95% CI]
CIFAR-10	ResNet-18	95.38 ± 0.10	96.02 ± 0.14	96.02 ± 0.09	$0.00\ [-0.10, 0.10]$
CIFAR-10	ResNet-34	95.58 ± 0.13	96.40 ± 0.12	96.36 ± 0.09	$-0.03\ [-0.14, 0.07]$
CIFAR-100	ResNet-18	78.39 ± 0.26	78.78 ± 0.25	78.78 ± 0.33	$-0.00\ [-0.29, 0.29]$
CIFAR-100	ResNet-34	78.55 ± 0.58	79.58 ± 0.33	79.56 ± 0.33	$-0.02\ [-0.31, 0.28]$
SST-2	TinyBERT	85.22 ± 0.49	85.62 ± 0.31	85.62 ± 0.32	$-0.00\ [-0.15, 0.15]$
SST-5	GPT-2	53.63 ± 0.46	52.29 ± 0.38	52.82 ± 0.45	$+0.53\ [0.30, 0.76]$

Table 2 reports final test accuracy over ten seeds. Its last column gives the paired TF-SAM–M-SAM final-accuracy gap and two-sided 95% t CI for each setting.

Finding 1: Removing zero-lag feedback incurs no consistent accuracy penalty. Across the six settings, the paired TF-SAM–M-SAM accuracy gaps are near zero in five cases and positive for GPT-2/SST-5. The 95% CIs for the five near-zero point estimates include zero, while the GPT-2/SST-5 interval is entirely positive; none is entirely negative. Best-accuracy results and the paired analysis are reported in Appendix D.6.

Finding 2: Delayed routing retains the observed SAM-associated accuracy in five settings. In the four image settings and SST-2/TinyBERT, both M-SAM and TF-SAM have higher mean accuracy than SGD, and their means differ by at most 0.03 percentage points (Table 2). The SAM-beneficial accuracy pattern therefore persists when each fresh correction first acts through momentum.

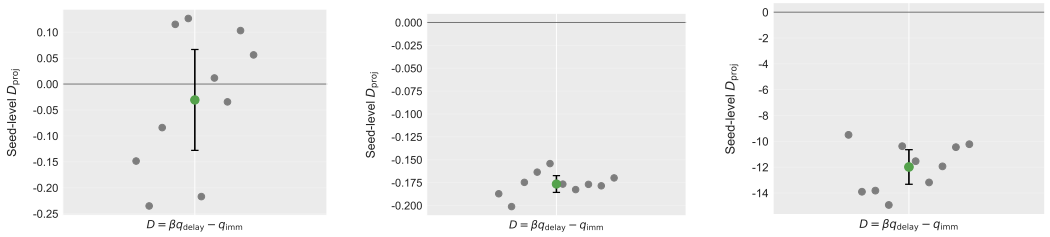
Finding 3: Delayed routing improves on M-SAM in the SAM-adverse GPT-2/SST-5 setting. M-SAM has lower mean accuracy than SGD, while TF-SAM exceeds M-SAM by 0.53 percentage

points, with a paired 95% CI of $[0.30, 0.76]$ (Table 2). TF-SAM also remains below SGD. This paired gain is consistent with mitigating an adverse immediate interaction.

4.2 TARGET-MATCHED ROUTING UTILITY DEPENDS ON OPTIMIZATION REGIME

The exact temporal-forcing identity replaces the fresh correction c_t with the momentum-weighted one-step-old correction βc_{t-1} at the same target update. The realized diagnostic in Eq. 8 compares both terms. The target parameter state, minibatch, clean gradient, and denominator are identical on both sides; only correction age changes.

Finding 1: CIFAR-10 favors immediate routing after the early phase. On CIFAR-10/ResNet-18, early q_{imm} varies across seeds while βq_{delay} stays near zero and slightly negative, so the ordering is mixed (Fig. 2). In the middle and late phases, q_{imm} is consistently positive while βq_{delay} remains near zero. The resulting negative D shows that delaying the correction attenuates a favorable fresh contribution. The larger normalized late-phase D is interpreted directionally because its denominator $\|g_t\|^2$ can shrink during training.



(a) Early. (b) Middle. (c) Late.

Figure 2: Target-matched routing difference on CIFAR-10/ResNet-18 for M-SAM, $D = \beta q_{\mathrm{delay}} - q_{\mathrm{imm}}$. Gray points show individual estimates and the green marker their mean. Negative values favor immediate routing at the same target update; positive values favor delayed routing. The middle and late phases show a negative ordering, while early estimates are mixed.

Finding 2: GPT-2 favors delaying an adverse fresh correction. On GPT-2/SST-5, both early projections are adverse, with $q_{\mathrm{imm}} \approx -0.16$ and $\beta q_{\mathrm{delay}} \approx -0.135$ (Fig. 9). In the middle phase, the fresh correction is strongly anti-aligned with the target gradient ($q_{\mathrm{imm}} \approx -0.48$), while the delayed projection is nearly neutral and slightly negative. This separation persists late in training ($q_{\mathrm{imm}} \approx -0.47$ and $\beta q_{\mathrm{delay}} \approx 0$). The positive D occurs for all ten seeds in each phase; after the early phase, delay largely removes the adverse same-target projection.

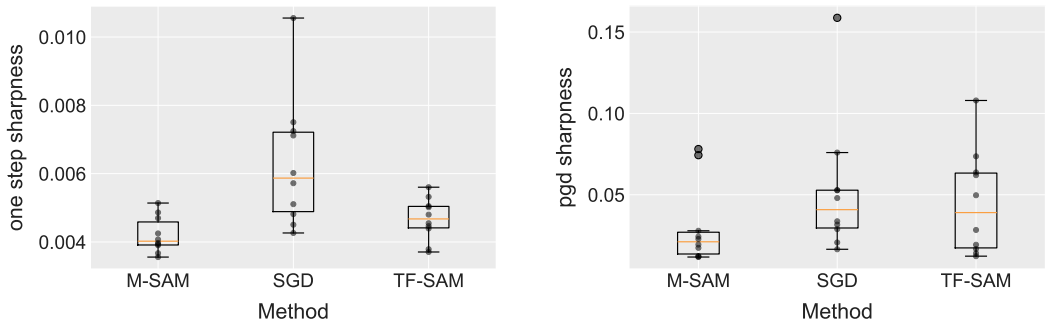
Finding 3: The first-order route preference reverses across regimes. Delaying the correction removes a favorable immediate contribution on CIFAR-10/ResNet-18 and an adverse one on GPT-2/SST-5. Final accuracy is similar for M-SAM and TF-SAM on CIFAR-10/ResNet-18, whereas TF-SAM improves over M-SAM on GPT-2/SST-5. The diagnostic compares first-order alignment at a fixed target update; these seed-level measurements do not establish a trajectory-level causal explanation for the accuracy gaps. Appendix D.7 gives the phase-wise protocol and projection plots.

4.3 LOCAL TRAINING-LOSS GEOMETRY

For CIFAR-10/ResNet-18, we pair per-seed local training-loss increases with a representative two-dimensional loss slice around the final checkpoints. The measurement protocol and paired-seed intervals are reported in Appendix D.4.

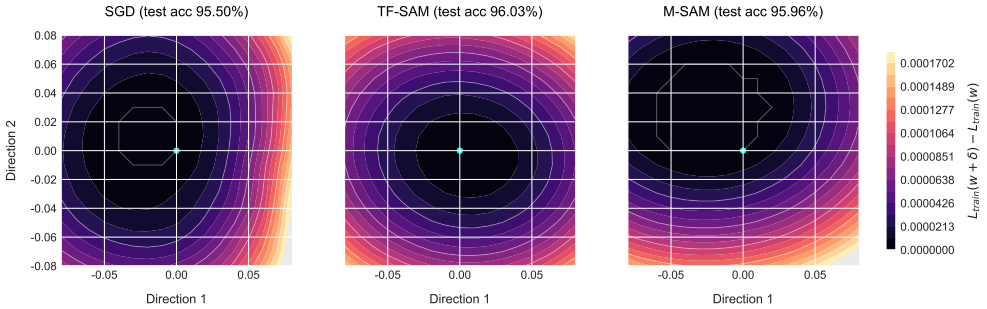
Finding 1: One-step scores are lower for both SAM variants than for SGD. For ResNet-18, the seed-paired one-step loss increases are lower for M-SAM and TF-SAM than for SGD, while the TF-SAM–M-SAM interval includes zero (Table 12).

Finding 2: The multi-start PGD search does not resolve an optimizer ordering. All three pairwise ResNet-18 PGD intervals include zero (Table 12).



(a) One-step sharpness.

(b) PGD sharpness.



(c) Training-loss slices (seed 42).

Figure 3: Local training-loss geometry at final CIFAR-10/ResNet-18 checkpoints. The top panels show per-seed one-step and multi-start PGD loss increases over ten matched seeds; one-step sharpness evaluates the normalized-gradient SAM perturbation. The bottom panel shows two-dimensional loss slices for the three optimizers along the same filter-normalized directions, coordinate range, contour levels, and color scale. The cyan point marks the unperturbed checkpoint, and colors show the loss increase relative to it. The slices visualize the displayed subspace and do not provide a global flatness ranking.

Finding 3: The loss slices show optimizer-specific local geometry. In the displayed seed-42 subspace, TF-SAM has smoother, more balanced contours than SGD, while M-SAM also has a broad low-loss region with a different directional shape. These slices are qualitative and do not rank flatness in the full parameter space or establish a generalization advantage.

4.4 Corruption Robustness

The corruption evaluation tests whether removing zero-lag sharpness feedback induces an aggregate robustness penalty relative to M-SAM. Final checkpoints from the main CIFAR experiments are evaluated under twelve fixed image corruptions. The NLP benchmarks are evaluated under thirteen fixed text corruptions. Table 3 reports per-corruption CIFAR-10 accuracy; the seed-level aggregate protocol and results across all six settings are reported in Appendix D.3.

The primary robustness comparison averages non-clean corruption or noise accuracy within each matched training seed and then computes the paired TF-SAM–M-SAM gap across ten seeds (Table 7).

Finding 1: Removing zero-lag feedback shows no consistent aggregate robustness penalty. Across the six image and text settings, five paired 95% CIs include zero, and no interval lies entirely below zero. This supports the absence of a consistent penalty in the tested settings; the five intervals containing zero do not establish equivalence.

Finding 2: SST-2/TinyBERT has a positive aggregate gap. TF-SAM exceeds M-SAM by 0.18 percentage points, with a paired 95% CI of $[0.02, 0.34]$. This is the only setting whose aggregate interval lies entirely above zero.

Table 3: Per-corruption CIFAR-10 test accuracy for ResNet-18 and ResNet-34. Entries are mean ± standard deviation in percent. The highest mean within each model and condition is shaded and bold. The vertical rule separates the two model groups.

Condition	ResNet-18			ResNet-34		
	SGD	M-SAM	TF-SAM	SGD	M-SAM	TF-SAM
Clean	95.38 ± 0.10	96.02 ± 0.14	96.02 ± 0.09	95.58 ± 0.13	96.40 ± 0.12	96.36 ± 0.09
Brightness	94.79 ± 0.14	95.52 ± 0.12	95.51 ± 0.06	95.03 ± 0.17	95.98 ± 0.16	95.93 ± 0.09
Contrast	94.98 ± 0.15	95.65 ± 0.11	95.65 ± 0.15	95.22 ± 0.14	96.08 ± 0.09	96.07 ± 0.12
Elastic	92.23 ± 0.23	93.34 ± 0.20	93.30 ± 0.22	92.67 ± 0.28	93.73 ± 0.18	93.89 ± 0.17
Gaussian blur	23.68 ± 1.68	24.08 ± 1.43	24.64 ± 1.12	24.31 ± 1.89	23.27 ± 1.72	23.39 ± 2.93
Gaussian noise	36.89 ± 3.39	38.07 ± 2.79	37.18 ± 2.60	42.12 ± 3.41	42.42 ± 3.14	40.65 ± 2.02
Heavy blur	18.93 ± 1.93	19.23 ± 1.49	19.85 ± 1.14	18.55 ± 1.89	17.59 ± 1.58	17.36 ± 1.67
Posterize	93.53 ± 0.21	94.39 ± 0.19	94.35 ± 0.28	93.83 ± 0.19	94.91 ± 0.15	94.90 ± 0.11
Rotation	89.16 ± 0.29	90.56 ± 0.23	90.44 ± 0.35	89.47 ± 0.39	90.93 ± 0.27	91.13 ± 0.25
Saturation	95.28 ± 0.13	95.93 ± 0.11	95.94 ± 0.08	95.50 ± 0.12	96.29 ± 0.12	96.28 ± 0.09
Sharpness	91.88 ± 0.47	92.79 ± 0.13	92.94 ± 0.34	92.49 ± 0.33	93.65 ± 0.18	93.63 ± 0.22
Shot noise	12.67 ± 0.95	13.15 ± 1.03	12.77 ± 0.96	13.40 ± 1.48	13.84 ± 1.16	13.46 ± 1.39
Solarize	62.43 ± 0.75	64.89 ± 0.66	65.01 ± 1.26	64.45 ± 0.91	67.39 ± 0.62	67.36 ± 1.29

Finding 3: Corruption-level differences depend on the condition. On CIFAR-10/ResNet-18, TF-SAM has higher mean accuracy under Gaussian blur and heavy blur, while M-SAM has higher means under Gaussian noise and shot noise (Table 3). These per-condition means are descriptive; the paired seed-level aggregate is the primary robustness comparison.

5 Discussion

5.1 Compatibility with scheduled and efficient SAM

Temporal routing acts after correction construction and is therefore conceptually composable with methods that change when or how sharpness corrections are generated, including scheduled and efficient SAM variants (Liu et al., 2022; Du et al., 2022; Zhou et al., 2025; Becker et al., 2025). Training-phase generation scheduling and post-generation action lag are distinct temporal questions. TF-SAM itself does not reduce SAM's additional gradient cost; combining delayed routing with sparse, approximate, or late-phase correction generation is left to future work.

5.2 Implications and scope

Across the evaluated settings, removing zero-lag feedback produces no consistent loss in predictive accuracy or aggregate corruption robustness. Target-matched measurements show that the relative first-order value of immediate and delayed routes varies by regime. Together, these results establish post-generation timing as a design choice and motivate target-matched adaptive routing. A candidate rule sets $\alpha_t = \phi(q_t^{\mathrm{imm}} - \beta q_t^{\mathrm{delay}})$ for an increasing gate $\phi : \mathbb{R} \to [0, 1]$. In the middle and late GPT-2/SST-5 phases, the smaller immediate projection points toward a lower zero-lag gain. Evaluating this rule is a natural next step.

These conclusions are limited to smooth, bounded-variance optimization and the evaluated image and text settings. At a fixed target gradient, the target-matched diagnostic captures first-order geometry between immediate and one-step-delayed corrections, not a full-trajectory counterfactual comparing M-SAM with TF-SAM. Mechanism, benchmark, and robustness estimates use ten seeds; some CIs remain wide. With time-varying η_t, first-order update contributions scale with η_t but normalized geometry q does not, so phase comparisons must account for update scale; learning-rate decay alone cannot establish phase trends in q. Finally, TF-SAM retains SAM's additional gradient computation and isolates routing timing rather than reducing cost.

AI USE STATEMENT

AI tools were used to assist with language polishing and to check for possible logical inconsistencies in theorem statements and proofs, such as mismatched notation, missing assumptions, or steps that did not follow from the stated conditions. They were not used as an authoritative source for mathematical results. The authors reviewed and edited the AI-assisted text, and are responsible for all scientific claims, mathematical statements, experiments, and final manuscript content.

ETHICS STATEMENT

This work studies optimization dynamics on standard image and text classification benchmarks. It introduces no new datasets or user-facing deployments. Broader concerns inherit from the underlying benchmark data and pretrained models.

REPRODUCIBILITY STATEMENT

The repository is designed for lightweight reproduction: dependencies are managed with `uv`, and the training, analysis, and table-generation entry points are exposed through `just` commands. Appendix D.2 documents the shared training setup, dataset and model coverage, seed counts, diagnostics, regeneration commands, and the Optuna-based hyperparameter search and optimizer-library components used in the experiments (Akiba et al., 2019; Kim, 2021).

REFERENCES

Atish Agarwala and Yann Dauphin. SAM operates far from home: eigenvalue regularization as a dynamical phenomenon. In *International Conference on Machine Learning*, volume 202, pp. 152–168, 2023.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, volume 162, pp. 247–257, 2022.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *International Conference on Knowledge Discovery and Data Mining*, 2019.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, volume 162, pp. 639–668, 2022.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47032–47051, 2023.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253, 2018.

Sanjeev Arora, Yuanzhi Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, volume 162, pp. 948–1024, 2022.

Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Annual Meeting of the Association for Computational Linguistics*, pp. 7360–7371, 2022.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Marlon Becker, Frederick Altrock, and Benjamin Risse. Momentum-sam: Sharpness aware minimization without computational overhead. In *Advances in Neural Information Processing Systems*, pp. 55336–55366, 2025.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67741–67770, 2023.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Khashayar Gatmiry, Zhiyuan Li, Sashank J Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. In *International Conference on Machine Learning*, pp. 15102–15129. PMLR, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt Kusner. When do flat minima optimizers work? In *Advances in Neural Information Processing Systems*, 2022.

Hyeongchan Kim. PyTorch_Optimizer: Optimizer & LR scheduler & loss function collections in PyTorch, 2021.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, volume 162, pp. 11148–11161, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5631–5640. IEEE, 2024.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024.

Thomas Möllenhoff and Mohammad Emtiyaz Khan. Sam as an optimal relaxation of bayes. In *International Conference on Machine Learning*, 2023.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17749–17761, 2021.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *International Conference on Learning Representations*, volume 2025, pp. 30071–30097, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(4):400–407, 1951.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. In *Advances in Neural Information Processing Systems*, 2023.

Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of SAM. In *International Conference on Machine Learning*, volume 267, pp. 55702–55719, 2025.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Jacob Springer, Vaishnavh Nagarajan, and Aditi Raghunathan. Sharpness-aware minimization enhances feature quality via balanced learning. In *International Conference on Learning Representations*, volume 2024, pp. 22534–22558, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Hailiang Ye, Xinyi Fang, Ming Li, and Feilong Cao. Ed-sam: Sharpness-aware minimization with energy-adjusted perturbations and direction-corrected updates. *Neural Networks*, 203:109185, 2026.

Yun Yue, Jiadi Jiang, Zhiling Ye, Ning Gao, Yongchao Liu, and Ke Zhang. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3185–3194, 2023.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, volume 2025, pp. 20949–20980, 2025.

Jinping Zou, Xiaoge Deng, and Tao Sun. Sharpness-aware minimization with adaptive regularization for training deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

Appendix Contents

A Related Work **13**

B Proofs of Temporal Routing and Alignment Criterion **13**

B.1 Original-iterate temporal routing 14

B.2 Lagged alignment and routing utility 14

C Full Temporal-Feedback Theory **15**

C.1 Sharpness Feedback as a Two-Channel Temporal System 15

C.2 Closed-Loop Variational Dynamics at Regular Points 17

D Additional Material **21**

D.1 Controlled temporal routing 21

D.2 Experimental configuration 22

D.3 Aggregate corruption robustness 23

D.4 Post-hoc training-set loss sharpness 24

D.5 Checkpoint loss-landscape visualizations 25

D.6 Paired-seed statistical analysis 26

D.7 Target-matched immediate versus delayed correction geometry 27

D.8 GPT-2 correction-magnitude sanity check 29

A Related Work

Correction construction and temporal routing. SAM replaces the clean update direction with a gradient evaluated at perturbed weights (Foret et al., 2021). Subsequent variants change the perturbation geometry, as in Adaptive Sharpness-Aware Minimization (ASAM) and Fisher SAM (Kwon et al., 2021; Kim et al., 2022), or modify how the resulting correction is interpreted and used (Möllenhoff & Khan, 2023; Yue et al., 2023; Zou et al., 2025; Ye et al., 2026). Broader empirical comparisons study SAM alongside other flat-minima optimizers across vision, language, and graph tasks (Kaddour et al., 2022). Momentum-SAM constructs its perturbation from the accumulated momentum direction, changing the signal used to generate sharpness information while avoiding SAM's additional gradient computation (Becker et al., 2025). The variants above change how the correction is constructed, weighted, or computed. In contrast, TF-SAM holds per-step SAM correction generation fixed and changes only its post-generation route: it removes the correction's zero-lag action while retaining its delayed momentum tail. This isolates whether an already generated correction acts on the current update or through later momentum updates.

Mechanisms, theory, and efficiency. Mechanistic studies connect SAM behavior to representation, implicit bias, language-model generalization, training phase, normalization, stochasticity, and stability (Andriushchenko & Flammarion, 2022; Andriushchenko et al., 2023; Springer et al., 2024; Zhou et al., 2025; Li et al., 2024; Bahri et al., 2022; Singh et al., 2025; Bartlett et al., 2023; Agarwala & Dauphin, 2023; Long & Bartlett, 2024; Dai et al., 2023; Si & Yun, 2023; Oikonomou & Loizou, 2025). Related theory connects sharpness minimization to simple feature structure in certain two-layer models (Gatmiry et al., 2024). Efficient and scheduled methods change when or how often sharpness information is generated or computed: LookSAM reuses perturbations, Efficient Sharpness-Aware Minimization (ESAM) uses stochastic selection, and late-phase SAM changes the training schedule (Liu et al., 2022; Du et al., 2022; Zhou et al., 2025). This decomposition complements these mechanistic and theoretical studies by separating the generated correction from its temporal route for M-SAM, and it distinguishes TF-SAM from efficient and scheduled methods: TF-SAM computes a correction at each SAM step but changes when it acts, from zero-lag application to a momentum-mediated delayed route. These choices act on orthogonal parts of the optimization procedure.

B Proofs of Temporal Routing and Alignment Criterion

This appendix proves the exact original-iterate routing identity and the lagged-alignment criterion used in Section 3. The alignment bound is a local first-order statement; it is not a convergence or generalization theorem.

Proof roadmap.

- Lemma 2 gives the exact original-iterate temporal forcing.
- Proposition 4 identifies the correction-attributable first-order term.

- Proposition 2 gives the immediate-versus-delayed criterion.

Recall that $F(w) = \mathbb{E}_\xi[\ell(w;\xi)]$ is the expected objective for the sampled loss. At step t, let $g_t = \nabla\ell(w_t;\xi_t)$ and, for normalized SAM, set

$$\epsilon_t = \begin{cases} \rho g_t / \|g_t\|, & g_t \neq 0, \\ 0, & g_t = 0, \end{cases} \qquad c_t = \nabla\ell(w_t + \epsilon_t;\xi_t) - \nabla\ell(w_t;\xi_t). \tag{13}$$

Assumption 1 (Smooth stochastic objective)**.** *The objective is lower bounded by F_*. The functions F and $\ell(\cdot;\xi)$ are L-smooth. With respect to the history $\mathcal{F}_t$ before sampling ξ_t, the stochastic gradient is unbiased and has bounded variance:*

$$\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla F(w_t), \qquad \mathbb{E}[\|g_t - \nabla F(w_t)\|^2 \mid \mathcal{F}_t] \le \sigma^2. \tag{14}$$

Since $\ell(\cdot;\xi)$ is L-smooth and $\|\epsilon_t\| \le \rho$, the correction is deterministically bounded by

$$\|c_t\| \le L\rho. \tag{15}$$

No unbiasedness assumption is made on c_t.

B.1 Original-iterate temporal routing

The following lemma records the exact original-iterate identity that preserves the distinction between the current and delayed correction.

Lemma 2 (Exact temporal-forcing representation)**.** *For $t \ge 1$, let $\Delta_t = w_t - w_{t-1}$. The update family in Eq. 3–4 satisfies*

$$\Delta_{t+1} = \beta\frac{\eta_t}{\eta_{t-1}}\Delta_t - \eta_t\left[g_t + \alpha c_t + \beta(1-\alpha)c_{t-1}\right]. \tag{16}$$

Proof. The update at step $t-1$ gives

$$\Delta_t = -\eta_{t-1}\left[\beta m_{t-2} + g_{t-1} + \alpha c_{t-1}\right]. \tag{17}$$

On the other hand, the momentum recursion gives

$$\beta m_{t-1} = \beta^2 m_{t-2} + \beta g_{t-1} + \beta c_{t-1}. \tag{18}$$

Substitute Eq. 18 into the step-t parameter update and use Eq. 17 to obtain

$$\begin{aligned} \Delta_{t+1} &= -\eta_t\left[\beta^2 m_{t-2} + \beta g_{t-1} + \beta c_{t-1} + g_t + \alpha c_t\right] \\ &= \beta\frac{\eta_t}{\eta_{t-1}}\Delta_t - \eta_t\left[g_t + \alpha c_t + \beta(1-\alpha)c_{t-1}\right], \end{aligned} \tag{19}$$

as required. □

B.2 Lagged alignment and routing utility

For $G_t = \nabla F(w_t)$ and the pre-minibatch history $\mathcal{F}_t$, define

$$A_{t,k} := \mathbb{E}[\langle G_t, c_{t-k}\rangle \mid \mathcal{F}_t], \qquad 0 \le k \le t. \tag{20}$$

For $k \ge 1$, c_{t-k} is $\mathcal{F}_t$-measurable; the conditional-expectation notation retains the same form for current and delayed corrections. Let

$$r_t^{(\alpha)} = \alpha c_t + \beta(1-\alpha)c_{t-1}, \qquad \mathcal{A}_t(\alpha) = \alpha A_{t,0} + \beta(1-\alpha)A_{t,1}. \tag{21}$$

Proposition 4 (Lagged-alignment contribution of temporal routing)**.** *Under Assumption 1 and constant $\eta_t = \eta$, for $t \ge 1$,*

$$\mathbb{E}[F(w_{t+1}) \mid \mathcal{F}_t] \le F(w_t) + \beta\langle G_t, \Delta_t\rangle - \eta\|G_t\|^2 - \eta\mathcal{A}_t(\alpha) + \mathcal{R}_t(\alpha), \tag{22}$$

where

$$\mathcal{R}_t(\alpha) := \frac{L}{2}\mathbb{E}\left[\left\|\beta\Delta_t - \eta(g_t + r_t^{(\alpha)})\right\|^2 \mid \mathcal{F}_t\right]. \tag{23}$$

Proof. By L-smoothness and Lemma 2,

$$F(w_{t+1}) \le F(w_t) + \left\langle G_t, \beta\Delta_t - \eta(g_t + r_t^{(\alpha)}) \right\rangle + \frac{L}{2}\left\| \beta\Delta_t - \eta(g_t + r_t^{(\alpha)}) \right\|^2. \tag{24}$$

Condition on $\mathcal{F}_t$. The vectors G_t, Δ_t, and c_{t-1} are $\mathcal{F}_t$-measurable, and Assumption 1 gives $\mathbb{E}[g_t \mid \mathcal{F}_t] = G_t$. Thus the conditional expectation of the linear term is

$$\begin{aligned}\beta\langle G_t, \Delta_t\rangle - \eta\|G_t\|^2 - \eta\alpha\mathbb{E}[\langle G_t, c_t\rangle \mid \mathcal{F}_t] & \\ - \eta\beta(1-\alpha)\mathbb{E}[\langle G_t, c_{t-1}\rangle \mid \mathcal{F}_t] & \\ = \beta\langle G_t, \Delta_t\rangle - \eta\|G_t\|^2 - \eta\mathcal{A}_t(\alpha). & \end{aligned} \tag{25}$$

The conditional expectation of the quadratic term is exactly $\mathcal{R}_t(\alpha)$, proving the result. □

Proof of Proposition 2. Proposition 4 identifies $-\eta\mathcal{A}_t(\alpha)$ as the correction-attributable linear term in the conditional descent bound. Differentiating Eq. 21 gives Eq. 7. Since $\mathcal{A}_t(\alpha)$ is affine and $\eta > 0$, its slope determines which endpoint of $[0, 1]$ maximizes first-order correction alignment; a zero slope makes all routes equal at this order. □

Descriptive multi-lag alignment. The target-matched comparison in the main text keeps the target update fixed. For a correction followed over several later updates, one can additionally log

$$q^{\parallel}_{t\to t+k} := \frac{\langle c_t, g_{t+k}\rangle}{\|g_{t+k}\|^2}, \qquad k \ge 0, \quad g_{t+k} \ne 0. \tag{26}$$

Here $k = 0$ is the same-step projection and $k = 1$ compares the correction to a future minibatch gradient. The state and minibatch can change with k, so this multi-lag profile is descriptive and is distinct from the same-target comparison in Eq. 8. Neither realized diagnostic is an unbiased estimator of the population alignment in Eq. 20.

C Full Temporal-Feedback Theory

Theory roadmap. This section separates the effect of temporal routing into open-loop and closed-loop components:

- **Open-loop response.** Treat the generated correction sequence as exogenous. Lemma 3 isolates the relative temporal placement of the clean-gradient and sharpness-correction channels. The analysis shows that TF-SAM is exactly a delayed and β-scaled correction response, while Proposition 6 shows that this transformation cannot be reproduced by any same-step scalar rescaling.
- **Closed-loop dynamics.** Couple correction generation to the evolving parameters. Lemma 1 identifies its Jacobian at a regular normalized-SAM trajectory point. Proposition 7 characterizes the frozen local variational system without assuming common eigenvectors. Lemma 5 gives an interpretable modal reduction, and Theorem 1 characterizes its admissible dynamics.

Together, the results distinguish a pure temporal transformation of an exogenous correction from the dynamical consequences that arise once correction generation is coupled back to the optimization trajectory.

C.1 Sharpness Feedback as a Two-Channel Temporal System

The update in Eq. 4 contains two temporally filtered signals: the clean gradient g_t and the adversarial correction c_t. The mechanism in Fig. 1 is therefore a relative-timing change rather than a deletion of sharpness information: the clean-gradient channel keeps the usual momentum response, whereas

the correction channel changes its zero-lag coefficient. The key question is not only how large c_t is, but where it is placed in time relative to the clean-gradient channel.

Using the z-transform, a causal response sequence $\{a_k\}_{k\geq 0}$ is represented by $A(z) = \sum_{k\geq 0} a_k z^{-k}$. Thus z is the transform variable and z^{-1} denotes a one-step delay in optimization time. Under the momentum recursion in Eq. 3, the clean-gradient channel has transfer function

$$G_\beta(z) = \frac{1}{1-\beta z^{-1}}. \tag{27}$$

This operator records how a clean gradient affects the current update and future updates through the momentum state.

Lemma 3 (Two-channel temporal factorization)**.** *Assume zero initial momentum. For the update family in Eq. 3–4, the update direction*

$$d_t = \beta m_{t-1} + g_t + \alpha c_t. \tag{28}$$

Let $D(z)$, $G(z)$, and $C_c(z)$ denote the unilateral z-transforms of d_t, g_t, and c_t, respectively, using $X(z) = \sum_{t\geq 0} x_t z^{-t}$. The update direction admits the following z-transform decomposition:

$$D(z) = G_\beta(z)G(z) + C_{\alpha,\beta}(z)C_c(z), \tag{29}$$

where

$$G_\beta(z) = \frac{1}{1-\beta z^{-1}}, \qquad C_{\alpha,\beta}(z) = \alpha + \frac{\beta z^{-1}}{1-\beta z^{-1}} = \frac{\alpha + \beta(1-\alpha)z^{-1}}{1-\beta z^{-1}}. \tag{30}$$

Consequently,

$$C_{\alpha,\beta}(z) = \underbrace{\left[\alpha + \beta(1-\alpha)z^{-1}\right]}_{P_{\alpha,\beta}(z)} G_\beta(z). \tag{31}$$

The factor $P_{\alpha,\beta}(z)$ is the relative temporal placement operator of the sharpness channel.

Proof. From Eq. 3,

$$m_{t-1} = \sum_{k=1}^{\infty} \beta^{k-1}(g_{t-k} + c_{t-k}) \tag{32}$$

with $g_j = c_j = 0$ for $j < 0$ by zero extension of the input sequences. Substituting into d_t gives

$$d_t^{(g)} = g_t + \sum_{k=1}^{\infty} \beta^k g_{t-k}, \qquad d_t^{(c)} = \alpha c_t + \sum_{k=1}^{\infty} \beta^k c_{t-k}. \tag{33}$$

Taking z-transforms yields Eq. 30, and Eq. 31 follows by division by $G_\beta(z)$. □

The relative placement operator isolates the structural difference between M-SAM and TF-SAM. For M-SAM, $\alpha = 1$, so

$$P_{\text{M-SAM}}(z) = 1. \tag{34}$$

Thus the clean-gradient and sharpness-correction channels are temporally co-located: both pass through the same momentum operator $G_\beta(z)$. For TF-SAM, $\alpha = 0$, so

$$P_{\text{TF}}(z) = \beta z^{-1}, \qquad C_{\text{TF}}(z) = \beta z^{-1} G_\beta(z). \tag{35}$$

TF-SAM therefore breaks the co-location by shifting only the sharpness channel by one step and scaling it by β. This is a temporal re-routing of the already constructed sharpness information, not a deletion of that information.

With the response-kernel notation introduced above, the correction-channel transfer function has coefficients $C_{\alpha,\beta}(z) = \sum_{k\geq 0} K_{\alpha,\beta}(k) z^{-k}$. For M-SAM and TF-SAM, these kernels are

$$K_{\text{M-SAM},\beta} = (1, \beta, \beta^2, \ldots), \qquad K_{\text{TF},\beta} = (0, \beta, \beta^2, \ldots). \tag{36}$$

Thus the kernel notation records the same temporal placement as the transfer functions, but directly in lag space. Its coefficients act on the update direction d_t; parameter increments multiply that direction by $-\eta$.

Proposition 5 (Open-loop delay equivalence). *Fix $\beta \in (0,1)$ and treat the correction sequence $\{c_t\}$ as exogenous. The TF-SAM correction channel is exactly a one-step delayed and β-scaled version of the M-SAM correction channel with the same momentum coefficient:*

$$C_{\mathrm{TF}}(z) = \beta z^{-1} C_{\text{M-SAM}}(z). \tag{37}$$

Equivalently, up to initialization transients,

$$d_t^{(c),\mathrm{TF}} = \beta d_{t-1}^{(c),\text{M-SAM}}. \tag{38}$$

Proof. For M-SAM,

$$C_{\text{M-SAM}}(z) = \frac{1}{1-\beta z^{-1}}. \tag{39}$$

For TF-SAM,

$$C_{\mathrm{TF}}(z) = \frac{\beta z^{-1}}{1-\beta z^{-1}}. \tag{40}$$

Combining the two expressions gives Eq. 37. Since multiplication by z^{-1} corresponds to a one-step delay, Eq. 38 follows in the time domain. □

The open-loop delay result identifies TF-SAM as an exact temporal transformation of the M-SAM correction response. The next result asks whether that transformation could nevertheless be reproduced by simply weakening the same-step correction.

Proposition 6 (Non-equivalence under same-step rescaling). *Assume $\beta > 0$. There exist no scalar $a \neq 0$ and no momentum coefficient $\tilde{\beta} \in [0,1)$ such that*

$$K_{\mathrm{TF},\beta}(k) = a K_{\text{M-SAM},\tilde{\beta}}(k) \qquad \text{for all } k \geq 0. \tag{41}$$

Therefore, TF-SAM cannot be obtained from M-SAM by a same-step scalar rescaling of the sharpness correction, including rescalings induced by the learning rate, perturbation radius, or correction magnitude.

Proof. A standard M-SAM correction response with momentum $\tilde{\beta}$ has impulse response $K_{\text{M-SAM},\tilde{\beta}} = (1, \tilde{\beta}, \tilde{\beta}^2, \ldots)$, while $K_{\mathrm{TF},\beta} = (0, \beta, \beta^2, \ldots)$. Equality at lag 0 in Eq. 41 requires $a = 0$. Then the scaled M-SAM response has lag-1 coefficient 0, contradicting $K_{\mathrm{TF},\beta}(1) = \beta > 0$. □

Together, the open-loop delay result and Proposition 6 distinguish temporal placement from correction strength: TF-SAM is exactly a delayed and scaled correction response, but it cannot be reproduced by any same-step rescaling of M-SAM.

C.2 Closed-Loop Variational Dynamics at Regular Points

The closed-loop analysis now allows correction generation to depend on the current parameters. At a regular point of the fixed minibatch gradient map, the main-text response matrix is the Jacobian of the actual normalized-SAM correction.

Proof of Lemma 1. For $u(w) = g(w)/\|g(w)\|$ and $\bar{g} \neq 0$, differentiating normalization gives

$$Du(\bar{w}) = \frac{1}{\|\bar{g}\|}(I_p - \bar{u}\bar{u}^\top) H_0. \tag{42}$$

Consequently $D\epsilon(\bar{w}) = \rho Du(\bar{w})$, and the chain rule for $c(w) = g(w+\epsilon(w)) - g(w)$ gives $Dc(\bar{w}) = H_+[I_p + D\epsilon(\bar{w})] - H_0$, which is Eq. 9. Differentiability gives the stated first-order expansion. □

Exact Hessian-drift decomposition. With $a = \rho/\|\bar{g}\|$ and $P_\perp = I_p - \bar{u}\bar{u}^\top$, regrouping Eq. 9 gives the identity

$$C_{\bar{w}} = \underbrace{aH_0P_\perp H_0}_{C_{\text{quad}}} + \underbrace{(H_+ - H_0)(I_p + aP_\perp H_0)}_{R_H}. \tag{43}$$

Here R_H is Hessian drift across the actual SAM perturbation segment, not a discarded Taylor remainder. Since H_0 is symmetric and $P_\perp$ is an orthogonal projection, $z^\top C_{\text{quad}} z = a\|P_\perp H_0 z\|^2 \geq 0$ for every z. If the minibatch loss has constant Hessian $H_+ = H_0 = H$ over the segment, then $R_H = 0$ and

$$C_{\bar{w}} = C_{\text{quad}} = \frac{\rho}{\|\bar{g}\|} H(I_p - \bar{u}\bar{u}^\top)H \succeq 0. \tag{44}$$

This PSD response need not commute with H: the projection generally does not share its eigenspaces. Consequently, PSD response does not imply a common eigenbasis or even a common invariant direction for the two matrices.

For the spectral operator norm, Eq. 43 and $\|P_\perp\|_2 \leq 1$ give

$$\|C_{\bar{w}} - C_{\text{quad}}\|_2 \leq \|H_+ - H_0\|_2 \left(1 + \frac{\rho}{\|\bar{g}\|}\|H_0\|_2\right). \tag{45}$$

If the Hessian is L_H-Lipschitz along the perturbation segment, then $\|H_+ - H_0\|_2 \leq L_H\rho$ and the right-hand side is at most $L_H\rho(1 + \rho\|H_0\|_2 / \|\bar{g}\|)$. Thus the constant-curvature specialization depends on Hessian variation and local gradient size, not on the nominal radius alone.

Signed response occurs for the exact normalized-SAM map even with positive clean curvature. In one dimension, let $F(w) = w + \frac{1}{2}w^2 - \frac{1}{6}w^3$ at $\bar{w} = 0$ and $0 < \rho < 1$. Then $\bar{g} = 1$, $H_0 = 1$, $H_+ = 1 - \rho$, and $P_\perp = 0$. Equation 43 gives $C_{\bar{w}} = -\rho < 0$, so the common mode has $(\lambda, \kappa) = (1, -\rho)$.

The regular-point condition is substantive. For a quadratic loss centered at $w^\star$ with $g(w^\star + x) = Hx$, fixed-radius normalized SAM gives

$$c(w^\star + x) = \rho\frac{H^2x}{\|Hx\|} \qquad (Hx \neq 0). \tag{46}$$

Along a direction with $Hx \neq 0$, this response need not tend to zero as $x \to 0$. It therefore generally has no expansion $c(w^\star + x) = Cx + o(\|x\|)$ at the exact zero-gradient point.

For a fixed minibatch sequence, compare a reference trajectory $(\bar{w}_t, \bar{m}_{t-1})$ with a neighboring one. Set $\delta s_t = (\delta w_t^\top, \delta m_{t-1}^\top)^\top$ and define $H_t = Dg_t(\bar{w}_t)$, $C_t = Dc_t(\bar{w}_t)$ whenever $g_t(\bar{w}_t) \neq 0$. No stationarity condition is imposed on the reference trajectory.

Lemma 4 (General local routed variational Jacobian)**.** *Let I_p be the $p \times p$ identity. At a regular reference state, the difference between neighboring trajectories has first-order evolution*

$$\delta s_{t+1} = \mathcal{J}_{\alpha,t}\delta s_t + o(\|\delta s_t\|), \qquad \mathcal{J}_{\alpha,t} = \begin{pmatrix} I_p - \eta_t(H_t + \alpha C_t) & -\eta_t\beta I_p \\ H_t + C_t & \beta I_p \end{pmatrix}. \tag{47}$$

For constant η and frozen H, C, the corresponding matrix $\mathcal{J}_\alpha$ is Schur stable exactly when $\text{spr}(\mathcal{J}_\alpha) < 1$*, where* spr *denotes spectral radius.*

Proof. Subtract the two updates in Eqs. 3 and 4. Differentiability at $\bar{w}_t$ gives $\delta g_t = H_t\delta w_t + o(\|\delta w_t\|)$ and $\delta c_t = C_t\delta w_t + o(\|\delta w_t\|)$. The reference gradient and correction values cancel. Thus $\delta m_t = \beta\delta m_{t-1} + (H_t + C_t)\delta w_t + o(\|\delta s_t\|)$ and $\delta w_{t+1} = [I_p - \eta_t(H_t + \alpha C_t)]\delta w_t - \eta_t\beta\delta m_{t-1} + o(\|\delta s_t\|)$. Stacking these equations proves Eq. 47. The spectral-radius statement applies to the frozen linear system. □

Along a time-varying reference trajectory, finite-horizon first-order perturbations propagate through products $\mathcal{J}_{\alpha,t+K-1}\cdots\mathcal{J}_{\alpha,t}$. The spectrum of one frozen Jacobian characterizes that local operating point; it does not alone characterize the stability of the full time-varying product.

Proposition 7 (General closed-loop routing polynomial). *For arbitrary real $p \times p$ matrices H and C, with no commutativity assumption, the Jacobian in Eq. 47 satisfies*

$$\det(rI_{2p} - \mathcal{J}_\alpha) = \det Q_\alpha(r), \qquad Q_\alpha(r) = r^2 I_p - rT_\alpha + D_\alpha, \tag{48}$$

where

$$T_\alpha = (1+\beta)I_p - \eta(H + \alpha C), \qquad D_\alpha = \beta[I_p + \eta(1-\alpha)C]. \tag{49}$$

Hence $\mathcal{J}_\alpha$ is Schur stable exactly when all roots of $\det Q_\alpha(r)$ lie in the open unit disk.

Proof. For $r \neq \beta$, take the Schur complement of the lower-right block in

$$rI_{2p} - \mathcal{J}_\alpha = \begin{pmatrix} (r-1)I_p + \eta(H + \alpha C) & \eta\beta I_p \\ -(H + C) & (r-\beta)I_p \end{pmatrix}. \tag{50}$$

Its determinant equals

$$\det((r-\beta)[(r-1)I_p + \eta(H + \alpha C)] + \eta\beta(H + C)), \tag{51}$$

which expands to Eq. 48. Both sides are polynomials in r, so the identity also holds at $r = \beta$. □

The coefficient T_α contains the zero-lag response αC, whereas D_α contains the delayed response $\beta(1-\alpha)C$. For noncommuting H, C, these matrix coefficients do not in general yield independent scalar quadratic factors. Equivalently, full-system Schur stability holds if and only if there is a symmetric positive definite P with

$$\mathcal{J}_\alpha^\top P \mathcal{J}_\alpha - P \prec 0. \tag{52}$$

Proposition 8 (Local eigenvalue sensitivity to temporal routing). *Let $r_i(\alpha)$ be a simple eigenvalue of $\mathcal{J}_\alpha$ with a differentiable local branch, right eigenvector $v_i = (v_{i,x}^\top, v_{i,m}^\top)^\top$, and left eigenvector $u_i = (u_{i,x}^\top, u_{i,m}^\top)^\top$, where $u_i^* v_i \neq 0$ and $*$ denotes conjugate transpose. Then*

$$\frac{dr_i}{d\alpha} = -\eta \frac{u_{i,x}^* C v_{i,x}}{u_i^* v_i}. \tag{53}$$

Proof. The block form in Eq. 47 gives

$$\frac{\partial \mathcal{J}_\alpha}{\partial \alpha} = \begin{pmatrix} -\eta C & 0 \\ 0 & 0 \end{pmatrix}. \tag{54}$$

For a simple differentiable eigenvalue branch, standard left–right eigenvalue differentiation yields

$$\frac{dr_i}{d\alpha} = \frac{u_i^* (\partial \mathcal{J}_\alpha / \partial \alpha) v_i}{u_i^* v_i}. \tag{55}$$

Substitution of the displayed block derivative proves Eq. 53. □

The matrix polynomial identifies where the two routes enter for arbitrary H, C. The sensitivity in Eq. 53 adds that the effect of increasing zero-lag routing on an individual eigenvalue generally has no fixed direction: it depends on C and the left–right geometry of that mode. The aligned-mode model below makes the coefficient separation explicit as a scalar stability region.

Lemma 5 (General local eigenmode dynamics). *At a frozen regular operating point, consider a real common mode v such that $Hv = \lambda v$ and $Cv = \kappa v$. This mode-wise analysis does not require H and C to be jointly diagonalizable globally; it characterizes any direction invariant under both response matrices. Along this mode, write $\delta w_t = \delta x_t v$ and $\delta m_{t-1} = \delta \mu_{t-1} v$, where δx_t and $\delta \mu_{t-1}$ are scalar coordinates. The variational dynamics have the scalar form*

$$\begin{pmatrix} \delta x_{t+1} \\ \delta \mu_t \end{pmatrix} = J_\alpha(\lambda, \kappa) \begin{pmatrix} \delta x_t \\ \delta \mu_{t-1} \end{pmatrix}, \qquad J_\alpha(\lambda, \kappa) = \begin{pmatrix} 1 - \eta(\lambda + \alpha\kappa) & -\eta\beta \\ \lambda + \kappa & \beta \end{pmatrix}. \tag{56}$$

The characteristic polynomial is

$$r^2 - \tau_\alpha r + \delta_\alpha = 0, \tag{57}$$

with

$$\tau_\alpha = 1 + \beta - \eta(\lambda + \alpha\kappa), \qquad \delta_\alpha = \beta[1 + \eta(1-\alpha)\kappa]. \tag{58}$$

Proof. Restricting Eq. 47 to perturbations along v gives

$$\delta\mu_t = \beta\delta\mu_{t-1} + (\lambda+\kappa)\delta x_t, \tag{59}$$

and

$$\delta x_{t+1} = \delta x_t - \eta[\beta\delta\mu_{t-1} + (\lambda+\alpha\kappa)\delta x_t]. \tag{60}$$

These two scalar recursions give Eq. 56; taking the trace and determinant gives Eq. 58. □

Equation 58 specializes the full matrix-polynomial separation to a common mode: immediate correction enters its trace, while delayed correction enters its determinant. More explicitly, $\alpha\kappa$ changes τ_α, whereas the momentum-mediated term $\beta(1-\alpha)\kappa$ changes δ_α. Therefore α is not an ordinary correction-magnitude parameter. At fixed κ, it changes the algebraic role of the correction in the closed loop, moving the system between a trace-controlled flip constraint and a determinant-controlled delayed-feedback constraint.

Theorem 1 (Exact aligned-mode Schur stability for signed response)**.** *Assume $\eta > 0$, $0 \le \beta < 1$, and real λ, κ. The frozen local mode in Eq. 56 is Schur stable if and only if all three inequalities hold:*

$$\lambda + [\alpha + \beta(1-\alpha)]\kappa > 0, \tag{61a}$$
$$2(1+\beta) > \eta[\lambda + \kappa((1+\beta)\alpha - \beta)], \tag{61b}$$
$$1 - \beta > \beta\eta(1-\alpha)\kappa. \tag{61c}$$

Proof of Theorem 1. For $p(r) = r^2 - \tau r + \delta$, Schur stability is equivalent to the Jury inequalities

$$1 - \tau + \delta > 0, \qquad 1 + \tau + \delta > 0, \qquad 1 - \delta > 0. \tag{62}$$

Substituting Eq. 58 gives

$$1 - \tau_\alpha + \delta_\alpha = \eta[\lambda + \kappa(\alpha + \beta(1-\alpha))], \tag{63}$$
$$1 + \tau_\alpha + \delta_\alpha = 2(1+\beta) - \eta[\lambda + \kappa((1+\beta)\alpha - \beta)], \tag{64}$$
$$1 - \delta_\alpha = 1 - \beta - \beta\eta(1-\alpha)\kappa. \tag{65}$$

Since $\eta > 0$, positivity of these expressions is equivalent to the stated conditions. □

The first identity is $p_\alpha(1) > 0$: equality gives an $r = +1$ root, where negative correction response can cancel positive clean restoring curvature. The second is $p_\alpha(-1) > 0$, the flip boundary. The third limits the root product δ_α; at $\delta_\alpha = 1$ with the other two strict, a complex-conjugate pair lies on the unit circle. For fixed $\kappa < 0$ and $\beta > 0$, increasing α relaxes the flip inequality but tightens both the determinant and $r = +1$ restoring inequalities. For $\kappa > 0$, increasing α tightens the flip inequality, relaxes the determinant inequality, and makes the restoring inequality easier to satisfy. When $\beta = 0$, the determinant inequality is independent of α.

Table 4: Schur boundaries for a frozen common mode with signed correction response. All three strict Jury inequalities are required.

Jury condition	Boundary	Local interpretation
$1 - \tau_\alpha + \delta_\alpha > 0$	$r = +1$	Restoring response
$1 + \tau_\alpha + \delta_\alpha > 0$	$r = -1$	Flip response
$1 - \delta_\alpha > 0$	$\delta_\alpha = 1$	Root-product constraint

Corollary 1 (Simultaneously diagonalizable system)**.** *If real H and C are simultaneously diagonalizable over the reals, let (λ_i, κ_i) be their paired eigenvalues. Then the frozen matrix $\mathcal{J}_\alpha$ is Schur stable if and only if every pair satisfies the three conditions of Theorem 1. In particular, real symmetric commuting H, C satisfy this decomposition.*

Proof. Transform both state coordinates by the common eigenvector matrix. After pairing each position coordinate with its momentum coordinate, the transformed Jacobian is the direct sum of the

blocks $J_\alpha(\lambda_i, \kappa_i)$ in Eq. 56. Its spectrum is their union, so the full system is Schur stable exactly when every block is. □

Corollary 2 (Positive-response common mode)**.** *If $\lambda > 0$ and $\kappa \geq 0$, the restoring condition in Eq. 61a holds automatically. The frozen mode is Schur stable exactly when Eqs. 61b and 61c hold.*

The constant-curvature normalized-SAM case $H_+ = H_0 = H$ yields $C_{\bar{w}} = C_{\text{quad}} \succeq 0$, so $\kappa \geq 0$ on any common invariant direction. Positive semidefiniteness alone does not supply such directions. Neither routing extreme is uniformly more stable even in this specialization.

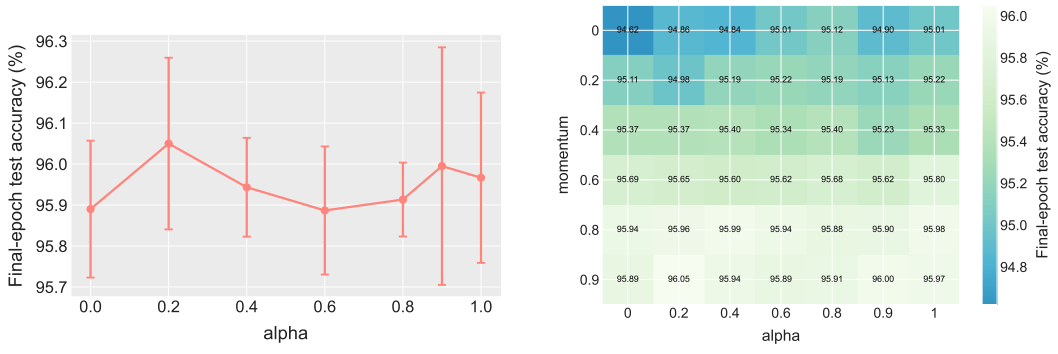
D Additional Material

D.1 Controlled temporal routing

The controlled-routing experiment measures how sensitivity to zero-lag sharpness routing changes with the momentum regime. At fixed momentum coefficient β, varying α changes the zero-lag contribution of the freshly generated correction while leaving the ordinary momentum coefficient unchanged. Let $A(\alpha, \beta)$ denote the mean final test accuracy at that routing pair. The resulting interaction is summarized by the endpoint penalty $P_{\text{zero}}(\beta) = A(1, \beta) - A(0, \beta)$ and the descriptive routing span $R(\beta) = \max_\alpha A(\alpha, \beta) - \min_\alpha A(\alpha, \beta)$ in the following summary (percentage points, ten seeds per setting):

- $\beta = 0.0$: $P_{\text{zero}} = +0.39$, $R = 0.50$; $\beta = 0.2$: $P_{\text{zero}} = +0.11$, $R = 0.24$.
- $\beta = 0.4$: $P_{\text{zero}} = -0.04$, $R = 0.17$; $\beta = 0.6$: $P_{\text{zero}} = +0.12$, $R = 0.20$.
- $\beta = 0.8$: $P_{\text{zero}} = +0.04$, $R = 0.10$; $\beta = 0.9$: $P_{\text{zero}} = +0.08$, $R = 0.11$.

When $\beta = 0$, no momentum-mediated correction pathway exists; at $\beta = 0.8$ and 0.9, momentum is high. In this regime comparison, the routing span is 0.50 percentage points at $\beta = 0$ and 0.10 and 0.11 points at $\beta = 0.8$ and 0.9, respectively; the corresponding endpoint penalties are $+0.39$, $+0.04$, and $+0.08$ points. **These results show that sensitivity to the zero-lag route is strongest when the momentum-mediated pathway is unavailable and becomes weak in the high-momentum regime.** At fixed momentum, the endpoint sweep gives 95.89 ± 0.17 for $\alpha = 0$ and 95.97 ± 0.21 for $\alpha = 1$, with non-monotone intermediate values (Fig. 4a). **Thus, the endpoint means alone do not capture a monotone benefit from increasing the zero-lag gain.** Because β also changes ordinary gradient momentum and the delayed correction tail, we interpret the grid as a structural routing interaction rather than a causal isolation of delayed sharpness feedback. The supporting no-momentum check, ablation, and full grid are reported in Appendix D.2.



(a) Fixed-momentum test-accuracy sweep.

(b) Full $\alpha \times \beta$ routing grid.

Figure 4: Controlled routing results on CIFAR-10/ResNet-18. The left panel varies α at fixed momentum; the right panel contains 42 settings with ten seeds per cell. Because β controls both ordinary momentum and the delayed correction tail, the grid is a structural interaction diagnostic rather than a causal isolation of delayed sharpness feedback.

Table 5: CIFAR-10/ResNet-18 routing ablation over ten seeds.

Method	Momentum	Final accuracy	Best accuracy
SGD	no	94.47 ± 0.11	94.52 ± 0.07
SAM	no	94.95 ± 0.12	95.10 ± 0.10
SGD	yes (0.9)	95.44 ± 0.14	95.50 ± 0.13
M-SAM	yes (0.9)	95.86 ± 0.05	95.94 ± 0.06
TF-SAM $(\alpha = 0)$	yes (0.9)	95.97 ± 0.08	96.05 ± 0.06
TF-SAM $(\alpha = 0.05)$	yes (0.9)	96.02 ± 0.11	96.09 ± 0.04

The no-momentum setting provides a structural anchor for the grid. When $\alpha = 0$ and $\beta = 0$, the parameter update is exactly SGD because neither the zero-lag correction nor a momentum-mediated tail remains. When $\alpha = 0$ and $\beta > 0$, the zero-lag route is removed but the correction can still act through the momentum tail; when $\alpha = 1$ and $\beta > 0$, both the immediate and delayed routes are present as in M-SAM. Thus the grid moves from no sharpness route, to delayed-only routing, to immediate-plus-delayed routing.

The structural anchor is also visible in Table 5. Removing both the zero-lag route and the momentum-mediated route reduces TF-SAM exactly to SGD by the update identity, whereas restoring momentum allows a sharpness correction to influence future updates even when $\alpha = 0$. The observed sequence from no-momentum SGD to momentum SGD, M-SAM, and TF-SAM is therefore consistent with the routing interpretation rather than a simple correction-size explanation.

D.2 Experimental configuration

This section gives the configuration details shared by the experiments. The anonymous repository linked in Section 4 contains the training, analysis, and table-generation entry points, with dependencies managed by `uv` and common commands exposed through `just`.

Compute hardware. Experiments were run on a system equipped with an Intel Core Ultra 7 270K Plus processor and two GPUs: an NVIDIA GeForce RTX 4070 Ti and an NVIDIA GeForce RTX 4090.

Datasets, models, and matched training budgets. The image experiments use CIFAR-10 and CIFAR-100 with ResNet-18 and ResNet-34 (He et al., 2016). The text experiments use TinyBERT on SST-2 and GPT-2 on SST-5 (Jiao et al., 2020; Radford et al., 2019). SST-2 is part of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019); SST-5 uses the five-class Stanford Sentiment Treebank setup (Socher et al., 2013). Table 6 lists the exact configurations used for the six main dataset–model pairs. Within each pair, SGD, M-SAM, and TF-SAM use the same model initialization protocol, data split and preprocessing, learning rate, momentum, weight decay, batch size, number of epochs, learning- rate schedule, and training seeds. Thus the main comparison changes the optimizer's sharpness-aware update rule while keeping the base optimization budget fixed. The values in the table are taken from the generated result table and the saved experiment configurations.

For image models, the scheduler is `CosineAnnealingLR` with a final learning rate of 10^{-6}. For NLP models, the scheduler is a step-wise cosine schedule with a linear warm-up over 10% of the total optimizer updates. The SAM perturbation radius is $\rho = 0.05$ in the main comparison, and TF-SAM uses $\alpha = 0$; these are the only additional optimizer-specific settings relevant to the comparison. M-SAM and TF-SAM use the same correction-generation procedure and differ in how the generated correction is routed through momentum.

Hyperparameter selection. For GPT-2/SST-5, the training data are partitioned into training and held-out validation subsets, and Optuna hyperparameter search uses only validation accuracy. The selected configurations are then fixed and used for the corresponding main runs, whose reported results are evaluated on the test set. The resulting base configurations are shared across SGD, M-SAM, and TF-SAM within each dataset–model pair.

Table 6: Training configurations for the main comparison. The entries in the first eight columns are shared by SGD, M-SAM, and TF-SAM within each dataset–model pair; ρ and α are optimizer-specific and are reported separately below. Here LR denotes learning rate and WD denotes weight decay. Image schedules are stepped once per epoch, whereas NLP schedules are stepped once per optimizer update.

Dataset	Model	LR	Momentum	WD	Batch	Epochs	Schedule	Seeds
CIFAR-10	ResNet-18	0.1	0.9	5×10^{-4}	128	200	Cosine, $\eta_{\min} = 10^{-6}$	10
CIFAR-10	ResNet-34	0.1	0.9	5×10^{-4}	128	200	Cosine, $\eta_{\min} = 10^{-6}$	10
CIFAR-100	ResNet-18	0.1	0.9	5×10^{-4}	128	200	Cosine, $\eta_{\min} = 10^{-6}$	10
CIFAR-100	ResNet-34	0.1	0.9	5×10^{-4}	128	200	Cosine, $\eta_{\min} = 10^{-6}$	10
SST-2	TinyBERT	10^{-4}	0.9	10^{-4}	32	10	10% linear warm-up + cosine	10
SST-5	GPT-2	0.0049	0.8	10^{-3}	16	10	10% linear warm-up + cosine	10

Table 7: Seed-level paired aggregate corruption/noise robustness. For each training seed, accuracy is first averaged over the non-clean corruption/noise conditions, after which the TF-SAM–M-SAM gap is computed between matched seeds. Positive values favor TF-SAM. CIs quantify uncertainty across training seeds and are not equivalence tests.

Setting	Conditions	Seeds	TF-SAM–M-SAM gap (95% CI)
CIFAR-10 / ResNet-18, $\rho = 0.05$	12	10	$0.35\ [-0.46, 1.16]$
CIFAR-10 / ResNet-34, $\rho = 0.05$	12	10	$0.05\ [-1.19, 1.29]$
CIFAR-100 / ResNet-18, $\rho = 0.05$	12	10	$-0.02\ [-0.34, 0.30]$
CIFAR-100 / ResNet-34, $\rho = 0.05$	12	10	$0.10\ [-0.54, 0.73]$
SST-2 / TinyBERT, $\rho = 0.05$	13	10	$0.18\ [0.02, 0.34]$
SST-5 / GPT-2, $\rho = 0.05$	13	10	$0.11\ [-0.12, 0.34]$

Evaluation protocol and coverage. Main benchmark results are means and standard deviations over ten independent training seeds (42, 10, 3407, 112966, 123456, 987654, 2021, 2022, 2023, and 2024), with both final-checkpoint and best-checkpoint accuracy reported in the generated result table. The target-matched geometry analyses on CIFAR-10/ResNet-18 and GPT-2/SST-5, the α sweep, and the focused momentum ablation each use ten seeds; each point in the target-matched plots is one seed-level estimate. The $\alpha \times \beta$ diagnostic covers 42 settings with ten seeds per cell. The six benchmark settings cover two image datasets, two image architectures, and two text architectures. This combination of matched training budgets, multiple seeds, heterogeneous architectures, and both benchmark and mechanism-level evaluations makes the central comparison controlled while testing whether the observed routing behavior transfers across model and task families.

Diagnostics and robustness. The correction diagnostics use logged checkpoints and report means with one-standard-deviation bands. Corruption robustness is evaluated from final checkpoints on twelve fixed image or thirteen fixed text corruptions using the ten main-comparison seeds; the main text reports the aggregate comparison by training seed.

D.3 Aggregate corruption robustness

The training seed is the inferential unit. For each matched seed, accuracy is first averaged over the non-clean corruption or noise conditions, and the paired aggregate gap between TF-SAM and M-SAM is then computed. This pre-defined aggregate is the robustness quantity used in the main claim.

Table 7 reports the seed-level aggregate analysis. The Conditions column gives the number of non-clean conditions averaged within each seed, and the CI is computed across the ten matched training-seed gaps. Across the six tested settings, five intervals include zero, and the only interval entirely above zero is SST-2/TinyBERT. No setting has an interval entirely below zero. Together, these results suggest that removing the zero-lag route does not produce a systematic aggregate robustness degradation in the tested regimes; they do not establish statistical equivalence or uniform superiority of either temporal routing.

Table 8: Per-corruption robustness on CIFAR-100 with ResNet-18 (accuracy in %).

Method	Clean	Brightness	Contrast	Elastic	Gaussian Blur	Gaussian Noise	Heavy Blur
SGD	78.40 ± 0.26	76.74 ± 0.32	77.11 ± 0.37	71.48 ± 0.37	11.83 ± 1.41	17.04 ± 0.63	4.61 ± 0.72
M-SAM	**78.78 ± 0.26**	**77.28 ± 0.29**	**77.70 ± 0.37**	71.97 ± 0.39	11.88 ± 0.78	17.32 ± 1.40	4.75 ± 0.64
TF-SAM	**78.78 ± 0.33**	77.18 ± 0.28	77.60 ± 0.26	**71.98 ± 0.57**	**12.20 ± 0.69**	**17.77 ± 1.01**	**4.88 ± 0.41**
Method	Posterize	Rotation	Saturation	Sharpness	Shot Noise	Solarize	
SGD	72.17 ± 0.39	65.66 ± 0.36	77.82 ± 0.32	66.99 ± 0.68	**3.52 ± 0.70**	22.39 ± 0.64	
M-SAM	72.98 ± 0.48	66.80 ± 0.63	**78.20 ± 0.32**	**67.79 ± 0.57**	3.50 ± 0.67	**22.77 ± 0.35**	
TF-SAM	**73.01 ± 0.27**	**66.82 ± 0.25**	78.20 ± 0.28	67.76 ± 0.62	3.48 ± 0.48	22.76 ± 0.65	

Table 10: Per-condition noise robustness on SST-2 with TinyBERT (accuracy in %).

Method	Clean	Char Delete	Char Insert	Char Substitute	Char Swap	Lowercase	Pad Noise
SGD	85.22 ± 0.49	68.29 ± 0.92	69.19 ± 0.24	64.29 ± 0.64	68.92 ± 0.46	85.22 ± 0.49	84.69 ± 0.38
M-SAM	**85.62 ± 0.31**	69.21 ± 0.45	71.17 ± 0.27	66.27 ± 0.26	69.97 ± 0.34	**85.62 ± 0.31**	85.16 ± 0.19
TF-SAM	**85.62 ± 0.32**	**69.50 ± 0.44**	**71.43 ± 0.36**	**66.40 ± 0.38**	**70.21 ± 0.29**	**85.62 ± 0.32**	**85.39 ± 0.35**
Method	Spelling Error	Truncate 50%	Truncate 80%	Uppercase	Word Delete	Word Insert	Word Swap
SGD	85.33 ± 0.60	**73.25 ± 0.36**	**62.50 ± 0.46**	85.22 ± 0.49	**83.60 ± 0.62**	84.72 ± 0.27	84.97 ± 0.23
M-SAM	85.78 ± 0.31	72.58 ± 0.24	60.53 ± 0.17	**85.62 ± 0.31**	82.80 ± 0.22	85.02 ± 0.34	**85.30 ± 0.14**
TF-SAM	**85.81 ± 0.24**	73.13 ± 0.20	60.67 ± 0.20	**85.62 ± 0.32**	82.96 ± 0.24	**85.19 ± 0.37**	85.28 ± 0.28

For completeness, Tables 8 and 9 report per-corruption accuracies on CIFAR-100, while Tables 10 and 11 report per-condition accuracies for the NLP settings. Entries give mean ± standard deviation over ten training seeds, in percentage points; shading marks the highest mean in each condition. The aggregate comparison above uses paired seed-level averages over each setting's non-clean conditions.

Table 9: Per-corruption robustness on CIFAR-100 with ResNet-34 (accuracy in %).

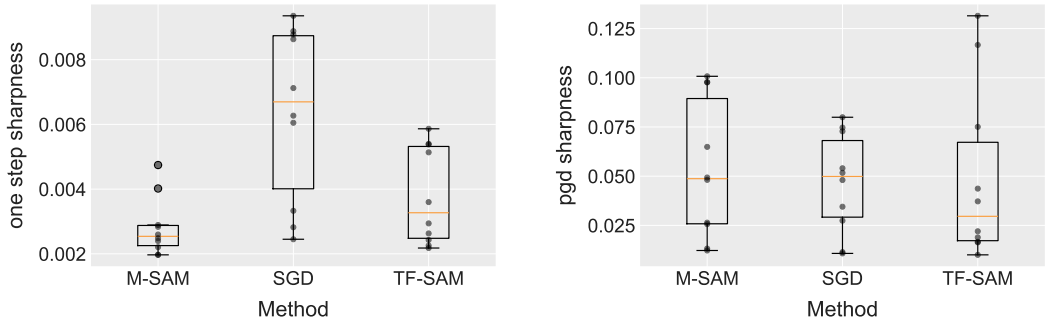
Method	Clean	Brightness	Contrast	Elastic	Gaussian Blur	Gaussian Noise	Heavy Blur
SGD	78.55 ± 0.58	77.03 ± 0.61	77.33 ± 0.61	72.14 ± 0.70	13.22 ± 1.60	21.07 ± 3.07	5.30 ± 0.97
M-SAM	**79.58 ± 0.33**	77.87 ± 0.46	**78.44 ± 0.38**	**73.09 ± 0.44**	**14.26 ± 1.68**	20.99 ± 2.10	**5.65 ± 0.82**
TF-SAM	79.56 ± 0.33	**77.96 ± 0.45**	78.37 ± 0.50	73.00 ± 0.48	13.58 ± 1.46	**21.27 ± 2.25**	5.51 ± 0.72
Method	Posterize	Rotation	Saturation	Sharpness	Shot Noise	Solarize	
SGD	73.01 ± 0.75	66.95 ± 0.46	78.07 ± 0.57	68.05 ± 0.96	3.62 ± 0.71	23.48 ± 0.73	
M-SAM	**74.07 ± 0.34**	67.91 ± 0.38	**79.08 ± 0.34**	**69.90 ± 0.96**	3.78 ± 0.54	**24.16 ± 0.67**	
TF-SAM	73.99 ± 0.36	**68.19 ± 0.45**	79.07 ± 0.33	69.65 ± 0.87	**4.00 ± 0.63**	23.91 ± 0.53	

D.4 Post-hoc training-set loss sharpness

A complementary geometry diagnostic compares per-seed local training-loss increases at final CIFAR-10/ResNet-34 checkpoints. Figure 5 summarizes the one-step and PGD scores; lower values indicate lower measured local sensitivity under these searches. The ResNet-18 boxplots are paired with its loss-landscape visualization in Fig. 3.

Table 11: Per-condition noise robustness on SST-5 with GPT-2 (accuracy in %).

Method	Clean	Char Delete	Char Insert	Char Substitute	Char Swap	Lowercase	Pad Noise
SGD	**53.63 ± 0.46**	41.24 ± 0.73	42.17 ± 0.65	37.46 ± 0.67	**40.15 ± 0.68**	**53.63 ± 0.46**	**53.16 ± 0.57**
M-SAM	52.29 ± 0.38	**42.00 ± 0.59**	**42.71 ± 0.53**	**38.94 ± 0.68**	39.87 ± 0.40	52.29 ± 0.38	51.63 ± 0.45
TF-SAM	52.82 ± 0.45	41.95 ± 0.50	42.43 ± 0.64	38.85 ± 0.61	39.94 ± 0.51	52.82 ± 0.45	52.31 ± 0.56
Method	Spelling Error	Truncate 50%	Truncate 80%	Uppercase	Word Delete	Word Insert	Word Swap
SGD	**53.19 ± 0.47**	36.66 ± 1.10	26.43 ± 0.73	**44.00 ± 0.77**	**51.06 ± 0.36**	**52.93 ± 0.41**	**52.03 ± 0.57**
M-SAM	52.07 ± 0.36	**38.72 ± 0.50**	**30.45 ± 0.48**	43.13 ± 0.73	50.14 ± 0.37	51.66 ± 0.46	51.39 ± 0.60
TF-SAM	52.58 ± 0.48	38.18 ± 0.52	29.47 ± 0.69	43.43 ± 0.70	50.30 ± 0.62	52.22 ± 0.28	51.74 ± 0.57



(a) ResNet-34, one-step.

(b) ResNet-34, PGD.

Figure 5: Per-seed local training-loss increases at final CIFAR-10/ResNet-34 checkpoints. Each boxplot summarizes ten training seeds. One-step sharpness evaluates the normalized-gradient SAM perturbation; PGD sharpness is the best increase found by the specified multi-start search.

For each of the ten matched training seeds, evaluation uses the same fixed subset of 1,280 training examples (ten batches of 128; evaluation seed 12345), with training and evaluation radii both set to $\rho = 0.05$. The one-step score is the loss increase after a single SAM perturbation in the normalized gradient direction within the parameter-space ℓ_2 ball. The PGD score is the largest loss increase found by ten projected gradient-ascent steps of size 0.0125 across five restarts (one zero start and four random starts), also taking the one-step SAM candidate into account. Both scores are nonnegative increases over the unperturbed loss on this subset.

Table 12 reports paired differences across training seeds. For both architectures, M-SAM and TF-SAM have lower one-step sharpness than SGD, with all paired 95% CIs below zero. The TF-SAM–M-SAM one-step difference is near zero for ResNet-18; for ResNet-34 it is small and positive, indicating a slightly larger measured loss increase under TF-SAM. Under the stronger PGD search, all three paired comparisons have CIs that include zero for both architectures, and the point estimates do not give a consistent ordering. Thus, the one-step result supports lower local sensitivity for the two SAM variants relative to SGD on this fixed training subset, while the multi-start result does not resolve differences between optimizers. These post-hoc training-loss measurements are not test-set sharpness or adversarial accuracy, and they do not establish that flatness mediates generalization or the effect of temporal routing.

Table 12: Paired seed-level differences in local training-set sharpness. Entries are mean differences with two-sided 95% t CIs over ten matched training seeds. Negative values indicate a smaller loss increase for the first method in the comparison.

Architecture	Metric	M-SAM − SGD	TF-SAM − SGD	TF-SAM − M-SAM
ResNet-18	One-step	$-0.00208\ [-0.00352, -0.00064]$	$-0.00162\ [-0.00275, -0.00048]$	$+0.00046\ [-0.00004, 0.00096]$
ResNet-18	PGD	$-0.02194\ [-0.04459, 0.00070]$	$-0.00719\ [-0.03708, 0.02270]$	$+0.01475\ [-0.00230, 0.03180]$
ResNet-34	One-step	$-0.00356\ [-0.00568, -0.00144]$	$-0.00259\ [-0.00499, -0.00019]$	$+0.00097\ [+0.00003, +0.00191]$
ResNet-34	PGD	$+0.00707\ [-0.02624, 0.04039]$	$+0.00228\ [-0.03512, 0.03968]$	$-0.00479\ [-0.02165, 0.01206]$

D.5 Checkpoint loss-landscape visualizations

Two-dimensional slices of the training-loss surface around representative final checkpoints complement the quantitative sharpness measurements above. For each architecture, SGD, TF-SAM, and M-SAM are evaluated along the same two filter-normalized parameter directions, over the shared coordinate range $[-0.08, 0.08]$, with common contour levels and color scale. The cyan marker denotes the unperturbed checkpoint, and the plotted value is the increase in training loss relative to that checkpoint. Both visualizations use seed 42, four training minibatches, and frozen batch-normalization statistics; the test accuracies printed in the panels are for these individual checkpoints.

The slices show distinct local geometries despite the similar predictive performance of these checkpoints. For ResNet-18, the SGD contours are strongly direction-dependent, whereas TF-SAM has a

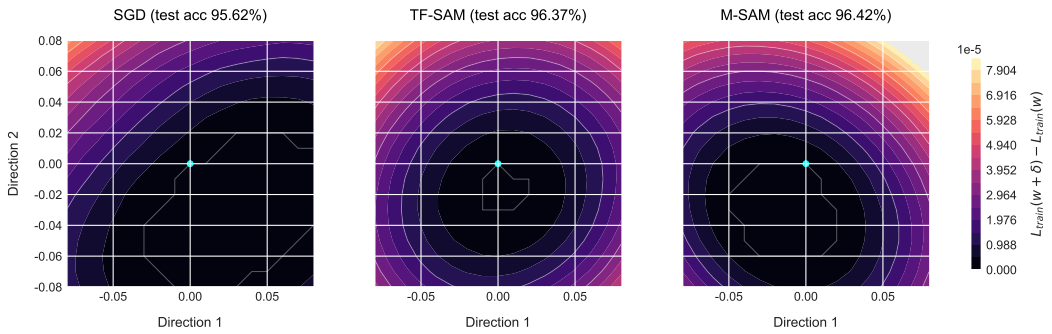


Figure 6: Two-dimensional training-loss slices around the final CIFAR-10/ResNet-34 checkpoints (seed 42), using the shared-direction protocol described in Fig. 3. The cyan point marks the unperturbed checkpoint; colors show the training-loss increase relative to it. The slice illustrates local anisotropy and should not be interpreted as a global flatness ranking.

Table 13: Paired-seed accuracy differences between TF-SAM and M-SAM. Entries are mean differences in percentage points (TF-SAM − M-SAM) and two-sided 95% t-based CIs. Positive values favor TF-SAM. The intervals quantify uncertainty in the paired effect and should not be interpreted as equivalence tests.

Setting	n	Final gap	95% CI	Best gap	95% CI
CIFAR-10 / ResNet-18	10	0.00	$[-0.10, 0.10]$	-0.02	$[-0.10, 0.06]$
CIFAR-10 / ResNet-34	10	-0.03	$[-0.14, 0.07]$	-0.04	$[-0.15, 0.08]$
CIFAR-100 / ResNet-18	10	-0.00	$[-0.29, 0.29]$	0.03	$[-0.24, 0.29]$
CIFAR-100 / ResNet-34	10	-0.02	$[-0.31, 0.28]$	-0.05	$[-0.34, 0.24]$
SST-2 / TinyBERT ($\rho = 0.05$)	10	-0.00	$[-0.15, 0.15]$	0.03	$[-0.06, 0.13]$
SST-5 / GPT-2 ($\rho = 0.05$)	10	0.53	$[0.30, 0.76]$	0.63	$[0.41, 0.85]$

smoother, more balanced contour pattern in this displayed subspace; M-SAM also has a broad low-loss region, with a different directional shape. For ResNet-34, SGD again has a visibly tilted and asymmetric slice, while TF-SAM has comparatively regular contours; the M-SAM slice is broad but more directionally structured. These observations are consistent with TF-SAM retaining a nontrivial local geometry after zero-lag feedback is removed, rather than reverting to an obviously SGD-like slice. The figures are qualitative: a two-dimensional subspace cannot rank flatness in the full parameter space, and visual contour shape does not establish a generalization advantage. Quantitative comparisons are given by the one-step and PGD sharpness measurements above.

D.6 Paired-seed statistical analysis

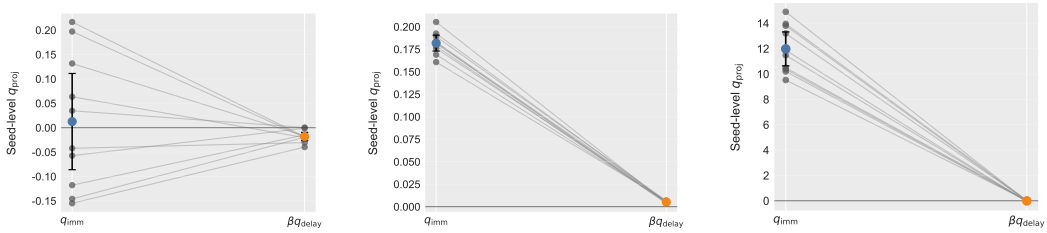
Means and standard deviations in the main results summarize each method separately, but the comparison between M-SAM and TF-SAM is paired by training seed. The paired effect is quantified for each shared seed s as

$$d_s = \mathrm{Acc}^{(s)}_{TF-SAM} - \mathrm{Acc}^{(s)}_{\text{M-SAM}}, \tag{66}$$

and report the mean gap together with a two-sided 95% t-based confidence interval,

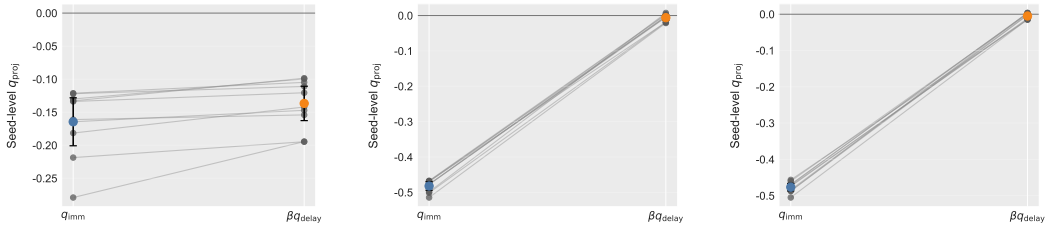
$$\bar{d} \pm t_{0.975,n-1}\frac{s_d}{\sqrt{n}}, \tag{67}$$

where $\bar{d} = n^{-1}\sum_{s=1}^{n} d_s$ is the mean paired gap, $n = 10$ is the number of shared seeds, and s_d is the sample standard deviation of the paired differences. Positive values favor TF-SAM. The analysis uses the recorded runs with matching hyperparameters for M-SAM and TF-SAM and is intended to quantify uncertainty in the paired effect, rather than to establish equivalence.



(a) Early. (b) Middle. (c) Late.

Figure 7: Phase-wise target-matched projections for M-SAM on CIFAR-10/ResNet-18. Gray segments pair the same seed, and colored markers show means across ten seeds. The immediate projection becomes positive in the middle and late phases, while the momentum-weighted delayed projection remains near zero.



(a) Early. (b) Middle. (c) Late.

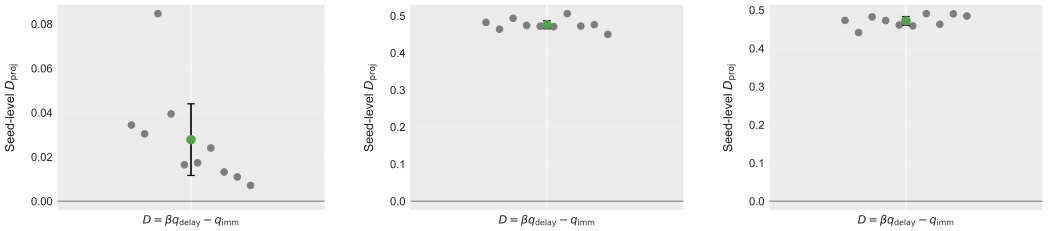
Figure 8: Phase-wise target-matched projections for M-SAM on GPT-2/SST-5. Gray segments pair the same training seed, and colored markers show means across ten seeds. The horizontal line marks zero; negative projections are adverse. The immediate projection is strongly negative in the middle and late phases, while the momentum-weighted delayed projection is nearly neutral.

Final-accuracy point estimates are near zero in five settings and positive for SST-5/GPT-2. The five near-zero final-accuracy intervals include zero, whereas the SST-5/GPT-2 interval is entirely positive; the same pattern holds for the best-accuracy intervals. The interval widths vary substantially, so the ten-seed experiments still cannot resolve small effects in some higher-variance regimes. Thus, the point estimates do not reveal a systematic penalty from removing the zero-lag route. The target-matched comparison in Fig. 9 is qualitatively consistent with a limited mitigation mechanism: at the same target gradient, the delayed correction is less adverse than the immediate correction while remaining negative. This association is suggestive rather than causal.

Taken together, these paired results do not establish statistical equivalence or a uniform advantage for either routing strategy. In particular, a CI that crosses zero is compatible with a positive, negative, or negligible paired effect. The supported conclusion is narrower: removing the zero-lag route does not produce a consistent accuracy penalty across the tested settings, while the direction and magnitude of the effect remain regime dependent.

D.7 Target-matched immediate versus delayed correction geometry

Equation 8 and Figs. 2 and 9 report the target-matched comparisons. The protocol divides training into the first, middle, and final epoch thirds, requires an exact one-step lag, and discards non-contiguous pairs. Each seed contributes 49 valid target updates in each phase. Within a seed and phase, each projection is aggregated as a ratio of sums; uncertainty is computed across the ten seed-level estimates. The training seed is the inferential unit, rather than the individual optimizer step.



(a) Early. (b) Middle. (c) Late.

Figure 9: Target-matched routing difference on GPT-2/SST-5 for M-SAM, $D = \beta q_{\text{delay}} - q_{\text{imm}}$. Gray points show individual estimates and the green marker their mean; positive values indicate that the delayed route is more favorable, or less adverse, at the same target update. The difference is modest in the early phase and substantially larger in the middle and late phases.

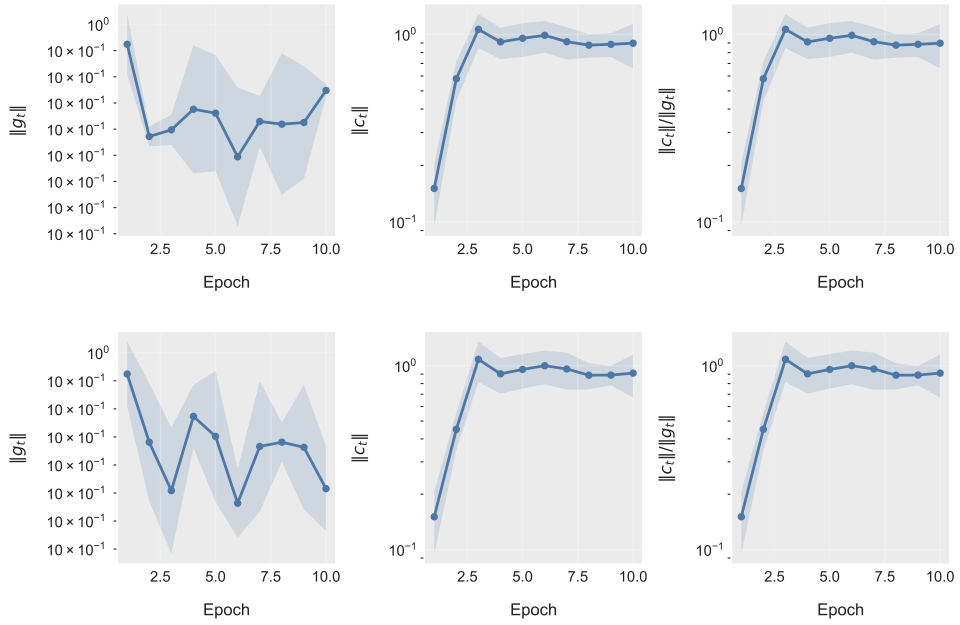


Figure 10: Clean-gradient and correction norms for GPT-2/SST-5 at $\rho = 0.05$. Top: M-SAM. Bottom: TF-SAM with $\alpha = 0$. Each panel reports mean $\pm$ standard deviation over ten seeds. The correction remains on the clean-gradient scale in both methods.

At a fixed target iterate, a correction r_t contributes $-\eta\langle g_t, r_t\rangle$ to the first-order change in minibatch loss. M-SAM uses $r_t = c_t$; the delayed route uses $r_t = \beta c_{t-1}$. Thus $q > 0$ is favorable to descent and $q < 0$ opposes it, while $D_t > 0$ favors the delayed correction. Both projections share the target gradient and denominator, isolating correction age in this local linear comparison.

For seed s and phase p, let $\mathcal{I}_{s,p}$ be the set of retained target updates. The phase-wise projections are

$$q_{\text{imm}}^{(s,p)} = \frac{\sum_{t\in\mathcal{I}_{s,p}}\langle c_t, g_t\rangle}{\sum_{t\in\mathcal{I}_{s,p}}\|g_t\|^2}, \qquad q_{\text{delay}}^{(s,p)} = \frac{\sum_{t\in\mathcal{I}_{s,p}}\langle c_{t-1}, g_t\rangle}{\sum_{t\in\mathcal{I}_{s,p}}\|g_t\|^2}. \tag{68}$$

The shared denominator makes each phase value a ratio of sums, which weights updates by clean-gradient energy. The phase-wise difference is defined as $D^{(s,p)} = \beta q_{\text{delay}}^{(s,p)} - q_{\text{imm}}^{(s,p)}$.

The phase-wise plots show distinct regimes. On CIFAR-10/ResNet-18, early immediate projections vary in sign across seeds, yielding mixed routing differences. In the middle and late phases, the immediate projection is positive while the delayed projection stays near zero, favoring the fresh

correction. Its larger late-phase normalized value reflects alignment relative to $\|g_t\|^2$ and does not establish a larger absolute loss change. On GPT-2/SST-5, both projections are adverse early, with delay less negative; midway and late, the immediate projection remains strongly negative while the delayed one is near zero, and all ten seed-level differences favor delay. Here, delay attenuates an adverse projection without making it strongly descent-aligned.

Overall, delaying removes part of the fresh projection, helping when it is adverse and removing a favorable contribution when it is positive. These plots compare corrections along the same observed M-SAM trajectory; they omit the quadratic remainder and do not simulate the counterfactual M-SAM and TF-SAM trajectories. They support a local geometric interpretation, not a causal explanation of final accuracy differences.

D.8 GPT-2 CORRECTION-MAGNITUDE SANITY CHECK

The GPT-2/SST-5 regime provides a complementary magnitude check. In the logged $\rho = 0.05$ runs, the correction norm rapidly reaches the same order as the clean-gradient norm for both M-SAM and TF-SAM.

This check supports the narrower mechanism claim that the GPT-2 pattern is not explained by a vanishing correction. It does not by itself establish a causal training-time explanation; the target-matched comparison remains the direct first-order evidence.