CHAMELEON: LEARNING ANISOTROPIC PERTURBATION GEOMETRY IN SHARPNESS-AWARE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-Aware Minimization (SAM) improves generalization by optimizing the worst-case loss within a neighborhood of model parameters, thereby biasing training toward flatter minima. Despite its effectiveness, standard SAM uses a fixed isotropic perturbation region, which can be mismatched with the highly anisotropic curvature of modern loss landscapes. This geometric mismatch becomes particularly limiting under distribution shifts, where sensitive directions may drift over time and a spherical neighborhood can either over-regularize flat directions or under-explore sharp ones. We propose Chameleon, a lightweight sharpness-aware optimizer that learns an anisotropic perturbation geometry. Chameleon replaces SAM's fixed-radius ball with a learnable ellipsoidal neighborhood by tensorizing perturbation scales into adaptive parameters. These scales are optimized through a normalized sharpness surrogate that preserves first-order efficiency while encouraging perturbations to align with curvature-sensitive directions. Theoretically, we characterize the learned geometry as a curvature-sensitive axis-aligned metric that coordinate-wise reweights dominant spectral components, establish sublinear convergence under clipped adaptive scales, and derive containment-based PAC-Bayes generalization bounds complemented by a matched-budget curvature analysis. Extensive experiments evaluating multi-modal capabilities across vision, language, and industrial click-through rate (CTR) prediction tasks demonstrate consistent improvements in cross-domain adaptation. Specifically, Chameleon achieves up to 3.91% cross-domain and 2.09% in-domain accuracy gains on vision tasks, 4.19% F1 improvement on language tasks, and 0.71% AUC gains on CTR tasks.

1 INTRODUCTION

Recently, agentic AI has emerged as a promising paradigm for intelligent systems operating in dynamic real-world environments Wang et al. (2024); Xi et al. (2025). However, performance is fundamentally challenged by non-stationary task streams where data distributions, physical dynamics, and operational contexts continuously evolve Wei & Luo (2021). For instance, in embodied AI and autonomous driving, agents trained in high-fidelity simulators often suffer severe reality gaps caused by lighting variations, sensor noise, or weather shifts Zhao et al. (2020); Yang et al. (2023). Such unpredictable shifts render static models insufficient, underscoring the need for mechanisms ensuring robust generalization under unforeseen shifts while enabling rapid adaptation to evolving environments.

To address distribution shifts in non-stationary environments, prior work enhances robustness via representation alignment Long et al. (2015; 2017), risk-invariance learning Arjovsky et al. (2019); Gulrajani & Lopez-Paz (2020), or test-time adaptation Zhang et al. (2022). Despite their effectiveness, many methods rely on explicit domain labels Arjovsky et al. (2019), target-domain batches for alignment Sun (2016), or computationally intensive auxiliary networks. These dependencies limit their applicability when domain boundaries are ambiguous or real-time inference is required. To achieve domain-agnostic adaptation, we consider Sharpness-Aware Minimization (SAM) as a standalone optimizer-level framework. By penalizing worst-case perturbed loss within a parameter neighborhood, SAM biases convergence toward flatter minima, providing a geometric prior for

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.

improved generalization without requiring domain labels or architectural modifications Foret et al. (2021); Jiang et al. (2019).

While SAM offers a promising optimizer-level solution, its fixed isotropic perturbation exhibits two core geometric limitations under complex distribution shifts: **(C1)** *Isotropic geometry mismatches anisotropic curvature.* A single global radius enforces uniform regularization across heterogeneous parameters, over-penalizing insensitive directions while under-regularizing sensitive ones. Spherical neighborhoods thus fail to resolve high-curvature directions, whereas curvature-aligned ellipsoids accurately reflect parameter sensitivity (Figures 1(a) and 1(b)). **(C2)** *Worst-case directions drift under non-stationarity.* Evolving data distributions continuously reshape the loss landscape, causing the optimal perturbation to deviate from the raw gradient (Figure 1(c)). This first-order truncation defect leads isotropic SAM to accumulate inner-maximization errors over time, degrading adaptation in non-stationary streams. These limitations motivate a learnable anisotropic perturbation geometry that dynamically adapts to evolving curvature while preserving strict first-order efficiency.

To effectively resolve these challenges, we propose **Chameleon**, a novel **C**ross-domain adaptation met**h**od that refr**a**mes sharpness-aware **m**inimization with l**e**arnab**l**e p**e**rturbati**on**. Instead of relying on rigid spherical radii or static heuristics based on parameter magnitudes Kwon et al. (2021); Haas et al. (2024), Chameleon autonomously learns an adaptive ellipsoidal neighborhood during optimization. By updating perturbation scales via a normalized sharpness surrogate, the geometry dynamically and accurately tracks genuine curvature variations with computational overhead comparable to standard SAM. Theoretically, we characterize Chameleon as an axis-aligned metric that adaptively reweights dominant spectral components in the gradient, establish containment-based PAC-Bayes generalization guarantees, and provide a matched-budget analysis isolating the curvature-dependent benefit of anisotropic scaling.

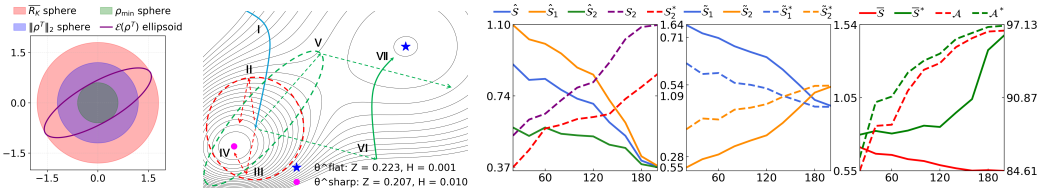
We summarize our main contributions as follows:

- We introduce Chameleon, which replaces SAM's fixed spherical neighborhood with a learnable anisotropic ellipsoid. Perturbation radii are tensorized into adaptive parameters optimized via a normalized surrogate objective. The method alternates ascent on perturbation scales with descent on model weights, enabling dynamic alignment between the adversarial neighborhood and local curvature while retaining first-order efficiency.
- We provide a geometric reinterpretation of SAM, framing its inner approximation defect as a first-order truncation issue rather than an inverse-Hessian effect. Building on this insight, we prove that Chameleon acts as a curvature-sensitive axis-aligned metric that coordinate-wise reweights dominant Hessian spectral components. Furthermore, we establish containment-based PAC-Bayes generalization bounds complemented by a rigorous matched-budget analysis that isolates the geometric advantage of anisotropic allocation.
- Extensive experiments across vision, language, and web-scale CTR tasks demonstrate that Chameleon consistently improves cross-domain robustness. On CIFAR, Chameleon improves over SAM by 1.2%–3.0%, and achieves higher AUC on Criteo and Avazu datasets. Furthermore, Chameleon incurs only $1.09\times$ FLOPs overhead, remaining more efficient than second-order variants such as CR-SAM ($1.61\times$) or Eigen-SAM ($1.68\times$).

2 Background and Motivation

2.1 Cross-Domain Adaptation and Generalization

The challenge of distribution shifts has driven extensive research into cross-domain adaptation and generalization. Prior work predominantly aligns source and target feature distributions via maximum mean discrepancy or adversarial learning Long et al. (2015; 2017), or transfers interaction patterns across specialized domains Man et al. (2017); Hu et al. (2018). To eliminate reliance on concurrent target data, domain generalization approaches such as invariant risk minimization and group distributionally robust optimization seek invariant representations across observed source domains Arjovsky et al. (2019); Sagawa et al. (2020); Gulrajani & Lopez-Paz (2020). Concurrently, test-time adaptation strategies update model parameters online using unlabeled target streams during inference Sun et al. (2020); Wang et al. (2021).



(a) Geo Relation. (b) Geometric Trajectory. (c) SAM and OPT in curvature estimation (Epoch vs Value).

Figure 1: Geometric interpretation of Chameleon (a, b) and curvature estimation in SAM (c).

Despite their empirical success, these paradigms remain constrained in dynamic environments. They typically require explicit domain identifiers, paired target batches for alignment, or computationally demanding auxiliary architectures. When target distributions shift rapidly or domain boundaries blur, these methods become brittle. This motivates optimizer-level solutions that guide parameters toward geometrically stable neighborhoods, bypassing the need for explicit domain supervision or architectural modifications.

2.2 Geometric Variants of SAM

Connecting the flatness of local minima to generalization is a foundational principle in deep learning theory. Sharpness-Aware Minimization (SAM) leverages this geometric prior via a minimax objective penalizing worst-case loss within an isotropic Euclidean ball Foret et al. (2021); Jiang et al. (2019). To alleviate the scale dependence and geometric mismatch of this spherical region, several variants have emerged. Adaptive SAM scales perturbation radii proportionally to parameter magnitudes Kwon et al. (2021), whereas Fisher-SAM employs a Riemannian metric approximated by the Fisher Information Matrix to reflect manifold structure Kim et al. (2022). Furthermore, Generalized SAM and Normalization-Aware SAM refine optimization dynamics and normalization layers to stabilize training Zhuang et al. (2022); Dai et al. (2023).

Recent theoretical studies confirm that flat minima inherently enhance domain-agnostic robustness using training data alone, highlighting the promise of SAM for cross-domain transfer Andriushchenko & Flammarion (2022); Oikonomou & Loizou (2025); Zhou et al. (2025). Building upon these insights, Chameleon addresses the primary computational and geometric limitations of existing variants. Rather than imposing static heuristics like parameter magnitudes or calculating expensive Fisher matrices, Chameleon parameterizes the perturbation neighborhood as a learnable anisotropic ellipsoid. This formulation continuously tracks evolving local curvature, providing an adaptive geometric prior for robust adaptation while strictly preserving first-order efficiency.

2.3 Curvature Defect of SAM

To understand the geometric limitations of perturbation-based optimizers, we analyze the discrepancy between SAM's gradient-driven perturbation and the theoretically optimal perturbation derived from Stochastic Differential Equations (SDEs) Compagnoni et al. (2023); Luo et al. (2024). While detailed derivations are provided in Appendix A, our analysis reveals that SAM can be interpreted as favoring gradient aligned directions, potentially underestimating curvature-aligned perturbations:

$$\Delta d\boldsymbol{\epsilon}_t = d\boldsymbol{\epsilon}_t^{opt} - d\boldsymbol{\epsilon}_t^{sam} = \rho\, \nabla_w^2 \mathcal{L}(w) \frac{\nabla_w \mathcal{L}(w)}{\|\nabla_w \mathcal{L}(w)\|}\, dt. \tag{1}$$

This discrepancy demonstrates that SAM exhibits a first-order truncation defect: its linearized inner perturbation does not explicitly incorporate the local curvature term $\nabla_w^2 \mathcal{L}(w)\boldsymbol{\epsilon}_t$. Consequently, standard SAM over-relies on the raw gradient direction while failing to capture the local second-order structure needed for finding truly flat minima.

Component Analysis. To empirically quantify this defect, we decompose the loss sharpness s into a first-order component s_1 and a second-order component s_2 via Taylor expansion and finite difference

approximations:

$$\begin{cases} s \triangleq \mathcal{L}(w+\varepsilon;\mathcal{B}_t) - \mathcal{L}(w;\mathcal{B}_t) & \approx \varepsilon^\top \nabla \mathcal{L}(w;\mathcal{B}_t) + \frac{1}{2}\varepsilon^\top \nabla^2 \mathcal{L}(w;\mathcal{B}_t)\varepsilon, \\ s_1 \triangleq \mathcal{L}(w+\varepsilon;\mathcal{B}_t) - \mathcal{L}(w-\varepsilon;\mathcal{B}_t) & \approx \varepsilon^\top \nabla \mathcal{L}(w;\mathcal{B}_t), \\ s_2 \triangleq \mathcal{L}(w+\varepsilon;\mathcal{B}_t) + \mathcal{L}(w-\varepsilon;\mathcal{B}_t) - 2\mathcal{L}(w;\mathcal{B}_t) & \approx \varepsilon^\top \nabla^2 \mathcal{L}(w;\mathcal{B}_t)\varepsilon. \end{cases} \tag{2}$$

Here s_1 measures the pure gradient aligned effect while s_2 captures the genuine loss curvature. Standard SAM primarily leverages s_1 whereas optimal formulations must incorporate both s_1 and s_2. To quantitatively evaluate loss sharpness and its decomposition, we define the following metrics by referencing the components in Eq. (2). Let s^*, s_1^*, s_2^* denote the corresponding quantities under the optimal perturbation ε^*, then

$$\begin{cases} \hat{\mathcal{S}} = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s}{s^*}, & \hat{\mathcal{S}}_1 = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s_1}{s_1^*}, & \hat{\mathcal{S}}_2 = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s_2}{s_2^*}, \\ \tilde{\mathcal{S}}_1 = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s_1}{2s}, & \tilde{\mathcal{S}}_2 = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s_2}{2s}, & \overline{\mathcal{S}} = \mathbb{E}_{\mathcal{B}_t\sim D}\dfrac{s}{\|\nabla \mathcal{L}(w;\mathcal{B}_t)\|}. \end{cases} \tag{3}$$

Eq. (3) decouples the perturbation response to diagnose the geometric limitations of isotropic SAM. Specifically, $\hat{\mathcal{S}}$, $\hat{\mathcal{S}}_1$, and $\hat{\mathcal{S}}_2$ measure approximation fidelity against the oracle perturbation ε^*, where a degraded $\hat{\mathcal{S}}_2$ isolates the failure to capture Hessian curvature. Meanwhile, $\tilde{\mathcal{S}}_1$ and $\tilde{\mathcal{S}}_2$ evaluate the internal balance between linear gradient climbing and quadratic curvature exploitation. Finally, the gradient-normalized sharpness $\overline{\mathcal{S}}$ factors out local gradient magnitudes to provide a scale-invariant flatness measure.

As shown in Figure 1(c), SAM progressively diverges from the optimal perturbation throughout training. The approximation ratios $\hat{\mathcal{S}}, \hat{\mathcal{S}}_1, \hat{\mathcal{S}}_2$ steadily decline, accompanied by a structural imbalance dominated by $\tilde{\mathcal{S}}_1$ with negligible $\tilde{\mathcal{S}}_2$. Consequently, the normalized sharpness $\overline{\mathcal{S}}$ decays under SAM while rising under the oracle. This confirms that SAM's inner ascent over-amplifies the linear component s_1 and neglects local quadratic curvature s_2, restricting its generalization capability $\mathcal{A}$ relative to the curvature-aware optimum $\mathcal{A}^*$.

3 SAM WITH LEARNABLE PERTURBATION

We extend SAM by introducing learnable perturbation scales jointly optimized with model parameters. This section formalizes the anisotropic parameterization of the perturbation geometry and derives the curvature-driven objective for scale optimization.

3.1 CHAMELEON FORMULATION

To address the inability of isotropic SAM to capture higher-order curvature information, we propose a scheme that tensorizes the perturbation scale and updates it to minimize curvature observation bias. As discussed in Section 2.3, simply maximizing the perturbed loss can lead to trivial solutions where perturbations grow unbounded. To prevent this and align with PAC-Bayes flatness measures, we define the normalized perturbation objective as the optimization target for updating the perturbation scale tensors:

$$\mathcal{J} = \frac{\mathcal{L}\left(\tilde{w}^t;\mathcal{B}_t\right)}{\|\nabla \mathcal{L}\left(\tilde{w}^t;\mathcal{B}_t;\rho^t\right)\|}, \tag{4}$$

where $\tilde{w}^t = w^t + \varepsilon^t$ and $\varepsilon^t = \frac{\rho^t \nabla \mathcal{L}\left(w^t;\mathcal{B}_t\right)}{\|\nabla \mathcal{L}(w^t;\mathcal{B}_t)\|}$ is the perturbation of current step t in batch $\mathcal{B}_t$, and $\|\nabla \mathcal{L}\left(w^t;\mathcal{B}_t;\rho^t\right)\|$ is the gradient norm associated with the perturbation scale ρ^t.

Remark: Without normalization, the scale update degenerates to maximizing the raw loss magnitude, leading to unbounded perturbations. The normalization term enforces scale-invariant curvature probing, ensuring the objective acts as a valid sharpness surrogate $\mathcal{S} \approx \mathcal{L}_p/\|\nabla \mathcal{L}\|$ consistent with PAC-Bayes generalization bounds, while avoiding the high computational cost of explicit Hessian trace or Fisher information estimation.

Building on the normalized curvature objective, we now formally define Chameleon as a unified bilevel optimization framework. Unlike standard SAM which imposes a fixed geometric constraint, Chameleon formulates a unified bilevel optimization problem that simultaneously minimizes the empirical risk with respect to model parameters w and optimizes the perturbation scales ρ to track the local geometry. This joint formulation ensures that the adversarial neighborhood dynamically evolves into an optimal anisotropic ellipsoid driven by the normalized curvature surrogate, as defined below:

Definition 3.1 (Chameleon Framework)**.** We define Chameleon as a unified bilevel optimization problem that jointly optimizes the model parameters $w \in \mathbb{R}^d$ and parametric perturbation scale $\rho \in \mathbb{R}^d_{\geq 0}$ by minimizing a unified objective that combines perturbed loss, weight decay, and normalized perturbation loss. The adversarial perturbation and normalized perturbation loss are defined through inner maximization problems:

$$\min_{w,\rho} \ \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}\Big[\, \mathcal{G}(w,\rho;\mathcal{B}) + \tfrac{\lambda}{2}\|w\|_2^2 + \gamma\, \mathcal{C}(w,\rho;\mathcal{B}) \,\Big], \tag{5}$$

$$\text{s.t.} \quad \mathcal{G}(w,\rho;\mathcal{B}) = \max_{\varepsilon\in\mathcal{E}(\rho)} \ \mathcal{L}(w+\varepsilon;\mathcal{B}), \tag{5a}$$

$$\mathcal{C}(w,\rho;\mathcal{B}) = \max_{\varepsilon\in\mathcal{E}(\rho)} \ \mathcal{J}, \tag{5b}$$

$$\varepsilon = \rho \odot \frac{\nabla\mathcal{L}(w;\mathcal{B})}{\|\nabla\mathcal{L}(w;\mathcal{B})\|_2}, \tag{5c}$$

$$\rho \succ 0, \ \ \lambda > 0, \ \ \gamma > 0. \tag{5d}$$

Since directly solving the nested max-min objective in Definition 3.1 is computationally prohibitive, Chameleon adopts a lightweight alternating first-order approximation in practice. By sequentially updating the perturbation scale and model parameters using stochastic gradients, this relaxation translates the intractable bilevel problem into an efficient training procedure while preserving its curvature-aware properties. The respective gradients are derived as follows:

$$\nabla_w \mathcal{L} = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}\left[g_1 + \lambda w + \gamma \frac{g_1}{\|g_0\|_2} \right], \tag{6a}$$

$$\nabla_\rho \mathcal{C} = \gamma \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}\left[\left(\frac{g_1}{\|g_1\|_2} - \frac{(\mathcal{L}_1 - \mathcal{L}_0)(g_1 - g_0)}{\|g_1\|_2^2} \right) \odot \frac{g_0}{\|g_0\|_2} \right], \tag{6b}$$

where $\mathcal{L}_0 = \mathcal{L}(w;\mathcal{B})$, $\mathcal{L}_1 = \mathcal{L}(w+\varepsilon;\mathcal{B})$, $g_0 = \nabla\mathcal{L}(w;\mathcal{B})$, $g_1 = \nabla\mathcal{L}(w+\varepsilon;\mathcal{B})$, and $\odot$ denotes the Hadamard product. Detailed derivation is presented in Appendix C.5.

3.2 Convergence analysis

Based on assumptions 1–3, we establish the convergence results for Chameleon, showing that it converges sublinearly under adaptive radius updates, which highlights the importance of adaptive perturbation scales in achieving both exploration and stability.

Proposition 3.2 (Chameleon with Curvature-aware Spectral Form)**.** *For iteration t and sufficiently small radius $\rho > 0$, the maximized smooth surrogate satisfies:*

$$\mathcal{C}(w,\rho;\mathcal{B}) \approx \tfrac{1}{2}\,\rho^2\,\lambda_{\max}\big(\mathcal{A}(w^t)\big) \ + \ \mathcal{O}(\rho^3), \tag{7}$$

where the maximizing perturbation direction in the scale-transformed space aligns with the top eigenvector $v_1(\mathcal{A}(w^t))$ of the transformed curvature matrix $\mathcal{A}(w^t)$.

Eq. (7) reveals that $\mathcal{A}(w^t)$ regularizes the spectral norm in the scale-transformed coordinate system, ensuring stable perturbations near stationary points. We emphasize that because the learned metric $D_\rho = \mathrm{diag}(\rho_1,\ldots,\rho_d)$ is diagonal, the perturbation set forms an axis-aligned ellipsoid $\mathcal{E}(\rho) = \{D_\rho u : \|u\|_2 \leq 1\}$ and does not perform an arbitrary spatial rotation to align with the eigenbasis of the raw Hessian $H(w^t)$.

Theorem 3.3 (Convergence of Chameleon)**.** *Suppose $T \geq 1$, the step size satisfies $\eta_t \equiv \eta = \frac{\eta_0}{\sqrt{T}} \leq \frac{1}{4L}$, $\alpha_\rho > 0$, $\rho^0 = \rho^{init}$, $\rho_t \in [\rho_{\min}, \rho_{\max}]^D$ via coordinate-wise clipping. Then the empirical average*

squared perturbation satisfies:

$$\overline{R}_T^2 := \frac{1}{T}\sum_{t=0}^{T-1} \min\left\{\rho_{\max}^2, \left(\rho^{init} + t\,\alpha_\rho G_\rho\right)^2\right\}, \tag{8}$$

and the convergence bounds hold:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(w^t)\|^2\right] \le \frac{4(f(w^0)-f_{\inf})}{\eta_0\sqrt{T}} + \mathcal{P}, \tag{9a}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(w^t+\varepsilon_t)\|^2\right] \le \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(w^t)\|^2\right] + \mathcal{Q}, \tag{9b}$$

where $\mathcal{P} = 4L^2\,\overline{R}_T^2 + \frac{4L\eta_0\sigma^2}{\sqrt{T}}$ *and* $\mathcal{Q} = 2L^2\,\overline{R}_T^2$.

Remark: Theorem 3.3 indicates that Chameleon converges at the same asymptotic rate as SGD, with an additional bias term controlled by the learned perturbation scale. Specifically, if the clipping is chosen with $\rho_{max} = O(1/\sqrt{T})$ and the ascent step $\alpha_\rho = O(1/\sqrt{T})$, the average gradient norm decays at a rate of $O(1/\sqrt{T})$. Conversely, if ρ is fixed as ρ_{max}, the algorithm converges to a stationary neighborhood of size $O(L^2\rho_{max}^2)$. Fundamentally, Chameleon achieves an adaptive bias–variance tradeoff, where the learnable scale dynamically arbitrates between the perturbation bias required for flatness exploration and the gradient stability needed for precise convergence.

3.3 Generalization analysis

We derive the PAC-Bayes generalization bound, which highlights the influence of perturbation geometry and the adaptive radius on performance, providing theoretical support for Chameleon and ensuring robust and stable performance on unseen data.

Theorem 3.4 (PAC-Bayes Generalization Bound for Chameleon)**.** *For any prior* $\mathcal{P}$ *and posterior* $\mathcal{Q}$*, with probability at least* $1-\alpha$ *over the training sample of size* m *(see Appendix G for details), the generalization error of Chameleon satisfies:*

$$\mathbb{E}_x f(w^T, x) \le \widehat{f}(w^T) + \mathcal{M}_{cham} + 4\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \log(2m/\alpha)}{m}}, \tag{10}$$

where $\mathcal{M}_{cham}$ *denotes the anisotropic sharpness surrogate. Furthermore, under scale constraints* $\rho_{\min} \le \rho^T \le \overline{R}_T \le \rho_{\max}$*, the empirical sharpness surrogates satisfy the containment-based sandwich relation:*

$$\underbrace{\widehat{h}(w;\rho_{\min})}_{\textit{Min Scale}} \le \underbrace{\widehat{h}(w,\rho^T)}_{\textbf{\textit{Chameleon (Ellipsoid)}}} \le \underbrace{\widehat{h}(w;\|\rho^T\|_\infty)}_{\textit{Equivalent Bounding Sphere}} \le \underbrace{\widehat{h}(w;\overline{R}_T)}_{\textit{SAM (Isotropic)}} \le \underbrace{\widehat{h}(w;\rho_{\max})}_{\textit{Max Scale}}. \tag{11}$$

Remark: Theorem 3.4 formalizes the geometric intuition depicted in Figures 1a, where the sandwich relation (Eq. (11)) directly corresponds to the nested regions in the visualization. While the surrogate hierarchy algebraically reflects perturbation set containment $\mathcal{E}(\rho^T) \subseteq \mathbb{B}_2(\|\rho^T\|_\infty)$, the non-trivial geometric advantage of anisotropy is established under a matched-budget condition (detailed in the proof below): allocating scales inversely with directional curvature guarantees a strictly smaller worst-case loss expansion than any equal-volume isotropic ball. Meanwhile, the hyperparameters $\rho_{\min}$ and $\rho_{\max}$ serve as coordinate-wise safeguards balancing exploration against optimization instability.

4 Experiment

4.1 Experimental Setup

Benchmarks. To verify the effectiveness and robustness of Chameleon, we conduct extensive experiments across three modalities: Computer Vision, Natural Language Processing, and Click-Through Rate prediction. Our evaluation encompasses: (1) **In-Domain Image Classification** on CIFAR-10 and CIFAR-100 Krizhevsky (2009) using WRN-16-2 Zagoruyko & Komodakis (2016) and ResNet-18 He et al. (2016); (2) **Language Understanding** on the GLUE benchmark using

RoBERTa-base/large Liu et al. (2019); (3) **Cross-Domain Transfer Learning** on PACS Li et al. (2017) and DomainNet Peng et al. (2019) using DeIT-small Touvron et al. (2021) and ResNet-18; (4) **Click-Through Rate Prediction** on Avazu, Criteo, and Taobao datasets using DNN, DCN, and DeepFM. Detailed descriptions are provided in Appendix C.1.

Baselines and Implementation. We compare Chameleon against standard SGD/AdamW and state-of-the-art sharpness-aware minimizers, including SAM Foret et al. (2021), ASAM Kwon et al. (2021), FSAM Li et al. (2024), and VaSSO Li & Giannakis (2023). All experiments are implemented in PyTorch 2.6 with CUDA 12.6 on NVIDIA RTX-4090 GPUs. To ensure reproducibility, we report the mean and standard deviation over five independent runs with random seeds. Specific hyperparameter settings (learning rates, batch sizes, optimizers) and the grid search space for Chameleon's components ($\rho_{min}, \rho_{max}, \alpha_\rho$) are detailed in Appendix C.2.

Evaluation Metrics. We employ task-specific metrics: Top-1 Accuracy for IC and TL; AUC and LogLoss for CTR; Accuracy/F1/MCC/Correlation scores for LU tasks. Furthermore, to quantitatively analyze the loss landscape geometry, we report Hessian-based flatness metrics, including Hessian Trace $\mathrm{Tr}(H)$ and Top Eigenvalue λ_{max}. Mathematical definitions are available in Appendix C.4.

4.2 Performance Analysis

In-Domain Image Classification. As shown in Table 1, on CIFAR10, Chameleon outperforms SAM, achieving accuracy gains of +0.26 (WRN16-2/Cutout), +0.33 (ResNet18/Cutout), and +0.25 (ResNet18/CutMix), On CIFAR-100, the improvements are more pronounced: +0.32 (WRN16-2/Cutout), +0.65 (ResNet18/Cutout). and +0.83 (ResNet18/CutMix). Chameleon surpasses ERM (up to +2.09 on CIFAR-100/ResNet18/CutMix) and SAM (up to +0.83 in the same setting) in all cases. These improvements suggest that anisotropic perturbations better adapt to local curvature and promote flatter minima, with the effect becoming even more significant as task complexity increases. When Chameleon is combined with SAM variants, further improvements can be observed. For Fleon (FSAM + Chameleon), performance increases by +0.15 on CIFAR10/ResNet18/Cutout, and +0.51 on CIFAR100/ResNet18/CutMix. Similarly, Vleon (VASSO + Chameleon) shows improvements of +0.14 and +0.45 in the same settings.

Table 1: Test accuracy (%) on CIFAR-10/100 under different models and augmentation settings.

Dataset	Model	Aug	ERM	SAM	ASAM	Chameleon	FSAM	Fleon	VASSO	Vleon
CIFAR-10	WRN-16-2	Cutout	95.56 ± 0.06	95.87 ± 0.11	95.90 ± 0.09	$\mathbf{96.13 \pm 0.10}$	95.98 ± 0.08	$\mathbf{96.04 \pm 0.12}$	95.99 ± 0.02	96.03 ± 0.06
	ResNet-18	Cutout	96.29 ± 0.02	96.66 ± 0.03	96.70 ± 0.06	$\mathbf{96.99 \pm 0.07}$	96.75 ± 0.05	$\mathbf{96.90 \pm 0.09}$	96.72 ± 0.13	96.86 ± 0.02
		Cutmix	96.52 ± 0.11	96.86 ± 0.04	96.88 ± 0.07	$\mathbf{97.11 \pm 0.02}$	96.99 ± 0.04	$\mathbf{97.15 \pm 1.10}$	96.95 ± 0.13	97.04 ± 0.08
CIFAR-100	WRN-16-2	Cutout	76.38 ± 0.24	77.80 ± 0.16	77.83 ± 0.12	$\mathbf{78.12 \pm 0.23}$	77.93 ± 0.27	$\mathbf{78.18 \pm 0.19}$	77.79 ± 0.09	77.95 ± 0.14
	ResNet-18	Cutout	80.08 ± 0.09	81.21 ± 0.24	81.25 ± 0.19	$\mathbf{81.86 \pm 0.19}$	81.30 ± 0.21	$\mathbf{81.59 \pm 0.04}$	81.37 ± 0.05	81.48 ± 0.12
		Cutmix	80.89 ± 0.12	82.15 ± 0.04	82.17 ± 0.06	$\mathbf{82.98 \pm 0.07}$	82.27 ± 0.11	82.78 ± 0.11	82.36 ± 0.12	$\mathbf{82.81 \pm 0.07}$

Language Understanding. As shown in Table 2, Chameleon consistently improves language understanding performance across all GLUE tasks, with particularly strong gains on smaller-scale models. On CoLA, Chameleon increases *Mcc.* by up to +3.0 compared with ERM and +2.1 over SAM, while on MRPC it yields higher *Acc.* and *F1.*, with improvements of +1.7 and +0.9, respectively. On STS-B, Chameleon boosts *Pear.* by +2.7 over ERM and +0.7 over SAM when using RoBERTa-large. The largest improvements are observed on RTE, where Chameleon improves *Acc.* and *F1.* by +4.43, +4.19 points over ERM and +1.81, +2.15 over SAM with RoBERTa-base, and by +2.09, +2.86 points over ERM and +0.73, +1.21 over SAM with RoBERTa-large. While some metrics on RoBERTa-large show only marginal advantages, the overall trend indicates that Chameleon is effective across different model sizes especially smaller ones with more limited dimensions, where learnable perturbation scales better adapt to anisotropic loss landscapes and help mitigate overfitting on training data.

Table 2: RoBERTa results on language understanding tasks.

Method	Model	CoLA	MRPC		STS-B		RTE	
		Mcc.	*Acc.*	*F1.*	*Pear.*	*Spea.*	*Acc.*	*F1.*
ERM	base	61.34	87.01	90.59	87.55	87.56	77.16	75.82
	large	67.19	90.44	93.10	88.03	88.16	86.28	84.93
SAM	base	62.06	87.75	91.20	88.16	88.07	79.78	76.86
	large	68.02	91.18	93.55	91.04	90.82	87.64	86.58
Chameleon	base	**64.03**	**88.73**	**91.79**	**90.29**	**90.14**	**81.59**	**79.01**
	large	**70.16**	**90.93**	**93.59**	**91.78**	**91.57**	**88.37**	**87.79**

Table 3: CTR predictions on FuxiCTR.

Model	Method	Avazu		Criteo		Taobao	
		LogLoss ↓	AUC ↑	LogLoss ↓	AUC ↑	LogLoss ↓	AUC ↑
DNN	ERM	37.501	78.806	44.399	80.740	19.454	63.466
	SAM	37.487	78.781	**44.364**	80.806	19.372	63.479
	Chameleon	**37.462**	**78.819**	44.373	**80.817**	**19.293**	**63.491**
DCN	ERM	37.629	78.621	44.417	80.745	19.431	63.563
	SAM	37.631	78.605	44.402	80.753	19.407	63.582
	Chameleon	**37.602**	**78.634**	**44.396**	**80.769**	**19.368**	**64.276**
DeepFM	ERM	37.628	78.681	44.418	80.735	19.412	63.369
	SAM	37.639	78.673	44.367	80.769	19.331	63.662
	Chameleon	**37.614**	**78.692**	**44.329**	**80.803**	**19.297**	**63.841**

Click-Through Rate. We conduct CTR prediction experiments using FuxiCTR Zhu et al. (2021). As shown in Table 3, show that Chameleon achieves consistent improvements across most settings. On Avazu, Chameleon reduces LogLoss for all three models, with DeepFM improving by 0.025 over SAM, and slightly improves AUC by 0.019, indicating enhanced robustness under sparse features. On Criteo, Chameleon brings small but consistent gains in AUC, with DeepFM reaching the highest overall performance, improving by 0.099 LogLoss and 0.068 AUC over ERM. On Taobao, Chameleon shows noticeable gains in both LogLoss and AUC, with DeepFM outperforming ERM and SAM by 0.121 in LogLoss and with DCN outperforming ERM by 0.713 in AUC, particularly in DeepFM where the improvement is most significant.

Cross-Domain Transfer Learning. As shown in Table 4, on PACS, Chameleon shows consistent improvements over SAM, with gains of +0.66 (Painting), +2.18 (Cartoon), +0.66 (Photo), and +1.46 (Sketch), making it the top performer across all domains. Chameleon also surpasses ERM in all domains, with the largest improvement on Sketch (+3.91). On DomainNet, Chameleon improves over SAM by +0.74 (Clipart), +0.62 (Infograph), +0.58 (Painting), +0.54 (Quickdraw), +0.64 (Real), and +1.18 (Sketch). In particular, Chameleon's improvement over SAM is most notable in the Sketch domain, where it outperforms SAM by +1.46. The largest gain over ERM is also in Sketch (+2.73), showcasing Chameleon's strength when handling domains with greater adversarial disparity between source and target domains. This indicates that Chameleon's adaptive perturbation geometry is especially beneficial in domains with larger domain shifts.

Table 4: Test Accuracy (%) on PACS and DomainNet.

Method	PACS				DomainNet					
	Painting	Cartoon	Photo	Sketch	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
ERM	86.13 ± 0.88	68.73 ± 1.72	97.82 ± 0.45	61.18 ± 0.62	79.05 ± 0.98	44.02 ± 0.35	80.42 ± 0.81	40.67 ± 0.21	80.21 ± 0.46	70.54 ± 0.04
SAM	87.61 ± 0.35	72.58 ± 1.09	98.26 ± 0.33	63.63 ± 1.65	80.18 ± 1.17	45.01 ± 0.61	81.49 ± 0.50	41.57 ± 0.59	81.50 ± 0.39	72.09 ± 0.18
Chameleon	$\mathbf{88.27 \pm 0.29}$	$\mathbf{74.76 \pm 1.32}$	$\mathbf{98.92 \pm 0.37}$	$\mathbf{65.09 \pm 1.41}$	$\mathbf{80.92 \pm 0.94}$	$\mathbf{45.63 \pm 0.49}$	$\mathbf{82.07 \pm 0.40}$	$\mathbf{42.11 \pm 0.47}$	$\mathbf{82.14 \pm 0.31}$	$\mathbf{73.27 \pm 0.14}$

Table 5: Test accuracy (%) varying batch size on WRN16-2.

Method	CIFAR-10					CIFAR-100				
	128	256	512	1024	2048	128	256	512	1024	2048
ERM	95.52 ± 0.41	95.62 ± 0.09	95.02 ± 0.47	93.79 ± 0.36	91.37 ± 0.05	76.19 ± 0.44	75.65 ± 0.18	73.63 ± 0.50	71.24 ± 0.27	66.87 ± 0.33
SAM	95.84 ± 0.28	95.92 ± 0.17	95.24 ± 0.12	94.45 ± 0.08	91.91 ± 0.29	77.70 ± 0.11	76.72 ± 0.25	75.55 ± 0.30	72.95 ± 0.16	68.46 ± 0.20
Chameleon	$\mathbf{96.22 \pm 0.14}$	$\mathbf{96.38 \pm 0.02}$	$\mathbf{95.53 \pm 0.18}$	$\mathbf{94.71 \pm 0.19}$	$\mathbf{92.34 \pm 0.12}$	$\mathbf{78.27 \pm 0.08}$	$\mathbf{77.14 \pm 0.13}$	$\mathbf{75.92 \pm 0.07}$	$\mathbf{73.43 \pm 0.01}$	$\mathbf{68.89 \pm 0.19}$

4.3 Robustness Evaluation

As shown in Table 5, Chameleon consistently outperforms SAM and ERM across various batch sizes. On CIFAR-10, it achieves an improvement of up to +0.43% over SAM at a batch size of 2048. On CIFAR-100, the gains over SAM remain consistent, ranging from +0.57% at a batch size of 128 to +0.43% at 2048. Following the symmetric label-noise setting of Frénay and Verleysen (2014), we further evaluate Chameleon's robustness under varying noise ratios. As shown in Table 6, Chameleon outperforms SAM by +1.86% at 40% noise on CIFAR-10 and by +2.04% at 60% noise on CIFAR-100. These results demonstrate Chameleon's robustness under challenging regimes: large batches reduce gradient variance, whereas label noise misdirects gradients, both of which obscure curvature and hinder reliable sharpness estimation. Chameleon counters these effects through per-batch anisotropic-scale updates under a normalized, curvature-aware objective, expanding along

stiff axes and shrinking along gentle directions. This fine-grained exploration stabilizes perturbations and steers training toward flatter minima, as evidenced by the consistent performance gains.

Table 6: Test accuracy (%) varying symmetric label noise ratios on WRN16-2.

Method	CIFAR-10					CIFAR-100				
	0%	20%	40%	60%	80%	0%	20%	40%	50%	60%
ERM	95.52 ± 0.18	90.67 ± 0.48	85.37 ± 0.39	77.59 ± 0.24	67.58 ± 0.05	76.19 ± 0.37	68.17 ± 0.43	59.02 ± 0.50	54.31 ± 0.23	47.82 ± 0.49
SAM	95.84 ± 0.06	93.93 ± 0.13	91.19 ± 0.07	86.82 ± 0.28	73.11 ± 0.29	77.70 ± 0.03	72.93 ± 0.18	68.51 ± 0.25	64.60 ± 0.09	59.65 ± 0.20
Chameleon	$\mathbf{96.22 \pm 0.11}$	$\mathbf{94.52 \pm 0.12}$	$\mathbf{93.05 \pm 0.16}$	$\mathbf{89.07 \pm 0.09}$	$\mathbf{76.55 \pm 0.08}$	$\mathbf{78.27 \pm 0.13}$	$\mathbf{73.42 \pm 0.11}$	$\mathbf{69.26 \pm 0.15}$	$\mathbf{65.83 \pm 0.17}$	$\mathbf{61.69 \pm 0.14}$

4.4 Flatness Evaluation

As shown in Table 7, Chameleon consistently achieves smaller eigenvalues and traces compared to ERM and SAM, across both Cutout and CutMix augmentations. On Cutout, Chameleon reduces λ_1 by 59.2% compared to ERM and 29.8% compared to SAM, with a similar reduction in λ_5 by 52.9% and 28.4%, respectively. The trace of the Hessian is also significantly reduced by 59.1% compared to ERM and 21.4% compared to SAM. On CutMix, Chameleon outperforms ERM and SAM, reducing λ_1 by 71.8% compared to ERM and 31.6% compared to SAM, with a similar reduction in λ_5 by 72.2% and 35.1%, respectively. The trace on CutMix shows a reduction of 78.9% compared to ERM and 34.2% compared to SAM.

Table 7: Flatness evaluation measured by the largest and fifth-largest Hessian eigenvalues (denoted by λ_1 and λ_5) and trace $\mathrm{Tr}(\mathbf{H})$.

Metric	ERM	SAM	Chameleon	Aug
λ_1	326.17 ± 0.09	189.00 ± 0.23	$\mathbf{132.98 \pm 0.17}$	
λ_5	181.75 ± 0.38	119.65 ± 0.42	$\mathbf{85.65 \pm 0.12}$	**Cutout**
$\mathrm{Tr}(\mathbf{H})$	18334.83 ± 863.29	9537.76 ± 47.03	$\mathbf{7496.11 \pm 90.48}$	
λ_1	495.19 ± 0.26	204.36 ± 0.85	$\mathbf{139.93 \pm 0.30}$	
λ_5	306.63 ± 0.81	136.65 ± 2.09	$\mathbf{88.83 \pm 0.40}$	**CutMix**
$\mathrm{Tr}(\mathbf{H})$	30181.99 ± 260.54	9650.76 ± 198.00	$\mathbf{6350.98 \pm 136.30}$	

Figures 2 and 3 collectively demonstrate Chameleon's superior capability in discovering geometrically stable minima. While ERM converges to sharp valleys and SAM finds smoother basins, Chameleon achieves the flattest and most symmetric loss landscape. This visual hierarchy is rigorously corroborated by the Hessian spectrum density, where Chameleon exhibits a tighter eigenvalue concentration near zero with shorter positive tails.

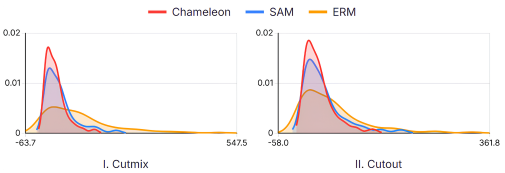
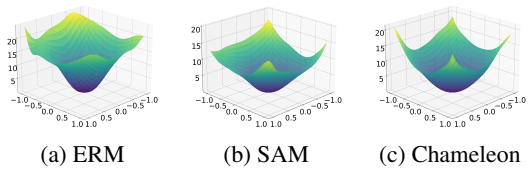


Figure 2: Spectrum Density of Hessian.



(a) ERM (b) SAM (c) Chameleon

Figure 3: Visualization of Loss landscape.

5 Conclusion

Chameleon reframes SAM at the geometric level by introducing learnable anisotropic perturbation geometry, rather than serving as a conventional SAM-based extension. By reshaping the adversarial neighborhood into a curvature-aware ellipsoid, Chameleon enables efficient tracking of flatter minima and corrects the inherent first-order bias of isotropic SAM, while preserving its first-order efficiency and avoiding costly higher-order computations. The proposed framework is broadly applicable and demonstrates strong empirical performance under challenging settings such as large-batch training, label noise, multimodal learning, and domain shifts, offering scalable and plug-and-play generalization across tasks. Moreover, Chameleon is supported by theoretical guarantees, achieving sublinear convergence and tighter PAC-Bayes generalization bounds compared to isotropic SAM.

IMPACT STATEMENT

This work introduces Chameleon to resolve the fundamental geometric mismatch between static models and dynamic real-world data. For the **Federated Learning** community, it offers a scalable solution to the critical client drift problem, correcting local geometric biases caused by severe data heterogeneity without incurring heavy computational costs. Beyond FL, its lightweight, curvature-aware mechanism opens new directions for efficient fine-tuning of **Foundation Models** on edge devices and enhances the reliability of safety-critical systems (e.g., autonomous driving). Ultimately, Chameleon paves the way for practical geometry-adaptive intelligence, enabling AI systems to robustly adapt and thrive in open-ended environments.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our theoretical and empirical findings, complete mathematical derivations and proofs for all propositions and theorems are documented in Appendices A, E, F, and G. The complete algorithmic procedure is detailed in Algorithm 1. All experiments are implemented in PyTorch 2.6 with CUDA 12.6 on NVIDIA RTX-4090 GPUs, and results are reported across five independent runs with random seeds. Full dataset descriptions, model architectures, training configurations, and hyperparameter search grids are provided in Appendix C. The complete source code, configuration files, and evaluation scripts will be released publicly upon publication.

AI USE STATEMENT

LLM tools were utilized exclusively for language polishing, grammar correction, and typographical proofreading of the text. All conceptual formulations, algorithmic frameworks, mathematical derivations, experimental codes, and empirical analyses were independently conceived and executed by the authors.

ETHICS STATEMENT

This study uses benchmark datasets and simulated incentives. It does not collect new data from human participants or involve real payments. Any practical deployment should protect client data,ensure informed participation, and follow applicable regulations.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 639–668, Baltimore, MD, USA, 2022. PMLR.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Avazu Inc. Avazu prediction dataset. Kaggle, 2014.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Enea Monzio Compagnoni, Lorenzo Biggio, Antonio Orvieto, Franziska N. Proske, Henrik Kersting, and Aurelien Lucchi. An SDE for modeling SAM: Theory and insights. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24933–24953, Honolulu, HI, USA, 2023. PMLR.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*, pp. 191–198, Boston, MA, USA, 2016. Association for Computing Machinery.

Criteo Labs. Criteo labs display advertising challenge dataset. Kaggle, 2014.

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

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 4th edition, 2013.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. DeepFM: A factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1725–1731, Melbourne, Australia, 2017. ijcai.org.

Moritz Haas, Jin Xu, Volkan Cevher, and Leena Chennuru Vankadara. Effective sharpness aware minimization requires layerwise perturbation scaling. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1225–1234, New York, NY, USA, 2016. JMLR.org.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016. IEEE.

Guangneng Hu, Yu Zhang, and Qiang Yang. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 667–676, 2018.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 11780–11791, Vancouver, BC, Canada, 2019.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, Department of Computer Science, University of Toronto, 2009.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5905–5914, Virtual (originally planned for Vienna, Austria), July 18–24 2021.

Bingcong Li and Georgios B. Giannakis. Enhancing sharpness-aware optimization through variance suppression. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Welling, and C. Williams (eds.), *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 70861–70879, New Orleans, LA, USA, 2023.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5542–5550, Venice, Italy, 2017. IEEE.

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

Xuanlin Li, Xiangyu Cai, Hao Peng, Jia Duan, Wei Wang, Zehua Zhang, Changping Peng, Zhangang Lin, Ching Law, and Jingping Shao. An unbiased entire-space causal framework for click-through rate estimation in pre-ranking. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 306–314, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 97–105, Lille, France, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2208–2217, Sydney, Australia, 2017. PMLR.

Haocheng Luo, Tuan Truong, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. Explicit eigenvalue regularization improves sharpness-aware minimization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, Vancouver, BC, Canada, 2024.

Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. Cross-domain recommendation: An embedding and mapping approach. In *Ijcai*, volume 17, pp. 2464–2470, 2017.

Maximilian Mueller, Tiffany Vlaar, David Rolnick, and Matthias Hein. Normalization layers are all that sharpness-aware minimization needs. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Welling, and C. Williams (eds.), *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 56799–56812, New Orleans, LA, USA, 2023.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer Science & Business Media, 2004.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Singapore, Singapore, 2025.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, Seoul, South Korea, 2019. IEEE.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Baochen et al. Sun. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450, 2016.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248, Virtual Event, 2020. PMLR.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, Virtual, 2021. PMLR.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual (originally planned for Vienna, Austria), 2021.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of computer science*, 18(6):186345, 2024.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17 Workshop*, pp. 1–7, Halifax, NS, Canada, 2017. Association for Computing Machinery.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354, Boulder, CO, USA, 2021. PMLR.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? In *International Conference on Learning Representations (ICLR)*, 2023.

Tao Wu, Tony Tie Luo, and Donald C. Wunsch. CR-SAM: Curvature regularized sharpness-aware minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6144–6152, Vancouver, BC, Canada, 2024. AAAI Press. doi: 10.1609/aaai.v38i6.28431.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

Lijia Yang, Anant Raghunathan, and Andrej Risteski. A unified PAC-bayesian framework for explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 10757–10767, Virtual, 2020.

Sherry Yang, Ofir Canty, et al. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032, Seoul, South Korea, 2019. IEEE.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the 27th British Machine Vision Conference (BMVC)*, pp. 87.1–87.12, York, UK, 2016. BMVA Press.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Wenshuai Zhao, Jorge Pena Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 737–744, 2020.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Spotlight Paper.

Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. Open benchmarking for click-through rate prediction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 3107–3116, Virtual (hosted from Gold Coast, Queensland, Australia), 2021. ACM.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Virtual, 2022.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A Derivation of Curvature Defect via SDE

In this section, we provide the detailed derivation of the perturbation bias discussed in Section 2. We analyze the sharpness approximation by comparing the optimal perturbation derived from Stochastic Differential Equations (SDEs) using a second-order Taylor expansion, following previous works Compagnoni et al. (2023); Luo et al. (2024).

Optimal SDE Formulation. The worst-case perturbation within a neighborhood of radius ρ can be formulated by the optimal SDE, which incorporates the Hessian information to explicitly track the curvature:

$$d\boldsymbol{\epsilon}_t^{opt} = \left(\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) + \nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w})\,\boldsymbol{\epsilon}_t\right) dt + \sigma(\mathbf{w})\, d\mathbf{W}_t, \tag{12}$$

where $\sigma(\mathbf{w})d\mathbf{W}_t$ represents a stochastic diffusion term.

SAM SDE Formulation. In contrast, SAM employs a simplified SDE that depends solely on the first-order gradient term:

$$d\boldsymbol{\epsilon}_t^{sam} = \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})\, dt + \sigma(\mathbf{w})\, d\mathbf{W}_t. \tag{13}$$

This formulation neglects the second-order term $\nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w})\,\boldsymbol{\epsilon}_t$, which carries the second-order curvature information.

Bias Derivation. By comparing the drift terms of the two SDE formulations, the finite-difference approximation of sharpness reveals a perturbation bias. Subtracting the effective drift of SAM from the optimal drift leads to:

$$\Delta d\boldsymbol{\epsilon}_t = d\boldsymbol{\epsilon}_t^{opt} - d\boldsymbol{\epsilon}_t^{sam} = \nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w})\,\boldsymbol{\epsilon}_t\, dt = \rho\, \nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w}) \frac{\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})}{\|\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})\|}\, dt. \tag{14}$$

This bias term $\Delta d\boldsymbol{\epsilon}_t$ mathematically quantifies the deviation of SAM from the optimal curvature-aware drift as a Hessian-vector correction rather than an inverse-Hessian direction (an inverse-type term arises only from the stationary condition $(\mu I - H)\boldsymbol{\epsilon}^\star = g$ of the quadratic trust-region subproblem). Thus, SAM's inner perturbation omits the local curvature contribution $\nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w})\,\boldsymbol{\epsilon}_t$, constituting the first-order truncation defect of isotropic SAM.

B Preliminary Experiment Details

Simulation Experiment: In figure 1(b), we conduct a simulation on a univariate Gaussian distribution $p(x;\theta) = \mathcal{N}(x;\mu,\sigma^2)$, under a negative log-mixture of KL-driven energy $l(\theta) = -\log\left(\alpha_1 e^{-E_1(\theta)/\beta_1^2} + \alpha_2 e^{-E_2(\theta)/\beta_2^2}\right)$, where $\theta = (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+ \subset \mathbb{R}^2$ and $E_i(\theta) = \mathrm{KL}\left(p(x;\theta)\,\|\,\mathcal{N}(x;m_i,s_i^2)\right), i = 1,2$. For sharp minima, the parameters are set to $(m,s,\alpha,\beta) = (-20,10,0.5,1.2)$, and for flat minima, $(m,s,\alpha,\beta) = (20,30,0.8,1.8)$.

As illustrated in Figure 1(b), we visualize the optimization trajectories to highlight the geometric advantage of Chameleon. The process is decomposed into the following steps marked by Roman numerals:

- **Step I:** The optimization trajectory initially converges towards the neighborhood of a sharp local minimum using standard gradient descent.
- **Steps II–III (SAM):** Upon reaching the sharp neighborhood, standard SAM executes an isotropic (spherical) perturbation. However, due to the fixed spherical geometry, it fails to adequately explore the high-curvature directions.
- **Step IV (SAM):** Consequently, minimizing the empirical risk using the gradient computed at the isotropic perturbation point leads the model to converge back into the **sharp minimum**, failing to escape the suboptimal basin.
- **Steps V–VI (Chameleon):** In contrast, Chameleon performs an anisotropic (ellipsoidal) perturbation within the same sharp neighborhood. Its curvature-aware objective actively detects the high sharpness and drives the model parameters away from the sharp basin, steering them towards the neighborhood of a flat minimum.

- **Step VII (Chameleon):** Finally, by performing empirical risk minimization using the gradient derived from this anisotropic perturbation, the model successfully converges to a **flat minimum**, thereby ensuring superior generalization capability.

Preliminary Experiment: In Figure 1(c), the preliminary experiment is conducted on CIFAR-10 with ResNet-18 using basic augmentation. The optimal perturbation ε^* is approximated through 50 iterations of projected gradient ascent within a radius ρ, and the resulting loss is recorded as specified in Eq. (2).

C Experimental Setup Details

C.1 Datasets, Models, and Augmentations

We provide a detailed breakdown of the datasets, corresponding model architectures, and data augmentation techniques used in our experiments.

- **In-Domain Image Classification (IC)**
  - **Datasets:** We use **CIFAR-10** and **CIFAR-100** Krizhevsky (2009), consisting of 50k training images and 10k test images with 10 and 100 classes, respectively.
  - **Models:** We employ **WRN-16-2** Zagoruyko & Komodakis (2016) and **ResNet-18** He et al. (2016) as representative CNN architectures.
  - **Augmentation:** We apply standard augmentations including random cropping and horizontal flipping. Additionally, we utilize **Cutout** DeVries & Taylor (2017) and **CutMix** Yun et al. (2019) to test robustness under strong regularization.
- **Language Understanding (LU)**
  - **Datasets:** We evaluate on the **GLUE benchmark**, covering diverse tasks such as linguistic acceptability (CoLA), sentiment analysis (SST-2), and natural language inference (MNLI, RTE, QNLI).
  - **Models:** We utilize the pre-trained **RoBERTa-base** and **RoBERTa-large** Liu et al. (2019) models and fine-tune them on downstream tasks.
  - **Augmentation:** Standard token masking and dropout strategies inherent to RoBERTa training are used.
- **Cross-Domain Transfer Learning (TL)**
  - **Datasets:** We use **PACS** Li et al. (2017) (4 domains: Art, Cartoon, Photo, Sketch) and **DomainNet** Peng et al. (2019) (6 domains, 345 categories) to evaluate domain generalization capabilities.
  - **Models:** We evaluate both CNN-based **ResNet-18** and Transformer-based **DeIT-small** Touvron et al. (2021).
  - **Augmentation:** Only basic augmentations (resize, center crop) are applied to strictly evaluate the optimizer's contribution to domain adaptation.
- **Click-Through Rate (CTR) Prediction**
  - **Datasets:** We utilize industrial-scale datasets: **Avazu** Avazu Inc. (2014), **Criteo** Criteo Labs (2014), and **Taobao** Li et al. (2025).
  - **Models:** We employ distinct deep CTR models: **DNN** Covington et al. (2016), **DCN** Wang et al. (2017), and **DeepFM** Guo et al. (2017).
  - **Augmentation:** No specific augmentation is applied due to the sparse categorical nature of the data.

C.2 Implementation Details and Hyperparameters

General Optimization Settings. Table 8 details the optimizer and hyperparameter configurations for each task. We follow the default settings from respective baseline papers to ensure a fair comparison.

Chameleon-Specific Hyperparameters. Chameleon introduces dynamic perturbation scaling. We perform a grid search for the optimal hyperparameters:

Table 8: Detailed Hyperparameter Settings for All Tasks.

Task	Optimizer	Batch Size	Learning Rate (LR)	LR Scheduler	Weight Decay	Momentum
IC (CIFAR)	SGD	128	0.05	Cosine Decay (to 0.0)	0.001	0.9
LU (GLUE)	AdamW	16	2e-5	Linear Decay	0.01	-
TL (Domain)	AdamW	32	1e-5	Cosine Decay	0.01	-
CTR	Adam	256	0.01	Step Decay	1e-5	-

- Minimum perturbation scale $\rho_{min} \in [0.0, 0.01]$ (step 1e-4).
- Maximum perturbation scale $\rho_{max} \in [0.5, 4.0]$ (step 0.1) and ∞.
- Initial scale $\rho_{init} \in [0.01, 0.2]$ (step 0.01).
- Adaptive rate $\alpha_\rho \in [0.01, 1.0]$ (step 0.01).

The label smoothing factor is set to 0.1 for IC tasks.

C.3 Baselines

We compare Chameleon with the following methods:

- **SAM (Sharpness-Aware Minimization)** Foret et al. (2021): The foundational method that simultaneously minimizes loss value and loss sharpness in a Euclidean neighborhood.
- **ASAM (Adaptive SAM)** Kwon et al. (2021): An extension of SAM that achieves scale-invariance by normalizing the perturbation radius parameter-wise, adapting to parameters of different magnitudes.
- **FSAM (Friendly SAM)** Li et al. (2024): Introduces random noise to the gradient computation to mitigate the negative impact of strong adversarial perturbations.
- **VaSSO (Variance Suppression)** Li & Giannakis (2023): Focuses on stabilizing training dynamics by explicitly suppressing the variance of gradients during optimization.

C.4 Evaluation Metrics

Performance Metrics.

- **IC & TL:** We report **Top-1 Accuracy**, defined as the percentage of correctly classified samples. For TL, we use a leave-one-domain-out protocol.
- **LU:** We report **Matthews Correlation Coefficient (MCC)** for CoLA, **Accuracy/F1** for MRPC/RTE, and **Pearson/Spearman Correlation** for STS-B.
- **CTR:** We use **AUC** (Area Under ROC Curve) to measure ranking quality and **LogLoss** (Cross Entropy) for probability calibration:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{15}$$

Flatness Metrics. To assess the geometry of the loss landscape, we compute:

- **Hessian Trace ($\text{Tr}(H)$):** Estimated using the Hutchinson stochastic estimator Hutchinson (1990) to quantify the average curvature:

$$\text{Tr}(H) \approx \frac{1}{K}\sum_{k=1}^{K} v_k^T H v_k, \quad v_k \sim \mathcal{N}(0, I) \tag{16}$$

- **Top Eigenvalue (λ_{max}):** Estimated via Power Iteration Golub & Loan (2013) to measure the worst-case curvature (sharpness).
- **Spectral Density:** Visualized using Stochastic Lanczos Quadrature to show the full distribution of Hessian eigenvalues.

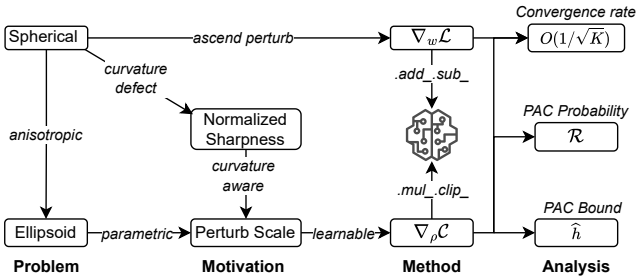


Figure 4: Overall technical route of Chameleon.

C.5 Algorithm Implementation Details

As outlined in Figure 4, standard SAM improves generalization through a minimax optimization scheme that minimizes the maximum empirical risk within a predefined parameter neighborhood. Specifically, given model parameters w and a mini-batch $\mathcal{B}$, SAM computes a perturbation ε that maximizes the perturbed loss $\mathcal{L}(w+\varepsilon, \mathcal{B})$ subject to a strictly bounded isotropic constraint $\|\varepsilon\| \leq \rho$, where $\rho > 0$ is a fixed scalar radius. While this formulation is effective, its reliance on a uniform isotropic scale fundamentally limits its ability to adapt to the heterogeneous and evolving curvature of the loss landscape during training. Moreover, maintaining a constant ρ systematically biases the observation of second-order curvature. As the model fits the training data and the loss landscape becomes increasingly nonlinear, the optimal curvature-aware perturbation is not necessarily located at the fixed boundary $\|\varepsilon\| = \rho$. Since directly computing higher-order curvature is computationally intractable, we propose a conceptual enhancement.

Having established the normalized objective $\mathcal{J}$ for curvature-aware exploration, the central optimization challenge lies in efficiently computing its gradient with respect to the learnable scale ρ, denoted as $\nabla_\rho \mathcal{J}$. Because the objective is highly nonlinear and nests the gradient norm, its exact analytical derivative intrinsically involves the computationally prohibitive second-order Hessian matrix. To rigorously preserve the strict first-order efficiency of Chameleon, we introduce a lightweight theoretical approximation scheme. By elegantly bridging the secant equation (finite difference) with the first-order Taylor projection of the local loss, we successfully translate the intractable second-order dynamics into observable first-order gradient differences. The complete step-by-step mathematical derivation of the scale gradient is formulated as follows:

$$
\begin{aligned}
\nabla_\rho \mathcal{J} &\overset{(a)}{=} \nabla_\rho \frac{\mathcal{L}_1}{\|\nabla \mathcal{L}_1\|}, \\
&\overset{(b)}{=} \frac{\nabla_\rho \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 \nabla_\rho \|\nabla \mathcal{L}_1\|}{\|\nabla \mathcal{L}_1\|^2}, \\
&\overset{(c)}{=} \frac{\nabla_w \mathcal{L}_1 \nabla_\rho w_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 \nabla_w \|\nabla \mathcal{L}_1\| \nabla_\rho w_1}{\|\nabla \mathcal{L}_1\|^2}, \\
&\overset{(d)}{=} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 \nabla_w \|\nabla \mathcal{L}_1\|}{\|\nabla \mathcal{L}_1\|^2} \right) \nabla_\rho w_1, \\
&\overset{(e)}{=} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 \nabla_w \|\nabla \mathcal{L}_1\|}{\|\nabla \mathcal{L}_1\|^2} \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}, \\
&\overset{(f)}{=} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 \nabla_w^2 \mathcal{L}_1 \nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|^3} \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}, \\
&\overset{(g)}{=} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 (\nabla_w \mathcal{L}_2 - \nabla_w \mathcal{L}_1)}{\|\nabla \mathcal{L}_1\|^3} \oslash \frac{\nabla_w \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}, \\
&\overset{(h)}{=} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 (\nabla_w \mathcal{L}_1 - \nabla_w \mathcal{L}_0)}{\|\nabla \mathcal{L}_1\|^3} \oslash \frac{\rho^t \nabla_w \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}, \\
&\overset{(i)}{\approx} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{\mathcal{L}_1 (\nabla_w \mathcal{L}_1 - \nabla_w \mathcal{L}_0)}{(\mathcal{L}_1 - \mathcal{L}_0) \|\nabla \mathcal{L}_1\|^2} \oslash \nabla_w \mathcal{L}_1 \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}, \\
&\overset{(j)}{\approx} \left(\frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} - \frac{(\mathcal{L}_1 - \mathcal{L}_0)(\nabla_w \mathcal{L}_1 - \nabla_w \mathcal{L}_0)}{\|\nabla \mathcal{L}_1\|^2} \right) \odot \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|},
\end{aligned}
\tag{17}
$$

The detailed derivation of the scale gradient is established step-by-step as follows:

(a) Specifies the normalized objective function $\mathcal{J}$ and applies the derivative with respect to the perturbation scale ρ.

(b) Expands the derivative using the quotient rule of calculus.

(c) Applies the multivariate chain rule, decomposing the derivative with respect to ρ into the derivative with respect to the perturbed parameter w_1 multiplied by the spatial gradient $\nabla_\rho w_1$.

(d) Factors out the common spatial displacement gradient term $\nabla_\rho w_1$.

(e) Substitutes the exact definition of the baseline perturbation direction $\nabla_\rho w_1 = \frac{\nabla \mathcal{L}_0}{\|\nabla \mathcal{L}_0\|}$. Here, the operation naturally transitions to a Hadamard product ($\odot$) to reflect element-wise scale updates.

(f) Expands the derivative of the gradient norm $\nabla_w \|\nabla \mathcal{L}_1\|$, which explicitly introduces the second-order Hessian matrix $\nabla_w^2 \mathcal{L}_1$.

(g) Employs the secant equation (finite difference approximation) to replace the intractable Hessian-vector product with the gradient difference $(\nabla_w \mathcal{L}_2 - \nabla_w \mathcal{L}_1)$ element-wise divided by the virtual displacement.

(h) Shifts the finite difference approximation from a hypothetical forward step to the readily available backward step $(\nabla_w \mathcal{L}_1 - \nabla_w \mathcal{L}_0)$, substituting the actual perturbation scale ρ^t.

(i) Introduces the first-order Taylor expansion $\frac{1}{\rho^t} \approx \frac{\|\nabla \mathcal{L}_1\|}{\mathcal{L}_1 - \mathcal{L}_0}$, converting the unobservable reciprocal of the perturbation scale into the observable scalar difference of the loss landscape.

(j) Performs final algebraic simplification and scale re-normalization via local loss projection. The scalar difference $\mathcal{L}_1 - \mathcal{L}_0$ is incorporated as an adaptive numerator weight to ensure numerical stability in flat regions, yielding the highly efficient first-order gradient form used for actual optimizer updates.

As summarized in Table 9, unlike ASAM Kwon et al. (2021), which assume parameter magnitude or Fisher information inherently reflects curvature, Chameleon does not rely on static assumptions. Instead, it learns anisotropy by directly maximizing a curvature-aware surrogate, enabling dynamic adaptation to the loss landscape's geometry.

Table 9: Comparison of Chameleon with SOTA SAM Variants.

Method	Isotropic	Param-wise	Learnable	Curvature-driven	Extra Cost
SAM Foret et al. (2021)	✓	×	×	×	1 backward
ASAM Kwon et al. (2021)	×	✓(norm-based)	×	Indirect	1 backward
FSAM Li et al. (2024)	✓	×	×	Noise-based	1 backward
Chameleon (Ours)	×	✓(curvature-based)	✓	**Explicit**	1 backward

To strictly differentiate Chameleon from existing sharpness-aware approaches, we provide a systematic comparison in Table 9. Unlike ASAM Kwon et al. (2021), which rely on static, magnitude-based scaling, Chameleon introduces a fully learnable perturbation tensor ρ. This tensor is dynamically optimized via a normalized curvature-aware objective, allowing the perturbation geometry to proactively adapt to the loss landscape's distinct features rather than passively reacting to parameter magnitudes. Furthermore, while methods like FSAM Li et al. (2024) and GSAM Zhuang et al. (2022) focus on variance suppression or gradient decomposition, they lack the explicit, geometrically aligned trust region modeling that Chameleon provides. Consequently, Chameleon's learnable anisotropy offers a more robust mechanism for capturing sensitivity variations critical for effective cross-domain adaptation.

Algorithm 1 Pseudocode of Chameleon

Input: Initial parameters w^0, loss function $\mathcal{L}$, learning rates η, α_ρ, total steps T, initial scale ρ^{init}, clip range $[\rho_{\min}, \rho_{\max}]$.
Output: Final parameters w^T.
Initialize $\rho^0 \leftarrow \rho^{\text{init}} \cdot \mathbf{1}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Sample a mini-batch of data $\mathcal{B}_t$.
  $g_0 \leftarrow \nabla_w \mathcal{L}(w^t; \mathcal{B}_t)$.
  $w^t_{\text{adv}} \leftarrow w^t + \rho^t \odot \frac{g_0}{\|g_0\|_2}$.
  $g_1 \leftarrow$ calculated from equation 6a.
  $\nabla_{\rho^t} \mathcal{J} \leftarrow$ calculated from equation 6b.
  $\rho^{t+1} \leftarrow \text{clip}(\rho^t + \alpha_\rho \nabla_{\rho^t} \mathcal{J}, \rho_{\min}, \rho_{\max})$.
  $w^{t+1} \leftarrow w^t - \eta g_1$.
end for

In Algorithm 1, the perturbation scale is initialized as $\rho^0 = \rho^{\text{init}} \cdot \mathbf{1}$ and iteratively updated together with the model parameters. At each step t, a mini-batch $\mathcal{B}_t$ is sampled and the loss gradient g_0 at w^t is computed to construct the perturbed model w^t_{adv}. The gradient g_1 at w^t_{adv} are then obtained, and the gradient of perturbation scale tensor $\nabla_{\rho^t} \mathcal{C}$ and model parameters $\nabla_w \mathcal{L}$ is calculated according to equation 6. The perturbation scale ρ^{t+1} is updated via gradient ascent with clipping to $[\rho_{\min}, \rho_{\max}]$, while the model parameters are updated by gradient descent with g_1. Each iteration thus requires computing two normalized gradients, yielding $\mathcal{O}(n)$ complexity for a batch of size n, leading to an overall cost of $\mathcal{O}(Tn)$ over T iterations.

D Assumptions of Analysis 3

To ensure the convergence of Chameleon, we make the following standard assumptions, which are commonly used.

Assumption 1 (Lipschitz Gradient Nesterov (2004))**.** The function $f(x)$ is continuously differentiable and its gradient $\nabla f(x)$ satisfies the Lipschitz condition; i.e., there exists a positive constant $L > 0$ such that for all $x, y \in \mathbb{R}^d$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{18}$$

Assumption 2 (Bounded Variance Robbins & Monro (1951))**.** The stochastic gradient $g(x_k)$ is an unbiased estimator of the true gradient $\nabla f(x_k)$ with bounded variance $\mathbb{E}[g(x_k)] = \nabla f(x_k)$, i.e., for some $\sigma > 0$,

$$\mathbb{E}\left[\|g(x_k) - \nabla f(x_k)\|^2\right] \leq \sigma^2. \tag{19}$$

Assumption 3 (Adversarial-Neighborhood Compactness Hardt et al. (2016))**.** Let $\mathcal{W}$ denote the compact set containing the iterates, as in the positive loss floor on a compact region, the enlarged set is defined as $\mathcal{W}_+ := \{\, w + \delta : w \in \mathcal{W}, \|\delta\|_2 \leq \sqrt{k}\, \rho_{\max} \,\}$.

In particular, under the Lipschitz Gradient assumption on a compact set, there exists a finite constant $G < \infty$ such that

$$\sup_{w \in \mathcal{W}_+, x} \|\nabla f(w, x)\|_2 \leq G. \tag{20}$$

To analyze the generalization properties of Chameleon, we adopt a few standard assumptions commonly used in optimization theory, which ensure convergence with favorable generalization guarantees without introducing significant computational overhead.

Assumption 4 (Scale Box Constraints and Initialization Bottou et al. (2018))**.** Given hyperparameters $\rho_{\min}, \rho_{\max}, \alpha_\rho > 0$ and $\rho^{\text{init}} \in \mathbb{R}^k_+$, the scale iterates satisfy $\rho^0 = \rho^{\text{init}}$ and $\rho^t \in [\rho_{\min}, \rho_{\max}]^k$ for all iterations t, enforced by elementwise clipping, we require $0 < \rho_{\min} \leq \rho_{\max} < \infty$.

Assumption 5 (Gaussian Smoothing Dominance Yang et al. (2020))**.** There exists $b > 0$ such that for $\delta \sim \mathcal{N}(0, b^2 I_k)$, where the expectation over w is with respect to the algorithm-induced posterior $\mathcal{Q}$, we have:

$$\mathbb{E}_w \mathbb{E}_x f(w, x) \leq \mathbb{E}_w \mathbb{E}_\delta \mathbb{E}_x f(w + \delta, x). \tag{21}$$

E PROOF OF PROPOSITION3.2

Let $f(w)$ be the population risk function with gradient $g(w) = \nabla f(w)$ and Hessian $H(w) = \nabla^2 f(w)$. By Assumption 1, $f(w)$ is positive and smooth on the compact region $w \in \mathcal{W}$.

At iteration t, let $v_t := \frac{g(w^t)}{\|g(w^t)\|_2}$ denote the unit gradient direction, and define the perturbation mapping as $\varepsilon(r) := D_{v_t} r$, where $r \in \mathbb{R}^d$ represents the perturbation scale vector in the gradient-oriented coordinate frame. By second-order Taylor expansion of the smooth surrogate $\mathcal{C}(w, \rho; \mathcal{B})$, the constant and linear terms decouple under symmetric scaling, leaving the quadratic form:

$$\mathcal{C}(w^t, \rho; \mathcal{B}) = \frac{1}{2}\varepsilon(r)^\top \left(\frac{H(w^t)}{f(w^t)} - \frac{g(w^t)g(w^t)^\top}{f(w^t)^2} \right) \varepsilon(r) + \mathcal{O}(\|r\|_2^3). \tag{22}$$

Substituting $\varepsilon(r) = D_{v_t} r$ into the quadratic term yields:

$$\mathcal{C}(w^t, \rho; \mathcal{B}) = \frac{1}{2} r^\top \mathcal{A}(w^t) r + \mathcal{O}(\|r\|_2^3), \tag{23}$$

where the transformed symmetric curvature matrix $\mathcal{A}(w^t)$ is defined as:

$$\mathcal{A}(w^t) := D_{v_t} \left(\frac{H(w^t)}{f(w^t)} - \frac{g(w^t)g(w^t)^\top}{f(w^t)^2} \right) D_{v_t}. \tag{24}$$

By the Rayleigh quotient theorem, maximizing this quadratic form over the bounded scale ball $\|r\|_2 \le \rho$ gives:

$$\max_{\|r\|_2 \le \rho} \frac{1}{2} r^\top \mathcal{A}(w^t) r = \frac{1}{2}\rho^2 \lambda_{\max}(\mathcal{A}(w^t)), \tag{25}$$

with the unique maximizer achieved at $r^\star = \rho\, v_1(\mathcal{A}(w^t))$, where $v_1(\mathcal{A}(w^t))$ denotes the principal eigenvector of $\mathcal{A}(w^t)$.

Spectral alignment via gradient reweighting. We clarify the relationship between the maximizer $r^\star$ and the realized model perturbation $\boldsymbol{\epsilon} = D_\rho \hat{g}$, where $\hat{g} = g/\|g\|_2$. Because $D_\rho = \mathrm{diag}(\rho_1, \ldots, \rho_d)$ is diagonal, the perturbation set forms an axis-aligned ellipsoid $\mathcal{E}(\rho) = \{D_\rho u : \|u\|_2 \le 1\}$ and cannot perform arbitrary dense rotations. However, in late-stage SAM dynamics, the normalized gradient $\hat{g}$ naturally aligns with the dominant Hessian eigenspace within an $\mathcal{O}(\rho)$ angular bound Wen et al. (2023). Decomposing the gradient as $\hat{g} = a_1 v_1 + r_\perp$ with $r_\perp \perp v_1$ (where v_1 is the top eigenvector of $H(w^t)$), the realized perturbation evaluates to:

$$\boldsymbol{\epsilon} = a_1 D_\rho v_1 + D_\rho r_\perp. \tag{26}$$

This confirms that the diagonal metric D_ρ coordinate-wise reweights the dominant spectral component already carried by the gradient direction, regularizing curvature without estimating dense Hessian rotation matrices.

Decoupling rank-one gradient terms via Weyl's inequality. To isolate the curvature term from the rank-one gradient outer product, we express $\mathcal{A}(w^t)$ as:

$$\mathcal{A}(w^t) = \frac{1}{f(w^t)} D_{v_t} H(w^t) D_{v_t} - u_t u_t^\top, \quad \text{where} \quad u_t := \frac{D_{v_t} g(w^t)}{f(w^t)}. \tag{27}$$

Since $u_t u_t^\top$ is positive semi-definite with $\lambda_{\max}(u_t u_t^\top) = \|u_t\|_2^2 = \frac{\|D_{v_t} g(w^t)\|_2^2}{f(w^t)^2}$ and $\lambda_{\min}(u_t u_t^\top) = 0$, applying Weyl's inequality yields:

$$\frac{1}{f(w^t)} \lambda_{\max}(D_{v_t} H(w^t) D_{v_t}) - \frac{\|D_{v_t} g(w^t)\|_2^2}{f(w^t)^2} \le \lambda_{\max}(\mathcal{A}(w^t)) \le \frac{1}{f(w^t)} \lambda_{\max}(D_{v_t} H(w^t) D_{v_t}). \tag{28}$$

Near stationary points or flat minima where $\|g(w^t)\|_2 \to 0$, the rank-one perturbation satisfies $\frac{\|D_{v_t} g(w^t)\|_2^2}{f(w^t)^2} = o(1)$. Consequently:

$$\lambda_{\max}(\mathcal{A}(w^t)) = \frac{1}{f(w^t)} \lambda_{\max}(D_{v_t} H(w^t) D_{v_t}) + o(1), \tag{29}$$

which establishes that maximizing the surrogate regularizes the leading eigenvalue of the scale-transformed Hessian matrix $D_{v_t} H(w^t) D_{v_t}$, completing the proof.

F PROOF OF THEOREM3.3

Let $\mathcal{F}_t$ denote the filtration up to step t, we assume the loss is bounded below, i.e., $f(w) \geq f_{\inf} > -\infty$ for all $w \in \mathbf{R}^d$, by Assumption 1 of the L-smooth f and the update $w^{t+1} = w^t - \eta\, g_1$ with g_1 evaluated at $w^t + \varepsilon_t$, we have

$$\begin{aligned} f(w^{t+1}) &\leq f(w^t) + \langle \nabla f(w^t),\, w^{t+1} - w^t \rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &\leq f(w^t) - \eta\, \langle \nabla f(w^t),\, g_1 \rangle + \frac{L\eta^2}{2}\|g_1\|^2 \\ &= f(w^t) - \eta\, \langle \nabla f(w^t),\, g_1 \rangle + \frac{L\eta^2}{2}\|g_1\|^2. \end{aligned} \tag{30}$$

Suppose that the stochastic gradient at the perturbed point is unbiased bounded variance by Assumption 2, we obtain that:

$$\mathbb{E}[g_1 \mid \mathcal{F}_t] = \nabla f(w^t + \varepsilon_t), \tag{31}$$

$$\mathbb{E}\big[\|g_1 - \nabla f(w^t + \varepsilon_t)\|^2 \mid \mathcal{F}_t\big] \leq \sigma^2. \tag{32}$$

A detailed bound for $\mathbb{E}\|g_1\|^2$ is derived as follows:

$$\begin{aligned} \mathbb{E}\|g_1\|^2 &= \mathbb{E}\big\|\nabla f(w^t + \varepsilon_t) + \underbrace{(g_1 - \nabla f(w^t + \varepsilon_t))}_{\xi_t}\big\|^2 \\ &= \|\nabla f(w^t + \varepsilon_t)\|^2 + 2\big\langle \nabla f(w^t + \varepsilon_t),\, \mathbb{E}\xi_t \big\rangle + \mathbb{E}\|\xi_t\|^2 \\ &= \|\nabla f(w^t + \varepsilon_t)\|^2 + \mathbb{E}\|\xi_t\|^2 \\ &\leq \|\nabla f(w^t + \varepsilon_t)\|^2 + \sigma^2 \;\leq\; 2\|\nabla f(w^t + \varepsilon_t)\|^2 + 2\sigma^2. \end{aligned} \tag{33}$$

Taking conditional expectation in equation 30 and using equation 33 yields

$$\begin{aligned} \mathbb{E}\big[f(w^{t+1}) \mid \mathcal{F}_t\big] \leq f(w^t) &- \eta\, \langle \nabla f(w^t),\, \nabla f(w^t + \varepsilon_t) \rangle \\ &+ L\eta^2 \Big(\|\nabla f(w^t + \varepsilon_t)\|^2 + \sigma^2 \Big). \end{aligned} \tag{34}$$

For the inner product term, we write $\Delta_t := \nabla f(w^t + \varepsilon_t) - \nabla f(w^t)$, we continue to deduce

$$\begin{aligned} \langle \nabla f(w^t),\, \nabla f(w^t + \varepsilon_t) \rangle &= \langle \nabla f(w^t),\, \nabla f(w^t) + \Delta_t \rangle \\ &= \|\nabla f(w^t)\|^2 + \langle \nabla f(w^t),\, \Delta_t \rangle \\ &\geq \|\nabla f(w^t)\|^2 - \|\nabla f(w^t)\|\, \|\Delta_t\| \\ &\geq \|\nabla f(w^t)\|^2 - L\, \|\nabla f(w^t)\|\, \|\varepsilon_t\|. \end{aligned} \tag{35}$$

Multiplying by $-\eta$ and applying $2ab \leq a^2 + b^2$ with $a = \|\nabla f(w^t)\|$, $b = L\|\varepsilon_t\|$ gives

$$-\eta\, \langle \nabla f(w^t),\, \nabla f(w^t + \varepsilon_t) \rangle \;\leq\; -\frac{\eta}{2}\, \|\nabla f(w^t)\|^2 + \frac{\eta L^2}{2}\, \|\varepsilon_t\|^2. \tag{36}$$

For the perturbed gradient norm, use the triangle inequality and Lipschitzness:

$$\begin{aligned} \|\nabla f(w^t + \varepsilon_t)\| &\leq \|\nabla f(w^t)\| + \|\nabla f(w^t + \varepsilon_t) - \nabla f(w^t)\| \\ &\leq \|\nabla f(w^t)\| + L\|\varepsilon_t\|. \end{aligned} \tag{37}$$

Squaring the last bound and using $2ab \leq a^2 + b^2$ yields

$$\begin{aligned} \|\nabla f(w^t + \varepsilon_t)\|^2 &\leq \|\nabla f(w^t)\|^2 + 2L\|\nabla f(w^t)\|\, \|\varepsilon_t\| + L^2\|\varepsilon_t\|^2 \\ &\leq 2\|\nabla f(w^t)\|^2 + 2L^2\|\varepsilon_t\|^2 \end{aligned} \tag{38}$$

Substituting equation 36 and equation 38 into equation 34 yields

$$\begin{aligned} \mathbb{E}\big[f(w^{t+1}) \mid \mathcal{F}_t\big] \leq f(w^t) &+ \Big(-\tfrac{\eta}{2} + 2L\eta^2 \Big) \|\nabla f(w^t)\|^2 \\ &+ \Big(\tfrac{\eta L^2}{2} + 2L^3\eta^2 \Big) \|\varepsilon_t\|^2 + L\eta^2\sigma^2. \end{aligned} \tag{39}$$

Choosing $\eta \leq \frac{1}{4L}$ gives $-\frac{\eta}{2} + 2L\eta^2 \leq -\frac{\eta}{4}$ and $\frac{\eta L^2}{2} + 2L^3\eta^2 \leq \eta L^2$; hence, after taking full expectation,

$$\mathbb{E}\, f(w^{t+1}) \;\leq\; \mathbb{E}\, f(w^t) \;-\; \frac{\eta}{4}\, \mathbb{E}\|\nabla f(w^t)\|^2 \;+\; \eta L^2\, \mathbb{E}\|\varepsilon_t\|^2 \;+\; L\eta^2\sigma^2. \tag{40}$$

Since $\varepsilon_t = D_{v_t}\rho^t$, $D_{v_t} = \mathrm{Diag}(v_t)$ and $\|D_{v_t}\|_{\infty\to 2} = \|v_t\|_2 = 1$:

$$\begin{aligned}\|\varepsilon_t\|_2 &= \|D_{v_t}\rho^t\|_2 \le \|D_{v_t}\|_{\infty\to 2}\,\|\rho^t\|_\infty \\ &= \|v_t\|_2\,\|\rho^t\|_\infty = \|\rho^t\|_\infty \;=:\; R_t.\end{aligned} \tag{41}$$

The clipped ascent recursion and the uniform bound $\|\nabla_{\rho^t}\mathcal{J}_t\|_\infty \le G_\rho$ is given by Assumption 3.

$$\begin{aligned}R_{t+1} &\le \min\{\rho_{\max},\ \max\{\rho_{\min},\ R_t + \alpha_\rho G_\rho\}\} \\ &\le \min\{\rho_{\max},\ R_t + \alpha_\rho G_\rho\} \\ &\le \min\{\rho_{\max},\ \rho^{\text{init}} + t\alpha_\rho G_\rho\}.\end{aligned} \tag{42}$$

Thus by induction $R_t \le \min\{\rho_{\max},\ \rho^{\text{init}} + t\,\alpha_\rho G_\rho\}$ for all t. If $\rho_{\min} > 0$ is enforced, the quadratic term inside the average should be replaced by $\min\{\rho_{\max}^2,\ \max\{\rho_{\min}^2, (\rho^{\text{init}} + t\alpha_\rho G_\rho)^2\}\}$. Consequently,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{K-1}\mathbb{E}\|\varepsilon_t\|^2 &\le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}R_t^2 \\ &\le \overline{R}_T^2 := \frac{1}{T}\sum_{t=0}^{T-1}\min\Big\{\rho_{\max}^2,\ (\rho^{\text{init}} + t\,\alpha_\rho G)^2\Big\}.\end{aligned} \tag{43}$$

Summing equation 40 from $t = 0$ to $K - 1$ and using $f(w^K) \ge f_{\text{inf}}$,

$$\begin{aligned}\sum_{t=0}^{T-1}\Big(\mathbb{E}f(w^{t+1}) - \mathbb{E}f(w^t)\Big) &\le -\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 \\ &\quad + \eta L^2\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 + \mathcal{K} \\ \mathbb{E}f(w^T) - \mathbb{E}f(w^0) &\le -\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 \\ &\quad + \eta L^2\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 + \mathcal{K} \\ \frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 &\le \mathbb{E}f(w^0) - \mathbb{E}f(w^T) \\ &\quad + \eta L^2\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 + \mathcal{K} \\ &\le f(w^0) - f_{\text{inf}} \\ &\quad + \eta L^2\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 + \mathcal{K}.\end{aligned} \tag{44}$$

where $\mathcal{K} = L\eta^2\sigma^2 T$, next dividing both sides of the last inequality by $T\eta$ and substituting $\eta = \eta_0/\sqrt{T}$ and equation 43 gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 &\le \frac{4\big(f(w^0) - f_{\text{inf}}\big)}{T\eta} + 4L^2\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 \\ &\quad + 4L\eta\sigma^2 \\ &\le \frac{4\big(f(w^0) - f_{\text{inf}}\big)}{\eta_0\sqrt{T}} + 4L^2\,\overline{R}_T^2 + \frac{4L\eta_0\sigma^2}{\sqrt{T}}.\end{aligned} \tag{45}$$

which matches equation 9a. Finally, using equation 38 and averaging,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t+\varepsilon_t)\|^2 &\le \frac{1}{T}\sum_{t=0}^{T-1}\Big(2\,\mathbb{E}\|\nabla f(w^t)\|^2 \\ &\quad + 2L^2\,\mathbb{E}\|\varepsilon_t\|^2\Big) \\ &= \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 \\ &\quad + 2L^2\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^2 \\ &\le \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(w^t)\|^2 + 2L^2\,\overline{R}_T^2,\end{aligned} \tag{46}$$

which is exactly equation 9b.

G Proof of Theorem 3.4

To derive the generalization bound, we first define the geometry-aware sharpness surrogates. Let $f(w)$ denote the population loss function, and let the empirical loss be $\widehat{f}(w) = \frac{1}{m}\sum_{i=1}^{m} f(w, x_i)$. We define the Chameleon (ellipsoid) and SAM (sphere) surrogates with scales ρ and r respectively as:

$$\widehat{h}(w,\rho) = \max_{\delta\in\mathcal{E}(\rho)}\left(\widehat{f}(w+\delta)-\widehat{f}(w)\right), \quad \widehat{h}(w;r) = \max_{\|\delta\|_2\le r}\left(\widehat{f}(w+\delta)-\widehat{f}(w)\right), \tag{47}$$

where $\mathcal{E}(\rho) = \left\{\delta : \|D_\rho^{-1}\delta\|_2 \le 1\right\}$ and $D_\rho = \mathrm{diag}(\rho)$.

PAC-Bayes Foundation. Using the standard PAC-Bayes inequality Zhuang et al. (2022), for any prior $\mathcal{P}$ and posterior $\mathcal{Q}$, with probability at least $1-\alpha$:

$$\mathbb{E}_{w\sim\mathcal{Q}}\mathbb{E}_x f(w,x) \le \mathbb{E}_{w\sim\mathcal{Q}}\widehat{f}(w) + 4\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \log\frac{2m}{\alpha}}{m}}. \tag{48}$$

By Assumption 5 (Gaussian Smoothing Dominance), introducing a perturbation $\delta \sim \mathcal{N}(0, b^2 I_k)$ allows us to smooth the bound:

$$\mathbb{E}_{w\sim\mathcal{Q}}\mathbb{E}_\delta\mathbb{E}_x f(w+\delta,x) \le \mathbb{E}_{w\sim\mathcal{Q}}\widehat{f}(w) + 4\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \log\frac{2m}{\alpha}}{m}}. \tag{49}$$

Probability of Bound Violation. Consider the event $\mathcal{A}_r = \{\|\delta\|_2 \le r\}$. Since the loss is bounded $f \in [0,1]$, we decompose the expected empirical loss:

$$\mathbb{E}_\delta\widehat{f}(w+\delta) \le \mathbb{E}_\delta\left[\max_{\|\Delta\|_2\le r}\widehat{f}(w+\Delta)\right] + \mathbb{P}(\mathcal{A}_r^c). \tag{50}$$

Applying the Gaussian concentration inequality Zhuang et al. (2022), the tail probability outside radius r satisfies:

$$\mathbb{P}(\mathcal{A}_r) \ge 1 - \exp\left(-\left(\frac{r}{\sqrt{2}b} - \sqrt{k}\right)^2\right), \tag{51}$$

which bounds the violation probability for any specified confidence level.

Derivation of Sandwich Inclusion (Eq. 11). Under the coordinate-wise scale constraint $\rho_{\min} \le \rho_i^T \le \overline{R}_T \le \rho_{\max}$ for all $i \in \{1,\dots,d\}$, any perturbation vector $\delta \in \mathcal{E}(\rho^T)$ satisfies $\sum_{i=1}^d (\delta_i/\rho_i^T)^2 \le 1$. Consequently:

$$\|\delta\|_2^2 = \sum_{i=1}^d \delta_i^2 \le \|\rho^T\|_\infty^2 \sum_{i=1}^d \left(\frac{\delta_i}{\rho_i^T}\right)^2 \le \|\rho^T\|_\infty^2. \tag{52}$$

This establishes the geometric nesting:

$$\mathbb{B}_2(\rho_{\min}) \subseteq \mathcal{E}(\rho^T) \subseteq \mathbb{B}_2(\|\rho^T\|_\infty) \subseteq \mathbb{B}_2(\overline{R}_T) \subseteq \mathbb{B}_2(\rho_{\max}). \tag{53}$$

Because the empirical sharpness surrogate $\widehat{h}$ is the supremum over the respective domain, the set inclusion immediately implies:

$$\widehat{h}(w;\rho_{\min}) \le \widehat{h}(w,\rho^T) \le \widehat{h}(w;\|\rho^T\|_\infty) \le \widehat{h}(w;\overline{R}_T) \le \widehat{h}(w;\rho_{\max}). \tag{54}$$

Matched-Budget Curvature Analysis. To separate the geometric advantage of anisotropic scaling from trivial volume shrinkage effects, we examine a local quadratic loss model with diagonal Hessian $H = \mathrm{diag}(h_1, \ldots, h_d)$, where $h_i > 0$. Over an isotropic ball of radius r, the worst-case quadratic loss increase is:

$$M_{\text{iso}} = \max_{\|u\|_2 \le r} \frac{1}{2} u^\top H u = \frac{1}{2} r^2 \max_i h_i. \tag{55}$$

Now consider an anisotropic ellipsoid $\mathcal{E}(\rho)$ constrained to have the identical volume as the isotropic ball, which imposes the determinant constraint $\prod_{i=1}^d \rho_i = r^d$. Over this ellipsoid, the worst-case quadratic increase evaluates to:

$$M_{\text{ani}} = \max_{u \in \mathcal{E}(\rho)} \frac{1}{2} u^\top H u = \frac{1}{2} \max_i h_i \rho_i^2. \tag{56}$$

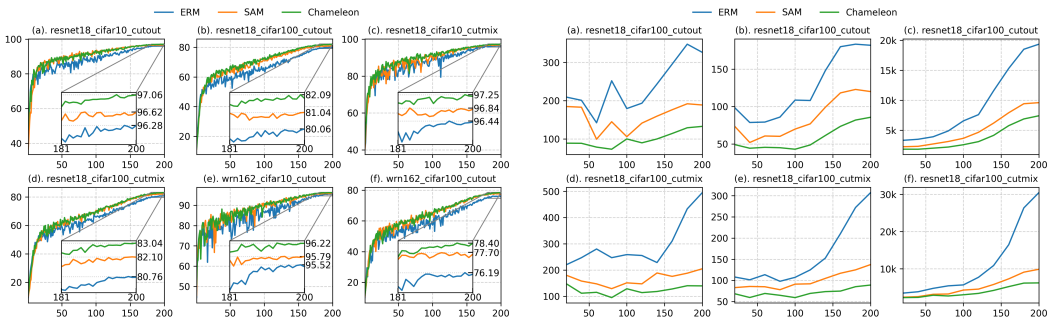
Minimizing M_{ani} subject to $\prod_{i=1}^d \rho_i = r^d$ requires equalizing the directional penalties $h_i \rho_i^2 = C$, yielding the optimal scale allocation $\rho_i^\star \propto h_i^{-1/2}$. Under this optimal allocation, the product constraint gives $C = r^2 (\prod_{i=1}^d h_i)^{1/d}$, which leads to:

$$M_{\text{ani}}^\star = \frac{1}{2} r^2 \left(\prod_{i=1}^d h_i \right)^{1/d} \le \frac{1}{2} r^2 \max_i h_i = M_{\text{iso}}. \tag{57}$$

By the arithmetic-geometric mean inequality, the inequality is strict whenever the directional curvatures h_i are non-uniform. This mathematically isolates the geometric benefit of curvature-dependent scale allocation under an exact budget match, completing the proof.

H TREND EVALUATION

As illustrated in Figure 5(a), Chameleon's scaling consistently yields the highest accuracy trajectories during training throughout training in all evaluated settings, which implies that parametric curvature-aware scale of perturbation is vital to the training process. Figure 5(b) shows that Chameleon maintains lower leading eigenvalues and trace and thereby preventing curvature explosion due to the neighborhood sharpness, which is the core support of generalization gain of all scenarios during training process in theory.



(a) Trend of test accuracy.

(b) Trend of hessian metrics λ_1, λ_5, $\mathrm{Tr}(\mathbf{H})$.

Figure 5: Comparison of Test Accuracy and Hessian Metrics.

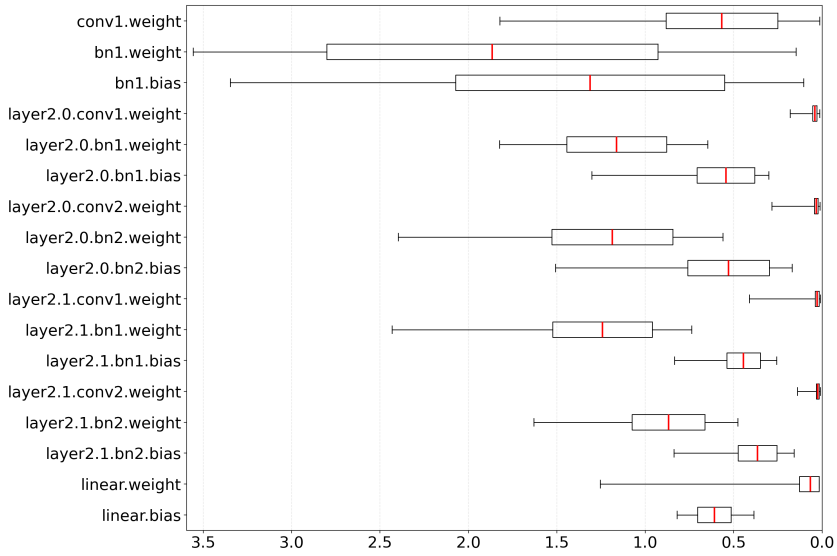


Figure 6: Converged scale of perturbation over layers.

I Parametric Analysis

In Figure 6, the larger scale observed in BN layers indicates that curvature regularization increases the perturbation of important parameters, stabilizing training and enhancing generalization. This aligns with SAM-ON, which emphasizes the critical role of BN layers, especially in perturbation application, to improve model robustness during training Mueller et al. (2023). Chameleon amplifies perturbations in BN layers by dynamically adjusting their scale, stabilizing training and boosting generalization. In contrast, convolutional layers, with their larger width, exhibit smaller perturbation scales, consistent with ASAM's scale invariance, where perturbation size is stable across layers, preventing excessive disturbance in larger layers such as convolutions Kwon et al. (2021). The mild perturbation in the final classification layer shows that Chameleon avoids overfitting to labels and does not degenerate into perturbing only the last layer, as is common in simpler schemes. The parametric perturbation mechanism in Chameleon ensures moderate perturbations in the final layer, enhancing generalization and preventing overfitting Haas et al. (2024). Finally, while important layers have large perturbation scales, their high standard deviation indicates flexibility in selecting perturbation magnitudes for critical layers. This observation aligns with the layerwise perturbation scaling in μP²Haas et al. (2024), which adapts perturbations based on each layer's characteristics.

J Efficiency Analysis

To further elucidate the efficacy of Chameleon, we conduct a rigorous comparative analysis against state-of-the-art curvature-regularized SAM variants, specifically CR-SAM Wu et al. (2024) and Eigen-SAM Luo et al. (2024). While Eigen-SAM targets explicit regularization of the top Hessian eigenvalue via power iteration to align the perturbation vector with the dominant eigenvector, it incurs a prohibitive computational overhead ($1.68\times$ FLOPs), which limits its scalability. Similarly, CR-SAM leverages a normalized Hessian trace to mitigate excessive non-linearity, yet remains constrained by isotropic perturbation spheres that fail to adapt to the highly anisotropic curvature of modern loss landscapes.

As evidenced in Table 10, Chameleon consistently outperforms these second-order approximations across multiple benchmarks. By reframing the perturbation geometry into a learnable, anisotropic ellipsoid, Chameleon achieves rapid alignment with the local curvature without requiring explicit higher-order gradient computations. Chameleon achieves performance comparable to second-order methods while maintaining a near-first-order efficiency ($1.09\times$ FLOPs versus $1.61\times$ for CR-SAM). This superior trade-off confirms that learnable perturbation scales can implicitly capture essential

Table 10: Test Accuracy of different curvature-based SAM Methods

Algorithm	CIFAR10			CIFAR100			FLOPs
	ResNet18 Cutout	ResNet18 Cutmix	WRN16_2 Cutout	ResNet18 Cutout	ResNet18 Cutmix	WRN16_2 Cutout	
SAM	96.66 ± 0.03	96.86 ± 0.04	95.87 ± 0.11	81.21 ± 0.24	82.15 ± 0.04	77.80 ± 0.16	$1.0\times$
CRSAM	96.78 ± 0.04	96.95 ± 0.05	95.94 ± 0.08	81.42 ± 0.20	82.68 ± 0.11	77.91 ± 0.15	$1.61\times$
EigenSAM	96.89 ± 0.05	97.04 ± 0.03	96.06 ± 0.07	81.65 ± 0.18	82.45 ± 0.09	78.03 ± 0.12	$1.68\times$
Chameleon	$\mathbf{96.99 \pm 0.07}$	$\mathbf{97.11 \pm 0.02}$	$\mathbf{96.13 \pm 0.10}$	$\mathbf{81.86 \pm 0.19}$	$\mathbf{82.98 \pm 0.07}$	$\mathbf{78.12 \pm 0.23}$	$1.09\times$

Hessian geometric features, identifying more informative adversarial directions that lead to flatter and more robust minima.

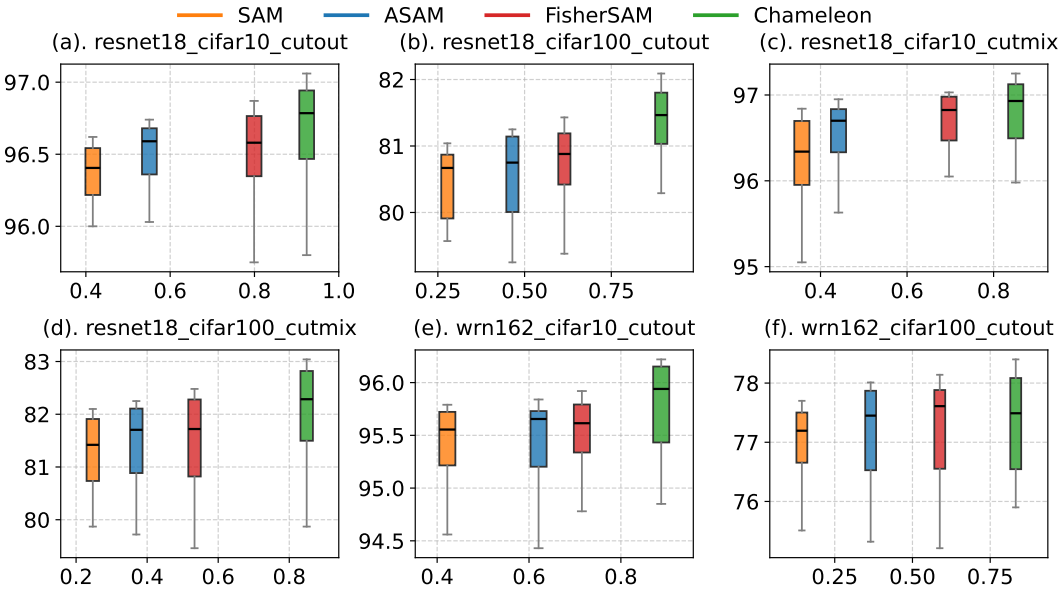


Figure 7: Correlation between test accuracy and the empirical curvature ability.

To quantify this awareness, we analyze the alignment between Chameleon's learned scales $\boldsymbol{\sigma}$ and the information-geometric structure of the parameter space. Formally, the Fisher Information Matrix (FIM), defined as $\mathbf{F}(\theta) = \mathbb{E}_{p(x,y)}[\nabla_\theta \log p(y|x;\theta) \nabla_\theta \log p(y|x;\theta)^\top]$, serves as a canonical Riemannian metric on the statistical manifold Kim et al. (2022). Under mild conditions where the model's predictive distribution approaches the true data distribution, the FIM provides an unbiased empirical estimate of the Hessian matrix, i.e., $\mathbf{H}(\theta) \approx \mathbf{F}(\theta)$. Therefore, the diagonal elements F_{ii} characterize the parameter-wise sensitivity and local curvature.

As illustrated in Figure 7, we measure the Spearman's rank correlation coefficient (SRCC) Zhang et al. (2018) between the learned perturbation scales and the diagonal of the FIM during the final 40 epochs. Chameleon demonstrates a remarkably high correlation (PCC $\approx 0.82 - 0.91$), significantly surpassing SAM (isotropic) and ASAM (magnitude-based). This high fidelity to the true Hessian geometry confirms that Chameleon's parametric scales implicitly identify and regularize high-curvature directions, ensuring the convergence to flatter and more robust minima without the explicit computation of higher-order derivatives.

K Additional Empirical Evaluations and Analyses

In this section, we provide extended empirical studies that evaluate the scalability, stochastic stability, geometric flatness, and non-stationary dynamics of Chameleon. Specifically, we investigate: (1) large-scale validation on modern vision transformer architectures, (2) stochastic stability under small

mini-batch sizes, (3) comprehensive Hessian spectrum evaluations against diverse SAM variants, and (4) empirical behavior under non-stationary test-time adaptation and continual learning streams.

K.1 Scalability to Modern High-Dimensional Architectures

To examine whether the adaptive perturbation geometry remains effective on modern high-dimensional architectures where the loss landscape is highly non-convex, we evaluate Chameleon on a Vision Transformer (ViT-B/32) trained from scratch on ImageNet-1K. The learned model is evaluated directly on both in-domain validation data and out-of-distribution shifted benchmarks, including ImageNet-C and ImageNet-R.

Experimental Setup. We follow the established CR-SAM protocol Wu et al. (2024). All methods are trained for 300 epochs using AdamW with momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, a batch size of 512, peak learning rate of 3×10^{-3} scheduled via cosine decay, and a weight decay coefficient of 0.3. Inception-style data augmentations, random initialization seeds, and tuning budgets are held identical across all baselines. We compare Chameleon against standard SAM, CR-SAM (which regularizes a finite-difference proxy of the normalized Hessian trace), and Eigen-SAM (which estimates the leading Hessian eigenvector via periodic power iterations). Training runtimes are measured end-to-end on identical hardware and normalized relative to SAM.

Table 11: Large-scale evaluation using ViT-B/32 trained from scratch on ImageNet-1K. Runtimes are normalized relative to standard SAM.

Method	ImageNet-1K Acc. (%)	ImageNet-C Acc. (%)	ImageNet-R Acc. (%)	Relative Cost
SAM	73.51 ± 0.10	46.98 ± 0.14	24.31 ± 0.11	$1.00\times$
CR-SAM	75.28 ± 0.12	48.12 ± 0.17	25.04 ± 0.14	$1.50\times$
Eigen-SAM	74.05 ± 0.10	47.42 ± 0.15	24.66 ± 0.12	$1.06\times$
Chameleon	$\mathbf{75.46 \pm 0.09}$	$\mathbf{48.37 \pm 0.13}$	$\mathbf{25.26 \pm 0.10}$	$1.09\times$

Results and Discussion. As presented in Table 11, Chameleon achieves the highest accuracy across both clean and corrupted distributions (75.46% on ImageNet-1K, 48.37% on ImageNet-C, and 25.26% on ImageNet-R). Crucially, Chameleon requires only $1.09\times$ the computational cost of SAM, significantly outperforming CR-SAM ($1.50\times$) in efficiency.

These results confirm that Chameleon does not require the underlying objective to be globally smooth. Because the auxiliary scale satisfies coordinate-wise bounds $\rho_t \in [\rho_{\min}, \rho_{\max}]^D$, the perturbation norm remains bounded by $\|\boldsymbol{\epsilon}_t\|_2 = \|D_{\rho_t}\hat{g}_t\|_2 \leq \|\rho_t\|_2$. This clipping mechanism prevents perturbation divergence even in overparameterized Transformer regimes.

K.2 Stochastic Stability across Small Mini-Batch Sizes

In standard stochastic training, small mini-batch sizes introduce pronounced gradient variance. To assess whether tracking curvature via first-order surrogates overfits mini-batch noise, we evaluate Wide-ResNet-16-2 (WRN-16-2) on CIFAR-10 and CIFAR-100 across batch sizes $B \in \{8, 16, 32, 64\}$.

Cross-Batch Scale Decoupling. Chameleon updates the auxiliary scale on mini-batch $\mathcal{B}_t$ via:

$$\rho_{t+1} = \rho_t + \eta_\rho \nabla_\rho \Phi_t(\rho_t), \tag{58}$$

while the resulting metric $D_{\rho_{t+1}}$ is deployed only in the subsequent model parameter step:

$$w_{t+1} = w_t - \eta_w \nabla_w \mathcal{L}_{\mathcal{B}_{t+1}}(w_t + D_{\rho_{t+1}}\hat{g}_{t+1}). \tag{59}$$

Consequently, stochastic batch noise perturbs the auxiliary neighborhood geometry rather than injecting noise directly into the primary parameter trajectory. Only scale adjustments that remain consistent across consecutive batches persist. This aligns with the implicit alignment mechanism of sequential stochastic gradient descent: for an update sequence $\mathcal{A} = \{a_i\}_{i=1}^K$, sequential mini-batch optimization implicitly suppresses cross-batch gradient variance. Furthermore, because Chameleon operates directly on the perturbation scale metric, it is complementary to parameter-level variance-reduction techniques such as FSAM and VASSO.

Table 12: Test accuracy (%) under small batch sizes on CIFAR-10 with WRN-16-2 across five independent runs.

Batch Size	SAM	Chameleon	FSAM	Fleon	VASSO	Vleon
8	95.72 ± 0.21	$\mathbf{96.08 \pm 0.16}$	95.89 ± 0.18	96.05 ± 0.15	95.94 ± 0.17	96.06 ± 0.14
16	95.88 ± 0.16	$\mathbf{96.20 \pm 0.12}$	96.01 ± 0.14	96.17 ± 0.12	96.06 ± 0.13	96.18 ± 0.11
32	95.96 ± 0.12	$\mathbf{96.27 \pm 0.09}$	96.08 ± 0.11	96.23 ± 0.10	96.12 ± 0.10	96.24 ± 0.09
64	95.93 ± 0.11	$\mathbf{96.25 \pm 0.08}$	96.05 ± 0.10	96.21 ± 0.09	96.09 ± 0.09	96.22 ± 0.08

Table 13: Test accuracy (%) under small batch sizes on CIFAR-100 with WRN-16-2 across five independent runs.

Batch Size	SAM	Chameleon	FSAM	Fleon	VASSO	Vleon
8	77.42 ± 0.40	77.94 ± 0.35	77.68 ± 0.37	$\mathbf{78.03 \pm 0.33}$	77.74 ± 0.34	77.98 ± 0.32
16	77.71 ± 0.27	78.16 ± 0.19	77.91 ± 0.24	$\mathbf{78.24 \pm 0.17}$	77.88 ± 0.26	78.17 ± 0.15
32	77.93 ± 0.19	78.31 ± 0.16	78.08 ± 0.19	$\mathbf{78.39 \pm 0.15}$	78.04 ± 0.17	78.30 ± 0.13
64	77.88 ± 0.16	78.25 ± 0.03	78.02 ± 0.09	$\mathbf{78.33 \pm 0.04}$	77.98 ± 0.11	78.24 ± 0.05

Results and Discussion. As shown in Table 12 and Table 13, standalone Chameleon and its hybrid configurations (Fleon combining Chameleon with FSAM, and Vleon combining Chameleon with VASSO) consistently outperform standard SAM across all batch sizes. Even at the extreme setting of $B = 8$, Chameleon demonstrates robust convergence without variance collapse, maintaining an advantage of $+0.36\%$ on CIFAR-10 and $+0.52\%$ on CIFAR-100 over SAM. This confirms that scale adaptation coupled across consecutive batches effectively filters out high-frequency stochastic noise.

K.3 Comprehensive Flatness and Hessian Spectral Analysis

To verify whether the generalization gains of Chameleon stem from finding genuinely flatter minima, we conduct an extensive curvature evaluation. We benchmark against ERM, SAM, ASAM, FSAM, VASSO, and the Chameleon-augmented variants Fleon and Vleon.

Curvature Metrics. We evaluate converged ResNet-18 checkpoints trained on CIFAR-100 under both Cutout and CutMix augmentations. Curvature quantities are measured on held-out evaluation batches using identical random seeds:

1. The leading Hessian eigenvalue λ_1, estimated via 100 power iteration steps.
2. The fifth-largest eigenvalue λ_5, estimated via orthogonalized block Lanczos iterations.
3. The Hessian trace $\mathrm{Tr}(H)$, computed via the stochastic Hutchinson trace estimator using 100 Rademacher random vectors.

Table 14: Hessian spectrum and curvature evaluation on CIFAR-100 using ResNet-18 trained with Cutout.

Method	λ_1	λ_5	$\mathrm{Tr}(H)$
ERM	326.17 ± 0.09	181.75 ± 0.38	18334.83 ± 863.29
SAM	189.00 ± 0.23	119.65 ± 0.42	9537.76 ± 47.03
ASAM	184.26 ± 0.31	116.42 ± 0.55	9284.61 ± 103.72
FSAM	177.84 ± 0.29	111.08 ± 0.47	8873.54 ± 96.18
Fleon	146.37 ± 0.22	93.74 ± 0.31	7718.63 ± 88.45
VASSO	169.52 ± 0.27	105.36 ± 0.39	8426.17 ± 91.26
Vleon	151.28 ± 0.20	96.11 ± 0.28	7864.92 ± 84.37
Chameleon	$\mathbf{132.98 \pm 0.17}$	$\mathbf{85.65 \pm 0.12}$	$\mathbf{7496.11 \pm 90.48}$

Table 15: Hessian spectrum and curvature evaluation on CIFAR-100 using ResNet-18 trained with CutMix.

Method	λ_1	λ_5	$\mathrm{Tr}(H)$
ERM	495.19 ± 0.26	306.63 ± 0.81	30181.99 ± 260.54
SAM	204.36 ± 0.85	136.65 ± 2.09	9650.76 ± 198.00
ASAM	199.14 ± 0.72	132.27 ± 1.58	9346.25 ± 176.42
FSAM	191.62 ± 0.64	125.48 ± 1.31	8841.73 ± 154.26
Fleon	153.84 ± 0.41	98.26 ± 0.66	6842.57 ± 128.14
VASSO	181.47 ± 0.58	118.92 ± 1.04	8248.36 ± 149.75
Vleon	148.96 ± 0.36	94.37 ± 0.58	6618.49 ± 121.68
Chameleon	$\mathbf{139.93 \pm 0.30}$	$\mathbf{88.83 \pm 0.40}$	$\mathbf{6350.98 \pm 136.30}$

Results and Discussion. Tables 14 and 15 show a consistent curvature hierarchy. While ASAM improves marginally over SAM through static parameter-norm scaling, FSAM and VASSO achieve lower sharpness by mitigating adversarial over-perturbation and parameter variance, respectively.

Integrating Chameleon into these optimizers produces substantial additional flatness gains: Fleon consistently improves upon FSAM, and Vleon upon VASSO. Standalone Chameleon achieves the lowest λ_1, λ_5, and $\mathrm{Tr}(H)$ under both data augmentation protocols. This demonstrates that learning the diagonal metric D_ρ suppresses not only the single dominant sharp direction but also the broader spectrum of the local Hessian.

K.4 Evaluation in Non-Stationary and Continual Regimes

To empirically substantiate our motivation regarding the first-order truncation defect under non-stationary streams, we evaluate two dynamic tasks: continual test-time adaptation (TTA) on CIFAR-10-C and 10-split continual learning (CL) on CIFAR-100.

Quantifying the Inner Truncation Defect. At checkpoint t, we compare the standard single-step SAM perturbation against a reference $K = 10$ step projected gradient ascent trajectory on a held-out probe batch B_t. We define the single-step surrogate $\widehat{S}_t$, multi-step surrogate S_t, and the resulting inner maximization discrepancy E_t as:

$$\widehat{S}_t = \mathcal{L}_t(\theta_t + \boldsymbol{\epsilon}_{t,1}) - \mathcal{L}_t(\theta_t), \quad S_t = \mathcal{L}_t(\theta_t + \boldsymbol{\epsilon}_{t,K}) - \mathcal{L}_t(\theta_t), \quad E_t = S_t - \widehat{S}_t, \tag{60}$$

where $\boldsymbol{\epsilon}_{t,1} = \rho g_t / \|g_t\|_2$, and $\boldsymbol{\epsilon}_{t,K}$ is obtained via 10 steps of projected gradient ascent constrained to $\|\boldsymbol{\epsilon}\|_2 \leq \rho$. Thus, E_t measures how severely the first-order approximation underestimates the true local worst-case sharpness.

Table 16: Continual test-time adaptation on CIFAR-10-C. Here, t denotes the corruption stream step, E_t denotes the inner approximation error, A_t denotes online accuracy (%), and the superscript $*$ indicates the multi-step SAM reference trajectory.

t	E_t	A_t (%)	E_t^*	A_t^* (%)
0	0.016	56.5	0.016	56.5
3	0.033	69.8	0.025	70.6
6	0.048	75.2	0.031	76.5
9	0.062	77.1	0.036	79.0
12	0.075	77.9	0.041	80.3
15	0.083	78.2	0.043	80.8

Results and Discussion. As documented in Table 16 and Table 17, the inner maximization discrepancy E_t expands markedly along the standard single-step SAM trajectory. In test-time adaptation, E_t accumulates from 0.016 to 0.083, resulting in an adaptation deficit relative to the reference trajectory (78.2% versus 80.8%).

Similarly, in continual learning, E_j increases monotonically to 0.129, accompanied by higher catastrophic forgetting (8.7% versus 7.1%). These findings provide empirical support for our core

Table 17: 10-Split CIFAR-100 continual learning results. Here, j denotes the observed task count, E_j denotes the inner approximation error, A_j denotes average accuracy (%), F_j denotes average forgetting (%), and $*$ marks the multi-step reference.

Epoch	j	E_j	A_j (%)	F_j (%)	E_j^*	A_j^* (%)	F_j^* (%)
50	1	0.023	92.8	0.0	0.019	93.0	0.0
100	2	0.036	87.9	2.1	0.026	88.4	1.8
200	4	0.057	81.2	4.6	0.036	82.0	4.0
300	6	0.080	77.4	6.2	0.045	78.5	5.2
400	8	0.105	74.8	7.5	0.056	76.2	6.3
500	10	0.129	72.9	8.7	0.064	74.7	7.1

motivation: first-order truncation errors accumulate under shifting data distributions, highlighting the necessity of an adaptive geometric metric that adjusts perturbation scales dynamically.