MOMENT TRANSPORT POISONING FOR ONLINE TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online test-time adaptation (TTA) updates a pretrained model on unlabeled test data, allowing poisoned inputs to influence benign predictions through shared normalization and model updates. We propose Moment Transport Poisoning (MTP), a grey-box attack that targets pre-BatchNorm activation statistics using a fixed source model. MTP maximizes standardized shifts in channel means and log-variances and combines this moment objective with non-ground-truth entropy through gradient-normalized projected updates. Our analysis connects moment displacement to feature distortion under mixed-batch normalization and to changes in the adaptation, and derives sufficient conditions for increased benign risk. Experiments on CIFAR-10-C, CIFAR-100-C, and four ImageNet-C subsets show that MTP outperforms DIA and RTTDP baselines in mean benign error. Five-seed evaluations on both CIFAR-C datasets show higher mean benign error at both perturbation budgets. Ablations show that moment transport alone produces higher benign error than NHE alone, while their combination achieves the highest benign error. Elevated error during subsequent benign-only adaptation and increased held-out probe loss further demonstrate effects beyond the poisoned batch.

1 INTRODUCTION

Test-time adaptation (TTA) improves robustness under distribution shift by adapting a pretrained model to unlabeled test data (Sun et al., 2020; Wang et al., 2021a; Niu et al., 2022; Wang et al., 2022; Niu et al., 2023). Incoming samples update normalization statistics, model parameters, or adaptation states, allowing predictions to track changes in the test distribution. This dependence on the test stream creates a poisoning channel: attacker-controlled samples can alter shared normalization and adaptation updates, degrading predictions on benign inputs. DIA demonstrated this vulnerability for benign samples sharing an adaptation batch with poisons (Wu et al., 2023). Subsequent work explored transferable poisoning (Cong et al., 2024), grey-box attacks (Su et al., 2025), label-free attacks (Rifat et al., 2025), and sample-wise targeted attacks (Nguyen & Nguyen, 2026).

A main challenge in grey-box poisoning is to construct effective poisons without observing the model's ongoing adaptation. In the RTTDP setting, the attacker has access to a fixed source model and labeled attacker-controlled samples, but cannot observe the current adapted parameters, benign samples in the adaptation batch, or future test inputs (Su et al., 2025). The attacker therefore needs an objective evaluated on the source model that transfers to the evolving target and disrupts its predictions on benign samples. Prediction-level losses and feature-level discrepancies provide such surrogate objectives. For normalization-based TTA, this motivates targeting feature statistics that directly participate in normalization and adaptation.

Pre-BatchNorm channel means and variances provide a natural target. These moments determine feature standardization, are used to update normalization statistics, and influence gradients for adapting normalization parameters (Schneider et al., 2020; Wang et al., 2021a; Mirza et al., 2022; Gong et al., 2022; Yuan et al., 2023). When poisoned and benign samples share an adaptation batch, changes in poison moments alter the mixed-batch statistics used to normalize benign features. Poisoned inputs can also change the adaptation update, affecting predictions beyond the current batch. Because pre-BatchNorm moments are observable and differentiable on the fixed source model, they offer a tractable objective for targeting these mechanisms under grey-box setting.

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.

We propose Moment Transport Poisoning (MTP), which maximizes standardized shifts in pre-BatchNorm channel means and log-variances relative to clean attacker-controlled inputs. The moment objective is motivated by the local geometry of feature normalization. We combine it with non-ground-truth entropy (NHE) (Su et al., 2025) to incorporate a prediction-level objective, normalize the two input gradients separately, and generate poisons through projected sign updates. The main contributions are as follows.

- We introduce MTP, a grey-box poisoning attack that targets pre-BatchNorm means and log-variances using a fixed source model. The attack combines a normalization-motivated moment objective with NHE through gradient-normalized projected updates.
- We analyze how moment displacement transfers across model states and propagates through mixed-batch normalization to distort benign features. We further decompose changes in adaptation update into a moment-based component and a residual, derive sufficient conditions for increased benign risk, and characterize the interaction between the two attack objectives.
- We evaluate MTP on CIFAR-10-C, CIFAR-100-C, and four ImageNet-C subsets. Five-seed CIFAR-C evaluations show higher mean benign error than the baselines at both perturbation budgets. Ablations establish the contribution of moment transport, while recovery and held-out probe experiments demonstrate effects beyond the poisoned batch.

2 Related Work

Test-time adaptation. TTA adapts pretrained models to unlabeled target data through normalization updates and parameter optimization. BatchNorm adaptation estimates statistics from target batches (Schneider et al., 2020), DUA updates these statistics online (Mirza et al., 2022), and TENT optimizes normalization parameters by minimizing prediction entropy (Wang et al., 2021a). EATA extends entropy-based adaptation with sample selection and anti-forgetting regularization (Niu et al., 2022). Other approaches use contrastive learning (Chen et al., 2022), augmentation consistency (Zhang et al., 2022b), or teacher–student updates with stochastic restoration (Wang et al., 2022). For changing test streams, NOTE combines instance-aware normalization with memory (Gong et al., 2022), RoTTA uses robust normalization (Yuan et al., 2023), and SoTTA and DeYO refine sample selection and confidence criteria (Gong et al., 2023; Lee et al., 2024). These methods couple model behavior to incoming samples, creating a channel through which poisoned inputs can affect benign predictions.

Test-time poisoning. Test-time poisoning exploits this dependence to degrade benign predictions. DIA optimizes malicious inputs through dynamic adaptation (Wu et al., 2023), while TePA studies poison transfer across adaptation mechanisms (Cong et al., 2024). RTTDP develops grey-box attacks using a fixed source model, introducing non-ground-truth entropy (NHE) and TTA-aware surrogate objectives (Su et al., 2025). FCA targets feature geometry without ground-truth labels (Rifat et al., 2025), and sample-wise attacks pursue targeted manipulation during online adaptation (Nguyen & Nguyen, 2026). MTP follows RTTDP's untargeted grey-box setting and targets pre-BatchNorm moments to disrupt shared normalization and adaptation updates.

Feature-level transferable attacks. Intermediate representations also provide objectives for transferable adversarial attacks. ILA amplifies perturbation directions at selected layers (Huang et al., 2019), FIA emphasizes object-relevant channels (Wang et al., 2021b), and NAA weights neurons by attribution scores (Zhang et al., 2022a). These methods improve adversarial transfer across fixed models. Online TTA introduces an additional pathway: adversarial inputs influence the model through adaptation. MTP targets this pathway using pre-BatchNorm means and variances, linking a feature-level attack objective to the statistics governing normalization and adaptation gradients.

3 Problem Formulation

Let θ_t denote the persistent adaptation state at time t, including model parameters, optimizer state, and running statistics. Let B_t be a benign batch, P_t^0 a clean attacker-controlled batch, and P_t its perturbed version. The mixed batch $X_t(P_t) = B_t \sqcup P_t$ induces the online update $\theta_{t+1}(P_t) =$

$\mathcal{A}(\theta_t, X_t(P_t))$, where $\mathcal{A}$ denotes the state-transition operator implemented by the TTA algorithm. The attacker seeks to increase prediction risk on the benign members of this batch. Writing this risk as $R_t^{\mathrm{cb}}(\theta, X; B_t)$, the poisoning objective is

$$\max_{\substack{\|P_t - P_t^0\|_\infty \le \epsilon \\ P_t \in [0,1]^{|P_t^0| \times d}}} R_t^{\mathrm{cb}}\left(\theta_t^{\mathrm{eval}}(P_t), X_t(P_t); B_t\right), \tag{1}$$

where d is the input dimension and $\theta_t^{\mathrm{eval}}(P_t)$ is the state used for prediction: either the incoming state or the state after the current update, depending on the TTA setting. A fixed held-out benign probe separately measures the effect of the updated state beyond the poisoned batch.

We adopt the RTTDP grey-box setting (Su et al., 2025). The attacker has access to the fixed source model f_{θ_0} and labels of attacker-controlled benchmark samples, while the current state θ_t, benign batch B_t, and future test inputs remain unobserved. For batch size m and poison ratio ρ, the attacker controls ρm samples. Each constrained by $\|\tilde{x} - x\|_\infty \le \epsilon$, $\tilde{x}$ is perturbed input and $\tilde{x} \in [0,1]^d$. The attack therefore uses a source-model surrogate for Eq. 1.

To analyze local update effects, let ϑ_t collect the parameters optimized during adaptation and let $A(\vartheta; B_t, P_t)$ be the unlabeled adaptation loss, with the remaining incoming state fixed. We study a single gradient step $\vartheta_{t+1}(P_t) = \vartheta_t - \gamma \nabla_\vartheta A(\vartheta_t; B_t, P_t)$, where $\gamma > 0$. This local model connects poison-induced changes in the adaptation gradient to parameter displacement and benign probe risk.

4 Moment Transport Poisoning

MTP combines moment displacement on the source model with NHE. We analyze its effects on normalization, adaptation, and benign risk, with proofs and extensions in Appendices A and C.

4.1 Moment Transport Objective and Normalization Geometry

Let $\phi_l(\cdot)$ denote the pre-BatchNorm activation tensor at layer l, with channel-wise means $\mu_l(\cdot)$ and variances $\sigma_l^2(\cdot)$. Given clean attacker-controlled inputs x, MTP maximizes the displacement of these moments under perturbed inputs $\tilde{x}$:

$$\begin{aligned}\mathcal{L}_{\mathrm{mom}}(\tilde{x}; x) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \Bigg[\frac{1}{C_l} \sum_{c=1}^{C_l} \frac{\left(\mu_{l,c}(\tilde{x}) - \mu_{l,c}(x)\right)^2}{\mathrm{sg}\Big(\sigma_{l,c}^2(x)\Big) + \tau} \\ + \frac{1}{4C_l} \sum_{c=1}^{C_l} \Big(\log\big(\sigma_{l,c}^2(\tilde{x}) + \tau\big) - \log\big(\sigma_{l,c}^2(x) + \tau\big)\Big)^2 \Bigg],\end{aligned} \tag{2}$$

where $\mathcal{L}$ is the set of selected layers, C_l is the number of channels, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and $\tau = 10^{-5}$. Clean reference moments remain fixed throughout the optimization. The two terms measure mean shifts standardized by the clean variance and changes in stabilized log-variance, respectively. The relative weighting follows from the local geometry of normalization. For a scalar channel activation Z with mean μ and variance v, let $s = \log(v + \tau)$ and define $\mathcal{N}_\tau(z; \mu, s) = e^{-s/2}(z - \mu)$.

Proposition 1 (Local geometry of normalization moments)**.** *For a perturbation $h = (a, b)$ of (μ, s), the expected squared directional derivative of $\mathcal{N}_\tau$ defines the local quadratic form $Q_{\mathrm{BN}}(h) = \frac{a^2}{v+\tau} + \frac{v}{4(v+\tau)} b^2$. Define the corresponding MTP quadratic as $Q_{\mathrm{MTP}}(h) = \frac{a^2}{v+\tau} + \frac{1}{4} b^2$. With $\kappa = v/(v+\tau)$, $\kappa Q_{\mathrm{MTP}}(h) \le Q_{\mathrm{BN}}(h) \le Q_{\mathrm{MTP}}(h)$. The two forms coincide when $\tau = 0$ and $v > 0$.*

Proposition 1 motivates the variance standardization and the $1/4$ log-variance weight in Eq. 2. The resulting objective measures moment displacement through a quadratic form tied to changes in normalized features. Appendix A gives the corresponding local expansion.

4.2 Transfer to the Adaptation Batch

The attack objective is computed on the source model, while poisoning acts on the adapting model through batches containing both benign and attacker-controlled samples. We analyze these two transitions in turn.

Transfer across model states. Fix a forward-pass configuration and batch construction across parameter states. Let $\mathcal{M}_\theta(X) \in \mathbb{R}^q$ collect the selected pre-BatchNorm means and stabilized log-variances of batch X under θ. For a clean poison batch P and perturbation ΔP, define $h_\theta(\Delta P) = \mathcal{M}_\theta(P + \Delta P) - \mathcal{M}_\theta(P)$. Let $J_\theta(X) = \nabla_X \mathcal{M}_\theta(X)$ be the input Jacobian of the moment map.

Theorem 1 (Local source-to-adapted moment transfer)**.** *Suppose $\mathcal{M}_\theta$ is continuously differentiable in X on a neighborhood of the segment $P_s = P + s\Delta P$, $s \in [0,1]$. If $\|J_{\theta_t}(P_s) - J_{\theta_0}(P_s)\|_{\mathrm{op}} \le L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2$ for every $s \in [0,1]$ and some finite L_{tr}, then $\|h_{\theta_t}(\Delta P) - h_{\theta_0}(\Delta P)\|_2 \le L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2\|\Delta P\|_2$. Consequently, $\|h_{\theta_t}(\Delta P)\|_2 \ge \|h_{\theta_0}(\Delta P)\|_2 - L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2\|\Delta P\|_2$. If the source moment map satisfies $\|h_{\theta_0}(\Delta P)\|_2 \ge \sigma_{\mathrm{src}}\|\Delta P\|_2$ for all $\Delta P \in \mathcal{U}$, where $\mathcal{U}$ is an admissible perturbation set and $\sigma_{\mathrm{src}} > 0$, then $\|h_{\theta_t}(\Delta P)\|_2 \ge (\sigma_{\mathrm{src}} - L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2)\,\|\Delta P\|_2$. Thus, nonzero perturbations in $\mathcal{U}$ retain nonzero moment displacement if $L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2 < \sigma_{\mathrm{src}}$.*

The transfer discrepancy scales with model drift, perturbation size, and the sensitivity of the moment Jacobian. Changes in forward-pass configuration or batch context contribute an additional term. Writing $e_{\mathrm{ctx}} = \|h_{\mathrm{target}} - h_{\theta_t}\|_2$, where h_{θ_t} uses the shared convention above, gives $\|h_{\mathrm{target}} - h_{\theta_0}\|_2 \le L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2\|\Delta P\|_2 + e_{\mathrm{ctx}}$. Comparisons in the weighted source and mixed-batch moment norms additionally depend on the eigenvalue bounds of their respective metrics.

Transport into mixed-batch statistics. The adaptation batch combines benign and poison activations. The identities below show how poison moments affect the mixed mean and variance.

Proposition 2 (Mixture-moment transport)**.** *Let the benign activation distribution have moments (μ_B, v_B), and let clean and perturbed poison activations have moments (μ_0, v_0) and (μ_1, v_1). For $M_i = (1-\rho)B + \rho P_i$, $\delta\mu = \mu_1 - \mu_0$, and $\delta v = v_1 - v_0$, $\mu_{M_1} - \mu_{M_0} = \rho\,\delta\mu$, and $v_{M_1} - v_{M_0} = \rho\,\delta v + 2\rho(1-\rho)(\mu_0 - \mu_B)\delta\mu + \rho(1-\rho)(\delta\mu)^2$.*

Poison mean displacement enters the mixed mean linearly in ρ and also changes the mixed variance through the between-group term. Appendix A extends these identities to deeper layers where benign activations vary with the batch context.

Distortion of benign features. Changes in mixed-batch moments alter the normalization of benign activations. Let (μ_i, s_i), $i \in \{0,1\}$, denote the clean and poisoned normalization states, with $e^{s_i} = v_i + \tau > 0$. Let Z_0 and Z_1 be paired pre-normalization activations of the same benign input under these two contexts.

Theorem 2 (Finite distortion of benign feature)**.** *Suppose Z_0, Z_1 have finite second moments. Set $\mu_B = \mathbb{E}Z_0$, $v_B = \mathrm{Var}(Z_0)$, $a = \mu_1 - \mu_0$, $b = s_1 - s_0$, and $r = e^{-b/2}$. Changing the normalization state at fixed Z_0 produces distortion*

$$\begin{aligned} D_{\mathrm{fix}} &:= \mathbb{E}\left[\left(\mathcal{N}_\tau(Z_0;\mu_1,s_1) - \mathcal{N}_\tau(Z_0;\mu_0,s_0)\right)^2\right] \\ &= e^{-s_0}\left[(r-1)^2 v_B + \big((r-1)(\mu_B - \mu_0) - ra\big)^2\right]. \end{aligned} \tag{3}$$

For the actual paired distortion $D_{\mathrm{act}} := \mathbb{E}\left[\left(\mathcal{N}_\tau(Z_1;\mu_1,s_1) - \mathcal{N}_\tau(Z_0;\mu_0,s_0)\right)^2\right]$, let $\zeta = e^{-s_1/2}\left(\mathbb{E}[(Z_1 - Z_0)^2]\right)^{1/2}$. Then

$$\left(\max\{0, \sqrt{D_{\mathrm{fix}}} - \zeta\}\right)^2 \le D_{\mathrm{act}} \le \left(\sqrt{D_{\mathrm{fix}}} + \zeta\right)^2. \tag{4}$$

If $Z_1 = Z_0$, $\mu_B = \mu_0$, and $v_B = v_0$, Eq. 3 has the local expansion $Q_{\mathrm{BN}}(a,b) + O(\|(a,b)\|_2^3)$.

Theorem 2 extends the moment objective's local geometry to finite benign feature changes. The term $\mu_B - \mu_0$ captures the gap between benign and mixed-batch means, with ζ accounting for upstream activation changes; the resulting prediction loss effect depends on the downstream model and labels.

4.3 From Moment Displacement to Adaptation and Benign Risk

Poisoned inputs also influence benign predictions through adaptation updates. Fix the incoming state and benign batch, vectorize the poison perturbation as u, and define $m(u) = \mathcal{M}_{\theta_t}\big(B_t \sqcup (P_t^0 + u)\big)$, $g_t(u) = \nabla_\vartheta A(\vartheta_t; B_t, P_t^0 + u)$. Let $h(u) = m(u) - m(0)$, $J_0 = Dm(0)$, $K_0 = Dg_t(0)$, $S_0 = K_0 J_0^\dagger$,

and $E_0 = K_0(I - J_0^\dagger J_0)$, where $J_0^\dagger$ is the Moore–Penrose pseudoinverse. This decomposition separates gradient changes represented by the moments from sensitivity outside their row space.

Theorem 3 (Moment representation and update residual)**.** *If m and g_t have Lipschitz Jacobians on a neighborhood of the segment from 0 to u, then the difference between poisoned and clean one-step updates satisfies*

$$\Delta\vartheta_t(u) = -\gamma\left[S_0 h(u) + E_0 u + r_t(u)\right], \qquad \|r_t(u)\|_2 \le C_t\|u\|_2^2, \tag{5}$$

where $\Delta\vartheta_t(u)$ is the difference between poisoned and clean one-step updates.

Here, $S_0 h(u)$ captures the moment-based component of the gradient change, $E_0 u$ is the first-order residual, and $r_t(u)$ contains higher-order terms. Appendix A characterizes the optimal residual and derives conditions for upper and lower bounds on the update norm.

Direction of the risk change. The effect of an update on benign loss depends on both its magnitude and direction. Let ϑ_{t+1}^c and ϑ_{t+1}^p denote the clean and poisoned one-step states, with $\Delta\vartheta_t = \vartheta_{t+1}^p - \vartheta_{t+1}^c$. Let R_t be a fixed held-out benign probe risk.

Theorem 4 (Directional criterion for increased benign loss)**.** *If ∇R_t is L_R-Lipschitz on the line segment between the clean and poisoned states, then $R_t(\vartheta_{t+1}^p) - R_t(\vartheta_{t+1}^c) = \nabla R_t(\vartheta_{t+1}^c)^\top \Delta\vartheta_t + r_R$, where $|r_R| \le \frac{L_R}{2}\|\Delta\vartheta_t\|_2^2$. For nonzero gradient and displacement, define $\chi_t = \frac{\nabla R_t(\vartheta_{t+1}^c)^\top \Delta\vartheta_t}{\|\nabla R_t(\vartheta_{t+1}^c)\|_2\|\Delta\vartheta_t\|_2}$. If $\chi_t > \frac{L_R\|\Delta\vartheta_t\|_2}{2\|\nabla R_t(\vartheta_{t+1}^c)\|_2}$, then $R_t(\vartheta_{t+1}^p) > R_t(\vartheta_{t+1}^c)$.*

Combining Theorem 4 with the moment decomposition gives the following sufficient condition for increased benign risk.

Corollary 1 (Sufficient condition for increased risk with a representation residual)**.** *Under Theorem 3, suppose R_t has an L_R-Lipschitz gradient on the line segment between the clean and poisoned states. Let $r_c = \nabla R_t(\vartheta_{t+1}^c)$ and $e = e_t(u) := \|E_0 u\|_2 + C_t\|u\|_2^2$. Then $R_t(\vartheta_{t+1}^p) - R_t(\vartheta_{t+1}^c) \ge -\gamma r_c^\top S_0 h(u) - \gamma\|r_c\|_2 e - \frac{L_R\gamma^2}{2}\left(\|S_0 h(u)\|_2 + e\right)^2$. In particular, benign risk increases if $-r_c^\top S_0 h(u) > \|r_c\|_2 e + \frac{L_R\gamma}{2}\left(\|S_0 h(u)\|_2 + e\right)^2$.*

This condition links the moment-based update component to increased benign risk when its harmful alignment exceeds the contributions of the residual and risk curvature.

4.4 Joint Objective and Projected Optimization

MTP combines moment transport with the NHE objective used in RTTDP (Su et al., 2025). For ground-truth label y, let q_y assign zero probability to class y and uniform probability to all remaining classes. The prediction-level objective is $\mathcal{L}_{\mathrm{NHE}}(\tilde{x}, y) = -D_{\mathrm{KL}}(q_y \| f_{\theta_0}(\tilde{x}))$. We normalize the two input gradients separately and combine them:

$$g = w_{\mathrm{NHE}}\frac{\nabla_{\tilde{x}}\mathcal{L}_{\mathrm{NHE}}}{\mathrm{mean}(|\nabla_{\tilde{x}}\mathcal{L}_{\mathrm{NHE}}|) + \eta} + w_{\mathrm{mom}}\frac{\nabla_{\tilde{x}}\mathcal{L}_{\mathrm{mom}}}{\mathrm{mean}(|\nabla_{\tilde{x}}\mathcal{L}_{\mathrm{mom}}|) + \eta}, \tag{6}$$

where the mean is taken over each image's coordinates, $w_{\mathrm{NHE}} = w_{\mathrm{mom}} = 1$, and $\eta = 10^{-12}$. The projected sign update is $\tilde{x}^{k+1} = \Pi_{B_\infty(x,\epsilon)\cap[0,1]^d}\left(\tilde{x}^k + \alpha\,\mathrm{sign}(g^k)\right)$.

Poison generation. Source parameters remain fixed, and clean and perturbed moments use the same source forward-pass configuration. Algorithm 1 caches the clean moments, initializes a feasible poison batch, and performs projected sign updates. Since both the moment loss and its input gradient vanish at $\tilde{x} = x$, moment-only optimization requires a nontrivial initialization or an initial step supplied by another objective.

Interaction between the objectives. The joint update combines prediction-level and moment-based directions. Their coordinate signs determine how much first-order moment gain is retained after combination and projection. Vectorize the current poison batch as $z \in \mathcal{C}$, where $\mathcal{C} = B_\infty(x, \epsilon) \cap [0,1]^d$ denotes the feasible box over all batch coordinates. Set $F(z) = \mathcal{L}_{\mathrm{mom}}(z; x)$ and $g_M = \nabla F(z)$. Let $a = w_{\mathrm{mom}}\hat{g}_{\mathrm{mom}}$ and $b = w_{\mathrm{NHE}}\hat{g}_{\mathrm{NHE}}$, with $w_{\mathrm{mom}} > 0$ and $w_{\mathrm{NHE}} \ge 0$. Define the moment-only displacement as $d_M = \Pi_{\mathcal{C}}(z + \alpha\,\mathrm{sign}(g_M)) - z$, and joint displacements as $d_J = \Pi_{\mathcal{C}}(z + \alpha\,\mathrm{sign}(a + b)) - z$.

Algorithm 1 Moment Transport Poisoning (MTP)

Require: Clean batch x, labels y, source model f_{θ_0}, layers $\mathcal{L}$, budget ϵ, step size α, steps K, weights $w_{\mathrm{NHE}}, w_{\mathrm{mom}}$, stabilizers τ, η
Ensure: Poison batch $\tilde{x}$

```
Cache clean moments {μ_l(x), σ_l^2(x)}_{l∈L}
Initialize x̃^0 ∈ B_∞(x, ε) ∩ [0,1]^d
for k = 0, ..., K − 1 do
    L_NHE ← −D_KL(q_y ‖ f_θ0(x̃^k))
    Compute {μ_l(x̃^k), σ_l^2(x̃^k)}_{l∈L} and evaluate L_mom using Eq. 2
    g_NHE ← ∇_x̃ L_NHE,   g_mom ← ∇_x̃ L_mom
    ĝ_NHE ← g_NHE / (mean(|g_NHE|) + η)
    ĝ_mom ← g_mom / (mean(|g_mom|) + η)
    g^k ← w_NHE ĝ_NHE + w_mom ĝ_mom
    x̃^{k+1} ← Π_{B_∞(x,ε)∩[0,1]^d} (x̃^k + α sign(g^k))
end for
return x̃^K
```

Proposition 3 (Moment trade-off in the projected sign step)**.** *Let* $\mathrm{sign}(0) = 0$*. For any* $z \in \mathcal{C}$ *and* $\alpha > 0$*,* $D_{\mathrm{step}} := g_M^\top(d_M - d_J) \geq 0$*. If both proposed steps are unclipped, define* $\mathcal{F} = \{j : g_{M,j}(a_j + b_j) < 0\}$ *and* $\mathcal{Z} = \{j : g_{M,j} \neq 0,\ a_j + b_j = 0\}$*. Then* $D_{\mathrm{step}} = 2\alpha \sum_{j\in\mathcal{F}} |g_{M,j}| + \alpha \sum_{j\in\mathcal{Z}} |g_{M,j}|$*. A coordinate with* $a_j \neq 0$ *belongs to* $\mathcal{F}$ *exactly when* $a_j b_j < 0$ *and* $|b_j| > |a_j|$*. If* ∇F *is* L_F*-Lipschitz on the segments from* z *to* $z + d_M$ *and* $z + d_J$*, then* $|F(z + d_J) - F(z + d_M) + D_{\mathrm{step}}| \leq \frac{L_F}{2}\left(\|d_J\|_2^2 + \|d_M\|_2^2\right)$*. The joint step increases* F *when* $g_M^\top d_J > (L_F/2)\|d_J\|_2^2$*, and decreases* F *when* $g_M^\top d_J < -(L_F/2)\|d_J\|_2^2$*.*

The moment-only step maximizes first-order moment gain over the feasible coordinatewise step box. Adding NHE reduces this gain where its gradient reverses or cancels the sign of the moment gradient. Proposition 3 quantifies this local interaction, while Theorem 4 relates the resulting adaptation displacement to benign risk. Together, they distinguish the optimization trade-off between the objectives from their downstream effect on benign predictions.

5 Experiments

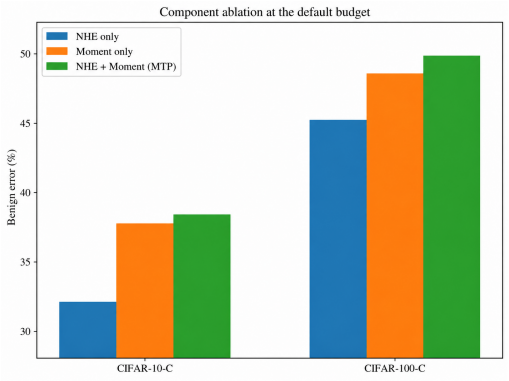
5.1 Experimental Setup

Datasets and Parameters. We evaluate MTP on CIFAR-10-C, CIFAR-100-C, and four ImageNet-C subsets (Hendrycks & Dietterich, 2019). The main CIFAR-C experiments follow the RTTDP setting with TENT adaptation, poison ratio 0.5, and batch size 200. Each corruption uses 400 examples, including 200 benign samples for evaluation. We test two perturbation budgets: the default setting with $\epsilon = 0.3$ and $\alpha = 0.01$, and a low-budget setting with $\epsilon = 8/255$ and $\alpha = 0.005$.

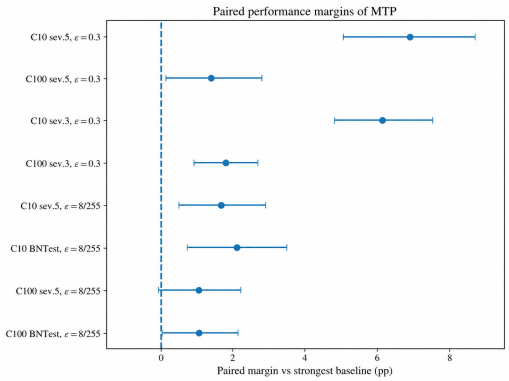
Baselines and Evaluation. We compare against DIA (Wu et al., 2023), RTTDP NHE, and RTTDP NHE+DR (Su et al., 2025). MTP uses 10 projected steps unless otherwise stated, whereas the baselines use 40. All comparisons use the same seed, corruption, poison ratio, perturbation budget, step size, and TTA configuration. We report benign classification error, with higher values indicating stronger poisoning. Appendix B provides full experimental settings and additional results, including evaluations under severity-3, BNTest, SAR, five-seed, and different optimization budgets.

5.2 Main Results

Table 1 summarizes the benign error across all 15 CIFAR-C corruptions at severity 5 over three seeds. Under the default budget, MTP reaches 38.40% benign error on CIFAR-10-C and 49.85% on CIFAR-100-C, which exceed those of the strongest baseline RTTDP NHE by 8.02 and 5.47 percentage points, respectively. At $\epsilon = 8/255$, MTP reaches 26.72% and 38.85% benign error on CIFAR-10-C and CIFAR-100-C, which exceed those of the strongest baseline DIA by 1.88 and 1.12 percentage points at this budget, respectively.



(a) Component ablation.



(b) Paired performance margins.

Figure 1: Component contributions and paired comparisons. (a) Ablation at $\epsilon = 0.3$, severity 5, TENT, and poison ratio 0.5. (b) Mean margins against the strongest baseline selected within each matched pair, with positive values favoring MTP.

Table 2 reports the mean and sample standard deviation across five seeds, with each seed evaluated on all 15 CIFAR-C corruptions. MTP retains the highest mean benign error across all four combinations of dataset and perturbation budget. At $\epsilon = 0.3$, MTP exceeds the strongest baseline by 7.79 percentage points on CIFAR-10-C and 2.57 points on CIFAR-100-C. At $\epsilon = 8/255$, the corresponding margins are 1.21 and 0.90 points. Overall, the five-seed evaluations show that MTP maintains a positive mean advantage on both datasets at both budgets.

Table 3 reports the benign error on ImageNet-C across four representative corruption subsets. MTP achieves the highest mean benign error on all four subsets. Compared with the strongest baseline on each subset, MTP achieves margins of 4.66, 3.00, 1.17, and 2.33 percentage points for contrast, defocus blur, Gaussian noise, and fog, respectively. These results demonstrate that MTP maintains effectiveness across diverse corruptions.

Table 4 reports benign error under SAR over three seeds and all 15 CIFAR-C corruptions at severity 5, with $\epsilon = 8/255$ and $\alpha = 0.005$, while batch size, poison ratio, sample counts, and continual sequences follow the main evaluation. MTP achieves the highest mean benign error, outperforming the strongest baseline DIA by 2.19 and 0.87 percentage points on CIFAR-10-C and CIFAR-100-C. Relative to the no-attack control, MTP increases benign error by 6.97 and 5.73 percentage points, respectively. These results demonstrate that MTP remains effective under SAR across both datasets.

Table 1: Mean benign error (%) across all 15 CIFAR-C corruptions at severity 5 over three seeds, with higher error indicating stronger poisoning.

Dataset	ϵ	MTP	DIA	RTTDP NHE	NHE+DR
CIFAR-10-C	0.3	**38.40**	26.96	30.38	25.63
CIFAR-100-C	0.3	**49.85**	41.36	44.38	36.72
CIFAR-10-C	8/255	**26.72**	24.84	21.99	21.51
CIFAR-100-C	8/255	**38.85**	37.73	35.11	34.00

5.3 Component Ablation and Robustness Diagnostics

Figure 1(a) compares NHE-only, Moment-only, and joint optimization under the default budget. Compared with NHE-only, Moment-only produces higher benign error on both datasets, demonstrating the contribution of the moment objective. Jointly optimizing the two objectives further increases the benign error from 37.73% to 38.40% on CIFAR-10-C and from 48.70% to 49.85% on CIFAR-100-C. Figure 1(b) compares MTP against the strongest baseline chosen from each matched pair. The results show that MTP maintains a positive mean performance margin across the evaluated datasets, budgets, severities, and adaptation settings. The corresponding component ablation and paired performance margins under low budgets are provided in Appendix B.

Table 2: Benign error (%) across all 15 CIFAR-C corruptions at severity 5, reported as mean ± sample standard deviation over five seeds. Higher error indicates stronger poisoning.

Dataset	ϵ	MTP	DIA	RTTDP NHE	NHE+DR
CIFAR-10-C	0.3	$\mathbf{38.41 \pm 0.70}$	27.77 ± 1.27	30.63 ± 1.02	25.97 ± 0.80
CIFAR-10-C	$8/255$	$\mathbf{26.69 \pm 0.49}$	25.47 ± 0.91	22.19 ± 1.04	21.71 ± 0.99
CIFAR-100-C	0.3	$\mathbf{47.12 \pm 2.72}$	41.86 ± 3.92	44.55 ± 3.14	36.93 ± 3.37
CIFAR-100-C	$8/255$	$\mathbf{39.15 \pm 3.02}$	38.25 ± 3.75	35.52 ± 3.32	34.61 ± 3.35

Table 3: Mean benign error (%) on four ImageNet-C corruption subsets over three seeds, with severity 3, $\epsilon = 8/255$, TENT, and poison ratio 0.5. Higher error indicates stronger poisoning.

Subset	MTP	DIA	RTTDP NHE	NHE+DR
Contrast	**46.83**	38.83	42.17	40.17
Defocus blur	**67.67**	64.00	64.67	62.67
Gaussian noise	**53.50**	52.33	42.67	43.67
Fog	**46.00**	41.00	43.67	40.00

Table 4: Benign error (%) under SAR at severity 5 and $\epsilon = 8/255$, reported as mean ± sample standard deviation over three seeds.

Method	CIFAR-10-C	CIFAR-100-C
No attack	20.33 ± 0.36	34.22 ± 3.75
MTP	27.30 ± 0.48	39.95 ± 3.85
DIA	25.11 ± 0.12	39.08 ± 4.75
RTTDP NHE	22.09 ± 0.84	36.36 ± 3.67
RTTDP NHE+DR	21.74 ± 0.51	35.24 ± 3.80

Figure 2 shows the results under different poison ratios, with $\rho \in 0.25, 0.50, 0.75$. On both CIFAR-C datasets, the benign error of MTP and DIA increases as the poison ratio increases. MTP exceeds DIA at every ratio, with a larger margin at higher ratios. Specifically, the margin grows from 4.18 points at $\rho = 0.25$ to 16.02 points at $\rho = 0.75$ on CIFAR-10-C. These results describe the effect of batch composition on attack performance and complement the mixture-moment analysis in Proposition 2.

5.4 Effects Beyond the Poisoned Batch

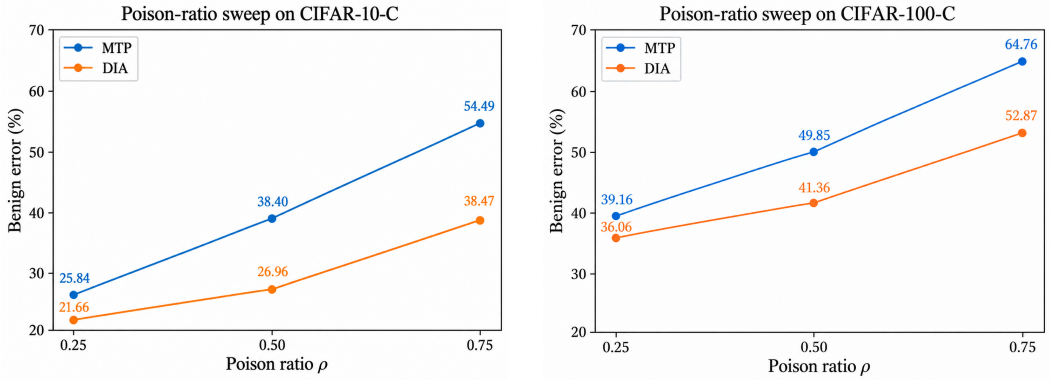
Table 5 reports recovery error and held-out probe loss after poisoning. Recovery measures the change in error on subsequent batches containing only benign samples. The probe compares clean and poisoned updates from identical incoming states using benign images excluded from adaptation and attack optimization. Relative to the no-attack control, MTP increases recovery error by 3.64 and 4.24 percentage points on CIFAR-10-C and CIFAR-100-C. Probe loss also increases on both datasets at both perturbation budgets. These results show that poisoning affects benign predictions beyond the poisoned batch. Detailed recovery protocols and state diagnostics appear in Appendix B.

Table 5: Effects beyond the poisoned batch. Recovery Δ denotes the benign error change relative to no attack at $\epsilon = 8/255$. Probe ΔL denotes the held-out benign loss change at $w_{\mathrm{NHE}} = 1$.

Dataset	Recovery Δ (pp)	Probe ΔL $(8/255)$	Probe ΔL (0.3)
CIFAR-10-C	+3.64	+0.02480	+0.04210
CIFAR-100-C	+4.24	+0.00594	+0.00839

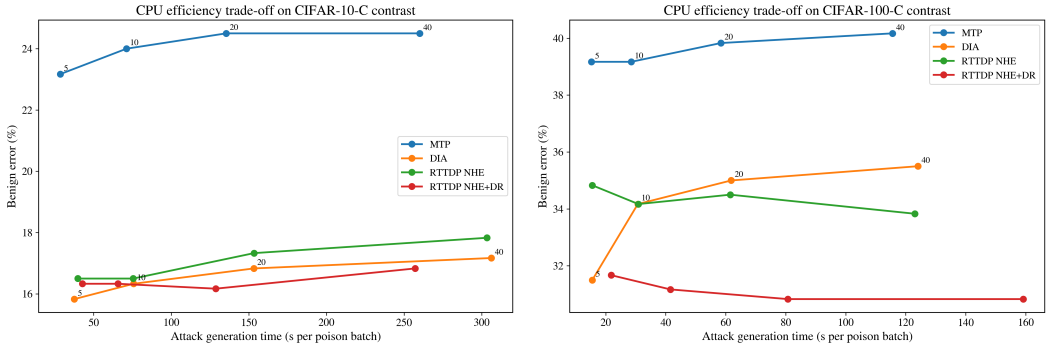
5.5 CPU Efficiency and Optimization Budget

Figure 3 compares the attack strength and CPU generation time of different methods under severity-5 contrast corruption, TENT, $\epsilon = 8/255$, and a poison ratio of 0.5. All methods are evaluated over three seeds with 5, 10, 20, and 40 attack optimization steps. The results show that MTP



(a) CIFAR-10-C (b) CIFAR-100-C

Figure 2: Effect of poison ratio at severity 5 with TENT, $\epsilon = 0.3$, and $\alpha = 0.01$. MTP yields higher benign error than DIA at all tested ratios, with a widening margin as the poison fraction increases.



(a) CIFAR-10-C (b) CIFAR-100-C

Figure 3: Benign error versus CPU attack-generation time on contrast at severity 5, $\epsilon = 8/255$, and $\alpha = 0.005$, with marker labels on the curves indicating attack steps. Higher error and lower time indicate stronger and faster poisoning.

achieves stronger poisoning with fewer optimization steps. Specifically, on CIFAR-10-C, MTP with 10 steps reaches 24.00% benign error in 71.20 seconds, whereas RTTDP NHE with 40 steps requires 303.33 seconds to reach 17.83%. On CIFAR-100-C, MTP with 10 steps achieves 39.17% in 28.50 seconds, while DIA with 40 steps requires 124.05 seconds to reach 35.50%. Overall, MTP delivers competitive poisoning with fewer optimization steps and lower CPU cost.

6 Conclusion

This paper proposes Moment Transport Poisoning (MTP), a grey-box attack on online TTA that combines pre-BatchNorm moment displacement with NHE using a fixed source model. Our analysis connects the objective to normalization distortion and adaptation updates, and derives conditions for increased benign risk. Experiments on CIFAR-C and four ImageNet-C subsets show higher mean benign error than DIA and RTTDP baselines, with degradation persisting after poisoning. These findings identify activation moments as an effective target for disrupting normalization-based TTA.

7 AI Use Statement

Generative AI tools were used to correct grammar and refine sentence wording.

REFERENCES

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Tianshuo Cong, Xinlei He, Yun Shen, and Yang Zhang. Test-time poisoning attacks against test-time adaptation models. In *2024 IEEE Symposium on Security and Privacy*, pp. 1306–1324, 2024. doi: 10.1109/SP54263.2024.00072.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, 2022.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. SoTTA: Robust test-time adaptation on noisy data streams. In *Advances in Neural Information Processing Systems*, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4733–4742, 2019. doi: 10.1109/ICCV.2019.00483.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

M. Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14765–14775, 2022.

Phuc Duc Nguyen and Quang Duc Nguyen. Sample-wise targeted adversarial attacks on test-time adaptation, 2026. URL `https://arxiv.org/abs/2605.23411`.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Shahriar Rifat, Jonathan Ashdown, Michael De Lucia, Ananthram Swami, and Francesco Restuccia. On the adversarial vulnerability of label-free test-time adaptation. In *International Conference on Learning Representations*, 2025.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yongyi Su, Yushu Li, Nanqing Liu, Kui Jia, Xulei Yang, Chuan-Sheng Foo, and Xun Xu. On the adversarial risk of test time adaptation: An investigation into realistic test-time data poisoning. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=7893vsQenk`.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022. doi: 10.1109/CVPR52688.2022.00706.

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7639–7648, 2021b. doi: 10.1109/ICCV48922.2021.00754.

Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. Uncovering adversarial risks of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37456–37495, 2023.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023. doi: 10.1109/CVPR52729.2023.01528.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14993–15002, 2022a.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022b. doi: 10.52202/068431-2799.

A Extended Theoretical Analysis

A.1 Local normalization and mixture expansions

The local geometry in Proposition 1 gives the second-order expansion

$$\mathbb{E}\Big[\big(\mathcal{N}_\tau(Z;\mu+a,s+b)-\mathcal{N}_\tau(Z;\mu,s)\big)^2\Big]=Q_{\mathrm{BN}}(h)+O(\|h\|_2^3), \tag{7}$$

where $h=(a,b)$. For the mixed-batch stabilized log-variance $s_M=\log(v_M+\tau)$, the variance displacement in Proposition 2 yields

$$s_{M_1}-s_{M_0}=\frac{v_{M_1}-v_{M_0}}{v_{M_0}+\tau}+O\big((v_{M_1}-v_{M_0})^2\big)\,. \tag{8}$$

These expansions connect changes in mixed-batch moments to the local normalization metric.

A.2 Context-dependent mixture transport

Batch-dependent benign activations. At deeper layers, upstream batch normalization makes benign activations depend on the poison context. Let B_i and P_i denote the benign and poison activation distributions from the same mixed forward pass under context i, with moments (μ_{B_i},v_{B_i}) and (μ_{P_i},v_{P_i}).

Proposition 4 (Context-dependent mixture transport)**.** *Set $M_i=(1-\rho)B_i+\rho P_i$, $d_0=\mu_{P_0}-\mu_{B_0}$, $a_B=\mu_{B_1}-\mu_{B_0}$, $a_P=\mu_{P_1}-\mu_{P_0}$, $b_B=v_{B_1}-v_{B_0}$, and $b_P=v_{P_1}-v_{P_0}$. Then*

$$\mu_{M_1}-\mu_{M_0}=(1-\rho)a_B+\rho a_P, \tag{9}$$

and

$$v_{M_1}-v_{M_0}=(1-\rho)b_B+\rho b_P+\rho(1-\rho)\left[2d_0(a_P-a_B)+(a_P-a_B)^2\right]. \tag{10}$$

Relative to the fixed-benign formula with the same a_P,b_P, the variance correction is

$$(1-\rho)b_B+\rho(1-\rho)\left[-2(d_0+a_P)a_B+a_B^2\right]. \tag{11}$$

Setting $a_B = b_B = 0$ recovers Proposition 2. The correction accounts for changes in benign variance and the difference between benign and poison mean. Both propositions use population variances, or empirical variances divided by the number of activation entries. Connecting these identities for mixed batch to the source objective requires measuring or bounding the discrepancy from source poison-only moments. Appendix C provides the proof.

A.3 Update representation and displacement bounds

The decomposition in Theorem 3 gives an optimal first-order representation of adaptation-gradient changes through the selected moments. With $J_0 = Dm(0)$, $K_0 = Dg_t(0)$, $S_0 = K_0 J_0^\dagger$, and $E_0 = K_0(I - J_0^\dagger J_0)$,

$$\inf_{S \in \mathbb{R}^{p \times q}} \|K_0 - S J_0\|_{\mathrm{op}} = \|E_0\|_{\mathrm{op}}. \tag{12}$$

An exact representation $K_0 = S J_0$ exists if and only if $\ker J_0 \subseteq \ker K_0$. To relate moment displacement to update magnitude, let $G_0 \succ 0$ be the metric on the selected mixed-batch moment and define $\|h\|_{G_0} = (h^\top G_0 h)^{1/2}$. Suppose

$$a_t \|h(u)\|_{G_0} \le \|S_0 h(u)\|_2 \le b_t \|h(u)\|_{G_0}, \qquad 0 < a_t \le b_t. \tag{13}$$

With $e_t(u) = \|E_0 u\|_2 + C_t \|u\|_2^2$, Theorem 3 yields

$$\gamma \left[a_t \|h(u)\|_{G_0} - e_t(u) \right] \le \|\Delta \vartheta_t(u)\|_2 \le \gamma \left[b_t \|h(u)\|_{G_0} + e_t(u) \right]. \tag{14}$$

If additionally $\|E_0 u\|_2 \le \delta_t \|h(u)\|_{G_0}$ and $\|u\|_2 \le \kappa_t \|h(u)\|_{G_0}$, with $0 \le \delta_t < a_t$, let $c_t = C_t \kappa_t^2$. Whenever $c_t \|h(u)\|_{G_0} \le (a_t - \delta_t)/2$,

$$\frac{\gamma(a_t - \delta_t)}{2} \|h(u)\|_{G_0} \le \|\Delta \vartheta_t(u)\|_2 \le \gamma \left[b_t + \delta_t + \frac{a_t - \delta_t}{2} \right] \|h(u)\|_{G_0}. \tag{15}$$

Thus, controlling the residual and higher-order terms gives two-sided bounds on update magnitude in terms of moment displacement.

B Complete Experimental Results

This appendix presents the full benchmark comparisons, component ablations, optimization-cost analysis, and post-poisoning diagnostics. CIFAR-C evaluations cover all 15 corruptions within each protocol. ImageNet-C results cover four subsets, grouped by data source.

B.1 Experimental Setup

We evaluate CIFAR-10-C, CIFAR-100-C, and ImageNet-C (Hendrycks & Dietterich, 2019). The main CIFAR-C experiments follow the RTTDP pipeline with TENT adaptation, poison ratio 0.5, and batch size 200. Each corruption uses 400 examples, including 200 benign samples for evaluation. We evaluate all 15 corruptions at severities 5 and 3. The ImageNet-C contrast and defocus-blur experiments use an ImageNet-pretrained ResNet-50.

The baselines are DIA (Wu et al., 2023), RTTDP NHE, and RTTDP NHE+DR (Su et al., 2025). Comparisons match the perturbation budget ϵ, step size α, poison ratio, seed, corruption, and TTA configuration. The default setting uses $\epsilon = 0.3$ and $\alpha = 0.01$; the low-budget setting uses $\epsilon = 8/255$ and $\alpha = 0.005$. For non-default budgets, we implement the released NHE and distribution-regularization objectives in a caller-controlled PGD wrapper, replacing the fixed image budget in the released NHE+DR estimate attack.

MTP uses 10 projected steps unless stated otherwise, and the baselines use 40. Additional evaluations examine BNTest and SAR adaptation. Throughout, higher benign classification error indicates stronger poisoning.

B.2 Main CIFAR-C Results

Table 6 summarizes performance across datasets, severities, perturbation budgets, and adaptation settings. At severity 5 and the default budget, MTP reaches 38.40% benign error on CIFAR-10-C

and 49.85% on CIFAR-100-C, exceeding RTTDP NHE, the baseline with the highest overall mean error, by 8.02 and 5.47 percentage points. At severity 3, the corresponding errors are 28.18% and 40.68%. At $\epsilon = 8/255$, MTP exceeds DIA by 1.88 points on CIFAR-10-C and 1.12 points on CIFAR-100-C. Under BNTest, these margins are 2.20 and 1.13 points.

Table 6: Mean benign error (%) across all 15 CIFAR-C corruptions in each setting. MTP-40 uses 40 attack steps.

Dataset	Setting	Method	Runs	Mean error
CIFAR-10-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	MTP	15	38.40
CIFAR-10-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	MTP-40	15	40.05
CIFAR-10-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE	15	30.38
CIFAR-10-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	DIA	15	26.96
CIFAR-10-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE+DR	15	25.63
CIFAR-100-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	MTP	15	49.85
CIFAR-100-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	MTP-40	15	50.17
CIFAR-100-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE	15	44.38
CIFAR-100-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	DIA	15	41.36
CIFAR-100-C	sev. 5, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE+DR	15	36.72
CIFAR-10-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	MTP	15	28.18
CIFAR-10-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE	15	20.80
CIFAR-10-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	DIA	15	20.12
CIFAR-10-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE+DR	15	17.73
CIFAR-100-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	MTP	15	40.68
CIFAR-100-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE	15	38.06
CIFAR-100-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	DIA	15	35.52
CIFAR-100-C	sev. 3, $\epsilon = 0.3$, $\alpha = 0.01$	RTTDP NHE+DR	15	31.51
CIFAR-10-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	MTP	15	26.72
CIFAR-10-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	DIA	15	24.84
CIFAR-10-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	RTTDP NHE	15	21.99
CIFAR-10-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	RTTDP NHE+DR	15	21.51
CIFAR-100-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	MTP	15	38.85
CIFAR-100-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	DIA	15	37.73
CIFAR-100-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	RTTDP NHE	15	35.11
CIFAR-100-C	sev. 5, $\epsilon = 8/255$, $\alpha = 0.005$	RTTDP NHE+DR	15	34.00
CIFAR-10-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	MTP	15	27.31
CIFAR-10-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	DIA	15	25.11
CIFAR-100-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	MTP	15	40.35
CIFAR-100-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	DIA	15	39.22

Table 7 selects the strongest baseline independently within each matched pair. MTP has a positive mean margin in every setting and wins in 12–15 of the 15 pairs. At the default budget, its paired margins are 6.90 and 6.21 points on CIFAR-10-C at severities 5 and 3, and 1.51 and 1.88 points on CIFAR-100-C.

B.3 Component and Protocol Diagnostics

We first isolate the contributions of moment transport and NHE, then examine sensitivity to attack steps and poison ratio. These experiments measure benign co-batch error; subsequent sections assess recovery and held-out loss.

Component contributions. Table 8 compares moment-only, NHE-only, and joint optimization. Moment transport is the stronger individual objective in all three settings. Combining the objectives further increases error by 0.67 points on default-budget CIFAR-10-C, 1.15 points on default-budget CIFAR-100-C, and 0.41 points on low-budget CIFAR-100-C. All main experiments use the joint objective with both weights equal to one.

Table 7: Paired margins of MTP against the strongest baseline selected independently within each matched pair. Margins are percentage points. CI denotes the half-width of a descriptive 95% confidence interval computed from paired runs using a normal approximation; W/L/T gives wins, losses, and ties.

Dataset	Setting	Pairs	Mean Δ	95% CI	W/L/T
CIFAR-10-C	severity 5, $\epsilon = 0.3$, $\alpha = 0.01$	15	+6.90	1.75	14/1/0
CIFAR-100-C	severity 5, $\epsilon = 0.3$, $\alpha = 0.01$	15	+1.51	1.38	13/2/0
CIFAR-10-C	severity 3, $\epsilon = 0.3$, $\alpha = 0.01$	15	+6.21	1.32	15/0/0
CIFAR-100-C	severity 3, $\epsilon = 0.3$, $\alpha = 0.01$	15	+1.88	0.90	13/2/0
CIFAR-10-C	severity 5, $\epsilon = 8/255$, $\alpha = 0.005$	15	+1.76	1.23	12/3/0
CIFAR-10-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	15	+2.20	1.39	13/2/0
CIFAR-100-C	severity 5, $\epsilon = 8/255$, $\alpha = 0.005$	15	+1.12	1.18	12/3/0
CIFAR-100-C	BNTest, $\epsilon = 8/255$, $\alpha = 0.005$	15	+1.13	1.09	13/2/0

Table 8: Component ablation at severity 5 with TENT and poison ratio 0.5. Each row averages five groups covering all 15 corruptions and three seeds (15 runs across seeds and corruption groups). Default and low-budget settings use $(\epsilon, \alpha) = (0.3, 0.01)$ and $(8/255, 0.005)$, respectively. Component variants use 10 steps; DIA uses 40.

Dataset	Variant	Runs	Mean error
Default budget: $\epsilon = 0.3$, $\alpha = 0.01$			
CIFAR-10-C	MTP: NHE + moment	15	38.40
CIFAR-10-C	Moment-only	15	37.73
CIFAR-10-C	NHE-only	15	31.94
CIFAR-100-C	MTP: NHE + moment	15	49.85
CIFAR-100-C	Moment-only	15	48.70
CIFAR-100-C	NHE-only	15	45.22
Low budget: $\epsilon = 8/255$, $\alpha = 0.005$			
CIFAR-100-C	MTP: NHE + moment	15	38.85
CIFAR-100-C	Moment-only	15	38.44
CIFAR-100-C	DIA	15	37.73
CIFAR-100-C	NHE-only	15	34.83

Optimization steps. Increasing MTP from 10 to 40 steps raises default-budget error from 38.40% to 40.05% on CIFAR-10-C and from 49.85% to 50.17% on CIFAR-100-C. At 40 steps, matching the baseline step count, its margins against the strongest baseline within each pair are 8.55 and 4.77 points (Table 9).

Table 9: Attack-step and poison-ratio comparisons at severity 5 with $\epsilon = 0.3$ and $\alpha = 0.01$. Each row covers all 15 corruptions and three seeds, aggregated into 15 matched pairs across seeds and corruption groups. The reference is the strongest baseline within each pair for 40-step comparisons and DIA for poison-ratio comparisons. Ratio experiments use 10-step MTP and 40-step DIA on CIFAR-10-C. Margins are computed before rounding

Diagnostic	Scope	Pairs	MTP mean	Reference mean	Margin
40-step MTP on CIFAR-10-C	all corruptions	15	40.05	31.50	+8.55
40-step MTP on CIFAR-100-C	all corruptions	15	50.17	45.40	+4.77
$\rho = 0.25$	all corruptions	15	25.84	21.66	+4.18
$\rho = 0.50$	all corruptions	15	38.40	26.96	+11.44
$\rho = 0.75$	all corruptions	15	54.49	38.47	+16.02

Poison ratio. On CIFAR-10-C, MTP exceeds DIA at every tested ratio, with margins increasing from 4.18 points at $\rho = 0.25$ to 16.02 points at $\rho = 0.75$. Each ratio yields 14/1/0 wins/losses/ties across the 15 matched pairs. Each corruption contains 400 total examples, so the three ratios leave 300, 200, and 100 benign examples for evaluation. Methods share the same inputs and protocol

within each ratio. The sweep measures attack effectiveness under changing batch compositions and complements the moment-transport identities in Proposition 2.

B.4 Five-Seed Extensions of the Main Comparisons

We extend both perturbation budgets to five seeds on both CIFAR-C datasets. Each seed covers all 15 corruptions at severity 5 using TENT and five corruption groups, with continual adaptation within each group. Batch size is 200, poison ratio is 0.5, and each corruption contains 400 examples, including 200 evaluated benign samples. MTP uses 10 steps and the baselines use 40. Tables report means and sample standard deviations across errors computed separately for each seed.

CIFAR-10-C at the low budget. Table 10 uses $\epsilon = 8/255$ and $\alpha = 0.005$. Seeds 1–3 come from the main evaluation; seeds 4–5 use the evaluator with additional logging with matching hashes of input batches across methods. Across five seeds, MTP exceeds DIA, RTTDP NHE, and RTTDP NHE+DR by 1.21, 4.49, and 4.98 points in mean benign error. It leads DIA in four seeds and trails by 0.47 points in seed 4, while exceeding both RTTDP variants in every seed. Its mean margin over DIA is 1.21 points, compared with 1.88 points in the three-seed evaluation.

Table 10: Five-seed CIFAR-10-C evaluation at $\epsilon = 8/255$ and $\alpha = 0.005$. Entries are benign error percentages across all 15 corruptions. The final column reports the seed mean and sample standard deviation.

Method	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ SD
MTP	27.23	26.00	26.93	26.37	26.90	26.69 ± 0.49
DIA	24.77	24.83	24.93	26.83	26.00	25.47 ± 0.91
RTTDP NHE	21.23	22.64	22.10	21.27	23.73	22.19 ± 1.04
RTTDP NHE+DR	21.10	21.80	21.63	20.70	23.30	21.71 ± 0.99

CIFAR-10-C at the default budget. Table 11 reports results for $\epsilon = 0.3$ and $\alpha = 0.01$. Seeds 1–3 come from the main evaluation. The 40 additional runs for seeds 4 and 5 use matching input-batch hashes across methods within each seed and group, and their perturbations satisfy the prescribed bound within numerical tolerance. MTP leads every baseline in every seed, with mean margins of 10.65, 7.79, and 12.44 points over DIA, RTTDP NHE, and RTTDP NHE+DR. RTTDP NHE is the strongest baseline in all five seeds; the margin over it ranges from 6.90 to 9.03 points. The default-budget protocol yields a larger margin than the low-budget protocol, with both the perturbation budget and step size changing between settings.

Table 11: Five-seed CIFAR-10-C evaluation at $\epsilon = 0.3$ and $\alpha = 0.01$. Entries are benign error percentages across all 15 corruptions. The final column reports the seed mean and sample standard deviation.

Method	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ SD
MTP	37.90	37.87	39.43	38.87	38.00	38.41 ± 0.70
DIA	26.07	27.04	27.76	29.17	28.80	27.77 ± 1.27
RTTDP NHE	28.87	30.77	31.50	30.90	31.10	30.63 ± 1.02
RTTDP NHE+DR	25.17	25.70	26.03	25.67	27.30	25.97 ± 0.80

CIFAR-100-C at both budgets. Table 12 extends both CIFAR-100-C settings to five seeds, using $\alpha = 0.005$ at $\epsilon = 8/255$ and $\alpha = 0.01$ at $\epsilon = 0.3$. At the default budget, MTP exceeds DIA, RTTDP NHE, and RTTDP NHE+DR by 5.26, 2.57, and 10.19 points. At the low budget, the corresponding margins are 0.90, 3.63, and 4.54 points. Figure 4 summarizes the positive mean margins across all four combinations of dataset and perturbation budget.

Table 12: Five-seed CIFAR-100-C evaluation under TENT at severity 5. Entries are mean benign error (%) across all 15 corruptions, reported as seed mean ± sample standard deviation.

Method	$\epsilon = 8/255$	$\epsilon = 0.3$
MTP	39.15 ± 3.02	47.12 ± 2.72
DIA	38.25 ± 3.75	41.86 ± 3.92
RTTDP NHE	35.52 ± 3.32	44.55 ± 3.14
RTTDP NHE+DR	34.61 ± 3.35	36.93 ± 3.37

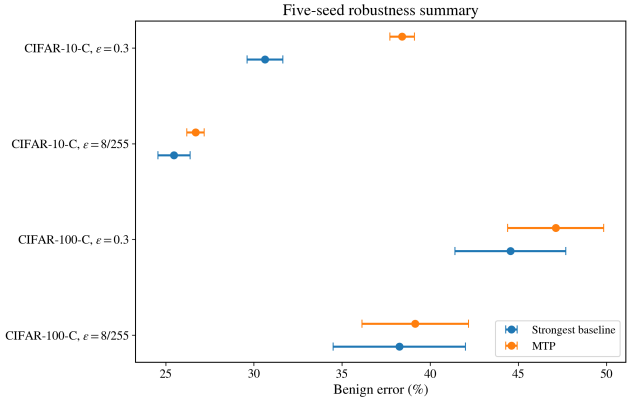


Figure 4: Five-seed comparisons at severity 5. Points show mean benign error and horizontal bars show one sample standard deviation across seeds. Each setting includes the baseline with the highest mean error.

B.5 Extended Evaluation

We further evaluate recovery after poisoning and examine the relationship between objective weights, adaptation updates, and held-out loss. Input-batch hashes match across methods within each protocol, and recorded perturbations satisfy the prescribed bounds within numerical tolerance.

Benign-stream recovery after poisoning. Recovery experiments measure whether degradation persists after poison injection ends. We use three matched seeds on Gaussian noise, defocus blur, and contrast at severity 5 for both CIFAR-C datasets, with TENT, $\epsilon = 8/255$, and $\alpha = 0.005$. Each corruption consists of two warmup batches, two mixed poison batches, and eight recovery batches. Mixed batches contain 200 images at poison ratio 0.5; warmup and recovery batches retain only the 100 benign images. Model state carries through the three-corruption sequence. The no-attack control uses unmodified attacker-controlled images during the mixed phase. Recovery inputs are identical across methods, with 2,400 recovery images per run. Table 13 reports mean recovery error and paired changes relative to no attack. MTP increases recovery error by 3.64 points on CIFAR-10-C and 4.24 points on CIFAR-100-C. RTTDP NHE produces the largest baseline increases, at 0.25 and 0.64 points. The elevated recovery error shows that MTP continues to affect benign predictions after poisoned inputs leave the stream.

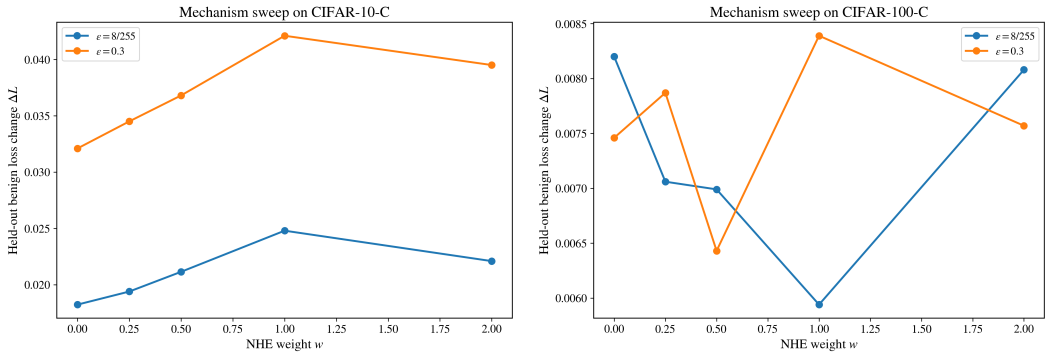
Table 13: Recovery after poisoning over three seeds. Error (%) and paired changes relative to no attack (pp) are averaged across seeds. Positive changes indicate persistent degradation.

Method	C10 error	C10 Δ	C100 error	C100 Δ
No attack	16.86	0.00	32.56	0.00
Random perturbation	16.89	0.03	32.51	-0.04
MTP	20.50	+3.64	36.80	+4.24
DIA	16.97	0.11	32.93	0.38
RTTDP NHE	17.11	0.25	33.19	0.64
RTTDP NHE+DR	16.97	0.11	33.08	0.53

Objective weights, state changes, and held-out loss. The weight sweep measures how combining NHE and moment transport changes moments, adaptation updates, and held-out benign loss. It covers both CIFAR datasets, both budgets, three seeds, and NHE weights $w \in \{0, 0.25, 0.5, 1, 2\}$, with moment weight one and 10 attack steps, for 60 runs. The corruptions are Gaussian noise, defocus blur, and contrast at severity 5.

Benign images from the first batch form a held-out probe excluded from adaptation and attack optimization. Each of two subsequent batches starts paired clean and poisoned updates from identical model and optimizer states; the reference model continues adapting on clean inputs between interventions. Each reported setting averages 18 interventions nested within three seeds.

Held-out loss changes are positive at every tested weight on both datasets (Figure 5), reaching $+0.04210$ on CIFAR-10-C and $+0.00839$ on CIFAR-100-C. The larger budget increases moment displacement at every weight, while held-out loss varies nonmonotonically with displacement. This pattern reflects the distinction between update magnitude and harmful alignment in Theorem 4. At $w = 1$, the mean cosine similarity between the objective gradients is positive in all four combinations of dataset and perturbation budget, indicating average alignment between the objectives at the weight used in the main experiments.



(a) CIFAR-10-C (b) CIFAR-100-C

Figure 5: Held-out benign loss change as the NHE weight varies over $\{0, 0.25, 0.5, 1, 2\}$, with moment weight fixed at one. Curves correspond to $\epsilon = 8/255$ and $\epsilon = 0.3$.

Tables 14 and 15 report the corresponding moment displacement, displacement of the adaptation update, held-out loss change, and gradient alignment. As the NHE weight increases, D_m generally decreases on both datasets and at both perturbation budgets, reflecting the local trade-off between the moment and NHE directions characterized in Proposition 3. In contrast, D_θ changes only moderately across the weight sweep, and the held-out loss change remains positive in every setting. The largest moment displacement therefore does not necessarily produce the largest benign loss increase. At the default choice $w_{\mathrm{NHE}} = 1$, the two objective gradients have positive mean cosine in all four combinations of dataset and perturbation budget, while the held-out loss remains elevated. These results support the distinction in Theorem 4 between the magnitude of the induced displacement and its alignment with the benign-risk gradient.

C Proofs of the Theoretical Results

C.1 Proof of Proposition 1

For $s = \log(v + \tau)$, the directional derivative of $\mathcal{N}_\tau(z; \mu, s) = e^{-s/2}(z - \mu)$ with respect to (μ, s) in direction $h = (a, b)$ is $D\mathcal{N}_\tau[h] = -e^{-s/2}a - \frac{1}{2}e^{-s/2}(Z - \mu)b$. Using $\mathbb{E}[Z - \mu] = 0$ and $\mathbb{E}[(Z - \mu)^2] = v$ gives

$$\mathbb{E}\big[(D\mathcal{N}_\tau[h])^2\big] = e^{-s}a^2 + \frac{1}{4}e^{-s}vb^2 = \frac{a^2}{v + \tau} + \frac{v}{4(v + \tau)}b^2, \tag{16}$$

which gives the expression for $Q_{\mathrm{BN}}(h)$. The second-order Taylor expansion is

$$\mathcal{N}_\tau(Z; \mu + a, s + b) - \mathcal{N}_\tau(Z; \mu, s) = D\mathcal{N}_\tau[h] + R(Z, h), \tag{17}$$

Table 14: CIFAR-10-C NHE weight sweep. D_m denotes displacement of mixed-batch moments, D_θ update displacement, ΔL held-out loss change, and cosine the alignment between the two objective gradients. Positive ΔL indicates increased benign loss.

Budget / w	D_m	D_θ	ΔL	Cosine
8/255 / 0	0.08276	0.14075	+0.01824	0.00463
8/255 / 0.25	0.08197	0.14073	+0.01940	+0.08821
8/255 / 0.5	0.08029	0.13981	+0.02115	+0.12540
8/255 / 1	0.07632	0.13877	+0.02480	+0.11547
8/255 / 2	0.06993	0.13630	+0.02210	+0.03151
0.3 / 0	0.16729	0.14416	+0.03210	0.03071
0.3 / 0.25	0.16644	0.14402	+0.03450	+0.08957
0.3 / 0.5	0.16394	0.14419	+0.03680	+0.08047
0.3 / 1	0.15796	0.14498	+0.04210	+0.10711
0.3 / 2	0.14981	0.14379	+0.03950	+0.05222

Table 15: CIFAR-100-C objective-weight sweep, using the same measures as Table 14.

Budget / w	D_m	D_θ	ΔL	Cosine
8/255 / 0	0.06998	0.13429	0.00820	0.03282
8/255 / 0.25	0.06968	0.13464	0.00706	+0.08767
8/255 / 0.5	0.06857	0.13491	0.00699	+0.13495
8/255 / 1	0.06519	0.13454	0.00594	+0.09915
8/255 / 2	0.06001	0.13328	0.00808	0.00259
0.3 / 0	0.11716	0.15254	0.00746	0.10515
0.3 / 0.25	0.11679	0.15224	0.00787	+0.07068
0.3 / 0.5	0.11610	0.15206	0.00643	+0.14253
0.3 / 1	0.11395	0.15107	0.00839	+0.15992
0.3 / 2	0.11153	0.15086	0.00757	+0.08448

where $\mathbb{E}[R(Z,h)^2] = O(\|h\|_2^4)$. By Cauchy–Schwarz, $\mathbb{E}[|D\mathcal{N}_\tau[h]R(Z,h)|] = O(\|h\|_2^3)$. Squaring and taking expectations yields Eq. 7. Finally, with $\kappa = v/(v+\tau) \in [0,1]$,

$$Q_{\mathrm{BN}}(h) = \frac{a^2}{v+\tau} + \frac{\kappa}{4}b^2, \qquad Q_{\mathrm{MTP}}(h) = \frac{a^2}{v+\tau} + \frac{1}{4}b^2. \tag{18}$$

Nonnegativity of both terms gives

$$\kappa Q_{\mathrm{MTP}}(h) \le Q_{\mathrm{BN}}(h) \le Q_{\mathrm{MTP}}(h). \tag{19}$$

When $\tau = 0$ and $v > 0$, $\kappa = 1$ and the two forms coincide.

C.2 Proof of Theorem 1

Along the path $P_s = P + s\Delta P$, the fundamental theorem of calculus gives

$$h_\theta(\Delta P) = \mathcal{M}_\theta(P+\Delta P) - \mathcal{M}_\theta(P) = \int_0^1 J_\theta(P_s)\Delta P\, ds. \tag{20}$$

Subtracting the source and adapted expressions,

$$h_{\theta_t}(\Delta P) - h_{\theta_0}(\Delta P) = \int_0^1 \left(J_{\theta_t}(P_s) - J_{\theta_0}(P_s)\right)\Delta P\, ds. \tag{21}$$

By the assumed Jacobian bound,

$$\|h_{\theta_t}(\Delta P) - h_{\theta_0}(\Delta P)\|_2 \le \int_0^1 \|J_{\theta_t}(P_s) - J_{\theta_0}(P_s)\|_{\mathrm{op}}\, \|\Delta P\|_2\, ds \le L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2\|\Delta P\|_2. \tag{22}$$

The reverse triangle inequality gives

$$\|h_{\theta_t}(\Delta P)\|_2 \ge \|h_{\theta_0}(\Delta P)\|_2 - \|h_{\theta_t}(\Delta P) - h_{\theta_0}(\Delta P)\|_2\,. \tag{23}$$

For $\Delta P \in \mathcal{U}$, the assumed lower bound on source moment displacement $\|h_{\theta_0}(\Delta P)\|_2 \geq \sigma_{\mathrm{src}}\|\Delta P\|_2$ implies

$$\|h_{\theta_t}(\Delta P)\|_2 \geq (\sigma_{\mathrm{src}} - L_{\mathrm{tr}}\|\theta_t - \theta_0\|_2)\,\|\Delta P\|_2. \tag{24}$$

Under the stated drift condition, the coefficient is positive, so every nonzero perturbation in $\mathcal{U}$ retains nonzero displacement.

C.3 Proof of Proposition 2

For $M_i = (1-\rho)B + \rho P_i$, total expectation gives $\mu_{M_i} = (1-\rho)\mu_B + \rho\mu_i$. Subtracting the two cases yields $\mu_{M_1} - \mu_{M_0} = \rho\,\delta\mu$. By total variance, $v_{M_i} = (1-\rho)v_B + \rho v_i + \rho(1-\rho)(\mu_B - \mu_i)^2$.

Substituting $\mu_1 = \mu_0 + \delta\mu$ and $v_1 = v_0 + \delta v$ gives

$$v_{M_1} - v_{M_0} = \rho\delta v + \rho(1-\rho)\left[(\mu_B - \mu_0 - \delta\mu)^2 - (\mu_B - \mu_0)^2\right] \tag{25}$$

$$= \rho\delta v + 2\rho(1-\rho)(\mu_0 - \mu_B)\delta\mu + \rho(1-\rho)(\delta\mu)^2. \tag{26}$$

The expansion

$$\log(x + \Delta x) = \log x + \frac{\Delta x}{x} + O((\Delta x)^2) \tag{27}$$

at $x = v_{M_0} + \tau > 0$ gives Eq. 8.

C.4 Proof of Proposition 4

Applying total expectation and total variance within each context gives

$$\mu_{M_i} = (1-\rho)\mu_{B_i} + \rho\mu_{P_i}, \qquad v_{M_i} = (1-\rho)v_{B_i} + \rho v_{P_i} + \rho(1-\rho)(\mu_{P_i} - \mu_{B_i})^2. \tag{28}$$

Subtracting the mean identities proves Eq. 9. The difference between benign and poison means changes from d_0 to $d_0 + a_P - a_B$, so $(d_0 + a_P - a_B)^2 - d_0^2 = 2d_0(a_P - a_B) + (a_P - a_B)^2$. Substitution into the variance difference gives Eq. 10. Subtracting its fixed-benign case, $a_B = b_B = 0$, leaves $(1-\rho)b_B + \rho(1-\rho)\left[-2(d_0 + a_P)a_B + a_B^2\right]$, which proves Eq. 11.

C.5 Proof of Theorem 2

Let $U = Z_0 - \mu_B$, so $\mathbb{E}U = 0$ and $\mathbb{E}U^2 = v_B$. Using $r = e^{-b/2}$, direct subtraction gives

$$\mathcal{N}_\tau(Z_0; \mu_1, s_1) - \mathcal{N}_\tau(Z_0; \mu_0, s_0) = e^{-s_0/2}\left[(r-1)U + (r-1)(\mu_B - \mu_0) - ra\right]. \tag{29}$$

Squaring and taking expectations eliminates the cross term in U, yielding

$$D_{\mathrm{fix}} = e^{-s_0}\left[(r-1)^2 v_B + \big((r-1)(\mu_B - \mu_0) - ra\big)^2\right]. \tag{30}$$

This proves Eq. 3. The actual difference in normalized features adds $e^{-s_1/2}(Z_1 - Z_0)$ to the fixed-activation difference. Its L_2 norm is ζ. The triangle and reverse-triangle inequalities therefore give

$$\max\{0, \sqrt{D_{\mathrm{fix}}} - \zeta\} \leq \sqrt{D_{\mathrm{act}}} \leq \sqrt{D_{\mathrm{fix}}} + \zeta. \tag{31}$$

Squaring proves Eq. 4. When $\mu_B = \mu_0$ and $v_B = v_0$, $D_{\mathrm{fix}} = e^{-s_0}\left[(r-1)^2 v_0 + r^2 a^2\right]$. Expanding $r = 1 - b/2 + O(b^2)$ gives

$$D_{\mathrm{fix}} = \frac{a^2}{v_0 + \tau} + \frac{v_0 b^2}{4(v_0 + \tau)} + O(\|(a,b)\|_2^3). \tag{32}$$

For $Z_1 = Z_0$, this also equals D_{act}, establishing the local statement.

C.6 Proof of Theorem 3

Let $\Pi_0 = J_0^\dagger J_0$, the orthogonal projector onto the row space of J_0. Then

$$K_0 = S_0 J_0 + E_0, \qquad E_0 = K_0(I - \Pi_0), \tag{33}$$

and hence $\|K_0 - SJ_0\|_{\mathrm{op}} \geq \|E_0\|_{\mathrm{op}}$. Equality holds at $S = S_0$, proving Eq. 12. An exact factorization requires K_0 to vanish on $\ker J_0$. Conversely, this kernel condition implies $K_0(I - \Pi_0) = 0$, giving $K_0 = S_0 J_0$.

Lipschitz continuity of the Jacobians yields

$$\begin{aligned} h(u) &= J_0 u + r_m(u), \quad \|r_m(u)\|_2 \leq \tfrac{1}{2} L_m \|u\|_2^2, \\ g_t(u) - g_t(0) &= K_0 u + r_g(u), \quad \|r_g(u)\|_2 \leq \tfrac{1}{2} L_g \|u\|_2^2. \end{aligned} \tag{34}$$

Substituting $K_0 = S_0 J_0 + E_0$ gives $g_t(u) - g_t(0) = S_0 h(u) + E_0 u + r_g(u) - S_0 r_m(u)$. Set $r_t(u) = r_g(u) - S_0 r_m(u)$ and $C_t = (L_g + \|S_0\|_{\mathrm{op}} L_m)/2$. Then $\|r_t(u)\|_2 \leq C_t \|u\|_2^2$. Multiplying by $-\gamma$ proves Eq. 5.

The triangle and reverse-triangle inequalities, together with Eq. 13, give Eq. 14. Under the additional residual and perturbation bounds, write $H = \|h(u)\|_{G_0}$. Since $e_t(u) \leq \delta_t H + c_t H^2$,

$$\gamma \left[(a_t - \delta_t) H - c_t H^2 \right] \leq \|\Delta \vartheta_t(u)\|_2 \leq \gamma \left[(b_t + \delta_t) H + c_t H^2 \right]. \tag{35}$$

Applying $c_t H \leq (a_t - \delta_t)/2$ proves Eq. 15, including $c_t = 0$.

C.7 Proof of Theorem 4

Write $\theta_c = \vartheta_{t+1}^c$ and $\Delta = \Delta \vartheta_t$. Lipschitz continuity of ∇R_t gives

$$\left| R_t(\theta_c + \Delta) - R_t(\theta_c) - \nabla R_t(\theta_c)^\top \Delta \right| \leq \frac{L_R}{2} \|\Delta\|_2^2. \tag{36}$$

For nonzero gradient and displacement, the definition of χ_t gives $\nabla R_t(\theta_c)^\top \Delta = \chi_t \|\nabla R_t(\theta_c)\|_2 \|\Delta\|_2$. Thus,

$$R_t(\theta_c + \Delta) - R_t(\theta_c) \geq \chi_t \|\nabla R_t(\theta_c)\|_2 \|\Delta\|_2 - \frac{L_R}{2} \|\Delta\|_2^2. \tag{37}$$

The right-hand side is positive when $\chi_t > L_R \|\Delta\|_2 / (2 \|\nabla R_t(\theta_c)\|_2)$, establishing the claimed risk increase. Applying the upper remainder bound similarly gives a strict risk decrease when $\chi_t < -L_R \|\Delta\|_2 / (2 \|\nabla R_t(\theta_c)\|_2)$.

C.8 Proof of Corollary 1

By Theorem 3,

$$\Delta \vartheta_t = -\gamma (S_0 h + v), \qquad v = E_0 u + r_t(u), \qquad \|v\|_2 \leq e. \tag{38}$$

The lower bound from the smoothness of R_t

$$R_t(\vartheta_{t+1}^p) - R_t(\vartheta_{t+1}^c) \geq -\gamma r_c^\top (S_0 h + v) - \frac{L_R \gamma^2}{2} \|S_0 h + v\|_2^2. \tag{39}$$

Using $r_c^\top v \leq \|r_c\|_2 e$ and $\|S_0 h + v\|_2 \leq \|S_0 h\|_2 + e$ yields

$$R_t(\vartheta_{t+1}^p) - R_t(\vartheta_{t+1}^c) \geq -\gamma r_c^\top S_0 h - \gamma \|r_c\|_2 e - \frac{L_R \gamma^2}{2} \left(\|S_0 h\|_2 + e \right)^2. \tag{40}$$

This lower bound is positive when $-r_c^\top S_0 h > \|r_c\|_2 e + \frac{L_R \gamma}{2} \left(\|S_0 h\|_2 + e \right)^2$, establishing the claimed risk increase.

C.9 Proof of Proposition 3

Write the feasible box as $\mathcal{C} = \prod_j [\ell_j, u_j]$. At $z \in \mathcal{C}$, the feasible displacement interval for coordinate j is $I_j = [\max\{-\alpha, \ell_j - z_j\}, \min\{\alpha, u_j - z_j\}]$. The moment-only displacement $d_{M,j}$ maximizes $g_{M,j} d_j$ over I_j. Since $d_{J,j} \in I_j$, summing over coordinates gives $D_{\mathrm{step}} = g_M^\top (d_M - d_J) \geq 0$. Without clipping, a sign reversal changes $\alpha \operatorname{sign}(g_{M,j})$ to its negative, losing $2\alpha |g_{M,j}|$ in first-order gain. Cancellation to zero loses $\alpha |g_{M,j}|$; all other coordinates contribute zero. Positive per-image

normalization and $w_{\mathrm{mom}} > 0$ preserve the sign of $g_{M,j}$ in a_j. For $a_j \neq 0$, reversal therefore occurs exactly when $a_j b_j < 0$ and $|b_j| > |a_j|$. Finally, Lipschitz continuity of ∇F gives

$$F(z+d) = F(z) + g_M^\top d + r(d), \qquad |r(d)| \leq \frac{L_F}{2}\|d\|_2^2. \tag{41}$$

Applying this expansion to d_J and d_M yields

$$F(z+d_J) - F(z+d_M) + D_{\mathrm{step}} = r(d_J) - r(d_M). \tag{42}$$

Taking absolute values and bounding both remainders establishes the finite-step bound. Applying the lower and upper Taylor bounds to d_J alone gives the stated sufficient conditions for an increase and a decrease in F.