FARSIGHTED CORRECTION OF SHORTCUT LEARNING: FROM DYNAMIC STARVATION TO CONFIDENCE-GAP OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

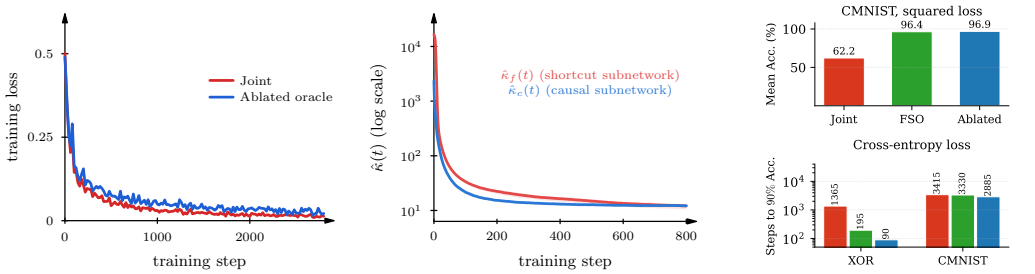
Deep neural networks frequently rely on easily accessible spurious correlations, driving a phenomenon known as shortcut learning. Remarkably, this reliance remains invisible on standard loss curves, which decrease smoothly as models build brittle representations. In this work, we uncover the mathematical root of this invisibility by analyzing optimization dynamics via the empirical Neural Tangent Kernel (eNTK) along the training trajectory. We establish the non-asymptotic Dynamic Starvation Theorem, proving that a dominant shortcut eigenvalue restricts causal parameter updates to a bounded regime $O(1/\sigma_f)$ in the worst case of unbounded shortcut dominance. We further characterize how this starvation mechanism manifests under different loss functions: under **squared loss**, joint and ablated training yield identical predictions in output space, trapping the causal subnetwork in a degenerate null space as loss vanishes; under **cross-entropy loss**, self-correction is merely logarithmic. To resolve these failure modes, we propose **Farsighted Signal Optimization** (FSO), which reinjects counterfactual causal gradients into the training objective. We extend this to model-agnostic settings via GAP-FSO, proving that inter-group confidence gaps serve as exact proxies for shortcut parameters without explicit architectural splits. Unlike prior theoretical treatments of shortcut learning, which validate their predictions on simple architectures, we test GAP-FSO on ResNet-50 and DistilBERT across six vision and language benchmarks. Results are competitive with established baselines overall and clearly outperform them on the NLP benchmark.

1 INTRODUCTION

Deep neural networks frequently achieve high training accuracy by relying on non-causal, spurious correlations, a phenomenon known as *shortcut learning* (Geirhos et al., 2020). For instance, a classifier trained to recognize cows often relies on green grass backgrounds rather than their physical features. Standard diagnostic tools typically identify these shortcuts only post-hoc via explicit visual saliency maps (Selvaraju et al., 2017; Smilkov et al., 2017; Petsiuk et al., 2018) or representation probes (Singla & Feizi, 2022; FEL et al., 2023). Crucially, the primary diagnostic monitored by every practitioner, the training loss curve, remains entirely uninformative, decreasing smoothly even as the model constructs brittle, shortcut-dependent representations (Fig. 1a). Existing theoretical accounts offer limited insight into this training-time behavior. In particular, analyses based on neural tangent kernel (NTK) assume a lazy-training regime that freezes the kernel at initialization, reducing these accounts to static snapshots rather than capturing how shortcut reliance dynamically emerges during training (Pezeshki et al., 2021; Hermann et al., 2024; Bombari & Mondelli, 2024; Lim et al., 2025). While recent work bypasses this limitation using Boolean-function complexity (Qiu et al., 2024) or evolutionary game theory (Li et al., 2026), both validate their models only on simple architectures (e.g., MLPs), leaving open whether their dynamics extend to scaled up models.

In this work, we analyze the dynamic trajectory of training through the empirical Neural Tangent Kernel (eNTK) $\boldsymbol{K}_{SS}(t) = \boldsymbol{J}_S(t)\boldsymbol{J}_S(t)^\top$ (Jacot et al., 2018), evaluated continuously along the realized optimization path. This dynamic perspective allows us to characterize shortcut learning as an active race, which we call *joint training*, between a causal subnetwork and a shortcut subnetwork competing for the same prediction error. Specifically, we formalize this mechanism in the **Dynamic**

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.



(a) Training Loss Curve (b) Rayleigh Activity (c) Toy-Benchmark Validation

Figure 1: **The loss curve conceals feature starvation, whereas the realized eNTK exposes it, and FSO/GAP-FSO recover the oracle on toy benchmarks. (a)–(b)** Evaluations on Colored MNIST under squared loss over the same training horizon. **(a)** Training loss trajectories for joint (starved) and ablated (oracle) models remain visually identical, despite the oracle achieving superior OOD accuracy (97% vs. 66%). **(b)** Continuous evolution of Rayleigh activities $\hat{\kappa}_f(t)$ and $\hat{\kappa}_c(t)$ under joint training captures active feature competition. Their later plateau stems from the near-complete dissipation of the shared error signal $\boldsymbol{g}(t)$ rather than a cessation of subnetwork competition. (Appendix F.2) **(c)** FSO demonstrates clear advantages on two key benchmarks: on XOR (LaBonte & Muthukumar, 2026), all methods reach 100% accuracy, but FSO significantly accelerates convergence; on Colored MNIST (CMNIST) (Arjovsky et al., 2019), while convergence speeds are comparable, FSO dramatically improves out-of-distribution (OOD) accuracy.

Starvation Theorem: when the shortcut subnetwork exhibits a persistently larger tangent-kernel eigenvalue σ_f^2, it rapidly absorbs the error signal, suppressing the causal subnetwork's parameter updates to a bounded regime that is $O(1/\sigma_f)$ in the worst case of unbounded shortcut dominance. Under squared loss, both joint training and *shortcut-ablated training* (a counterfactual condition that trains only the causal subnetwork) converge to identical outputs on the training set. Hence, the training loss reaches zero while the causal subnetwork under joint training remains permanently undertrained. Under cross-entropy (CE) loss, this suppression is not permanent, but self-correction progresses logarithmically, requiring an impractical compute budget to recover causal features. (Fig. 2)

To resolve these dynamic failure modes, we introduce **Farsighted Signal Optimization (FSO)**. At each optimization step, FSO computes the counterfactual gradient that the causal subnetwork would have received in the absence of the shortcut, injecting this missing signal directly into its parameter update. FSO mathematically prevents the two subnetworks from settling into a shared gradient under squared loss and sharply accelerates convergence under CE loss (Figure 1c), matching the performance of an unstarved causal oracle (where no shortcut feature is present), as Fig. 2a illustrates.

Despite FSO's theoretical promises, real-world architectures fail to explicitly segregate causal and shortcut pathways. Nevertheless, we prove that this separation is unnecessary when we have a group label (indicating alignment between the target label and candidate shortcut cue) for each example: the inter-group prediction gap serves as an exact output-level proxy for the internal shortcut parameter. Building on this equivalence, we propose **GAP-FSO**, an algorithm that suppresses shortcut reliance through a targeted output-space penalty without architectural modifications.

To bridge our linear theoretical derivations with realistic deep architectures, we evaluate GAP-FSO across six vision and language benchmarks using ResNet-50 (He et al., 2016) and DistilBERT (Sanh et al., 2019), empirically confirming that the dynamics predicted by our theory persist under deep neural networks. While baselines calibrated for a specific loss or modality fluctuate sharply outside that regime (e.g., IRM's CelebA accuracy swings 11.4% between loss functions, and DFR's accuracy drops 5.2% moving from vision to text datasets), GAP-FSO achieves 87.8% overall out-of-distribution (OOD) performance with markedly consistent results across loss functions and modalities. Our main contributions are summarized below:

- We pioneer the theoretical analysis of shortcut learning via trajectory-realized eNTK dynamics, establishing the Dynamic Starvation Theorem to explain why standard loss curves mask shortcut reliance under the two most common loss functions, squared and cross-entropy loss.

- We introduce FSO, a gradient-injection framework that restores causal learning signals to achieve oracle convergence under squared loss while accelerating cross-entropy dynamics. Building on this, GAP-FSO uses inter-group confidence gaps as exact proxies for shortcut parameters, eliminating physical subnetwork splits.
- To the best of our knowledge, we are the first to empirically validate dynamic theory on real-world transformer and convolutional architectures for classification: performance is consistent and competitive with established baselines overall, and clearly ahead of them on the NLP benchmark.

The remainder of this paper is structured as follows: Section 2 provides preliminaries, Section 3 presents the Dynamic Starvation Theorem, Sections 4 and 5 introduce FSO and GAP-FSO, Section 6 describes our experimental design, and Sections 7 and 8 offers discussion and conclusion.

2 PRELIMINARIES

2.1 DYNAMICS VIA THE TRAJECTORY-REALIZED eNTK

To model feature learning beyond the frozen-kernel limit (Chizat et al., 2019; Yu et al., 2025), we track the empirical Neural Tangent Kernel (eNTK) continuously along the realized training trajectory. Let $\boldsymbol{u}(\boldsymbol{w}) \in \mathbb{R}^{nm}$ denote the model predictions across n training examples (m outputs each) given parameters $\boldsymbol{w}$. Then $\boldsymbol{K}_{SS}(t) \triangleq \boldsymbol{K}_{SS}(\boldsymbol{w}(t)) = \boldsymbol{J}_S(t)\boldsymbol{J}_S(t)^\top$, where $\boldsymbol{w}(t)$ denotes parameter state at time t and $\boldsymbol{J}_S(\boldsymbol{w}) \triangleq D_{\boldsymbol{w}}\boldsymbol{u}(\boldsymbol{w})$ is the output Jacobian (Jacot et al., 2018). For a convex C^2 loss Φ_S with empirical risk $L_S = \Phi_S \circ \boldsymbol{u}$, let $\boldsymbol{g}(t) \triangleq \nabla_{\boldsymbol{u}}\Phi_S(\boldsymbol{u}(t))$ be the output error signal and $\boldsymbol{B}(t) \triangleq \nabla^2\Phi_S(\boldsymbol{u}(t))$ be the loss Hessian. Under gradient flow, the output-space dynamics follow

$$\partial_t \boldsymbol{u}(t) = -\boldsymbol{K}_{SS}(t)\boldsymbol{g}(t), \quad \partial_t \boldsymbol{g}(t) = -\boldsymbol{B}(t)\boldsymbol{K}_{SS}(t)\boldsymbol{g}(t), \quad \tfrac{d}{dt}\Phi_S(\boldsymbol{u}(t)) = -\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}(t)\boldsymbol{g}(t).$$

While these general dynamics hold exact for arbitrary architectures, they do not explicitly isolate which features drive the error signal's decay. To make shortcut learning legible, we next decompose the network into a shortcut and a causal subnetwork competing for this shared error signal $\boldsymbol{g}(t)$.

2.2 SUBNETWORK DECOMPOSITION AND COMPETITION

For theoretical tractability, we decompose the network parameters into a *candidate shortcut subnetwork* f and a *causal complement* c, such that $\boldsymbol{w} = (\boldsymbol{w}_f, \boldsymbol{w}_c)$. Each subnetwork adopts a scalar readout, $u_f = \sigma_f a_f b_f$ and $u_c = \sigma_c a_c b_c$, where $a_f, b_f, a_c, b_c \in \mathbb{R}$ are trainable bilinear factors, and fixed constants $\sigma_f, \sigma_c > 0$ control the relative strength of the shortcut and causal pathways. These capacity constants serve as theoretical proxies for physical properties (such as width, depth, or initialization scale) that grant one subnetwork a persistently larger tangent-kernel eigenvalue than the other. This formulation yields subnetwork kernels $\boldsymbol{K}_{SS}^{(f)}(t) = \sigma_f^2\big(a_f(t)^2 + b_f(t)^2\big)$ (and analogously for c). Unlike in frozen-kernel regimes, these magnitudes evolve dynamically throughout training. This provides a concrete instance of the general *Rayleigh activity* (Eq. 2), whose continuous trajectory we empirically measure on trained Colored MNIST classification task (Fig. 1b).

Assumption 2.1 (Additive Readout & Isotropic Curvature)**.** We assume: ***(A1) Additive Readout:*** $\boldsymbol{u}(\boldsymbol{w}) = \boldsymbol{u}^{(f)}(\boldsymbol{w}_f) + \boldsymbol{u}^{(c)}(\boldsymbol{w}_c) + \text{const}$, so the kernel decomposes as $\boldsymbol{K}_{SS} = \boldsymbol{J}_S^{(f)}\boldsymbol{J}_S^{(f)\top} + \boldsymbol{J}_S^{(c)}\boldsymbol{J}_S^{(c)\top} =: \boldsymbol{K}_{SS}^{(f)} + \boldsymbol{K}_{SS}^{(c)} \succeq 0$; and ***(A2) Isotropic Curvature:*** $\boldsymbol{B}(t) = \beta(t)\boldsymbol{I}$, which holds naturally under squared loss.

Under Assumption 2.1, both subnetworks compete for the *identical* error signal vector $\boldsymbol{g}(t)$:

$$\dot{\boldsymbol{w}}_f = -\boldsymbol{J}_S^{(f)\top}\boldsymbol{g}(t), \qquad \dot{\boldsymbol{w}}_c = -\boldsymbol{J}_S^{(c)\top}\boldsymbol{g}(t). \tag{1}$$

Here, the capacity constants σ_f and σ_c dictate the relative rates at which each subnetwork dissipates $\boldsymbol{g}(t)$, as $\boldsymbol{J}_S^{(f)} = \sigma_f(b_f, a_f)$ scales linearly with σ_f. Consequently, when the shortcut subnetwork f exhibits a larger capacity ($\sigma_f > \sigma_c$), it rapidly absorbs the error signal, driving it to zero and thereby starving the causal subnetwork c of parameter updates. To isolate this competitive mechanism, we compare the joint flow against a counterfactual *ablated flow* driven solely by the unstarved error signal $\boldsymbol{g}^{(\neg f)}(t) \triangleq \nabla\Phi_S(\boldsymbol{u}^{(c)}(\boldsymbol{w}_c(t)))$, where c is trained in isolation ($\boldsymbol{w}_f \equiv \boldsymbol{0}$).

3 DYNAMIC STARVATION THEOREM

Section 3.1 establishes a general Dynamic Starvation Theorem, while Section 3.2 examines its loss-dependent mechanics under squared and cross-entropy regimes.

3.1 THE STARVATION BOUND

We derive the explicit parameter bound governing causal starvation. By characterizing the rate of error dissipation attributed to each subnetwork, we show that the decay rate of $\boldsymbol{g}(t)$ directly constrains the integral of causal parameter updates. This translates the error signal's effective lifetime into a firm bound on the causal subnetwork's total parameter displacement, regardless of training duration.

According to Eq. 1, both subnetworks' updates are driven by the shared error signal $\boldsymbol{g}(t)$: starvation occurs whenever one subnetwork dissipates this signal so quickly that little cumulative signal remains for the other. To quantify this, we measure each subnetwork's instantaneous *Rayleigh activity*, defined as its error dissipation rate (the Rayleigh quotient of $\boldsymbol{K}_{SS}^{(f)}(t)$ with respect to $\boldsymbol{g}(t)$, evaluated continuously along the optimization trajectory):

$$\hat{\kappa}_f(t) \triangleq \frac{\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(f)}(t)\boldsymbol{g}(t)}{\|\boldsymbol{g}(t)\|^2}, \qquad \hat{\kappa}_c(t) \triangleq \frac{\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(c)}(t)\boldsymbol{g}(t)}{\|\boldsymbol{g}(t)\|^2}. \tag{2}$$

$\hat{\kappa}_f(t)$ and $\hat{\kappa}_c(t)$ represent the exact rate of error signal dissipation driven by the subnetworks at time t, respectively. Under Assumption 2.1, they directly dictate $\boldsymbol{g}(t)$'s own decay rate via $\frac{d}{dt}\|\boldsymbol{g}(t)\|^2 = -2\beta(t)(\hat{\kappa}_f(t) + \hat{\kappa}_c(t))\|\boldsymbol{g}(t)\|^2$, which integrates to the exact decaying envelope

$$\|\boldsymbol{g}(t)\| = \|\boldsymbol{g}(0)\| \exp\Big(-\int_0^t \beta(\tau)\big(\hat{\kappa}_f(\tau) + \hat{\kappa}_c(\tau)\big)\, d\tau\Big). \tag{3}$$

The larger the combined activity $\hat{\kappa}_f(\tau) + \hat{\kappa}_c(\tau)$, the faster $\boldsymbol{g}(t)$ is driven to zero (effectively shortening the error signal's lifetime). Because subnetwork c's parameters move only in response to $\boldsymbol{g}(t)$ (Eq. 1), this decay directly bounds its total displacement:

Theorem 3.1 (Starvation Bound)**.** *Fix an arbitrary horizon $T > 0$, and let $\Delta\boldsymbol{w}_c(0,T) \triangleq \boldsymbol{w}_c(T) - \boldsymbol{w}_c(0)$ denote the causal subnetwork c's net parameter displacement over $[0,T]$. Under Assumption 2.1, suppose both subnetworks exhibit* persistent excitation *on $[0,T]$: $\hat{\kappa}_f(\tau) \geq \kappa_f^{\min}$, $\hat{\kappa}_c(\tau) \geq \kappa_c^{\min}$, $\beta(\tau) \geq \beta_{\min}$, and $\bar{\sigma}_c(\tau) \triangleq \|\boldsymbol{J}_S^{(c)}(\tau)\|_{\mathrm{op}} \leq \bar{\sigma}_c$, for constants $\kappa_f^{\min}, \kappa_c^{\min}, \beta_{\min}, \bar{\sigma}_c > 0$ independent of T. Substituting Eq. 3 into $\Delta\boldsymbol{w}_c(0,T) = -\int_0^T \boldsymbol{J}_S^{(c)}(\tau)^\top \boldsymbol{g}(\tau)\, d\tau$ and applying these bounds, $\Delta\boldsymbol{w}_c(0,T)$ is bounded uniformly in T:*

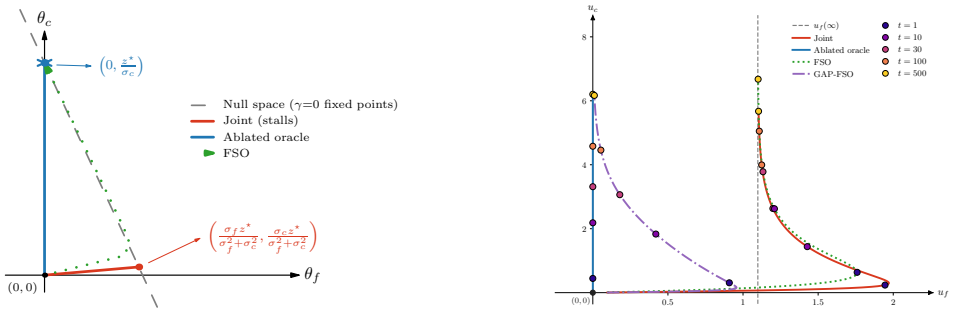
$$\|\Delta\boldsymbol{w}_c(0,T)\| \;\leq\; \bar{\sigma}_c\|\boldsymbol{g}(0)\| \int_0^T \exp\Big(-\int_0^\tau \beta(s)\big(\hat{\kappa}_f(s) + \hat{\kappa}_c(s)\big)ds\Big)d\tau \;\leq\; \frac{\|\boldsymbol{g}(0)\|\,\bar{\sigma}_c}{\beta_{\min}\big(\kappa_f^{\min} + \kappa_c^{\min}\big)}. \tag{4}$$

In words, Eq. 4 says the causal subnetwork's parameters change only negligibly over $[0,T]$ under persistent excitation, no matter how long training continues: the middle (exact) term shows accumulated displacement is capped by the error signal's own decaying envelope (Eq. 3), and the right-hand (uniform) bound shows this cap never grows with T once $\boldsymbol{g}(t)$'s dissipation rate is bounded away from zero (full derivation in Appendix C.1).

The core insight of Eq. 4 resides in its denominator $\kappa_f^{\min} + \kappa_c^{\min}$: a larger guaranteed activity floor for subnetwork f imposes a strictly tighter bound on subnetwork c's displacement, confining any parameter movement to a brief initial phase and making prolonged training virtually ineffective for c. Proposition 3.2 instantiates this concretely on the bilinear bottleneck construction, deriving the same starvation phenomenon identified by Pezeshki et al. (2021) directly from the trajectory-realized eNTK rather than their frozen-kernel approximation.

3.2 LOSS-DEPENDENT FAILURE REGIMES

Despite reaching identical vanishing training loss as the ablated flow, subnetwork c under joint training remains frozen at initialization as $\sigma_f \to \infty$ (Corollary C.1). Standard training and validation



(a) Squared loss

(b) Cross-entropy loss

Figure 2: **Trajectory dynamics under different loss functions**. **(a) Squared loss** (scalar linear-readout model, Eq. 6): Joint training gets trapped in degenerate equilibria, whereas FSO matches the ablated oracle in reaching the causal equilibrium (Proposition 3.2). **(b) CE loss:** Unlike squared loss, u_c grows unboundedly across all methods. Under Joint training and FSO, u_f converges to $\frac{1}{2}\ln\frac{p}{1-p}$, with FSO accelerating this convergence (Proposition 4.1). Only GAP-FSO structurally redirects u_f toward 0 (Proposition 5.2). The ablated oracle serves as the shortcut-free baseline.

losses are thus intrinsically blind to whether a model learns causal mechanisms or exploits shortcuts. Whether subnetwork c recovers past this point depends on how the loss function shapes the error signal $\boldsymbol{g}(t)$, yielding two distinct failure regimes under a dynamic bilinear readout (u_f, u_c):

Proposition 3.2 (Permanent Trapping under Squared Loss)**.** *Under squared loss with target output $z^\star \in \mathbb{R}$ and generic initialization ($a_f(0)^2 \neq b_f(0)^2$, $a_f(0)b_f(0) \neq 0$), gradient flow converges globally to $u_f(\infty)+u_c(\infty) = z^\star$, while subnetwork c's displacement remains permanently bounded by $O(1/\sigma_f)$ in the limit $\sigma_f \to \infty$.*

Intuitively, Proposition 3.2 shows that while joint training reaches the global target $z^\star$, it relies almost entirely on f: as σ_f grows, the displacement of c remains bounded by $O(1/\sigma_f)$, keeping it stuck near initialization. Consequently, joint training (red in Fig. 2a) stalls near a degenerate stationary point, whereas the ablated oracle (blue) moves c to its causal-only equilibrium.

Proposition 3.3 (Escape under Cross-Entropy Loss)**.** *Under CE loss with an aligned shortcut probability $p \in (\frac{1}{2}, 1)$, where shortcuts agree with labels in aligned examples (e.g., cow on grass) and disagree in conflicting ones (e.g., cow on beach), $u_f(t)$ converges to a finite constant $\frac{1}{2}\ln\frac{p}{1-p}$, whereas $u_c(t) \to \infty$ monotonically as $t \to \infty$.*

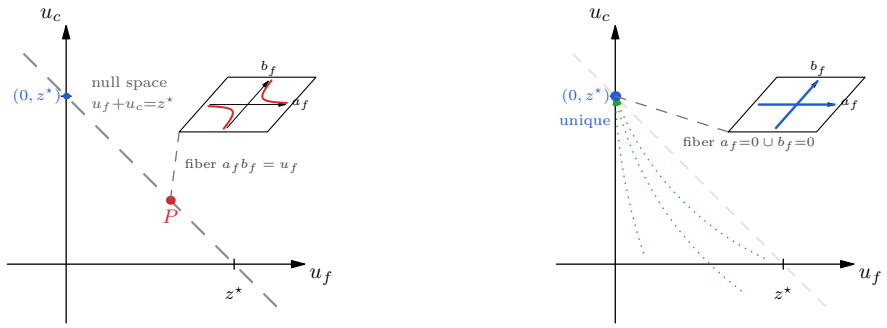
Here, $u_f(t)$ and $u_c(t)$ denote each subnetwork's contribution to the model's output $z = u_f + u_c$. Unlike Proposition 3.2, which evaluates parameter displacement, we track these output contributions directly. This shift is necessary because CE loss lacks a finite equilibrium, leaving no terminal reference point for measuring displacement. While $u_c(t)$ climbs infinitely and eventually dominates the near-stationary shortcut offset $u_f(t)$ in the model's output even though $u_f(t)$ itself never vanishes (Appendix C.4), this theoretical escape is practically unattainable within finite training budgets. Specifically, as high model confidence diminishes loss curvature (Soudry et al., 2018; Hui & Belkin, 2021), $u_c(t)$ creeps at a logarithmic rate of $\ln t + \ln\ln t + \mathcal{O}(1)$, rendering the escape mechanism exceedingly inefficient (Corollary C.3; Appendix C.4).

4 Farsighted Signal Optimization

This section presents the FSO mechanism and its null-space geometry (Section 4.1), followed by theoretical convergence guarantees and empirical validations across both loss functions (Section 4.2).

4.1 The FSO Mechanism and Null-Space Geometry

Due to the competition, a high shortcut Rayleigh activity $\hat{\kappa}_f(t)$ absorbs the shared error signal, rendering the causal subnetwork's update negligible (Corollary C.1: an objective that only reduces



(a) $\gamma = 0$: whole line is stationary

(b) $\gamma > 0$: unique fixed output

Figure 3: **The bilinear analogue of null-space dynamics, plotted in output space (u_f, u_c) for $u_f = \sigma_f a_f b_f$. (a)** At $\gamma = 0$, every point on $u_f + u_c = z^\star$ is a fixed point. The inset shows the parameter-space fiber over P ($a_f b_f = u_f$). Points along this line represent genuinely different functions with varying shortcut reliance, constituting a *harmful, loss-invisible redundancy*. **(b)** Green dotted lines denote the optimization trajectories under $\gamma > 0$. At $\gamma > 0$, the set collapses to the single output $(0, z^\star)$ with fiber $a_f{=}0 \cup b_f{=}0$. Although representational freedom remains, every parameterization realizes $u_f = 0$ identically. FSO explicitly eliminates the behavior-altering degree of freedom while retaining only *harmless representational redundancy*.

instantaneous loss provides no signal to protect subnetwork c's future trajectory). **Farsighted Signal Optimization (FSO)** addresses this by decoupling subnetwork c's update from the shared error signal $\boldsymbol{g}(t)$: it identifies subnetwork f online as the dominant channel draining $\boldsymbol{g}(t)$, tracked by the early domination of $\hat{\kappa}_f(t)$ (Fig. 1b, 7) and aligning with $\boldsymbol{K}_{SS}(t)$'s top eigendirection, then injects subnetwork c's counterfactual gradient contribution directly into its own update

$$\dot{\boldsymbol{w}}_c \;\leftarrow\; \dot{\boldsymbol{w}}_c \;-\; \gamma\, \boldsymbol{J}_S^{(c)}(t)^\top \boldsymbol{g}^{(\neg f)}(t), \qquad \boldsymbol{g}^{(\neg f)}(t) \triangleq \nabla \Phi_S\big(\boldsymbol{u}(t) - \boldsymbol{u}^{(f)}(t)\big). \tag{5}$$

This modification requires only one additional backward pass without second-order differentiation, while leaving subnetwork f's update unmodified. Crucially, update Eq. 5 is exactly gradient flow on a modified static objective, $\tilde{\Phi}_S(\boldsymbol{w}) \triangleq \Phi_S(\boldsymbol{u}(\boldsymbol{w})) + \gamma\, \Phi_S(\boldsymbol{u}^{(\neg f)}(\boldsymbol{w}_c))$. Differentiating the second term with respect to $\boldsymbol{w}_c$ yields the injected counterfactual term by the chain rule, whereas its derivative with respect to $\boldsymbol{w}_f$ vanishes identically because $\boldsymbol{u}^{(\neg f)}$ is independent of $\boldsymbol{w}_f$.

Applying Eq. 5 to the *scalar linear model* $z = \sigma_f \theta_f + \sigma_c \theta_c$ (with $\theta_f, \theta_c \in \mathbb{R}$ each subnetwork's scalar parameter) yields a gradient flow that is exactly linear in $\boldsymbol{\theta} = (\theta_f, \theta_c)$; writing it compactly as $\dot{\boldsymbol{\theta}} = -A_\gamma \boldsymbol{\theta} + \boldsymbol{b}_\gamma$, we have

$$A_\gamma \triangleq \begin{pmatrix} \sigma_f^2 & \sigma_f \sigma_c \\ \sigma_f \sigma_c & (1+\gamma)\sigma_c^2 \end{pmatrix}, \qquad \boldsymbol{b}_\gamma \triangleq \begin{pmatrix} \sigma_f z^\star \\ (1+\gamma)\sigma_c z^\star \end{pmatrix}, \qquad \det A_\gamma = \gamma\, \sigma_f^2 \sigma_c^2. \tag{6}$$

At $\gamma = 0$ (joint training), $\det A_0 = 0$, rendering A_0 rank-one with a non-trivial null space $\mathrm{span}\{(\sigma_c, -\sigma_f)\}$. Every point on the line $\sigma_f \theta_f + \sigma_c \theta_c = z^\star$ is a fixed point reproducing $z^\star$ perfectly, making these equilibria intrinsically indistinguishable to the loss function. Because A_0 is symmetric, its null space is orthogonal to its range, so the component of $\boldsymbol{\theta}(t)$ along the null direction $(\sigma_c, -\sigma_f)$ receives zero force and is exactly conserved at its initial value for all t. Starting from the origin, this component remains zero throughout training, pinning gradient flow to the unique equilibrium on the line with zero null-space component: $\big(\sigma_f z^\star/(\sigma_f^2+\sigma_c^2),\ \sigma_c z^\star/(\sigma_f^2+\sigma_c^2)\big)$. As σ_f/σ_c grows, an ever-increasing share of $z^\star$ is attributed to subnetwork f at this selected point.

As illustrated in Fig. 2a, this null-space blind spot drives shortcut reliance, as the ablated oracle equilibrium $(0, z^\star/\sigma_c)$ shares the exact same line of equilibria. Injecting FSO ($\gamma > 0$) forces $A_\gamma \succ 0$, making A_γ invertible and collapsing the degenerate line to the single point $(0, z^\star/\sigma_c)$ for every $\gamma > 0$, which is exactly the ablated oracle's equilibrium, rather than a mere approximation as $\gamma \to \infty$. Additionally, since $A_\gamma \succ 0$ is symmetric positive-definite for all $\gamma > 0$, gradient flow converges to this unique point exponentially fast from any initialization.

This dimensional collapse extends directly to the bilinear readouts $u_f = \sigma_f a_f b_f$ by distinguishing between *harmful* and *harmless* degrees of freedom. At $\gamma = 0$, the uncorrected 3D stationary

Table 1: **Experimental results for a single shared network under CE (white columns) and MSE loss (gray columns).** Reported values represent Train / Test accuracy (% worst-group accuracy for real-world datasets; OOD accuracy for Colored MNIST; J=1/J=2 denote pooled/stratified GAP-FSO). **Bold** and underline mark the best and second-best *Test* values per loss per dataset (Oracle is excluded as an upper bound). "—" denotes not applicable.

	Colored-MNIST		**Waterbirds**		**Dogs-vs-Cats**		**CelebA**		**CivilComments**	
Method	CE	MSE	CE	MSE	CE	MSE	CE	MSE	CE	MSE
Oracle	97.5/97.5	96.9/96.9	—	—	—	—	—	—	—	—
ERM	98.1/19.5	97.9/10.0	89.9/78.4	91.6/77.6	98.3/96.7	98.4/96.4	95.8/76.3	95.9/75.7	73.0/45.7	73.1/45.1
IRM	98.5/77.5	96.0/**87.6**	91.8/84.3	93.9/81.1	99.2/98.8	99.1/**98.6**	92.8/**90.9**	94.8/79.5	79.6/64.8	80.9/67.7
V-REx	98.1/80.2	96.7/87.5	91.8/84.0	93.6/81.3	99.3/**98.9**	98.9/98.3	95.1/83.3	95.5/81.9	79.6/65.1	80.7/67.7
SD	97.9/10.1	97.9/10.0	92.2/78.7	91.4/77.6	98.3/96.8	98.4/96.3	95.7/75.4	95.8/75.6	73.7/47.1	73.1/43.9
GroupDRO	98.4/76.3	97.3/83.9	91.8/85.3	93.3/80.4	99.2/**98.9**	99.1/**98.6**	95.4/82.4	95.3/83.0	80.4/69.4	80.9/70.6
DFR	84.5/73.3	79.7/68.1	92.0/**86.3**	90.2/**82.5**	98.7/97.6	98.7/97.6	90.9/89.4	91.6/**90.4**	80.9/65.9	80.3/66.4
GAP-FSO (J=1)	95.8/**89.3**	89.8/83.4	94.1/83.5	94.3/81.5	99.3/**98.9**	99.1/98.5	95.1/79.3	95.7/79.7	80.2/68.7	80.4/**71.3**
GAP-FSO (J=2)	95.1/87.3	91.1/82.1	92.9/84.3	93.1/80.6	99.2/98.7	99.1/98.5	95.0/87.5	94.8/87.5	79.7/**71.1**	79.7/70.7

manifold allows u_f to vary continuously; different points yield zero training loss yet correspond to functionally distinct models with varying shortcut reliance (Fig. 3a). Activating FSO ($\gamma > 0$) shrinks the stationary manifold by exactly one dimension, locking $u_f = 0$. The remaining 2D freedom (the fiber $a_f{=}0 \cup b_f{=}0$) is entirely harmless, representing pure scale-invariant redundancy that yields identical outputs across all inputs (Fig. 3b).

4.2 Convergence Guarantees and Empirical Validation

Building on this geometric intuition, we establish exact convergence guarantees under both loss functions and validate these theoretical predictions empirically.

Proposition 4.1 (FSO Recovers the Ablated Equilibrium)**.** *Instantiate Eq. 5 on the bottleneck construction from generic initialization. Under squared loss with target $z^\star \in \mathbb{R}$, for any $\gamma > 0$, the trajectory satisfies $\big(u_f(t), u_c(t)\big) \to (0, z^\star)$, recovering the ablated equilibrium regardless of γ and f's initialization. Under CE loss, FSO adds a strictly positive acceleration term to c while leaving f's updates untouched, ensuring $u_c(t) \to \infty$ and $u_f(t) \to \frac{1}{2} \ln \frac{p}{1-p}$ with accelerated convergence.*

Proposition 4.1 reveals that FSO operates through distinct loss-dependent mechanisms (Appendix C.6). Under squared loss, FSO acts as a *geometric unlock*: it eliminates the null-space degeneracy, collapsing the line of equilibria to the unique oracle point. Empirical PCA projections on Colored MNIST confirm this transition (Fig. 4): joint training's final representations scatter along a line-like manifold (total spread 8.42), whereas FSO concentrates them into a tight cluster (spread 1.54). Under CE loss, FSO acts as a *computational accelerator*: leaving f's equilibrium unchanged, it directly injects gradient energy into c to bypass the sluggish logarithmic regime. These analytical predictions translate directly into empirical performance gains (Figure 1c). On the XOR benchmark, FSO converts sluggish logarithmic trajectories into rapid convergence, yielding a $13.2\times$ speedup under MSE loss and a $7.0\times$ speedup under CE. On Colored MNIST with reversed OOD color correlations, standard joint training collapses to 62.2% accuracy under MSE loss due to shortcut reliance, whereas FSO recovers 96.4% accuracy, fully matching the oracle trained without shortcut cues.

Thus, resolving signal starvation requires no direct modification to the shortcut subnetwork. Injecting counterfactual causal gradients eliminates squared loss trapping and accelerates cross-entropy convergence via a single extra backward pass. Next section we will show how this mechanism extends to overparameterized networks without explicit subnetwork splits using only group labels.

5 GAP-FSO

5.1 Confidence Gap as an Exact Shortcut Proxy

Practical architectures do not provide explicit Jacobian splits $\boldsymbol{J}_S^{(f)}$ to compute Eq. 5 directly, as causal and spurious representations are entangled throughout deep layers. **GAP-FSO** eliminates this structural requirement, replacing physical subnetwork isolation with a strictly weaker dependency on group labels which we define below.

Proposition 5.1 (Confidence Gap Identifies the Shortcut Subnetwork)**.** *Partition the dataset into* aligned *examples (e.g., cow on grass) and* conflicting *examples (e.g., cow on beach). Define the signed confidence gap as* $\Delta(t) \triangleq \mathbb{E}[yz(t) \mid \textit{aligned}] - \mathbb{E}[yz(t) \mid \textit{conflicting}]$*. For every* $t \geq 0$*, along any training trajectory and under any loss function, the exact identity* $\Delta(t) = 2u_f(t)$ *holds.*

The intuition behind Proposition 5.1 lies in algebraic symmetry. Because the causal contribution u_c tracks only the true label y, its value is invariant across both groups and cancels out upon subtraction. In contrast, the shortcut contribution flips sign between groups (u_f for aligned examples, $-u_f$ for conflicting), so group-wise subtraction strips away the causal baseline and doubles the shortcut signal to exactly $2u_f(t)$ across all t and any loss function (full derivation in Appendix C.8).

Thus, the scalar $\Delta(t)$ (computable solely from output logits) acts as a proxy for the hidden shortcut strength $u_f(t)$. GAP-FSO leverages this by adding an auxiliary penalty $\mathcal{L}_{\text{aux}} = \gamma\Delta(t)^2 = 4\gamma u_f(t)^2$. Standard backpropagation through $\mathcal{L}_{\text{aux}}$ directly suppresses the parameters driving shortcut reliance, bypassing the need to identify physical subnetwork parameters. To theoretically characterize how this penalty resolves causal starvation, we analyze its effect under explicit loss functions.

Proposition 5.2 (Convergence Dynamics of GAP-FSO)**.** *Consider the setup of Proposition 3.3 with class-conditional errors* $q_+ \triangleq \sigma(-(u_f+u_c))$ *and* $q_- \triangleq \sigma(u_f-u_c)$*, where* σ *is the sigmoid function. Under GAP-FSO with* $\Delta(t) = 2u_f(t)$*, the dynamics converge to* $(u_f, u_c) \to (0, 1)$ *for any* $\gamma \geq 0$ *under squared loss. Under cross-entropy loss, the modified dynamics obey* $\dot{u}_f = \boldsymbol{K}_{SS}^{(f)}(t)\big([pq_+ - (1-p)q_-] - 8\gamma u_f\big)$ *while* $\dot{u}_c$ *is unperturbed, driving* $u_f(t) \to 0$ *for all* $\gamma > 0$*.*

The convergence paths under GAP-FSO reveal a clear dichotomy between loss functions. Under squared loss, even without any intervention, gradient flow converges globally to the causal-only solution $u_f = 0$ whenever $p < 1$. The augmented Hessian $\left(\begin{smallmatrix} 1+8\gamma & \mu \\ \mu & 1 \end{smallmatrix}\right)$, with $\mu \triangleq 2p - 1$, has determinant $1 - \mu^2 + 8\gamma > 0$ at $\gamma = 0$, as conflicting examples ($p < 1$) break the null-space degeneracy of the perfectly aligned case ($p = 1$). Because real-world datasets exhibit probabilistic rather than perfect shortcut correlations, asymptotic self-correction is the generic case. GAP-FSO makes this self-correction practical by accelerating convergence, as γ enters the Hessian of the augmented objective only through the u_f-direction curvature $1+8\gamma$, sharpening convexity precisely along the direction that needs correcting (full derivation in Appendix C.9).

Conversely, under CE loss, uncorrected dynamics permanently stall at a non-zero shortcut offset $\frac{1}{2}\ln\frac{p}{1-p}$, where the causal signal u_c cancels out of u_f's fixed-point condition (Proposition 3.3). $\mathcal{L}_{\text{aux}}$ plays a essential role: as primary gradients vanish ($u_c \to \infty$), the non-decaying auxiliary gradient $-8\gamma u_f$ takes over to enforce $u_f(t) \to 0$. Thus, $\mathcal{L}_{\text{aux}}$ provides a structural fix rather than a mere speedup, eliminating an otherwise permanent shortcut offset (Fig. 2b).

6 Experimental Design

We evaluate GAP-FSO against six representative group-robustness baselines across four paradigms: standard empirical risk minimization (**ERM**), invariance-based training (**IRM** (Arjovsky et al., 2019), **V-REx** (Krueger et al., 2021)), spectral regularization (**Spectral Decoupling** (**SD**) (Pezeshki et al., 2021)), and group-aware adaptation (**Group DRO** (Sagawa et al., 2020), **DFR** (Kirichenko et al., 2023)). All baselines are tuned on worst-group validation accuracy (Appendix E.2). When a shortcut correlates with the label in opposite directions across classes, pooling all classes into a single confidence gap can cancel to near zero even though shortcut reliance is large within each class, hiding exactly the signal GAP-FSO is meant to suppress. We therefore also evaluate a per-class stratified (J=2) form, alongside the standard pooled (J=1) form, that penalizes the confidence gap separately within each class; see Appendix D.2 for details.

Our empirical evaluation spans six vision and language benchmarks. Synthetic validation using *XOR* (LaBonte & Muthukumar, 2026) and *Colored MNIST* (Arjovsky et al., 2019) confirms our theoretical predictions under both explicit subnetwork splits (Figure 1c) and model-agnostic setups (Table 1). For real-world vision shortcuts, we test *Waterbirds* (Sagawa et al., 2020), *CelebA* (Liu et al., 2015), and *Dogs-vs-Cats* (Byungju et al., 2019), following the color-bias construction and experimental protocol of Lim et al. (2025), on an ImageNet-pretrained ResNet-50 backbone. For textual shortcuts, we evaluate *CivilComments-WILDS* (Koh et al., 2021) on DistilBERT, mitigating

class imbalance via the group-balanced subsampling of Idrissi et al. (2022). Requiring no architectural modifications, these deep non-linear validations confirm that our core mechanisms extend directly to practice, bridging the linear readout assumption and real-world benchmarks. Hardware details and compute budgets are documented in Appendix E.1.

7 Discussion and Concluding Remarks

In this section, we analyze the mechanistic behavior, cross-domain dynamics, and structural boundaries of GAP-FSO. Extended discussions on the FSO/GAP-FSO equivalence, stratification ($J > 1$) and failure-mode boundaries appear in Section D.

RQ1: How Does GAP-FSO Transfer Across Vision and NLP Modalities? Cross-modal evaluation reveals significant domain asymmetries when assessing the optimal stratification setting per Table 1. On CivilComments, GAP-FSO strictly dominates all six baselines under both loss functions, outperforming uncorrected methods (ERM and SD) by 24.0% to 27.4%, invariance-based objectives (IRM and V-REx) by 3.6% to 6.3%, and group-aware baselines (DFR and Group DRO) by 0.7% to 5.2%. On real-world vision benchmarks (Waterbirds, Dogs-vs-Cats, CelebA), GAP-FSO leads ERM, SD, and V-REx on average by 1.8% to 6.3%, while performing comparably to Group DRO on average ($+1.6\%$) despite minor trade-offs on Waterbirds under CE and Dogs-vs-Cats under MSE loss. Against vision specialists DFR and IRM, GAP-FSO matches performance on Waterbirds and Dogs-vs-Cats within 2%, whereas CelebA stands as an outlier discussed next.

RQ2: What Are the Structural Boundaries and Mechanistic Distinctions of GAP-FSO? Performance trade-offs depend directly on where spurious correlations reside, with GAP-FSO's output-level mechanism proving strongest when shortcuts act as clean, low-dimensional confounds. DFR retrains only the classification head over a frozen backbone; while this succeeds on vision tasks with localized background cues (achieving 86.3% CE accuracy on Waterbirds compared to 84.3% for GAP-FSO), it fails on text, trailing GAP-FSO on CivilComments by 5.2% (65.9% vs. 71.1%). This performance gap may arise because the causal representations within the frozen backbone had already been starved prior to head retraining. IRM enforces representation-level invariance and maintains an edge on CelebA under CE (90.9% vs. 87.5% for GAP-FSO), likely because spurious features are non-additively entangled within embeddings. Crucially, this advantage is loss-specific rather than structural: under MSE loss on the same benchmark, GAP-FSO reverses the trend, leading IRM by 8.0% (87.5% vs. 79.5%). IRM's own CelebA accuracy fluctuates by 11.4% across losses (90.9% to 79.5%), whereas GAP-FSO maintains an identical 87.5% accuracy under both; however, this stability requires stratification: the pooled ($J{=}1$) penalty reaches just 79.3% (CE) / 79.7% (MSE) on this benchmark, since `Heavy_Makeup`'s sign-heterogeneous correlation across classes cancels under pooling, while per-class stratification ($J{=}2$) recovers the missing 8.0% without sacrificing the equilibrium guarantee of Corollary C.4 (Appendix D.2). On CivilComments, where the shortcut is a clean output-level confound, GAP-FSO surpasses IRM by 6.3% without relying on IRM's penalty-annealing schedule or SD's indiscriminate logit suppression.

8 Conclusion

We present an eNTK-based dynamical analysis of shortcut learning on a bilinear bottleneck model, establishing the Dynamic Starvation Theorem: under squared loss, the causal subnetwork remains permanently trapped as shortcut dominance grows unbounded, while under cross-entropy loss, it escapes only logarithmically slowly. To resolve this, we introduce FSO, which leverages counterfactual gradients to restore oracle convergence under squared loss and accelerate escape under cross-entropy loss. We then generalize this to real-world architectures via GAP-FSO, which uses a confidence-gap identity derived from logits and group labels as an exact proxy for shortcut reliance. Unlike prior theoretical treatments of shortcut learning, which validate their predictions only on simple architectures, we extend this theory to real data and complex models such as ResNet-50 and DistilBERT, where it performs competitively with established baselines overall and clearly outperforms them on the transformer-based NLP benchmark. (See Appendix A for extended discussion of related work.) Promising directions include relaxing group annotations and extending dynamical guarantees to entangled representations (Appendix D.2 and D.3).

AI USE STATEMENT

We utilized generative AI assistants (Claude and Gemini) as drafting, editing, and mathematical verification aids throughout this work. Specifically, AI tools were used to refine LaTeX exposition and ensure notation consistency, as well as to assist with preliminary algebraic derivations and step-by-step mathematical verification of our theoretical framework. All theoretical claims, formal propositions, and proofs were independently checked, verified by hand, and finalized by the authors, who take full responsibility for the technical accuracy and originality of the content.

ETHICS STATEMENT

This work is methodological: it studies why and how deep networks come to rely on spurious correlations and proposes a training-time correction, evaluated on established public benchmarks (Waterbirds, CelebA, a synthetic Dogs-vs-Cats construction, CivilComments-WILDS) under their existing licenses and terms of use, with no new human-subjects data collection. CivilComments-WILDS pairs toxicity labels with the presence of demographic identity terms; our method's goal there is specifically to reduce a model's reliance on that correlation rather than to build a system that uses demographic attributes for any downstream decision, but we note that GAP-FSO, like most of the group-robustness baselines we compare against (all except ERM and SD), requires access to the sensitive or group-defining attribute at training time, which carries its own data-collection and privacy considerations in any deployment context beyond the benchmark setting studied here. We are not aware of other conflicts of interest, sponsorship, or dual-use concerns specific to this work.

REPRODUCIBILITY STATEMENT

The full derivation of every proposition in Sections 3 to 5 is provided in Appendix C, with one subsection per result. Appendix E.2 discloses the complete hyperparameter grid and selected values for every baseline, dataset, and loss combination in Table 1, while Appendix E.3 reports all headline metrics with error bars across random seeds, ensuring that no result in the main text relies on undisclosed settings. Complete training code to reproduce Table 1 is provided in the supplementary material.

REFERENCES

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

Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and ntk features. In *ICML*, 2024. URL https://openreview.net/forum?id=o6N1Bqay0k.

Kim Byungju, Kim Hyunwoo, Kim Kyungsu, Kim Sungjin, and Kim Junmo. Learning not to learn: Training deep neural networks with biased data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9004–9012, 06 2019. doi: 10.1109/cvpr.2019.00922. URL https://cir.nii.ac.jp/crid/1360580628209447680.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Thomas FEL, Victor Boutin, Louis Béthune, Remi Cadene, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=MziFFGjpkb.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tj3xLVuE9f.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hsFN92eQEla.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Tyler LaBonte and Vidya Muthukumar. Sgd provably prioritizes a shortcut spurious feature in the xor model. *arXiv preprint arXiv:2606.30444*, 2026.

Xiayang Li, Kuo Gai, and Shihua Zhang. Deciphering shortcut learning from an evolutionary game theory perspective. *arXiv preprint arXiv:2605.02658*, 2026.

Jinwoo Lim, Suhyun Kim, and Soo-Mook Moon. Shortcut features as top eigenfunctions of NTK: A linear neural network case and more. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TYroQXu6X0.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 151. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence of spurious correlations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0tuwdgBiSN.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017. URL https://doi.org/10.1109/ICCV.2017.74.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XVPqLyNxSyh.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Zixiong Yu, Songtao Tian, and Guhan Chen. Divergence of neural tangent kernel in classification problems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VEJzjAvaIy.

A Related Work

NTK-based accounts of shortcut learning. Prior theoretical treatments of shortcut learning built on the neural tangent kernel (NTK) inherit that theory's lazy-training assumption, which fixes the kernel at (or near) initialization. Pezeshki et al. (2021) formalize gradient starvation using a kernel frozen at initialization; Bombari & Mondelli (2024) analyze memorization of spurious features via random-feature and NTK regression, likewise treating the kernel as static; Lim et al. (2025) identify shortcut features as the top eigenfunctions of a frozen NTK, validated up to ResNet-18 on vision-only benchmarks; Hermann et al. (2024) instead characterize the converged solution as a function of feature predictivity and availability, without modeling the trajectory that reaches it. Each of these accounts therefore yields a static snapshot (namely, the frozen-kernel eigenstructure or the converged solution) rather than a description of how shortcut reliance emerges over the course of training.

Theoretical perspectives beyond NTK regimes. A smaller line of work targets this evolution directly, but concludes the lazy-training NTK framework cannot support it. Li et al. (2026) argue explicitly that analyzing shortcut-learning dynamics "lies beyond the scope of classical theoretical frameworks such as the neural tangent kernel (NTK), because the NTK framework treats $\nabla_\theta f$ as a frozen random feature," and instead model the evolution of tangent features via evolutionary game theory, characterizing gradient descent (SD) and stochastic gradient descent (SGD) as converging to two distinct stochastically stable states. Qiu et al. (2024) similarly forgo kernel theory altogether, analyzing feature-learning dynamics under spurious correlations through Boolean-function (Fourier) complexity, with formal results restricted to parity features, CE loss, and a one-hidden-layer ReLU network.

This work. We show the conclusion that NTK theory cannot capture shortcut-learning dynamics is specific to the frozen-kernel assumption these accounts inherit, not to NTK-based analysis itself. Working with the empirical NTK (eNTK) evaluated exactly as realized along the training trajectory, rather than frozen at initialization, we derive a non-asymptotic account of how shortcut reliance emerges and, unlike the synthetic or shallow-architecture validations above, confirm the resulting theory remains predictive when applied to production-scale transformer and deep vision architectures (Section 6).

B Notation

Table 2 summarizes all mathematical notation introduced in this paper, ordered by section: the eNTK dynamics (Section 2), the subnetwork decomposition (Assumption 2.1), the Dynamic Starvation Theorem (Section 3), the CE extension (Corollary C.3), Farsighted Signal Optimization (Section 4), and the confidence gap / GAP-FSO mechanism, including its multi-stratum extension (Section 5).

In the "Space / dimension" column, n is the number of training examples, m is the output dimension per example ($\boldsymbol{u} \in \mathbb{R}^{nm}$ is the stacked output vector), d is the total parameter count, and d_f, d_c are the parameter counts of subnetworks f and c ($d = d_f + d_c$).

Table 2: **Mathematical objects used in this paper.** Bold roman symbols ($\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{g}, \boldsymbol{J}$) are vectors or matrices in output/parameter space; plain italic symbols ($\theta_f, \theta_c, \sigma_f, \sigma_c, z$) belong to the scalar two-subnetwork reduction used for every closed-form result from Section 3 onward.

Symbol	Space / dimension	Meaning
eNTK dynamics (Section 2).		
$\boldsymbol{w},\ \boldsymbol{w}(t)$	$\mathbb{R}^d$	Trainable parameters; their value along the realized trajectory at time t
t	$\mathbb{R}_{\geq 0}$	Training time (gradient-flow index)
$\boldsymbol{u}(\boldsymbol{w}),\ \boldsymbol{u}(t)$	$\mathbb{R}^{nm}$	Stacked training output; its value along the trajectory at time t
$\boldsymbol{J}_S(\boldsymbol{w})$	$\mathbb{R}^{nm \times d}$	Output Jacobian $D_{\boldsymbol{w}}\boldsymbol{u}(\boldsymbol{w})$

Continued on next page.

Table 2 (continued).

Symbol	Space / dimension	Meaning
$\boldsymbol{K}_{SS},\ \boldsymbol{K}_{SS}(t)$	$\mathbb{R}^{nm\times nm},\succeq 0$	Empirical NTK (eNTK) $\boldsymbol{J}_S\boldsymbol{J}_S^\top$; its trajectory-realized value at time t
$\Phi_S,\ L_S$	$\mathbb{R}^{nm}\to\mathbb{R}$	Convex C^2 training loss on outputs; empirical risk $L_S=\Phi_S\circ\boldsymbol{u}$
$\boldsymbol{g}(t),\ \boldsymbol{B}(t)$	$\mathbb{R}^{nm},\ \mathbb{R}^{nm\times nm}$	Output error signal $\nabla_{\boldsymbol{u}}\Phi_S(\boldsymbol{u}(t))$ and Hessian $\nabla^2\Phi_S(\boldsymbol{u}(t))$
Subnetwork decomposition (Assumption 2.1).		
$\boldsymbol{w}_f,\ \boldsymbol{w}_c$	$\mathbb{R}^{d_f},\ \mathbb{R}^{d_c}$	Subnetwork f's and c's parameter blocks, $\boldsymbol{w}=(\boldsymbol{w}_f,\boldsymbol{w}_c)$
$\boldsymbol{u}^{(f)}(t),\ \boldsymbol{u}^{(c)}(t)$	$\mathbb{R}^{nm}$	Subnetwork f's and c's own output contribution, $\boldsymbol{u}=\boldsymbol{u}^{(f)}+\boldsymbol{u}^{(c)}+\text{const}$
$\boldsymbol{J}_S^{(f)},\ \boldsymbol{J}_S^{(c)}$	$\mathbb{R}^{nm\times d_f},\ \mathbb{R}^{nm\times d_c}$	Per-subnetwork output Jacobians
$\boldsymbol{K}_{SS}^{(f)},\ \boldsymbol{K}_{SS}^{(c)}$	$\mathbb{R}^{nm\times nm},\succeq 0$	Per-subnetwork kernels, $\boldsymbol{K}_{SS}=\boldsymbol{K}_{SS}^{(f)}+\boldsymbol{K}_{SS}^{(c)}$ (A1)
$\beta(t)$	$\mathbb{R}_{>0}$	Isotropic curvature scalar, $\boldsymbol{B}(t)=\beta(t)\boldsymbol{I}$ (A2)
$\boldsymbol{g}^{(\neg f)}(t)$	$\mathbb{R}^{nm}$	Ablated flow's error signal, subnetwork f deleted throughout
$u_f,\ u_c$	$\mathbb{R}$	Bilinear bottleneck readout, $u_f=\sigma_f a_f b_f$, $u_c=\sigma_c a_c b_c$
a_f, b_f, a_c, b_c	$\mathbb{R}$	The two subnetworks' bottleneck parameters
$\sigma_f,\ \sigma_c$	$\mathbb{R}_{>0}$	Capacity constants scaling subnetwork f's/c's Jacobian, $\boldsymbol{J}_S^{(f)}=\sigma_f(b_f,a_f)$
$\theta_f(t),\ \theta_c(t)$	$\mathbb{R}$	Scalar subnetwork parameters, $\theta_f(0)=\theta_c(0)=0$
Dynamic Starvation Theorem (Section 3).		
$\hat{\kappa}_f(t),\ \hat{\kappa}_c(t)$	$\mathbb{R}_{\geq 0}$	Instantaneous Rayleigh activity of f and c, $\boldsymbol{g}^\top\boldsymbol{K}_{SS}^{(f)}\boldsymbol{g}/\|\boldsymbol{g}\|^2$ and $\boldsymbol{g}^\top\boldsymbol{K}_{SS}^{(c)}\boldsymbol{g}/\|\boldsymbol{g}\|^2$
$\kappa_f^{\min},\ \kappa_c^{\min}$	$\mathbb{R}_{>0}$	Persistent-excitation lower bounds on $\hat{\kappa}_f(\tau),\hat{\kappa}_c(\tau)$ in Theorem 3.1; instantiated on the bilinear bottleneck as $\kappa_f^{\min}=\sigma_f^2\lvert a_f(0)^2-b_f(0)^2\rvert$ (analogously for c) via the conservation law of Section C.6, the same conserved quantity underlying κ_0 in Corollary C.4
T	$\mathbb{R}_{>0}$	Arbitrary, fixed training horizon (Theorem 3.1)
T	$\mathbb{R}_{>0}$	Arbitrary, fixed training horizon (Theorem 3.1)
$\Delta\boldsymbol{w}_c(0,T)$	$\mathbb{R}^{d_c}$	Subnetwork c's net parameter displacement, $\boldsymbol{w}_c(T)-\boldsymbol{w}_c(0)$
$\beta_{\min},\ \bar{\sigma}_c$	$\mathbb{R}_{>0}$	The remaining two persistent-excitation constants of Theorem 3.1, alongside σ_f,σ_c above: curvature-rate lower bound $\beta(\tau)\geq\beta_{\min}$, and causal-subnetwork Jacobian operator-norm upper bound $\bar{\sigma}_c(\tau)\triangleq\|\boldsymbol{J}_S^{(c)}(\tau)\|_{\text{op}}\leq\bar{\sigma}_c$
$z(t),\ z^\star$	$\mathbb{R}$	Shared scalar output $\sigma_f\theta_f(t)+\sigma_c\theta_c(t)$; squared-loss target
Cross-entropy extension (Proposition 3.3 and Corollary C.3).		
$p\in(0,1)$	$\mathbb{R}$	Shortcut-label correlation probability (aligned vs. conflicting population)
$q_+,\ q_-$	$(0,1)$	Class-conditional error probabilities, $q_+\triangleq\sigma(-(u_f+u_c))$ and $q_-\triangleq\sigma(u_f-u_c)$, where σ is the sigmoid function; drive $\dot{u}_f,\dot{u}_c$ under CE loss (Propositions 3.3 and 5.2)
FSO (Section 4).		
γ	$\mathbb{R}_{\geq 0}$	Injection / penalty strength, the one free hyperparameter of both FSO and GAP-FSO

Continued on next page.

Table 2 (continued).

Symbol	Space / dimension	Meaning
$\boldsymbol{\theta},\ A_\gamma,\ \boldsymbol{b}_\gamma$	$\mathbb{R}^2,\ \mathbb{R}^{2\times 2},\ \mathbb{R}^2$	$\boldsymbol{\theta} = (\theta_f, \theta_c)$; FSO-corrected linear system $\dot{\boldsymbol{\theta}} = -A_\gamma \boldsymbol{\theta} + \boldsymbol{b}_\gamma$, with $A_\gamma = \begin{pmatrix} \sigma_f^2 & \sigma_f\sigma_c \\ \sigma_f\sigma_c & (1+\gamma)\sigma_c^2 \end{pmatrix}$, $\boldsymbol{b}_\gamma = \begin{pmatrix} \sigma_f z^\star \\ (1+\gamma)\sigma_c z^\star \end{pmatrix}$ (Eq. 6)
Confidence gap / GAP-FSO (Section 5).		
$y,\ z$	$\{-1, 1\},\ \mathbb{R}$	Task label; shared scalar output (model's decision value)
$\Delta(t)$	$\mathbb{R}$	Signed confidence gap, $\mathbb{E}[yz \mid \text{aligned}] - \mathbb{E}[yz \mid \text{conflicting}]$; $= 2u_f(t)$ exactly, any loss
$\mathcal{L}_{\text{aux}}$	$\mathbb{R}_{\geq 0}$	Auxiliary penalty added to the task objective, $\gamma\Delta(t)^2$
$J,\ w_j$	$\mathbb{N},\ \mathbb{R}_{>0}$	Number of strata; stratum j's weight, $\sum_j w_j = 1$ (multi-stratum extension)
$\Delta^{(j)}(t),\ u_f^{(j)}$	$\mathbb{R}$	Stratum j's confidence gap and shortcut output contribution
$\bar{\Phi},\ \bar{\Phi}^\star$	$\mathbb{R}^2 \to \mathbb{R},\ \mathbb{R}$	GAP-FSO's augmented squared-loss population objective (Case 1, Appendix C.9); $\bar{\Phi}^\star$ its value at the unique critical point $(u_f, u_c) = (0, 1)$
μ	$(0, 1)$	$\mu \triangleq 2p - 1$, enters $\bar{\Phi}$'s Hessian as the shortcut-causal cross term
$\boldsymbol{H}_{\bar{\Phi}},\ \lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})$	$\mathbb{R}^{2\times 2},\ \mathbb{R}_{>0}$	$\bar{\Phi}$'s constant Hessian $\begin{pmatrix} 1+8\gamma & \mu \\ \mu & 1 \end{pmatrix}$; its smallest eigenvalue $(1 + 4\gamma) - \sqrt{16\gamma^2 + \mu^2}$, strictly increasing in γ toward ceiling 1
κ_0	$\mathbb{R}_{>0}$	Persistent-excitation lower bound $\min\{\boldsymbol{K}_{SS}^{(f)}(t), \boldsymbol{K}_{SS}^{(c)}(t)\} \geq \kappa_0$, computable from initial conditions (Cor. C.4)

C PROOFS

This section provides self-contained proofs for all theoretical results in the main text (Theorems 3.1, Corollary C.1, and Propositions 3.2–5.1), organized with cross-references to Sections 3–5.

C.1 PROOF OF THE STARVATION BOUND (THEOREM 3.1)

Proof of Theorem 3.1 (Gradient Dynamic Starvation Theorem). We break the proof down into three explicit, self-contained algebraic steps.

Step 1: Exact Differential Identity for the Error-Signal Norm. Recall from Section 2 that under continuous-time gradient flow on squared loss, the error signal vector $\boldsymbol{g}(t) \in \mathbb{R}^N$ evolves according to the linear dynamical system:

$$\dot{\boldsymbol{g}}(t) = -\boldsymbol{B}(t)\boldsymbol{K}_{SS}(t)\boldsymbol{g}(t).$$

By Assumption (A2) (Isotropic Loss Curvature), the loss Hessian operator on the output space reduces to a scalar-weighted identity: $\boldsymbol{B}(t) = \beta(t)\boldsymbol{I}$, where $\beta(t) > 0$. Substituting this isotropic structure into Eq. C.1 yields:

$$\dot{\boldsymbol{g}}(t) = -\beta(t)\boldsymbol{K}_{SS}(t)\boldsymbol{g}(t).$$

Now consider the time derivative of the squared Euclidean norm $y(t) \triangleq \|\boldsymbol{g}(t)\|^2 = \boldsymbol{g}(t)^\top \boldsymbol{g}(t)$. Applying the product rule for vector inner products:

$$\begin{aligned} \frac{d}{dt}\|\boldsymbol{g}(t)\|^2 &= 2\boldsymbol{g}(t)^\top \dot{\boldsymbol{g}}(t) \\ &= 2\boldsymbol{g}(t)^\top \big(-\beta(t)\boldsymbol{K}_{SS}(t)\boldsymbol{g}(t)\big) \\ &= -2\beta(t)\, \boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}(t)\boldsymbol{g}(t). \end{aligned} \tag{7}$$

By Assumption (A1) (Block Orthogonal/Additive Jacobian Structure), the global eNTK decomposes additively into the shortcut and causal components:

$$\boldsymbol{K}_{SS}(t) = \boldsymbol{K}_{SS}^{(f)}(t) + \boldsymbol{K}_{SS}^{(c)}(t),$$

where $\boldsymbol{K}_{SS}^{(f)}(t) \triangleq \boldsymbol{J}_S^{(f)}(t)\boldsymbol{J}_S^{(f)}(t)^\top$ and $\boldsymbol{K}_{SS}^{(c)}(t) \triangleq \boldsymbol{J}_S^{(c)}(t)\boldsymbol{J}_S^{(c)}(t)^\top$. Substituting this operator additive decomposition into the quadratic form Eq. 7 gives:

$$\frac{d}{dt}\|\boldsymbol{g}(t)\|^2 = -2\beta(t)\left[\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(f)}(t)\boldsymbol{g}(t) + \boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(c)}(t)\boldsymbol{g}(t)\right]. \tag{8}$$

Recalling the Rayleigh quotients $\hat{\kappa}_f(t)$ and $\hat{\kappa}_c(t)$ defined in Eq. 2:

$$\hat{\kappa}_f(t) \triangleq \frac{\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(f)}(t)\boldsymbol{g}(t)}{\|\boldsymbol{g}(t)\|^2}, \quad \text{and} \quad \hat{\kappa}_c(t) \triangleq \frac{\boldsymbol{g}(t)^\top \boldsymbol{K}_{SS}^{(c)}(t)\boldsymbol{g}(t)}{\|\boldsymbol{g}(t)\|^2}.$$

Multiplying and dividing both quadratic forms in Eq. 8 by $\|\boldsymbol{g}(t)\|^2$ yields an exact, scalar linear ODE in $y(t) \triangleq \|\boldsymbol{g}(t)\|^2$:

$$\frac{dy(t)}{dt} = -2\beta(t)\big(\hat{\kappa}_f(t) + \hat{\kappa}_c(t)\big)y(t). \tag{9}$$

To integrate Eq. 9 analytically, define the explicit integrating factor $\mu(t)$:

$$\mu(t) \triangleq \exp\left(2\int_0^t \beta(\tau)\big(\hat{\kappa}_f(\tau) + \hat{\kappa}_c(\tau)\big)d\tau\right).$$

Differentiating the product $\mu(t)y(t)$ with respect to t using the Leibniz rule:

$$\begin{aligned}\frac{d}{dt}\big[\mu(t)y(t)\big] &= \dot{\mu}(t)y(t) + \mu(t)\dot{y}(t) \\ &= \big(2\beta(t)\big(\hat{\kappa}_f(t) + \hat{\kappa}_c(t)\big)\mu(t)\big)\, y(t) + \mu(t)\left(-2\beta(t)\big(\hat{\kappa}_f(t) + \hat{\kappa}_c(t)\big)y(t)\right) \\ &= 0 \quad \text{identically for all } t \geq 0.\end{aligned}$$

Because its derivative vanishes identically, $\mu(t)y(t) = \mu(0)y(0) = 1 \cdot \|\boldsymbol{g}(0)\|^2 = \|\boldsymbol{g}(0)\|^2$. Solving for $y(t) = \|\boldsymbol{g}(t)\|^2$ gives the exact closed-form error signal trajectory:

$$\|\boldsymbol{g}(t)\|^2 = \|\boldsymbol{g}(0)\|^2 \exp\left(-2\int_0^t \beta(\tau)\big(\hat{\kappa}_f(\tau) + \hat{\kappa}_c(\tau)\big)d\tau\right).$$

Taking the positive square root of both sides yields the exact error signal norm identity:

$$\|\boldsymbol{g}(t)\| = \|\boldsymbol{g}(0)\| \exp\left(-\int_0^t \beta(\tau)\big(\hat{\kappa}_f(\tau) + \hat{\kappa}_c(\tau)\big)d\tau\right). \tag{10}$$

Step 2: Derivation of the Parameter Displacement Upper Bound. Under continuous-time gradient descent, the time-derivative of the causal subnetwork $\boldsymbol{w}_c \in \mathbb{R}^{d_c}$ is driven exclusively by its parameter Jacobian transpose $\boldsymbol{J}_S^{(c)}(t)^\top$:

$$\dot{\boldsymbol{w}}_c(t) = -\boldsymbol{J}_S^{(c)}(t)^\top \boldsymbol{g}(t).$$

Integrating both sides over the finite time window $[0, T]$:

$$\boldsymbol{w}_c(T) - \boldsymbol{w}_c(0) = \int_0^T \dot{\boldsymbol{w}}_c(\tau)d\tau = -\int_0^T \boldsymbol{J}_S^{(c)}(\tau)^\top \boldsymbol{g}(\tau)d\tau.$$

Taking the Euclidean norm on both sides and applying Minkowski's continuous triangle inequality for vector-valued integrals:

$$\|\Delta\boldsymbol{w}_c(0,T)\| = \|\int_0^T \boldsymbol{J}_S^{(c)}(\tau)^\top \boldsymbol{g}(\tau)d\tau\| \leq \int_0^T \|\boldsymbol{J}_S^{(c)}(\tau)^\top \boldsymbol{g}(\tau)\|d\tau. \tag{11}$$

By the definition of the operator (spectral) norm $\|\boldsymbol{J}_S^{(c)}(\tau)\| = \bar{\sigma}_c(\tau)$, we have the pointwise vector norm inequality $\|\boldsymbol{J}_S^{(c)}(\tau)^\top \boldsymbol{g}(\tau)\| \leq \bar{\sigma}_c(\tau)\|\boldsymbol{g}(\tau)\|$. Substituting this into Eq. 11:

$$\|\Delta\boldsymbol{w}_c(0,T)\| \leq \int_0^T \bar{\sigma}_c(\tau)\|\boldsymbol{g}(\tau)\|d\tau. \tag{12}$$

Substituting the exact analytical error signal expression Eq. 10 into Eq. 12 yields:

$$\|\Delta\boldsymbol{w}_c(0,T)\| \leq \|\boldsymbol{g}(0)\| \int_0^T \bar{\sigma}_c(\tau) \exp\left(-\int_0^\tau \beta(s)\big(\hat{\kappa}_f(s) + \hat{\kappa}_c(s)\big)ds\right)d\tau. \tag{13}$$

Step 3: Application of Persistent-Excitation Lower Bounds. Now suppose the trajectory satisfies the uniform lower bounds $\hat{\kappa}_f(\tau) \geq \kappa_f^{\min}$, $\hat{\kappa}_c(\tau) \geq \kappa_c^{\min}$, $\beta(\tau) \geq \beta_{\min} > 0$, and the upper bound $\bar{\sigma}_c(\tau) \leq \bar{\sigma}_c$ for all $\tau \in [0, T]$. Under these conditions, the exponent inside Eq. 13 is upper-bounded as:

$$-\int_0^\tau \beta(s)\big(\hat{\kappa}_f(s) + \hat{\kappa}_c(s)\big)ds \leq -\int_0^\tau \beta_{\min}\big(\kappa_f^{\min} + \kappa_c^{\min}\big)ds = -\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})\tau.$$

Because the exponential function is strictly monotonically increasing, we obtain the pointwise integrand bound:

$$\bar{\sigma}_c(\tau)\exp\left(-\int_0^\tau \beta(s)\big(\hat{\kappa}_f(s) + \hat{\kappa}_c(s)\big)ds\right) \leq \bar{\sigma}_c \exp\big(-\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})\tau\big).$$

Inserting this bound back into Eq. 13:

$$\|\Delta \boldsymbol{w}_c(0,T)\| \leq \|\boldsymbol{g}(0)\|\bar{\sigma}_c \int_0^T \exp\big(-\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})\tau\big)d\tau.$$

Since the integrand is strictly positive everywhere, extending the upper limit of integration from T to $+\infty$ can only strictly increase the value of the integral:

$$\begin{aligned}
\|\Delta \boldsymbol{w}_c(0,T)\| &\leq \|\boldsymbol{g}(0)\|\bar{\sigma}_c \int_0^\infty \exp\big(-\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})\tau\big)d\tau \\
&= \|\boldsymbol{g}(0)\|\bar{\sigma}_c \left[\frac{-1}{\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})}\exp\big(-\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})\tau\big)\right]_0^\infty \\
&= \|\boldsymbol{g}(0)\|\bar{\sigma}_c \left(0 - \frac{-1}{\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})}\right) \\
&= \frac{\bar{\sigma}_c\|\boldsymbol{g}(0)\|}{\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})}.
\end{aligned}$$

This completes the proof of Theorem 3.1. □

C.2 Loss Invisibility of Shortcut Dominance (Corollary C.1)

Corollary C.1 (Loss Invisibility of Shortcut Dominance)**.** *If the ablated flow converges to a global minimizer of Φ_S, both joint and ablated flows yield identical empirical risk trajectories. Under squared loss, $\boldsymbol{u}(t), \boldsymbol{u}^{(\neg f)}(t) \to \boldsymbol{y}$. Yet as $\sigma_f \to \infty$, $\|\Delta \boldsymbol{w}_c(0,T)\| \to 0$, so subnetwork c's learning vanishes in the shortcut-dominant limit despite zero training loss.*

Proof of Corollary C.1 (Loss-Invisibility and Representational Collapse). We prove the divergence at the representation level between joint training and ablated training in two distinct steps.

Step 1: Uniqueness of the Global Output Minimizer. Under squared loss $\mathcal{L}(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{y}\|^2$ defined over the prediction space $\mathbb{R}^{nm}$, the loss functional is strictly convex. Therefore, there exists a unique global output minimizer $\boldsymbol{u}^\star = \boldsymbol{y}$. By assumption, the ablated flow (training subnetwork c alone without subnetwork f) successfully optimizes the task loss, implying:

$$\lim_{t\to\infty} \boldsymbol{u}^{(\neg f)}(t) = \boldsymbol{y}.$$

For the joint flow, the prediction space $\boldsymbol{u}(t) = \boldsymbol{u}^{(f)}(t) + \boldsymbol{u}^{(c)}(t)$ contains the prediction space of the ablated flow as a subspace ($\boldsymbol{u}^{(f)} = \boldsymbol{0}$). Because gradient flow on a continuous convex objective with a non-empty set of minimizers converges to the global minimum, the joint flow output predictions must converge to the same unique global point:

$$\lim_{t\to\infty} \boldsymbol{u}(t) = \boldsymbol{y}.$$

Thus, $\lim_{t\to\infty} \mathcal{L}(\boldsymbol{u}(t)) = \lim_{t\to\infty} \mathcal{L}(\boldsymbol{u}^{(\neg f)}(t)) = 0$. The loss trajectories are functionally indistinguishable.

Step 2: Divergence in Parameter Space under Shortcut Dominance. Now analyze the parameter movement of subnetwork c under both setups. For the ablated flow, subnetwork c must traverse a finite Euclidean distance in parameter space to alter $\boldsymbol{u}^{(c)}$ from its initialization $\boldsymbol{u}^{(c)}(0)$ to the target $\boldsymbol{y}$; this parameter displacement is strictly positive:

$$\|\Delta \boldsymbol{w}_c^{(\neg f)}(0,\infty)\| > 0.$$

For the joint flow, instantiate the bilinear bottleneck of Assumption 2.1 for subnetwork f: $u_f = \sigma_f a_f b_f$, with gradient flow $\dot{a}_f = -g(t)\sigma_f b_f(t)$, $\dot{b}_f = -g(t)\sigma_f a_f(t)$. Then $\frac{d}{dt}\big(a_f(t)^2 - b_f(t)^2\big) = 2a_f\dot{a}_f - 2b_f\dot{b}_f = -2\sigma_f g(t)a_f b_f + 2\sigma_f g(t)a_f b_f = 0$, so $a_f(t)^2 - b_f(t)^2 = a_f(0)^2 - b_f(0)^2$ is conserved for all $t \geq 0$. By the elementary inequality $a_f(t)^2 + b_f(t)^2 \geq |a_f(t)^2 - b_f(t)^2|$, this conserved quantity lower-bounds subnetwork f's Rayleigh activity uniformly in time, with no assumption on the trajectory beyond its starting point:

$$\hat{\kappa}_f(t) = \sigma_f^2\big(a_f(t)^2 + b_f(t)^2\big) \geq \sigma_f^2|a_f(0)^2 - b_f(0)^2| =: \kappa_f^{\min} > 0 \qquad \text{for all } t \geq 0,$$

strictly positive under generic initialization ($a_f(0)^2 \neq b_f(0)^2$ and $a_f(0)b_f(0) \neq 0$), the same condition already used for κ_0 in Corollary C.4. The identical argument applied to subnetwork c ($a_c(t)^2 - b_c(t)^2$ is likewise conserved) gives $\hat{\kappa}_c(t) \geq \kappa_c^{\min} := \sigma_c^2|a_c(0)^2 - b_c(0)^2| \geq 0$ for all $t \geq 0$. Now consider the regime of extreme shortcut dominance where $\sigma_f \to \infty$ at fixed initialization $a_f(0), b_f(0), \sigma_c, a_c(0)$ and $b_c(0)$. In Theorem 3.1, instantiate the persistent-excitation floor for f with these $\kappa_f^{\min}, \kappa_c^{\min}$. Crucially, $\boldsymbol{g}(0) = u_f(0) + u_c(0) - z^\star = \sigma_f a_f(0)b_f(0) + u_c(0) - z^\star$ itself grows linearly in σ_f at this fixed initialization, so $\|\boldsymbol{g}(0)\| = \Theta(\sigma_f)$, not a constant. The parameter displacement bound Eq. 4 simplifies to:

$$\begin{aligned}
\|\Delta \boldsymbol{w}_c(0,\infty)\| &\leq \frac{\bar{\sigma}_c\|\boldsymbol{g}(0)\|}{\beta_{\min}(\kappa_f^{\min} + \kappa_c^{\min})} \\
&= \frac{\bar{\sigma}_c\left|\sigma_f a_f(0)b_f(0) + u_c(0) - z^\star\right|}{\beta_{\min}\left(\sigma_f^2|a_f(0)^2 - b_f(0)^2| + \kappa_c^{\min}\right)} \\
&= \frac{\bar{\sigma}_c|a_f(0)b_f(0)|}{\beta_{\min}|a_f(0)^2 - b_f(0)^2|} \cdot \frac{1}{\sigma_f} + O\left(\frac{1}{\sigma_f^2}\right).
\end{aligned}$$

Taking the limit as $\sigma_f \to \infty$, $\|\Delta \boldsymbol{w}_c(0,\infty)\| = O(1/\sigma_f) \to 0$; Proposition 3.2 sharpens this to the exact leading-order rate. Therefore, under joint training, subnetwork c's parameters remain vanishingly displaced from initialization as $\sigma_f \to \infty$ ($\boldsymbol{w}_c(\infty) \approx \boldsymbol{w}_c(0)$), whereas under ablated training, subnetwork c's parameters undergo significant displacement ($\boldsymbol{w}_c(\infty) \neq \boldsymbol{w}_c(0)$). Both models achieve 100% training accuracy, proving that gradient starvation is completely invisible in the loss trajectory but fully determinative of the learned internal representations. □

C.3 Proof of Permanent Trapping under Squared Loss (Proposition 3.2)

Proof of Proposition 3.2 (Dynamics under Squared Loss). Recall the two-layer linear bottleneck construction:

$$z = u_f + u_c = \sigma_f a_f b_f + \sigma_c a_c b_c,$$

where $a_f, b_f, a_c, b_c \in \mathbb{R}$ are single scalar parameters. The task loss is $\mathcal{L} = \frac{1}{2}(z - z^\star)^2$, yielding the scalar error signal $g(t) = z(t) - z^\star$.

Step 1: Derivation of the Sum and Difference Dynamics. By the chain rule, the gradient flow equations for subnetwork f's parameters are:

$$\dot{a}_f(t) = -\frac{\partial \mathcal{L}}{\partial a_f} = -(z - z^\star)\frac{\partial z}{\partial a_f} = -g(t)\sigma_f b_f(t), \tag{14}$$

$$\dot{b}_f(t) = -\frac{\partial \mathcal{L}}{\partial b_f} = -(z - z^\star)\frac{\partial z}{\partial b_f} = -g(t)\sigma_f a_f(t). \tag{15}$$

Adding Eq. 14 and Eq. 15:

$$\frac{d}{dt}\big(a_f(t) + b_f(t)\big) = -\sigma_f g(t)\big(a_f(t) + b_f(t)\big).$$

Subtracting Eq. 15 from Eq. 14:

$$\frac{d}{dt}\big(a_f(t) - b_f(t)\big) = +\sigma_f g(t)\big(a_f(t) - b_f(t)\big).$$

Define the accumulated error signal integral $\rho(t) \triangleq \int_0^t g(\tau)d\tau$. Integrating these two uncoupled linear homogeneous scalar ODEs directly:

$$\begin{aligned}
(a_f + b_f)(t) &= \Sigma_f e^{-\sigma_f \rho(t)}, \quad \text{where } \Sigma_f \triangleq a_f(0) + b_f(0), \\
(a_f - b_f)(t) &= D_f e^{+\sigma_f \rho(t)}, \quad \text{where } D_f \triangleq a_f(0) - b_f(0).
\end{aligned}$$

Repeating the exact algebraic steps for subnetwork c gives:

$$\begin{aligned}
(a_c + b_c)(t) &= \Sigma_c e^{-\sigma_c \rho(t)}, \quad \text{where } \Sigma_c \triangleq a_c(0) + b_c(0), \\
(a_c - b_c)(t) &= D_c e^{+\sigma_c \rho(t)}, \quad \text{where } D_c \triangleq a_c(0) - b_c(0).
\end{aligned}$$

Step 2: Expressing Outputs $u_f(t)$ and $u_c(t)$ in terms of $\rho(t)$. Using the algebraic identity $ab = \frac{1}{4}\left[(a+b)^2 - (a-b)^2\right]$, express $u_f(t) = \sigma_f a_f(t) b_f(t)$ as:

$$\begin{aligned} u_f(t) &= \frac{\sigma_f}{4}\left[(a_f(t)+b_f(t))^2 - (a_f(t)-b_f(t))^2\right] \\ &= \frac{\sigma_f}{4}\left[\Sigma_f^2 e^{-2\sigma_f \rho(t)} - D_f^2 e^{+2\sigma_f \rho(t)}\right]. \end{aligned} \tag{16}$$

Symmetrically, for subnetwork c:

$$u_c(t) = \frac{\sigma_c}{4}\left[\Sigma_c^2 e^{-2\sigma_c \rho(t)} - D_c^2 e^{+2\sigma_c \rho(t)}\right]. \tag{17}$$

Step 3: Autonomous ODE Reduction and Global Convergence. Differentiating $\rho(t)$ yields $\dot{\rho}(t) = g(t) = z(t) - z^\star = u_f(t) + u_c(t) - z^\star$. Substituting equation 16 and equation 17 into this derivative produces a single autonomous scalar ODE for $\rho(t)$:

$$\dot{\rho}(t) = Q(\rho(t)), \tag{18}$$

where $Q(\rho) \triangleq \frac{\sigma_f}{4}\left[\Sigma_f^2 e^{-2\sigma_f \rho} - D_f^2 e^{2\sigma_f \rho}\right] + \frac{\sigma_c}{4}\left[\Sigma_c^2 e^{-2\sigma_c \rho} - D_c^2 e^{2\sigma_c \rho}\right] - z^\star$. Differentiating $Q(\rho)$ with respect to ρ gives:

$$Q'(\rho) = -\frac{\sigma_f^2}{2}\left[\Sigma_f^2 e^{-2\sigma_f \rho} + D_f^2 e^{2\sigma_f \rho}\right] - \frac{\sigma_c^2}{2}\left[\Sigma_c^2 e^{-2\sigma_c \rho} + D_c^2 e^{2\sigma_c \rho}\right].$$

Since exponential terms are strictly positive and initializations are generic ($\Sigma_f, D_f, \Sigma_c, D_c \neq 0$), $Q'(\rho) < 0$ strictly for all $\rho \in \mathbb{R}$, establishing that $Q(\rho)$ is strictly monotonically decreasing on $\mathbb{R}$. Combined with the asymptotic limits $\lim_{\rho \to +\infty} Q(\rho) = -\infty$ and $\lim_{\rho \to -\infty} Q(\rho) = +\infty$, the Intermediate Value Theorem guarantees a unique root $\rho^\star \in \mathbb{R}$ such that $Q(\rho^\star) = 0$.

Standard phase-line analysis for 1D autonomous ODEs confirms that $\rho^\star$ is a globally asymptotically stable fixed point: $\dot{\rho}(t) > 0$ whenever $\rho(t) < \rho^\star$, and $\dot{\rho}(t) < 0$ whenever $\rho(t) > \rho^\star$. Thus, from the initial condition $\rho(0) = 0$, $\rho(t)$ converges monotonically to $\rho^\star$ as $t \to \infty$. Consequently, $\lim_{t\to\infty} \dot{\rho}(t) = Q(\rho^\star) = 0$, proving that $\lim_{t\to\infty} g(t) = 0$ and $z(\infty) = z^\star$ exactly.

Step 4: Asymptotic Rate of Subnetwork c's Displacement as $\sigma_f \to \infty$. Fix $\sigma_c, \Sigma_c, D_c, z^\star, \Sigma_f, D_f$ and let $\sigma_f \to \infty$ (the shortcut-dominant limit motivated in Section 2). Because $u_f(\rho) = \frac{\sigma_f}{4}[\Sigma_f^2 e^{-2\sigma_f \rho} - D_f^2 e^{2\sigma_f \rho}]$ carries an unbounded prefactor σ_f, keeping $Q(\rho^\star) = 0$ satisfiable as $\sigma_f \to \infty$ forces $\rho^\star \to 0$ at the matching rate $\rho^\star = \Theta(1/\sigma_f)$. Writing $\rho^\star = \epsilon/\sigma_f$ and substituting into $u_f(\rho^\star) + u_c(\rho^\star) = z^\star$, the leading-order balance requires

$$\Sigma_f^2 e^{-2\epsilon} - D_f^2 e^{2\epsilon} = 0 \quad \Longrightarrow \quad \epsilon = \frac{1}{2}\ln\left|\frac{\Sigma_f}{D_f}\right|, \quad \text{so} \quad \rho^\star = \frac{1}{2\sigma_f}\ln\left|\frac{\Sigma_f}{D_f}\right| + O(1/\sigma_f^2).$$

Since $\rho^\star \to 0$, Taylor-expanding $u_c(\rho)$ from Eq. 17 around $\rho = 0$ gives $u_c(\rho^\star) - u_c(0) = u_c'(0)\rho^\star + O((\rho^\star)^2)$, where $u_c'(0) = -\frac{\sigma_c^2}{2}(\Sigma_c^2 + D_c^2)$. Substituting $\rho^\star$:

$$u_c(\infty) - u_c(0) = -\frac{\sigma_c^2(\Sigma_c^2 + D_c^2)}{4\sigma_f}\ln\left|\frac{\Sigma_f}{D_f}\right| + O(1/\sigma_f^2) = O(1/\sigma_f). \tag{19}$$

The same expansion applied to $a_c(\rho) = \frac{1}{2}[\Sigma_c e^{-\sigma_c \rho} + D_c e^{\sigma_c \rho}]$ and $b_c(\rho) = \frac{1}{2}[\Sigma_c e^{-\sigma_c \rho} - D_c e^{\sigma_c \rho}]$ gives $\Delta a_c = a_c(\infty) - a_c(0) = -\sigma_c b_c(0)\rho^\star + O((\rho^\star)^2)$ and $\Delta b_c = -\sigma_c a_c(0)\rho^\star + O((\rho^\star)^2)$, so the parameter-space displacement $\|\Delta \boldsymbol{w}_c(0,\infty)\| = \sqrt{\Delta a_c^2 + \Delta b_c^2}$ is likewise $O(1/\sigma_f)$, confirming Proposition 3.2's stated rate. □

C.4 Proof of Logarithmic Escape under Cross-Entropy Loss (Proposition 3.3)

Proof of Proposition 3.3 (Dynamics under Cross-Entropy Loss). Without loss of generality, fix target class $y = 1$. The binary CE loss over aligned ($z_+ = u_f + u_c$) and conflicting ($z_- = -u_f + u_c$) splits with alignment probability $p \in (\frac{1}{2}, 1)$ is:

$$\mathcal{L}_{\mathrm{CE}} = -p\ln\sigma(u_f + u_c) - (1-p)\ln\sigma(-u_f + u_c). \tag{20}$$

Using $\frac{d}{dx}[-\ln\sigma(x)] = 1 - \sigma(x)$, define the class-conditional error probabilities $q_+ \triangleq 1 - \sigma(z_+) = \sigma(-z_+)$ and $q_- \triangleq 1 - \sigma(z_-) = \sigma(-z_-)$, where σ is the sigmoid function.

Step 1: Gradient Trajectories and Dynamic eNTK. By the chain rule, the parameter ODEs for subnetwork f are:

$$\dot{a}_f = \sigma_f b_f G_f(t), \quad \dot{b}_f = \sigma_f a_f G_f(t), \quad \text{where } G_f(t) \triangleq pq_+(t) - (1-p)q_-(t).$$

Similarly, for subnetwork c:

$$\dot{a}_c = \sigma_c b_c G_c(t), \quad \dot{b}_c = \sigma_c a_c G_c(t), \quad \text{where } G_c(t) \triangleq pq_+(t) + (1-p)q_-(t).$$

Integrating these ODEs using the exact sum/difference transformation from Proposition 3.2 with $\phi_f(t) \triangleq \int_0^t G_f(\tau)d\tau$ and $\phi_c(t) \triangleq \int_0^t G_c(\tau)d\tau$:

$$u_f(t) = \frac{\sigma_f}{4}\left[\Sigma_f^2 e^{2\sigma_f\phi_f(t)} - D_f^2 e^{-2\sigma_f\phi_f(t)}\right], \tag{21}$$

$$u_c(t) = \frac{\sigma_c}{4}\left[\Sigma_c^2 e^{2\sigma_c\phi_c(t)} - D_c^2 e^{-2\sigma_c\phi_c(t)}\right]. \tag{22}$$

Differentiating $u_f(t)$ with respect to t:

$$\begin{aligned}\dot{u}_f(t) &= \frac{\sigma_f}{4}\left[2\sigma_f\Sigma_f^2 e^{2\sigma_f\phi_f(t)} + 2\sigma_f D_f^2 e^{-2\sigma_f\phi_f(t)}\right]\dot{\phi}_f(t)\\ &= \frac{\sigma_f^2}{2}\left[\Sigma_f^2 e^{2\sigma_f\phi_f(t)} + D_f^2 e^{-2\sigma_f\phi_f(t)}\right]G_f(t).\end{aligned}$$

Using $\Sigma_f^2 e^{2\sigma_f\phi_f} = \frac{4u_f}{\sigma_f} + D_f^2 e^{-2\sigma_f\phi_f}$ from Eq. 21, we simplify the coefficient to:

$$\dot{u}_f(t) = \boldsymbol{K}_{SS}^{(f)}(t)G_f(t), \quad \text{where } \boldsymbol{K}_{SS}^{(f)}(t) = 2\sigma_f u_f(t) + \sigma_f^2 D_f^2 e^{-2\sigma_f\phi_f(t)} \geq 0. \tag{23}$$

Symmetrically, for subnetwork c:

$$\dot{u}_c(t) = \boldsymbol{K}_{SS}^{(c)}(t)G_c(t), \quad \text{where } \boldsymbol{K}_{SS}^{(c)}(t) = 2\sigma_c u_c(t) + \sigma_c^2 D_c^2 e^{-2\sigma_c\phi_c(t)} \geq 0. \tag{24}$$

Here, $\boldsymbol{K}_{SS}^{(f)}(t)$ and $\boldsymbol{K}_{SS}^{(c)}(t)$ represent the scalar instantiations of the subnetwork eNTKs $\boldsymbol{K}_{SS}^{(f)}(t)$ and $\boldsymbol{K}_{SS}^{(c)}(t)$ introduced in Section 2. Because the gradient $\boldsymbol{g}(t)$ is scalar in this 1D setup, the Rayleigh activity ratio Eq. 2 reduces directly to $\hat{\kappa}_f(t) = \boldsymbol{K}_{SS}^{(f)}(t)$ and $\hat{\kappa}_c(t) = \boldsymbol{K}_{SS}^{(c)}(t)$.

Step 2: Proof that $u_c(t) \to \infty$. Because $q_+, q_- \in (0,1)$ strictly for all finite inputs, $G_c(t) = pq_+ + (1-p)q_- > 0$ strictly for all $t \geq 0$. By Eq. 24, $\dot{u}_c(t) > 0$, so $u_c(t)$ is strictly monotonically increasing. We now prove $u_c(t) \to \infty$ by contradiction. Suppose $u_c(t)$ converges to a finite limit $u_c^\infty < \infty$. If $u_c(t)$ is bounded, q_+ and q_- remain bounded away from 0 uniformly in t, so there exists $\epsilon > 0$ such that $G_c(t) \geq \epsilon$ for all t. This forces $\phi_c(t) = \int_0^t G_c(\tau)d\tau \to +\infty$ as $t \to \infty$. As $\phi_c(t) \to \infty$, $e^{-2\sigma_c\phi_c(t)} \to 0$, so $D_c^2 e^{-2\sigma_c\phi_c(t)} \to 0$. From Eq. 24, this implies $\boldsymbol{K}_{SS}^{(c)}(t) \to 2\sigma_c u_c^\infty$.

- If $u_c^\infty > 0$, then $\boldsymbol{K}_{SS}^{(c)}(t) \to 2\sigma_c u_c^\infty > 0$, so for sufficiently large t, $\dot{u}_c(t) = \boldsymbol{K}_{SS}^{(c)}(t)G_c(t) > \sigma_c u_c^\infty \epsilon > 0$, contradicting the assumption that $u_c(t)$ converges to a finite limit.
- If $u_c^\infty \leq 0$: since $D_c^2 e^{-2\sigma_c\phi_c(t)} \to 0$, Eq. 22 forces $\Sigma_c^2 e^{2\sigma_c\phi_c(t)} \to \frac{4u_c^\infty}{\sigma_c} \leq 0$. But Σ_c is a fixed nonzero constant (generically, excluding the degenerate initialization $a_c(0) = -b_c(0)$, under which $(a_c + b_c)(t) \equiv 0$ and $u_c(t) = -\sigma_c a_c(t)^2 \leq 0$ for all t) and $\phi_c(t) \to \infty$, so $\Sigma_c^2 e^{2\sigma_c\phi_c(t)} \to +\infty$, a direct contradiction.

Thus, $u_c(t) \to \infty$ unconditionally.

Step 3: Asymptotic Equilibrium of $u_f(t)$. As $u_c(t) \to \infty$, we evaluate the asymptotic limits of q_+ and q_-:

$$q_+ = \sigma(-u_f - u_c) = \frac{1}{1 + e^{u_f + u_c}} \approx e^{-u_f - u_c},$$

$$q_- = \sigma(u_f - u_c) = \frac{1}{1 + e^{-u_f + u_c}} \approx e^{u_f - u_c}.$$

Substitute these asymptotic expressions into $G_f(t)$:

$$G_f(t) = pq_+ - (1-p)q_- \approx e^{-u_c(t)}\left[pe^{-u_f(t)} - (1-p)e^{u_f(t)}\right].$$

Define the function $H(u_f) \triangleq pe^{-u_f} - (1-p)e^{u_f}$. Note that $H'(u_f) = -pe^{-u_f} - (1-p)e^{u_f} < 0$, so H is strictly monotonically decreasing. Setting $H(u_f) = 0$:

$$pe^{-u_f} = (1-p)e^{u_f} \implies e^{2u_f} = \frac{p}{1-p} \implies u_f^\star = \frac{1}{2}\ln\left(\frac{p}{1-p}\right).$$

Because $\boldsymbol{K}_{SS}^{(f)}(t) > 0$ (generically, excluding the degenerate trajectory $a_f \equiv b_f \equiv 0$) and $e^{-u_c(t)} > 0$, the sign of $\dot{u}_f(t)$ asymptotically matches the sign of $H(u_f(t))$ as $u_c(t) \to \infty$:

- If $u_f(t) < u_f^\star$, then $H(u_f) > 0 \implies \dot{u}_f(t) > 0$.
- If $u_f(t) > u_f^\star$, then $H(u_f) < 0 \implies \dot{u}_f(t) < 0$.

Therefore, $u_f(t) \to \frac{1}{2}\ln\left(\frac{p}{1-p}\right)$ as $t \to \infty$, establishing the non-zero CE shortcut equilibrium. □

C.5 Proof of the Logarithmic Growth Rate

In Proposition 3.3 establishes the convergence of the causal parameter $u_c(t)$ under CE loss, but leaves its precise asymptotic growth rate unspecified. To provide a complete and rigorous characterization, this section derives the sharp two-term asymptotic expansion $u_c(t) = \ln t + \ln\ln t + O(1)$.

We first establish a self-contained lemma on the asymptotic behavior of the exponential integral function $\mathrm{Ei}(x)$, followed by the detailed proof of the main result.

Lemma C.2 (Asymptotic Expansion of the Exponential Integral)**.** *Let* $\mathrm{Ei}(x) \triangleq \int_{-\infty}^{x} \frac{e^s}{s}\,ds$ *(principal value for* $x > 0$*) denote the exponential integral. As* $x \to +\infty$*,*

$$\mathrm{Ei}(x) = \frac{e^x}{x}\big(1 + O(1/x)\big) = \frac{e^x}{x}\big(1 + o(1)\big). \tag{25}$$

Here $O(1/x)$ is the sharper statement, giving an explicit decay rate for the correction term; $o(1)$ only records that it vanishes, without specifying how fast. We prove the stronger $O(1/x)$ form below; Corollary C.3 only ever invokes the weaker $o(1)$ form that it implies.

Proof. By the Fundamental Theorem of Calculus, Ei is differentiable on $(0,\infty)$ with $\mathrm{Ei}'(x) = e^x/x$. Integrating by parts once, and using $e^s/s \to 0$ as $s \to -\infty$,

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^s}{s}\,ds = \Big[\frac{e^s}{s}\Big]_{-\infty}^{x} - \int_{-\infty}^{x} e^s\Big(-\frac{1}{s^2}\Big)\,ds = \frac{e^x}{x} + \int_{-\infty}^{x} \frac{e^s}{s^2}\,ds.$$

Integrating by parts a second time on the remaining integral,

$$\int_{-\infty}^{x} \frac{e^s}{s^2}\,ds = \Big[\frac{e^s}{s^2}\Big]_{-\infty}^{x} - \int_{-\infty}^{x} e^s\Big(-\frac{2}{s^3}\Big)\,ds = \frac{e^x}{x^2} + O\Big(\frac{e^x}{x^3}\Big),$$

where the last integral is itself $O(e^x/x^3)$ by the identical argument applied once more (each further integration by parts divides the leading term by one additional factor of x). Combining the two steps,

$$\mathrm{Ei}(x) = \frac{e^x}{x} + \frac{e^x}{x^2} + O\Big(\frac{e^x}{x^3}\Big) = \frac{e^x}{x}\Big(1 + \frac{1}{x} + O\big(\tfrac{1}{x^2}\big)\Big) = \frac{e^x}{x}\big(1 + O(1/x)\big),$$

proving Eq. 25. □

Corollary C.3 (Logarithmic growth rate of the causal subnetwork under GAP-FSO's bottleneck construction)**.** *Under Proposition 3.3's balanced sub-case (*$D_f = D_c = 0$*), the causal subnetwork's parameter grows as*

$$u_c(t) = \ln t + \ln\ln t + O(1) \qquad (t \to \infty). \tag{26}$$

Proof. Write $h(\theta) \triangleq pe^{-\theta} + (1-p)e^{\theta}$ and $v(t) \triangleq u_c(t)$.

Step 1: Properties of h. By the AM-GM inequality, for every $\theta \in \mathbb{R}$,

$$h(\theta) = pe^{-\theta} + (1-p)e^{\theta} \;\geq\; 2\sqrt{(pe^{-\theta})\cdot((1-p)e^{\theta})} \;=\; 2\sqrt{p(1-p)},$$

with equality iff $pe^{-\theta} = (1-p)e^{\theta}$, i.e. $\theta = \frac{1}{2}\ln\frac{p}{1-p} = u_f(\infty)$. Moreover $h''(\theta) = pe^{-\theta} + (1-p)e^{\theta} = h(\theta) > 0$ everywhere, so h is strictly convex, confirming $u_f(\infty)$ is h's unique global minimizer, with minimum value $h(u_f(\infty)) = 2\sqrt{p(1-p)}$.

Step 2: Upper bound. Since $\sigma(-z) \leq e^{-z}$ for every $z \in \mathbb{R}$, $q_+ \leq e^{-u_f - v} \leq e^{-v}$ and $q_- \leq e^{u_f - v}$, so

$$\dot{v}(t) = 2\sigma_c v(t)\big[pq_+ + (1-p)q_-\big] \;\leq\; 2\sigma_c v(t) e^{-v(t)} \underbrace{\big[pe^{-u_f(t)} + (1-p)e^{u_f(t)}\big]}_{=\,h(u_f(t))}.$$

Fix $\varepsilon > 0$. Since $u_f(t) \to u_f(\infty)$ (Proposition 3.3) and h is convex with its minimum at $u_f(\infty)$ (Step 1), there is t_ε with $h(u_f(t)) \le h(u_f(\infty) + \varepsilon) =: C(\varepsilon)$ for all $t \ge t_\varepsilon$. Writing $K \triangleq 2\sigma_c C(\varepsilon)$, this gives the differential inequality

$$\dot{v}(t) \;\le\; K\, v(t) e^{-v(t)}, \qquad t \ge t_\varepsilon. \tag{27}$$

Since $v(t) > 0$, dividing Eq. 27 by $v(t)e^{-v(t)} > 0$ gives $\frac{e^{v(t)}}{v(t)}\dot{v}(t) \le K$; by the chain rule, the left side is exactly $\frac{d}{dt}\mathrm{Ei}(v(t))$ (Lemma C.2's definition of Ei). Integrating from t_ε to t,

$$\mathrm{Ei}(v(t)) - \mathrm{Ei}(v(t_\varepsilon)) \;\le\; K(t - t_\varepsilon), \qquad \text{i.e.} \qquad \mathrm{Ei}(v(t)) \;\le\; Kt + O(1),$$

where the $O(1)$ absorbs the (once ε is fixed) constant $\mathrm{Ei}(v(t_\varepsilon)) - Kt_\varepsilon$. Since $v(t) \to \infty$ (Proposition 3.3), Lemma C.2 applies to the left side:

$$\frac{e^{v(t)}}{v(t)}\big(1 + o(1)\big) \;\le\; Kt + O(1).$$

Taking $\ln(\cdot)$ of both sides and simplifying ($\ln(1 + o(1)) = o(1)$; $\ln(Kt + O(1)) = \ln t + O(1)$),

$$v(t) - \ln v(t) \;\le\; \ln t + O(1). \tag{28}$$

We solve Eq. 28 for $v(t)$ by a two-step bootstrap. First, since $v(t) \to \infty$, $\ln v(t) = o(v(t))$, so Eq. 28 alone already gives the coarse bound $v(t) = O(\ln t)$. Substituting this coarse bound back into $\ln v(t)$ sharpens it to $\ln v(t) = \ln\big(\ln t + O(1)\big) = \ln\ln t + O(1)$; feeding this refined estimate back into Eq. 28,

$$v(t) \;\le\; \ln t + \ln v(t) + O(1) \;=\; \ln t + \ln\ln t + O(1).$$

Step 3: Lower Bound. We now establish a matching lower bound for $v(t)$. Recall from Proposition 3.3 that $u_f(t)$ converges, hence remains bounded, while $v(t) \to \infty$ as $t \to \infty$. Consequently, there exists a sufficiently large threshold $t_0 > 0$ such that $z_+(t), z_-(t) \ge 0$ for all $t \ge t_0$.

For any $z \ge 0$, we have the elementary exponential lower bound on the sigmoid function:

$$\sigma(-z) = \frac{1}{1 + e^z} \ge \frac{1}{2} e^{-z}.$$

Applying this bound to both terms q_+ and q_-, and utilizing the global minimum $h(\theta) \ge 2\sqrt{p(1-p)}$ established in Step 1, the time derivative of $v(t)$ satisfies

$$\begin{aligned}\dot{v}(t) &\ge \sigma_c v(t) e^{-v(t)} h(u_f(t)) \\ &\ge 2\sigma_c \sqrt{p(1-p)}\, v(t) e^{-v(t)}, \quad \forall t \ge t_0.\end{aligned}$$

Let $K' \triangleq 2\sigma_c\sqrt{p(1-p)} > 0$. Rearranging the differential inequality yields

$$\frac{e^{v(t)}}{v(t)}\dot{v}(t) \ge K', \quad \forall t \ge t_0.$$

Integrating both sides with respect to time from t_0 to t via the change of variables formula and Lemma C.2 (definition of Ei), we obtain

$$\mathrm{Ei}(v(t)) - \mathrm{Ei}(v(t_0)) \ge K'(t - t_0),$$

which simplifies to $\mathrm{Ei}(v(t)) \ge K't + O(1)$ as $t \to \infty$, where the constant term $\mathrm{Ei}(v(t_0)) - K't_0$ is absorbed into $O(1)$.

Finally, invoking the asymptotic expansion from Lemma C.2 along with the identical two-step bootstrap argument employed in Step 2, we conclude that

$$v(t) \ge \ln t + \ln\ln t + O(1) \quad \text{as } t \to \infty.$$

Step 4: Conclusion. Steps 2 and 3 sandwich $v(t)$ between two bounds that agree at both leading orders: the multiplicative constants $K \ne K'$ enter only as $\ln K, \ln K'$ inside the $O(1)$ term (since $\ln(Kt) = \ln t + \ln K$), never affecting the $\ln t$ or $\ln\ln t$ coefficients themselves. By the squeeze theorem,

$$u_c(t) = v(t) = \ln t + \ln\ln t + O(1) \qquad (t \to \infty),$$

proving Eq. 26. □

C.6 Proof of FSO Recovers the Ablated Equilibrium (Proposition 4.1)

Proof of Proposition 4.1 (Dynamics under FSO). FSO modifies the parameter updates by injecting the counterfactual gradient exclusively into subnetwork c.

Case 1: Squared Loss. FSO corresponds to joint gradient flow on the augmented energy functional:

$$\tilde{\Phi}_S(u_f, u_c) \triangleq \frac{1}{2}(u_f + u_c - z^\star)^2 + \frac{\gamma}{2}(u_c - z^\star)^2.$$

Let $A \triangleq u_f + u_c - z^\star$ and $B \triangleq A + \gamma(u_c - z^\star)$. The parameter update equations are given by:

$$\dot{\boldsymbol{w}}_f = -{\boldsymbol{J}_S^{(f)}}^\top A \implies \dot{a}_f = -\sigma_f b_f A, \quad \dot{b}_f = -\sigma_f a_f A,$$
$$\dot{\boldsymbol{w}}_c = -{\boldsymbol{J}_S^{(c)}}^\top B \implies \dot{a}_c = -\sigma_c b_c B, \quad \dot{b}_c = -\sigma_c a_c B.$$

As in Section C.3, this bilinear form conserves $a_f^2 - b_f^2$ and $a_c^2 - b_c^2$ exactly: $\frac{d}{dt}(a_f^2 - b_f^2) = 2a_f\dot{a}_f - 2b_f\dot{b}_f = -2\sigma_f a_f b_f A + 2\sigma_f a_f b_f A = 0$, and likewise for c. Taking the time derivative of $\tilde{\Phi}_S$ along the trajectory:

$$\frac{d}{dt}\tilde{\Phi}_S = \left(\frac{\partial\tilde{\Phi}_S}{\partial \boldsymbol{w}_f}\right)^\top \dot{\boldsymbol{w}}_f + \left(\frac{\partial\tilde{\Phi}_S}{\partial \boldsymbol{w}_c}\right)^\top \dot{\boldsymbol{w}}_c = -\|\nabla_{\boldsymbol{w}_f}\tilde{\Phi}_S\|^2 - \|\nabla_{\boldsymbol{w}_c}\tilde{\Phi}_S\|^2 \le 0.$$

Thus, $\tilde{\Phi}_S(t) \le \tilde{\Phi}_S(0)$, which uniformly bounds $A(t)$ and $(u_c(t) - z^\star)$, and consequently bounds $u_f(t)$ and $u_c(t)$ for all $t \ge 0$. Combined with the conserved quantities above, this bounds $a_f(t), b_f(t), a_c(t), b_c(t)$ individually (a bounded product together with a fixed difference of squares pins both factors to a bounded range), so the trajectory stays in a compact subset of $\mathbb{R}^4$, as required by LaSalle's Invariance Principle below. By LaSalle's Invariance Principle, trajectories bounded in a compact set converge to the largest invariant subset where $\frac{d}{dt}\tilde{\Phi}_S = 0$, corresponding to critical points where $\nabla_{\boldsymbol{w}_f}\tilde{\Phi}_S = \mathbf{0}$ and $\nabla_{\boldsymbol{w}_c}\tilde{\Phi}_S = \mathbf{0}$. The stationarity conditions require:

$$(A = 0 \text{ or } a_f = b_f = 0) \quad \text{and} \quad (B = 0 \text{ or } a_c = b_c = 0)\,.$$

Under generic initialization of subnetwork c ($a_c(0)^2 \neq b_c(0)^2$), the conserved quantity $a_c^2 - b_c^2 \neq 0$ for all t, so subnetwork c never approaches the origin $(a_c, b_c) = (0, 0)$. Thus $B = 0$ must hold. Setting $A = 0$ and $B = A + \gamma(u_c - z^\star) = 0$ simultaneously yields:

$$\gamma(u_c - z^\star) = 0 \implies u_c(\infty) = z^\star, \quad \text{and} \quad u_f(\infty) + z^\star - z^\star = 0 \implies u_f(\infty) = 0.$$

Because $\tilde{\Phi}_S$ is strictly convex in (u_f, u_c) with Hessian determinant $\gamma > 0$, the target $(0, z^\star)$ is the unique global attractor. This aligns with the geometric intuition in Fig. 3b: while subnetwork f's stationarity condition ($A = 0$ or $a_f = b_f = 0$) does not uniquely determine a single branch, both scenarios consistently yield $u_f = \sigma_f a_f b_f = 0$. Consequently, although the critical set forms the fiber $a_f{=}0 \cup b_f{=}0$, it universally collapses onto the unique output state $(u_f, u_c) = (0, z^\star)$.

Case 2: Cross-Entropy Loss. FSO leaves subnetwork f's update equation untouched ($\dot{u}_f = \boldsymbol{K}_{SS}^{(f)}(t)G_f(t)$) while adding a positive counterfactual term to subnetwork c's update:

$$\dot{u}_c(t) = \boldsymbol{K}_{SS}^{(c)}(t)\left([pq_+ + (1-p)q_-] + \gamma\sigma(-u_c)\right).$$

Since $\gamma\sigma(-u_c) > 0$ strictly, $G_c(t) + \gamma\sigma(-u_c)$ is strictly positive. Following the exact contradiction proof in Step 2 of Proposition 3.3, $u_c(t) \to \infty$ continues to hold unconditionally. Once $u_c(t) \to \infty$, $G_f(t) \approx e^{-u_c(t)}H(u_f(t))$. Applying Step 3 of Proposition 3.3, $u_f(t) \to \frac{1}{2}\ln\left(\frac{p}{1-p}\right)$ unchanged. □

C.7 Empirical Validation of the Null-Space Geometry (Colored MNIST)

To test whether Proposition 4.1's null-space collapse extends beyond the scalar toy model of Section 4 to a genuine trained network, we instantiate subnetworks f and c as the BranchF/BranchC architecture of Colored MNIST (a single linear unit on a color-difference feature and a two-layer MLP on luminance, respectively) and train under squared loss with $\gamma = 0$ (joint) and $\gamma = 8$ (FSO), 20 random weight initializations each, sharing identical colorized training data across all 40 runs so only the initialization differs – mirroring the toy model's "same A_0/A_γ, different starting point" setup. For each trained run, we evaluate (u_f, u_c) on a fixed subset of the *training* data and concatenate them into a single representation vector: the null-space line is precisely an invariance of $u_f + u_c$ (equivalently, the training loss) under the training objective, with the f/c split varying freely, so this is the distribution on which the degeneracy is actually defined. A single joint PCA (via SVD) across all 40 runs projects these vectors to 2D so every run's landing point is directly comparable. Fig. 4 shows the result: Joint training's landing points scatter along an extended, line-like manifold (total spread 8.42), mirroring the degenerate equilibrium line of the scalar toy model, while FSO's landing points cluster far more tightly (total spread 1.54, $\sim 5.5\times$ smaller), consistent with the collapse to a near-unique equilibrium predicted by Proposition 4.1.

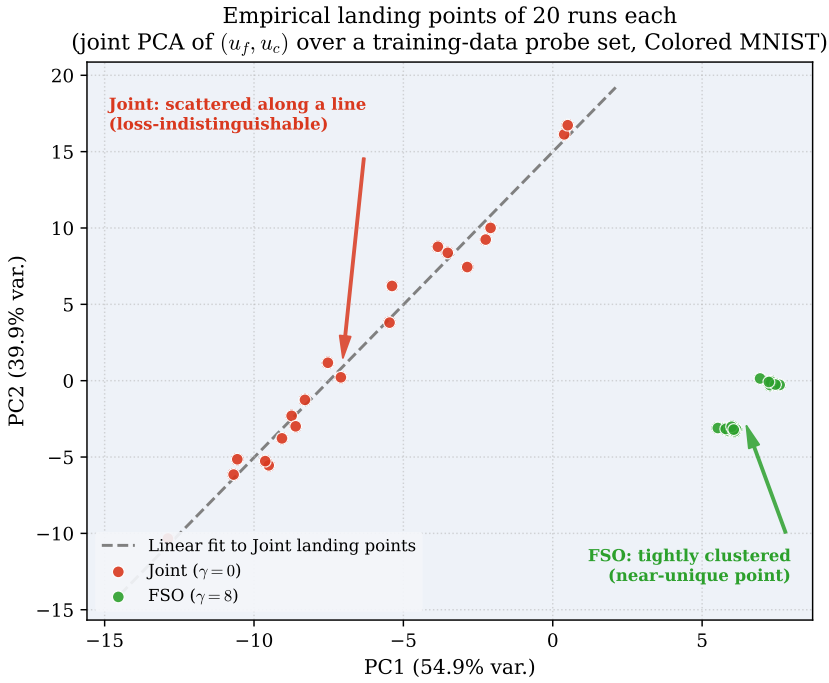


Figure 4: **Empirical null-space collapse on Colored MNIST.** Landing points of 20 runs each for Joint and FSO training, projected via a single joint PCA on (u_f, u_c) evaluated over a fixed training-data probe set. The dashed line is an empirical least-squares fit through the Joint points, not a closed-form null-space direction (unavailable for this non-bilinear-toy network). Joint training's runs scatter along a line-like manifold (total spread 8.42); FSO's cluster tightly (total spread 1.54), $\sim$ **5.5**$\times$ smaller.

C.8 Proof of Confidence Gap Identifies the Shortcut Subnetwork (Proposition 5.1)

Proof of Proposition 5.1 (Exactness of the Confidence Gap). By Assumption 2.1(A1), the output decomposes additively as $z = u_f + u_c$, with $u_c(t)\, y$ tracking only the true label and $u_f(t)\, y$ carrying the shortcut cue's sign on the aligned/conflicting split. Then

$$yz = yu_f + yu_c = \pm u_f(t) + u_c(t),$$

using $y^2 = 1$: every aligned example has $yz = u_f(t) + u_c(t)$ and every conflicting example has $yz = -u_f(t) + u_c(t)$, identically within each group at the current $(u_f(t), u_c(t))$, so the group expectations in the definition of $\Delta(t)$ are exact. Subtracting the conflicting-group expectation from the aligned-group expectation cancels the shared $u_c(t)$ term and doubles the $\pm u_f(t)$ term, giving $\Delta(t) = 2u_f(t)$ exactly. □

C.9 Proof of Convergence Dynamics of GAP-FSO (Proposition 5.2)

Proof of Proposition 5.2 (Dynamics under GAP-FSO). GAP-FSO adds the explicit confidence gap penalty $\mathcal{L}_{\text{aux}} = \gamma\Delta(t)^2 = 4\gamma u_f(t)^2$ directly to the population loss.

Case 1: Squared Loss. Without loss of generality, fix the target class $y = 1$ so the aligned/conflicting population targets are both 1. The total augmented loss functional over the two subpopulations is:

$$\bar{\Phi}(u_f, u_c) = \frac{1}{2}p(u_f + u_c - 1)^2 + \frac{1}{2}(1-p)(-u_f + u_c - 1)^2 + 4\gamma u_f^2.$$

Let $\mu \triangleq 2p - 1 \in (0, 1)$ since $p \in (\frac{1}{2}, 1)$. Expanding and collecting terms:

$$\bar{\Phi}(u_f, u_c) = \frac{1}{2}(1 + 8\gamma)u_f^2 + \frac{1}{2}u_c^2 + \mu u_f u_c - \mu u_f - u_c + \frac{1}{2}.$$

We compute the Hessian matrix $\boldsymbol{H}_{\bar{\Phi}}$ of $\bar{\Phi}$ with respect to (u_f, u_c):

$$\boldsymbol{H}_{\bar{\Phi}} = \begin{pmatrix} \frac{\partial^2\bar{\Phi}}{\partial u_f^2} & \frac{\partial^2\bar{\Phi}}{\partial u_f \partial u_c} \\ \frac{\partial^2\bar{\Phi}}{\partial u_c \partial u_f} & \frac{\partial^2\bar{\Phi}}{\partial u_c^2} \end{pmatrix} = \begin{pmatrix} 1 + 8\gamma & \mu \\ \mu & 1 \end{pmatrix}.$$

The determinant of the Hessian is:

$$\det(\boldsymbol{H}_{\bar{\Phi}}) = (1 + 8\gamma)(1) - \mu^2 = 1 - \mu^2 + 8\gamma.$$

Since $\mu \in (0, 1)$, we have $1 - \mu^2 > 0$, ensuring $\det(\boldsymbol{H}_{\bar{\Phi}}) > 0$ for any $\gamma \geq 0$. Combined with the positive top-left entry $(1 + 8\gamma > 0)$, Sylvester's Criterion guarantees that $\boldsymbol{H}_{\bar{\Phi}} \succ 0$ everywhere. Consequently, $\bar{\Phi}(u_f, u_c)$ is strictly convex and coercive. Setting the gradient $\nabla\bar{\Phi} = \mathbf{0}$:

$$\frac{\partial\bar{\Phi}}{\partial u_f} = (1 + 8\gamma)u_f + \mu u_c - \mu = 0, \tag{29}$$

$$\frac{\partial\bar{\Phi}}{\partial u_c} = u_c + \mu u_f - 1 = 0 \implies u_c = 1 - \mu u_f. \tag{30}$$

Substituting Eq. 30 into Eq. 29:

$$(1 + 8\gamma)u_f + \mu(1 - \mu u_f) - \mu = 0 \implies (1 + 8\gamma - \mu^2)u_f = 0.$$

Since $1 - \mu^2 + 8\gamma > 0$, this forces $u_f^\star = 0$. Substituting $u_f^\star = 0$ into Eq. 30 gives $u_c^\star = 1$. By the chain rule (as in Section C.6's Case 1), the induced dynamics satisfy $\dot{u}_f = -\boldsymbol{K}_{SS}^{(f)}(t)\frac{\partial\bar{\Phi}}{\partial u_f}$ and $\dot{u}_c = -\boldsymbol{K}_{SS}^{(c)}(t)\frac{\partial\bar{\Phi}}{\partial u_c}$ with $\boldsymbol{K}_{SS}^{(f)}(t), \boldsymbol{K}_{SS}^{(c)}(t) \geq 0$ as before, so $\bar{\Phi}$ is a valid Lyapunov function decreasing monotonically along the trajectory. Coercivity of $\bar{\Phi}$ together with the conserved quantities $a_f^2 - b_f^2$, $a_c^2 - b_c^2$ (Section C.6) place the trajectory in a compact set, as required by LaSalle's Invariance Principle. The largest invariant subset where $\frac{d}{dt}\bar{\Phi} = 0$ is exactly the critical point $(u_f^\star, u_c^\star) = (0, 1)$ found above, so gradient flow converges globally to $(u_f(t), u_c(t)) \to (0, 1)$ for any $\gamma \geq 0$.

Corollary C.4 (Exponential Convergence Rate under GAP-FSO, Squared Loss)**.** *Under generic initialization $(a_f(0)^2 \neq b_f(0)^2$ and $a_c(0)^2 \neq b_c(0)^2)$, and writing $\bar{\Phi}^\star \triangleq \bar{\Phi}(u_f^\star, u_c^\star) = \bar{\Phi}(0, 1)$ for the value of $\bar{\Phi}$ at the unique critical point found in Case 1, the augmented objective converges exponentially,*

$$\bar{\Phi}(t) - \bar{\Phi}^\star \;\leq\; \big(\bar{\Phi}(0) - \bar{\Phi}^\star\big)\exp\big(-2\kappa_0\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\, t\big),$$

$$\textit{where} \quad \kappa_0 \triangleq \min\big\{\sigma_f^2|a_f(0)^2 - b_f(0)^2|,\ \sigma_c^2|a_c(0)^2 - b_c(0)^2|\big\} > 0.$$

Here, $\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) = (1 + 4\gamma) - \sqrt{16\gamma^2 + \mu^2}$ is strictly increasing in γ, from $1 - \mu$ at $\gamma = 0$ to a ceiling of 1 as $\gamma \to \infty$. The guaranteed convergence rate therefore strictly improves with γ, saturating at $2\kappa_0$.

Proof. Recall from Case 1 that $\bar{\Phi}$'s Hessian $\boldsymbol{H}_{\bar{\Phi}}$ is constant, and that the induced dynamics satisfy $\dot{u}_f = -\boldsymbol{K}_{SS}^{(f)}(t)\,\partial\bar{\Phi}/\partial u_f$ and $\dot{u}_c = -\boldsymbol{K}_{SS}^{(c)}(t)\,\partial\bar{\Phi}/\partial u_c$.

Step 1: Establishing Lower Bounds for the Kernels. By the sum and difference conservation laws established in Section C.6, we have $\frac{d}{dt}(a_f^2 - b_f^2) = 0$, which implies $a_f(t)^2 - b_f(t)^2 = a_f(0)^2 - b_f(0)^2$ for all $t \geq 0$, and analogously for the causal dimension c.

Applying the algebraic inequality $x^2 + y^2 \geq |x^2 - y^2|$ to $x = a_f(t)$ and $y = b_f(t)$, we bound the shortcut kernel $\boldsymbol{K}_{SS}^{(f)}(t)$ from below:

$$\boldsymbol{K}_{SS}^{(f)}(t) \triangleq \sigma_f^2\big(a_f(t)^2 + b_f(t)^2\big) \geq \sigma_f^2|a_f(t)^2 - b_f(t)^2| = \sigma_f^2|a_f(0)^2 - b_f(0)^2|.$$

Similarly, for the causal kernel $\boldsymbol{K}_{SS}^{(c)}(t)$, we obtain $\boldsymbol{K}_{SS}^{(c)}(t) \geq \sigma_c^2|a_c(0)^2 - b_c(0)^2|$. Under generic initialization ($a_f(0)^2 \neq b_f(0)^2$ and $a_c(0)^2 \neq b_c(0)^2$), both lower bounds are strictly positive. Defining

$$\kappa_0 \triangleq \min\big\{\sigma_f^2|a_f(0)^2 - b_f(0)^2|,\ \sigma_c^2|a_c(0)^2 - b_c(0)^2|\big\} > 0,$$

it follows that $\min\{\boldsymbol{K}_{SS}^{(f)}(t), \boldsymbol{K}_{SS}^{(c)}(t)\} \geq \kappa_0 > 0$ holds uniformly for all $t \geq 0$.

Step 2: Deriving the Polyak–Łojasiewicz (PL) Inequality. Since $\bar{\Phi}$ is a quadratic objective with a constant, strictly positive-definite Hessian $\boldsymbol{H}_{\bar{\Phi}} \succeq \lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\mathbf{I}$, we define the error vector $v \triangleq (u_f, u_c)^\top - (u_f^\star, u_c^\star)^\top$. The gradient and objective suboptimality are given by:

$$\nabla\bar{\Phi} = \boldsymbol{H}_{\bar{\Phi}} v, \qquad \bar{\Phi} - \bar{\Phi}^\star = \frac{1}{2} v^\top \boldsymbol{H}_{\bar{\Phi}} v.$$

Because $\boldsymbol{H}_{\bar{\Phi}}$ and $\boldsymbol{H}_{\bar{\Phi}} - \lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\mathbf{I}$ are simultaneously diagonalizable and both positive semi-definite (PSD), their product is also PSD. This yields:

$$\|\nabla\bar{\Phi}\|^2 = v^\top \boldsymbol{H}_{\bar{\Phi}}^2 v \geq \lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\, v^\top \boldsymbol{H}_{\bar{\Phi}} v = 2\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\big(\bar{\Phi} - \bar{\Phi}^\star\big).$$

Step 3: Differential Inequality and Grönwall Integration. Let $P \triangleq \partial\bar{\Phi}/\partial u_f$ and $Q \triangleq \partial\bar{\Phi}/\partial u_c$, so that $\|\nabla\bar{\Phi}\|^2 = P^2 + Q^2$. The time evolution of the objective function satisfies:

$$\dot{\bar{\Phi}} = -\boldsymbol{K}_{SS}^{(f)}(t)P^2 - \boldsymbol{K}_{SS}^{(c)}(t)Q^2 \leq -\kappa_0(P^2 + Q^2) = -\kappa_0\|\nabla\bar{\Phi}\|^2.$$

Substituting the PL inequality from Step 2 yields:

$$\dot{\bar{\Phi}}(t) \leq -2\kappa_0\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\big(\bar{\Phi}(t) - \bar{\Phi}^\star\big).$$

Applying Grönwall's inequality to $y(t) \triangleq \bar{\Phi}(t) - \bar{\Phi}^\star$ gives the exact exponential rate bound:

$$\bar{\Phi}(t) - \bar{\Phi}^\star \leq \big(\bar{\Phi}(0) - \bar{\Phi}^\star\big)\exp\big(-2\kappa_0\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}})\, t\big).$$

Step 4: Monotonicity with Respect to γ. Differentiating $\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) = (1 + 4\gamma) - \sqrt{16\gamma^2 + \mu^2}$ with respect to γ gives:

$$\frac{d}{d\gamma}\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) = 4 - \frac{16\gamma}{\sqrt{16\gamma^2 + \mu^2}}.$$

For any non-zero shortcut-causal correlation parameter $\mu \neq 0$, we have $\sqrt{16\gamma^2 + \mu^2} > 4\gamma$, which implies $\frac{16\gamma}{\sqrt{16\gamma^2+\mu^2}} < 4$. Hence, $\frac{d}{d\gamma}\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) > 0$, proving that the guaranteed convergence rate is strictly monotonically increasing in γ.

Step 5: Limiting Behavior and Rate Saturation. At $\gamma = 0$ (no GAP-FSO penalty), the rate coefficient reduces to $\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) = 1 - \mu$. As $\gamma \to \infty$, applying a first-order Taylor expansion yields:

$$\sqrt{16\gamma^2 + \mu^2} = 4\gamma\sqrt{1 + \frac{\mu^2}{16\gamma^2}} = 4\gamma + \frac{\mu^2}{8\gamma} + \mathcal{O}(\gamma^{-3}).$$

Taking the limit as $\gamma \to \infty$:

$$\lim_{\gamma\to\infty}\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}}) = \lim_{\gamma\to\infty}\left[(1 + 4\gamma) - \left(4\gamma + \frac{\mu^2}{8\gamma}\right)\right] = 1.$$

Thus, the exponential rate strictly increases with γ and saturates at $2\kappa_0$. □

Extension to Multi-Stratum Data. We extend Case 1's squared-loss argument to J strata sharing the same shortcut parameter u_f: stratum j occupies population fraction π_j ($\sum_j \pi_j = 1$), has its own alignment probability $p_j \in (\frac{1}{2}, 1)$, and its shortcut contribution is $u_f^{(j)} = s_j u_f$ for a stratum-specific sign $s_j \in \{\pm 1\}$ (the same shared parameter u_f entering with a class-dependent sign, exactly as in the CelebA example of Section D.2). The multi-stratum augmented objective directly generalizes Case 1's $\bar{\Phi}(u_f, u_c)$, to which it reduces at $J = 1$:

$$\bar{\Phi}^{(J)}(u_f, u_c) = \sum_{j=1}^{J} \pi_j \Big[\tfrac{1}{2} p_j (s_j u_f + u_c - 1)^2 + \tfrac{1}{2}(1 - p_j)(-s_j u_f + u_c - 1)^2 \Big] + 4\gamma u_f^2 \textstyle\sum_{j=1}^{J} w_j.$$

Expanding each stratum's task-loss term using $s_j^2 = 1$, summing with $\sum_j \pi_j = 1$ and $\sum_j w_j = 1$, gives

$$\bar{\Phi}^{(J)}(u_f, u_c) = \tfrac{1}{2}(1 + 8\gamma) u_f^2 + \tfrac{1}{2} u_c^2 + \bar{\mu}\, u_f u_c - \bar{\mu} u_f - u_c + \tfrac{1}{2}, \qquad \bar{\mu} \triangleq \sum_{j=1}^{J} \pi_j s_j (2p_j - 1),$$

which is algebraically *identical* to Case 1's expanded $\bar{\Phi}(u_f, u_c)$ above, with the single alignment parameter μ replaced by the population-weighted, sign-adjusted average $\bar{\mu}$. Since each $|s_j(2p_j - 1)| = 2p_j - 1 < 1$ and $\bar{\mu}$ is a convex combination of these terms ($\pi_j \geq 0$, $\sum_j \pi_j = 1$), the triangle inequality gives $|\bar{\mu}| \leq \sum_j \pi_j (2p_j - 1) < \sum_j \pi_j = 1$, so $\bar{\mu} \in (-1, 1)$ exactly as μ was. Case 1's entire argument therefore carries over verbatim with $\mu \to \bar{\mu}$: $\boldsymbol{H}_{\bar{\Phi}^{(J)}} = \left(\begin{smallmatrix} 1+8\gamma & \bar{\mu} \\ \bar{\mu} & 1 \end{smallmatrix}\right) \succ 0$ for any $\gamma \geq 0$, the unique critical point remains $(u_f^\star, u_c^\star) = (0, 1)$ regardless of J, $\{\pi_j\}, \{p_j\}, \{s_j\}$, or the weights w_j, and Corollary C.4's exponential-rate guarantee applies unchanged with $\lambda_{\min}(\boldsymbol{H}_{\bar{\Phi}^{(J)}}) = (1 + 4\gamma) - \sqrt{16\gamma^2 + \bar{\mu}^2}$. This confirms the claim of Section D.2: stratification only reweights the effective cross term ($\mu \to \bar{\mu}$), which modulates the guaranteed convergence rate, while the target equilibrium $u_f^{(j),\star} = 0$ for all j is unaffected.

Case 2: Cross-Entropy Loss. Under CE loss, the auxiliary penalty $4\gamma u_f^2$ contributes the additive gradient term $\frac{\partial}{\partial u_f}[4\gamma u_f^2] = 8\gamma u_f$ exclusively to subnetwork f's ODE:

$$\dot{u}_f(t) = \boldsymbol{K}_{SS}^{(f)}(t) \left([pq_+ - (1-p)q_-] - 8\gamma u_f(t) \right), \tag{31}$$

$$\dot{u}_c(t) = \boldsymbol{K}_{SS}^{(c)}(t) [pq_+ + (1-p)q_-]. \tag{32}$$

Because Eq. 32 is identical to the uncorrected CE update, $u_c(t) \to \infty$ holds unconditionally (by the proof of Proposition 3.3). Independently, $u_f(t)$ remains bounded for all t: the task-gradient term $[pq_+ - (1-p)q_-]$ is strictly confined to $(-(1-p), p) \subset (-1, 1)$ regardless of $u_c(t)$'s trajectory. Thus, the bounded forcing coupled with the linear restoring term $-8\gamma u_f(t)$ in Eq. 31 prevents $u_f(t)$ from diverging.

As $u_c(t) \to \infty$ at the rate $u_c(t) = \ln t + \ln \ln t + O(1)$ from Corollary C.3, the task-gradient term satisfies $g(t) \triangleq pq_+ - (1-p)q_- = \Theta\big(e^{-u_c(t)}\big) = \Theta\big(\frac{1}{t \ln t}\big)$, assuming $u_f(t)$ remains bounded. This decay is exponential in u_c, but since $u_c(t)$ itself grows only logarithmically in t, it is merely polynomial, $\Theta(1/(t \ln t))$, in t.

We can write Eq. 31 as

$$\dot{u}_f(t) = \boldsymbol{K}_{SS}^{(f)}(t) \big(g(t) - 8\gamma u_f(t) \big).$$

Under generic (nondegenerate) initialization, $\boldsymbol{K}_{SS}^{(f)}(t) = \sigma_f^2 (a_f(t)^2 + b_f(t)^2)$ is bounded away from zero and converges to a positive constant $K_\infty > 0$ as $u_f(t) \to 0$. Using the integrating factor

$$\mu(t) = \exp\left(8\gamma \int_0^t \boldsymbol{K}_{SS}^{(f)}(s)\, ds \right) \sim e^{8\gamma K_\infty t},$$

the variation-of-parameters solution for $u_f(t)$ is given by

$$u_f(t) = \mu(t)^{-1} \left[u_f(0) + \int_0^t \mu(s)\, \boldsymbol{K}_{SS}^{(f)}(s)\, g(s)\, ds \right].$$

Here, the homogeneous term $\mu(t)^{-1} u_f(0)$ decays exponentially. For the integral term, since $\mu(s)$ grows exponentially while $g(s)$ decreases polynomially, the integrand is dominated by values near $s \approx t$. Applying integration by parts yields

$$\int_0^t \mu(s) K_\infty g(s)\, ds \sim \frac{\mu(t) g(t)}{8\gamma}.$$

Substituting this approximation back into the solution gives the asymptotic behavior as $t \to \infty$:

$$u_f(t) \sim \frac{g(t)}{8\gamma} = \Theta\left(\frac{1}{\gamma\, t \ln t} \right).$$

This indicates that $u_f(t) \to 0$ as $t \to \infty$ for any $\gamma > 0$, showing that shortcut reliance under CE loss vanishes asymptotically at the polynomial rate $\Theta(1/(t \ln t))$, in contrast to the exponential decay rate observed under squared loss in Corollary C.4. This completes the proof. □

D Extended Discussion: Stratification and Failure-Mode Analysis

This section provides the full quantitative results and detailed baseline comparisons for research questions 3, 4 and 5, expanding upon the discussion in Section 7.

D.1 Equivalence Between FSO and GAP-FSO Under Squared vs. Cross-Entropy Loss (RQ3)

Under the additive split $z = u_f + u_c$ (Assumption 2.1), restoring the starved causal gradient (FSO's action on c) and erasing the excess spurious gradient (GAP-FSO's action on f) are equivalent only under squared loss, where both interventions force the unique fixed point to $u_f = 0$ (Propositions 4.1 and 5.2). Under CE loss, this equivalence breaks down: because u_f's fixed point is independent of u_c (Proposition 3.3), accelerating c leaves u_f permanently pinned at offset $\frac{1}{2} \ln \frac{p}{1-p}$ (Proposition 4.1), whereas directly suppressing f shifts this equilibrium to 0 (Proposition 5.2). This reveals a striking role reversal: **FSO acts structurally under squared loss but merely accelerates under CE, whereas GAP-FSO acts structurally under CE but merely accelerates under squared loss.**

D.2 Detailed Impact of Multi-Stratum Optimization (RQ4)

The Pooling Failure Mode under Sign-Heterogeneous Shortcuts. When shortcut reliance varies in sign across J distinct strata (e.g., due to class-opposing attribute correlations), global pooling creates a fundamental failure mode. An unstratified statistic computes an unweighted average over per-class parameters, $\bar{u}_f \approx \frac{1}{2}\big(u_f^{(1)} + u_f^{(2)}\big)$. If a shortcut exhibits opposite orientations across classes, for instance, $u_f^{(1)} = +3$ in one class and $u_f^{(2)} = -3$ in another—the pooled mean vanishes ($\bar{u}_f = 0$). Consequently, signed gaps $\Delta^{(j)}(t)$ with opposing drift directions cancel out, causing the global penalty to detect zero shortcut reliance despite substantial reliance within each individual class.

Multi-Stratum Extension and Theoretical Guarantee. To prevent cross-stratum cancellation, we extend GAP-FSO by penalizing a weighted sum of squared gaps, $\mathcal{L}_{\text{aux}} = \sum_{j=1}^{J} 4\gamma w_j (\Delta^{(j)}(t))^2$ with stratum weights $w_j > 0$ ($\sum w_j = 1$). Summing *squared* terms ensures that $\sum_j w_j \big(u_f^{(j)}\big)^2 = 0$ if and only if $u_f^{(j)} = 0$ for all j (e.g., $w_1(3)^2 + w_2(3)^2 > 0$), preventing opposing strata from neutralizing corrective gradients.

Crucially, this stratification does not compromise our core convergence theory. In Section C.9, we show that the multi-stratum objective across J strata sharing the shortcut parameter u_f remains algebraically identical to Case 1's (the squared-loss case of Proposition 5.2's proof), merely replacing the alignment parameter μ with a population-weighted, sign-adjusted average $\bar{\mu} \in (-1, 1)$. Consequently, the unique equilibrium, Hessian positive-definiteness, and exponential convergence rate (Corollary C.4) carry over verbatim. Thus, dividing the data into finer strata resolves the cancellation problem without invalidating any theoretical guarantees.

Empirical Analysis and Trade-offs (J=1 vs. J=2). This mathematical distinction predicts that stratification ($J > 1$) provides substantial gains whenever sign-heterogeneous shortcut reliance is prevalent, whereas global pooling ($J = 1$) remains lossless when shortcut reliance is nearly uniform.

When shortcut reliance is symmetric across classes, pooling ($J = 1$) is weakly preferable: while theoretical guarantees hold for any J, each stratum's finite-batch estimate $\hat{\Delta}^{(j)}(t)$ in $J > 1$ is computed from fewer examples, introducing estimation noise without adding new information. However, real-world shortcuts are rarely symmetric, and empirical results in Table 3 confirm this prediction:

- On `CelebA`, where the `Heavy_Makeup` correlate is highly asymmetric across underlying classes, moving from $J = 1$ to $J = 2$ yields remarkable improvements in worst-group accuracy (+8.2% under CE loss and +7.8% under squared loss)—the largest swing observed across all hyperparameter ablations.
- On the remaining three datasets with more uniform shortcut distributions, the difference between $J = 1$ and $J = 2$ is marginal and inconsistent in sign, behaving primarily as estimation noise.

Table 3: **GAP-FSO worst-group accuracy, pooled (J=1) vs. stratified (J=2), all four real datasets and both losses.** "CE" refers to CE loss while "MSE" indicates squared loss. Δ is J=2 minus J=1.

Dataset	Loss	J=1	J=2	Δ
Waterbirds	CE	83.5%	84.3%	+0.8
Waterbirds	MSE	81.5%	80.6%	−0.9
Dogs-vs-Cats	CE	98.9%	98.7%	−0.2
Dogs-vs-Cats	MSE	98.5%	98.5%	0.0
CelebA	CE	79.3%	87.5%	+8.2
CelebA	MSE	79.7%	87.5%	+7.8
CivilComments	CE	68.7%	71.1%	+2.4
CivilComments	MSE	71.3%	70.7%	−0.7

Summary. In summary, multi-stratum optimization (J=2) acts as a safeguard against gradient cancellation under sign-heterogeneous shortcuts, fully preserving GAP-FSO's theoretical convergence properties while delivering substantial empirical gains on highly asymmetric real-world benchmarks.

D.3 Comprehensive Failure Mode Analysis and Baseline Comparison (RQ5)

We identify three concrete boundary conditions for GAP-FSO. Rather than treating these limitations as unique to our method, we examine their underlying mechanisms and contrast them with other group-supervised baselines.

1. **Entangled or mislabeled shortcuts:** Proposition 5.1 relies on correct group labels and a near-additive output split; non-additive representational entanglement violates both. Mechanistically, this vulnerability stems from relying on output-level signal separation: every group-supervised baseline (IRM, V-REx, Group DRO, DFR) shares this dependency because their penalty objectives implicitly require the group label to track a distinct, factorized shortcut feature. Only SD circumvents this by avoiding group labels altogether, though at the expense of failing to distinguish the shortcut from the true causal signal.
2. **Discrete optimization instability:** Although continuous-time convergence is guaranteed for any $\gamma > 0$, large values of γ induce discrete stiffness along the shortcut direction (Proposition 5.2), requiring gradient clipping to maintain stability; in practice we found momentum specifically destabilizing at every γ we tried. This instability is driven by a general optimization mechanism: introducing a high-gain auxiliary penalty sharply steepens the local loss curvature, narrowing the stable step-size window for discrete gradient descent. This same landscape ill-conditioning affects IRM and V-REx, which require extreme penalty scaling (reaching up to 10,000) and complex warm-up schedules (Appendix E.2) to avoid collapsing training outright.
3. **Sparse stratum sampling:** As discussed in Appendix D.2, stratification ($J > 1$) degrades when per-stratum sample sizes become too small to yield reliable estimates of $\hat{\Delta}^{(j)}(t)$, so a shortcut whose reliance genuinely varies across many small, sparsely-populated strata is a setting where neither $J = 1$ nor $J = 2$ is likely to be the right answer without further work. This bottleneck arises from a statistical mechanism common to finite-sample estimation: high variance in stratum-level empirical risk estimators propagates directly into the penalty gradient. Group DRO's own minimax reweighting is well known to degrade the same way under small groups, and DFR's group-balanced retraining step similarly needs enough per-group representation to fit a reliable linear head — small, sparsely-populated groups are a shared weak point across every group-supervised method here, not one unique to stratified GAP-FSO.

Beyond these three boundary conditions, our correction targets the shortcut-driving term specifically; whether techniques for accelerating the softmax's own saturating gradient under high confidence (e.g., normalized gradient descent or logit-adjustment methods (Soudry et al., 2018; Hui & Belkin, 2021)) can be combined with GAP-FSO to further speed up overall convergence remains an orthogonal question left to future work.

In summary, GAP-FSO trades the complex representational flexibility of methods like IRM for an exact, hyperparameter-light, and provably convergent correction wherever its structural assumption holds. The analysis above delineates precisely where this trade-off is advantageous (asymmetric, well-populated strata with a clean output-level confound) and where it reaches its limits (entangled representations, or extreme finite-sample/optimization regimes shared with every other group-supervised baseline).

E EXPERIMENTAL SETUP DETAILS

Seed and draw structure. XOR and Colored MNIST results are means over 5 training seeds. Waterbirds, CelebA, and Dogs-vs-Cats results are means over 3 training seeds, on a single fixed dataset construction. CivilComments is the one exception: the raw comment-toxicity labels are severely class-imbalanced, so training requires the group-balanced subsampling of Idrissi et al. (2022), which downsamples the majority class randomly and thereby introduces its own variance on top of the model's training seed. We treat this as a bootstrap problem: CivilComments results are instead means over 10 independent draws of the balanced subsample (train, validation, and test together), each trained with a single fixed model seed. Reported error bars for CivilComments are therefore not directly comparable in kind to the other five datasets' seed-variance error bars, though both are reported the same way (mean $\pm$ std). Hyperparameter tuning for every swept method (Appendix E.2) was conducted via a grid sweep at a single fixed seed.

XOR (LaBonte & Muthukumar, 2026). Binary classification on $x \in \{-1, +1\}^D$, $D = 24$. The label is a pure XOR of two coordinates, $y = -\text{sign}(x_1 x_2)$; the remaining $D - 2 = 22$ coordinates are independent, uninformative noise that the causal subnetwork must search past to find the two that matter. This search difficulty (rather than the XOR nonlinearity in isolation) is the actual source of shortcut vulnerability in this construction. The spurious feature is a separate scalar that agrees with the label with probability $p = 0.98$. Both split (architectural, Table 1c) and no-split (single shared network, Tables 1) conditions use this identical data model. Training is online minibatch SGD, batch size 64, 8000 training examples, 40 epochs.

Colored MNIST (Arjovsky et al., 2019). Binary classification, "is the digit ≥ 5." 6000 training examples, 4000 test examples. The spurious feature is a per-image color summary statistic, correlated with the label with probability $p_{\text{train}} = 0.98$ at training time; the in-distribution test split keeps $p_{\text{test,ID}} = 0.98$, the out-of-distribution split reverses it to $p_{\text{test,OOD}} = 0.1$. In the split (architectural) condition, the causal subnetwork reads only color-invariant luminance (a small MLP) and the spurious subnetwork reads only the color statistic (a single linear unit), combined additively, exactly satisfying the paper's subnetwork assumption with no approximation. The no-split condition instead reads raw pixels through a single shared CNN with no architectural separation at all.

Waterbirds (Sagawa et al., 2020). WILDS' official distribution. Binary classification, waterbird vs. landbird, spuriously correlated with background (water vs. land) at 95% agreement in the training split, the standard construction in this literature. ImageNet-pretrained ResNet-50, fine-tuned end to end, 100 epochs, Adam (lr $= 10^{-5}$), gradient clipping $\|g\| \leq 5$, 3 seeds. The group label used throughout is the official WILDS `background` metadata; no reconstruction is needed.

CelebA (Liu et al., 2015). Binary classification on the attribute `Blond_Hair`, spuriously correlated with a second attribute. We deliberately do *not* use the `male` attribute common in this literature: on this exact data, (blond, male) is only 0.85% of the training split, asymmetric enough that the aligned/conflicting indicator $\mathbb{1}[a = y]$ has its $y = 1$ branch backwards (majority and minority are reversed within that class). ImageNet-pretrained ResNet-50, 10 epochs, same optimizer/clipping/seed settings as Waterbirds.

Dogs-vs-Cats (Byungju et al., 2019). Binary cat-vs-dog classification with a synthetic spurious feature: image darkness (brightness reduced to 0.3). During training, 95% of cats are darkened and 95% of dogs remain bright (5% conflicting per class); during validation and test, this correlation is set to 50%. The ground-truth group assignment is exactly defined by construction. Evaluation uses worst-group accuracy. Architecture: ImageNet-pretrained ResNet-50 trained for 20 epochs, using identical optimizer, clipping, and seed settings as Waterbirds.

CivilComments (Koh et al., 2021). Binary comment toxicity classification. The spurious indicator $s \in \{0, 1\}$ denotes whether any of eight demographic identity attributes (`male`, `female`, `LGBTQ`, `christian`, `muslim`, `other_religions`, `black`, `white`) exceeds 0.5 annotator consensus (Idrissi et al., 2022). Subcategories `LGBTQ` and `other_religions` are aggregated via element-wise maximum over their respective raw sub-identity columns following WILDS conventions. Real training statistics yield $P(s = 1 \mid y = 0) = 0.377$ and $P(s = 1 \mid y = 1) = 0.581$. Architecture: DistilBERT-base-uncased (max sequence length 300 tokens, batch size 32) trained for 20 epochs with the same optimizer and clipping settings as the vision datasets.

E.1 COMPUTE

Table 4 presents the empirical GPU-hour consumption for the Tier-3 real-world datasets (Waterbirds, CelebA, Dogs-vs-Cats, and CivilComments-WILDS across three seeds, four datasets, and two loss functions). All compute metrics were logged programmatically per run. The calibration row reflects the hyperparameter grid sweeps (single seed, reduced epoch budget) used to select the final configurations in Appendix E.2. `DFR`

exhibits substantially lower compute requirements because it reuses the trained ERM backbone (Kirichenko et al., 2023) rather than retraining from scratch. Synthetic datasets (XOR and Colored MNIST) are omitted as each run completes in under one minute, contributing negligibly to the overall footprint.

Table 4: Empirical GPU-hour consumption measured on three NVIDIA RTX PRO 6000 GPUs. (3 seeds $\times$ 4 datasets)

Method	Cross-entropy Loss	Squared Loss	Total GPU Hours
ERM	3.1	3.2	6.3
IRM	3.1	3.0	6.1
V-REx	3.1	3.0	6.1
SD	3.5	3.1	6.6
Group DRO	3.3	3.4	6.7
DFR	0.1	0.1	0.2
GAP-FSO	4.3	3.5	7.8

E.2 Hyperparameter Grid

Every condition shares an identical optimization setup: Adam, learning rate 10^{-5}, gradient clipping $\|g\| \leq 5$, batch size 128 for vision datasets and 32 for CivilComments/DistilBERT, model selection by best validation worst-group accuracy, and three dataset-construction seeds. IRM and V-REx additionally anneal their penalty weight from 1 to its target value over the first 20% of epochs to avoid collapsing training; Group DRO ($\eta = 0.01$) and DFR ($\ell_2 = 1$, 10 restarts) use their originating papers' own defaults and are not swept. Every other method's hyperparameter was chosen from the grid below by a held-out sweep at a reduced epoch budget, selecting, among values within an accuracy floor of the sweep's best point, the smallest average-to-worst-group gap: $\lambda_{\text{IRM}} \in \{1, 10, 100, 1000, 10000\}$, $\beta_{\text{V-REx}} \in \{0.1, 1, 10, 100, 1000\}$, $\lambda_{\text{SD}} \in \{0, 0.001, 0.01, 0.1, 1\}$, and GAP-FSO's γ swept per dataset and reported alongside each result in Tables 1 and 5.

E.3 Full Experimental Results, All Six Datasets

Table 5 collects every headline result in one place: XOR and Colored MNIST (no architectural subnetwork split, so directly comparable against baselines that were never given one either; the split-subnetwork FSO numbers in Table 1c are a separate condition and are not repeated here), plus the four real-world benchmarks. *Test/Mean Acc.* is overall test accuracy; *Worst-Group/OOD Acc.* is worst-group accuracy for the four real datasets and out-of-distribution accuracy under the reversed/shifted test condition for XOR and Colored MNIST, the closest analogue available in that synthetic setup. Oracle (ablated, shortcut signal removed entirely) exists only for XOR and Colored MNIST, where the causal-only subnetwork can be isolated; none of the four real datasets admit an architectural ablation, so no oracle row is reported for them. All results are 5-seed means for XOR/Colored MNIST, 3-seed means for the three real image datasets, and 10-draw means for CivilComments.

Method	Hyperparam.	Test/Mean Acc.	Worst-Group/OOD Acc.
XOR, CE			
ERM	—	100.0%	100.0%
IRM	$\lambda = 1$	100.0%	100.0%
V-REx	$\beta = 2$	100.0%	100.0%
SD	—	100.0%	100.0%
GroupDRO	—	100.0%	100.0%
DFR	—	100.0%	$100.0\% \pm 0.1\%$
GAP-FSO	$\gamma = 0.1$	100.0%	100.0%
Oracle	—	100.0%	100.0%
XOR, MSE			
ERM	—	100.0%	100.0%
IRM	$\lambda = 0.3$	$99.8\% \pm 0.2\%$	$89.5\% \pm 10.7\%$
V-REx	$\beta = 1$	100.0%	100.0%
SD	—	100.0%	100.0%
GroupDRO	—	100.0%	100.0%
DFR	—	100.0%	100.0%
GAP-FSO	$\gamma = 0.25$	100.0%	100.0%
Oracle	—	100.0%	100.0%
Colored MNIST, CE			
ERM	—	98.1%	$19.5\% \pm 5.1\%$
IRM	$\lambda = 1$	$98.5\% \pm 0.4\%$	$77.5\% \pm 3.2\%$
V-REx	$\beta = 2$	$98.1\% \pm 0.1\%$	$80.2\% \pm 3.6\%$
SD	—	97.9%	$10.1\% \pm 0.6\%$
GroupDRO	—	$98.4\% \pm 0.5\%$	$76.3\% \pm 2.8\%$
DFR	—	$84.5\% \pm 3.0\%$	$73.3\% \pm 2.8\%$
GAP-FSO (J=1)	$\gamma = 0.4$	95.8%	$89.3\% \pm 2.9\%$
GAP-FSO (J=2)	$\gamma = 0.4$	95.1%	$87.3\% \pm 3.1\%$
Oracle	—	97.5%	97.5%
Colored MNIST, MSE			
ERM	—	97.9%	$10.0\% \pm 0.6\%$
IRM	$\lambda = 0.3$	$96.0\% \pm 0.7\%$	$87.6\% \pm 2.2\%$
V-REx	$\beta = 1$	$96.7\% \pm 0.7\%$	$87.5\% \pm 2.5\%$
SD	—	97.9%	$10.0\% \pm 0.6\%$
GroupDRO	—	$97.3\% \pm 0.9\%$	$83.9\% \pm 2.1\%$
DFR	—	$79.7\% \pm 2.2\%$	$68.1\% \pm 5.4\%$
GAP-FSO (J=1)	$\gamma = 2.0$	89.8%	$83.4\% \pm 1.7\%$
GAP-FSO (J=2)	$\gamma = 2.0$	91.1%	$82.1\% \pm 2.5\%$
Oracle	—	96.9%	96.9%
Waterbirds, CE			
ERM	—	$89.86\% \pm 1.26\%$	$78.41\% \pm 1.51\%$
IRM	1	$91.82\% \pm 0.52\%$	$84.26\% \pm 1.27\%$
V-REx	1	$91.81\% \pm 0.73\%$	$83.96\% \pm 1.09\%$
SD	1	$92.21\% \pm 0.22\%$	$78.71\% \pm 1.04\%$
GroupDRO	—	$91.77\% \pm 0.94\%$	$85.33\% \pm 0.71\%$
DFR	—	$92.03\% \pm 0.54\%$	$86.28\% \pm 1.17\%$
GAP-FSO (J=1)	1	$94.14\% \pm 0.40\%$	$83.49\% \pm 2.70\%$
GAP-FSO (J=2)	0.5	$92.90\% \pm 0.91\%$	$84.32\% \pm 0.74\%$
Waterbirds, MSE			
ERM	—	$91.55\% \pm 0.72\%$	$77.57\% \pm 1.77\%$

continued on next page

Method	Hyperparam.	Test/Mean Acc.	Worst-Group/OOD Acc.
IRM	1	$93.93\% \pm 0.55\%$	$81.10\% \pm 0.63\%$
V-REx	1	$93.61\% \pm 0.45\%$	$81.26\% \pm 1.34\%$
SD	0.1	$91.40\% \pm 0.70\%$	$77.57\% \pm 1.53\%$
GroupDRO	—	$93.27\% \pm 0.38\%$	$80.43\% \pm 1.54\%$
DFR	—	$90.21\% \pm 0.68\%$	$82.45\% \pm 1.25\%$
GAP-FSO (J=1)	5	$94.32\% \pm 0.73\%$	$81.52\% \pm 1.30\%$
GAP-FSO (J=2)	4	$93.10\% \pm 0.59\%$	$80.63\% \pm 0.63\%$
Dogs-vs-Cats, CE			
ERM	—	$98.31\% \pm 0.16\%$	$96.74\% \pm 0.16\%$
IRM	100	$99.17\% \pm 0.05\%$	$98.84\% \pm 0.17\%$
V-REx	10	$99.25\% \pm 0.06\%$	$98.94\% \pm 0.06\%$
SD	1	$98.33\% \pm 0.16\%$	$96.78\% \pm 0.44\%$
GroupDRO	—	$99.17\% \pm 0.10\%$	$98.89\% \pm 0.13\%$
DFR	—	$98.68\% \pm 0.14\%$	$97.60\% \pm 0.09\%$
GAP-FSO (J=1)	3	$99.27\% \pm 0.10\%$	$98.89\% \pm 0.13\%$
GAP-FSO (J=2)	4	$99.23\% \pm 0.05\%$	$98.65\% \pm 0.18\%$
Dogs-vs-Cats, MSE			
ERM	—	$98.41\% \pm 0.32\%$	$96.38\% \pm 0.80\%$
IRM	1	$99.07\% \pm 0.10\%$	$98.55\% \pm 0.23\%$
V-REx	1000	$98.92\% \pm 0.11\%$	$98.29\% \pm 0.31\%$
SD	0.001	$98.43\% \pm 0.23\%$	$96.33\% \pm 0.73\%$
GroupDRO	—	$99.11\% \pm 0.08\%$	$98.55\% \pm 0.18\%$
DFR	—	$98.72\% \pm 0.14\%$	$97.57\% \pm 0.32\%$
GAP-FSO (J=1)	12	$99.11\% \pm 0.16\%$	$98.50\% \pm 0.36\%$
GAP-FSO (J=2)	128	$99.07\% \pm 0.08\%$	$98.50\% \pm 0.09\%$
CelebA (Heavy_Makeup), CE			
ERM	—	$95.80\% \pm 0.09\%$	$76.33\% \pm 1.12\%$
IRM	100	$92.84\% \pm 0.63\%$	$90.88\% \pm 0.29\%$
V-REx	100	$95.10\% \pm 0.21\%$	$83.27\% \pm 1.35\%$
SD	0.01	$95.66\% \pm 0.22\%$	$75.40\% \pm 2.01\%$
GroupDRO	—	$95.40\% \pm 0.06\%$	$82.38\% \pm 0.74\%$
DFR	—	$90.86\% \pm 0.54\%$	$89.42\% \pm 0.80\%$
GAP-FSO (J=1)	20	$95.06\% \pm 0.14\%$	$79.34\% \pm 1.75\%$
GAP-FSO (J=2)	4	$94.97\% \pm 0.13\%$	$87.54\% \pm 0.56\%$
CelebA (Heavy_Makeup), MSE			
ERM	—	$95.93\% \pm 0.03\%$	$75.73\% \pm 0.98\%$
IRM	100	$94.82\% \pm 0.18\%$	$79.45\% \pm 0.41\%$
V-REx	10	$95.47\% \pm 0.07\%$	$81.88\% \pm 1.09\%$
SD	1	$95.80\% \pm 0.11\%$	$75.62\% \pm 2.58\%$
GroupDRO	—	$95.30\% \pm 0.12\%$	$83.03\% \pm 0.49\%$
DFR	—	$91.57\% \pm 0.24\%$	$90.39\% \pm 0.52\%$
GAP-FSO (J=1)	8	$95.67\% \pm 0.05\%$	$79.74\% \pm 1.73\%$
GAP-FSO (J=2)	16	$94.80\% \pm 0.28\%$	$87.54\% \pm 1.71\%$
CivilComments (balanced), CE			
ERM	—	$73.04\% \pm 1.55\%$	$45.73\% \pm 4.78\%$
IRM	1	$79.59\% \pm 1.50\%$	$64.80\% \pm 5.58\%$
V-REx	0.1	$79.62\% \pm 1.56\%$	$65.08\% \pm 3.60\%$
SD	0.001	$73.66\% \pm 1.94\%$	$47.12\% \pm 5.61\%$
GroupDRO	—	$80.37\% \pm 1.66\%$	$69.35\% \pm 2.63\%$

continued on next page

Method	Hyperparam.	Test/Mean Acc.	Worst-Group/OOD Acc.
DFR	—	$80.86\% \pm 0.84\%$	$65.92\% \pm 4.00\%$
GAP-FSO (J=1)	1	$80.23\% \pm 2.16\%$	$68.67\% \pm 6.18\%$
GAP-FSO (J=2)	48	$79.67\% \pm 2.17\%$	$71.10\% \pm 5.10\%$
CivilComments (balanced), MSE			
ERM	—	$73.07\% \pm 1.62\%$	$45.10\% \pm 2.96\%$
IRM	10000	$80.86\% \pm 1.44\%$	$67.70\% \pm 6.40\%$
V-REx	0.1	$80.72\% \pm 1.62\%$	$67.67\% \pm 6.02\%$
SD	0.1	$73.05\% \pm 1.85\%$	$43.90\% \pm 5.92\%$
GroupDRO	—	$80.91\% \pm 1.69\%$	$70.55\% \pm 5.27\%$
DFR	—	$80.30\% \pm 0.94\%$	$66.38\% \pm 2.42\%$
GAP-FSO (J=1)	20	$80.41\% \pm 1.67\%$	$71.33\% \pm 3.93\%$
GAP-FSO (J=2)	128	$79.71\% \pm 2.05\%$	$70.65\% \pm 2.65\%$

Table 5: **Full experimental results across all six datasets.** GAP-FSO (J=1) denotes the pooled confidence-gap penalty, and GAP-FSO (J=2) denotes its per-class stratified generalization (Section 5).

F Loss Curves and Rayleigh Activity

This section presents empirical trajectories of training losses, Rayleigh activity dynamics, and parameter evolution to substantiate the theoretical findings in the main text.

F.1 Training Loss Invisibility Across Baselines

Fig. 5 and 6 display the training loss trajectories of all baselines and both GAP-FSO variants across all six datasets under CE and MSE loss, respectively.

Across every benchmark and loss function, the empirical curves reinforce the central message of Fig. 1: methods with vastly divergent final worst-group accuracy (Table 1) exhibit training loss profiles that are nearly indistinguishable throughout training. While the loss curve confirms that numerical optimization is proceeding smoothly, it provides no diagnostic signal regarding feature reliance or shortcut exploitation.

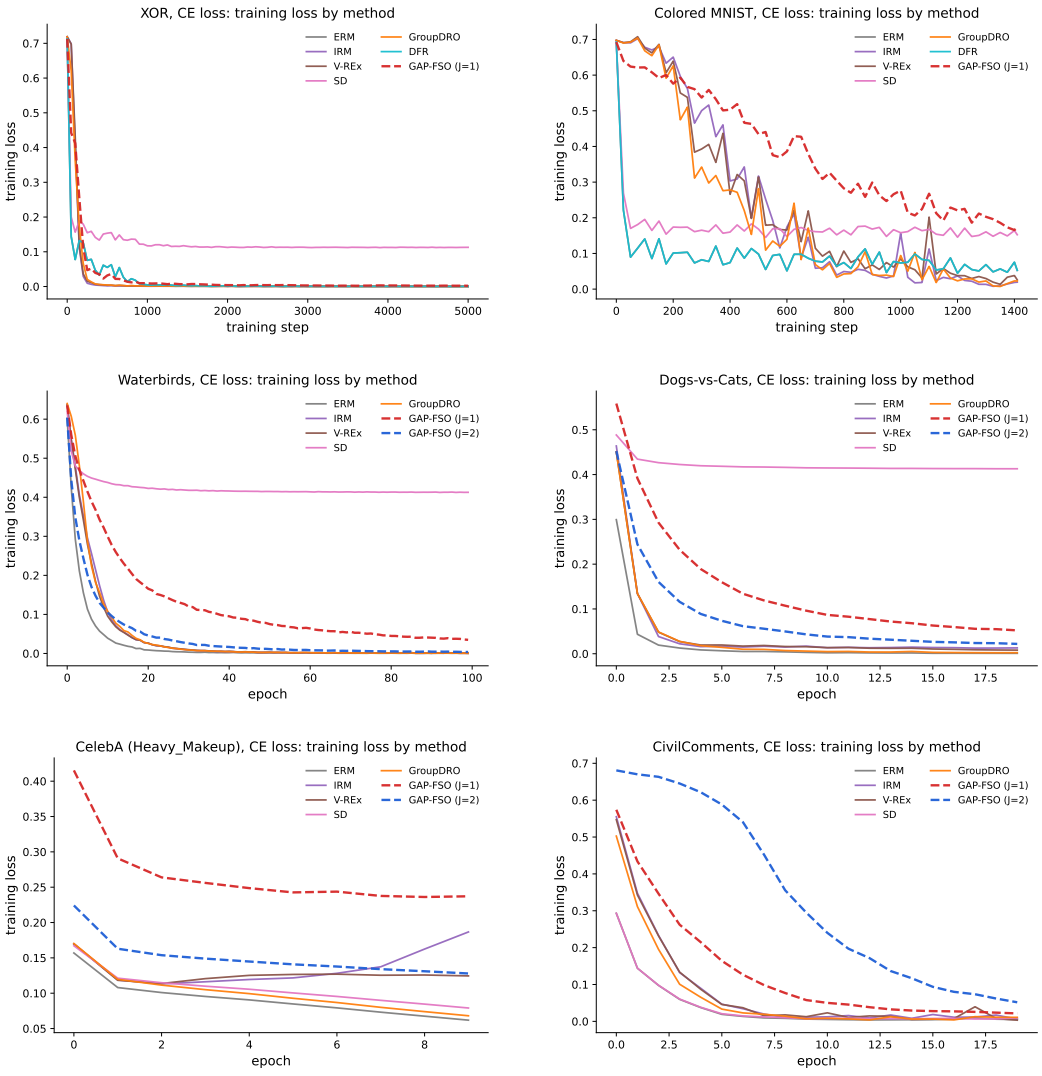


Figure 5: Training loss by method, CE loss, all six datasets (no architectural subnetwork split; same runs as Table 1).

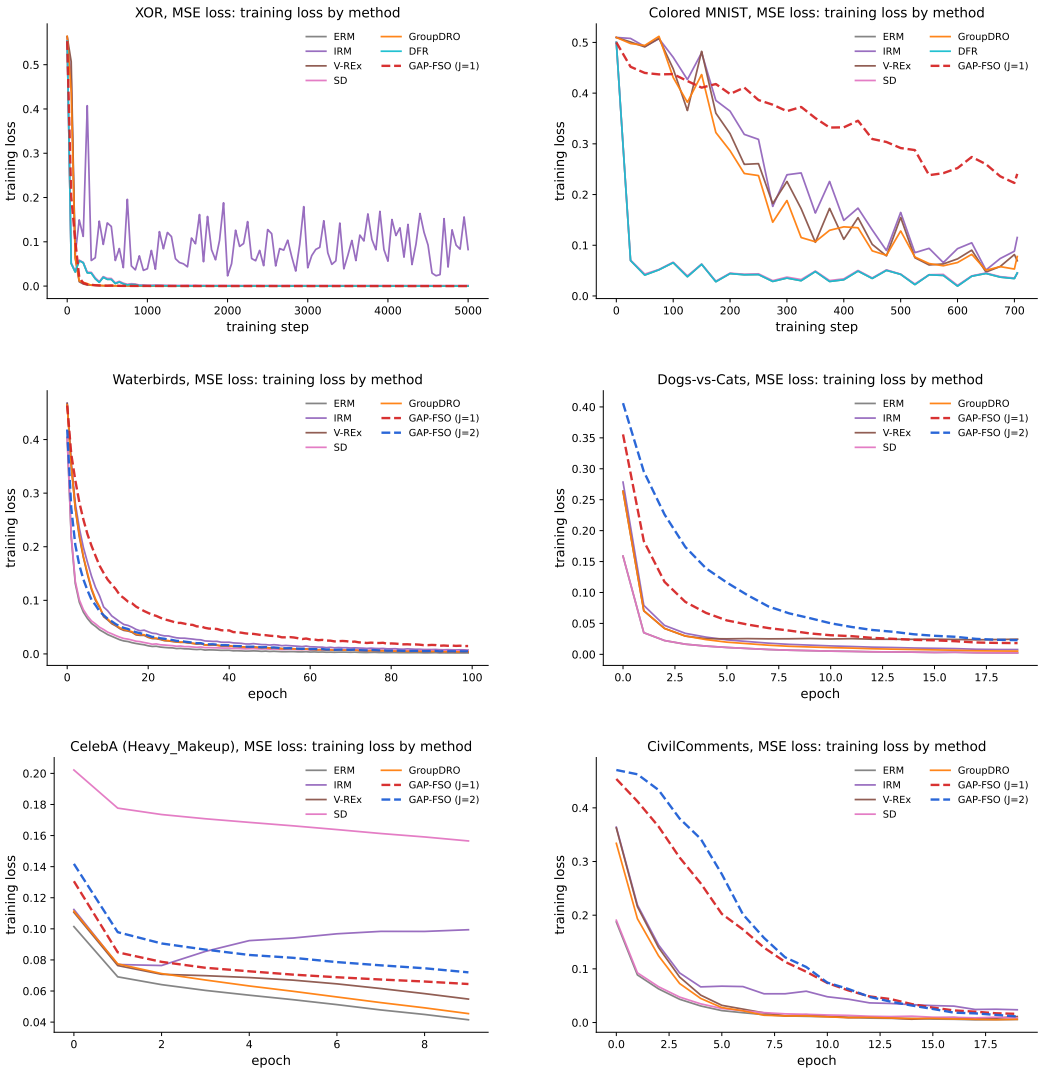
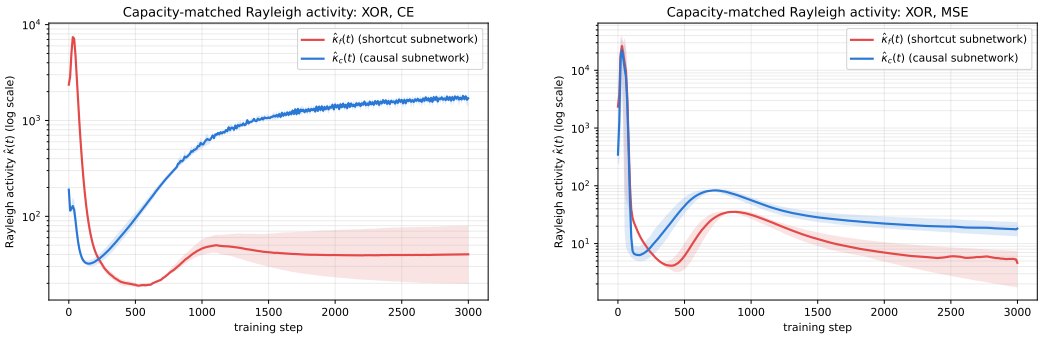


Figure 6: Training loss by method, MSE loss, all six datasets.

F.2 Rayleigh Activity Dynamics: Early Starvation and Loss Divergence

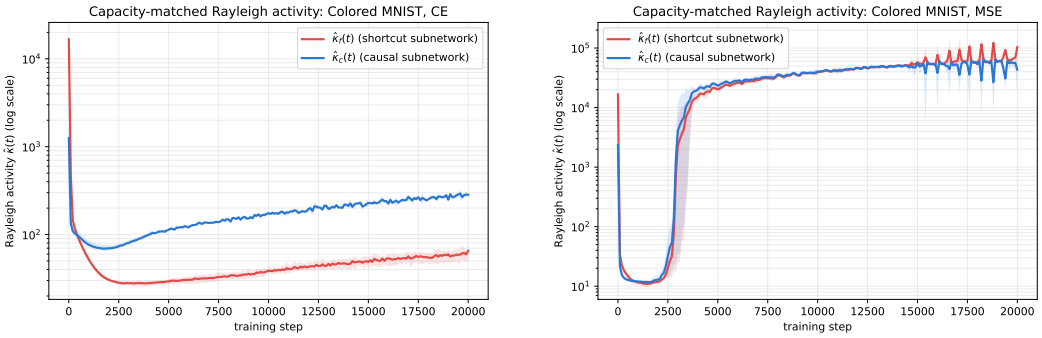
To uncover what standard loss curves conceal, Fig. 7 tracks the trajectory-realized Rayleigh activities $\hat{\kappa}_f(t)$ and $\hat{\kappa}_c(t)$ directly on the XOR and Colored MNIST benchmarks under both loss regimes. This complements the single panel in Fig. 1b by providing a complete dataset $\times$ loss cross-comparison. Two dynamical patterns emerge:

- **Early Shortcut Dominance:** Across all datasets and loss functions, the shortcut subnetwork's activity $\hat{\kappa}_f(t)$ dominates the initial training phase. This validates Theorem 3.1: the subnetwork associated with the larger tangent-kernel eigenvalue rapidly absorbs the shared error signal.
- **Late-Stage Divergence across Losses:** Under CE loss (Fig. 7a,c), $\hat{\kappa}_c(t)$ steadily surpasses $\hat{\kappa}_f(t)$ and remains dominant, with $\hat{\kappa}_f(t)$ settling an order of magnitude lower (confirming the asymptotic regime in Proposition 3.3). Under MSE loss, the initial shortcut spike decays as its error signal contribution is absorbed, allowing the causal subnetwork to govern the remaining dissipation (observed cleanly on XOR, Fig. 7b). Early shortcut dominance also induces a subsequent shared V- or U-shaped trough: as the shortcut circuit rapidly consumes the initial error signal before the causal circuit reorganizes to exploit the remainder, both gradient signals temporarily quiet down. This trough represents a transient hand-off gap rather than an independent mechanism. Colored MNIST under MSE loss (Fig. 7d) exhibits noisy, comparable-magnitude oscillations beyond step $\gtrsim 14000$. This late-stage oscillation does not indicate optimization failure: training loss decreases monotonically below 0.003, ruling out numerical instability from a vanishing denominator. Because the overparameterized network achieves 100% training accuracy by step ≈ 5000, MSE loss admits a continuum of zero-loss parameter configurations. The eNTK continues to evolve along this near-flat manifold after loss plateauing, where transient reorganization appears as fluctuations in the $\hat{\kappa}_f/\hat{\kappa}_c$ ratio.



(a) XOR, CE

(b) XOR, MSE



(c) Colored MNIST, CE

(d) Colored MNIST, MSE

Figure 7: **Realized Rayleigh activity $\hat{\kappa}_f(t)$ (shortcut) vs. $\hat{\kappa}_c(t)$ (causal) under a capacity-matched construction across both datasets and losses.** Subnetworks share an identical Linear→ReLU→Linear architecture to isolate dynamics from parameter count. Inputs are dimension-matched per dataset: for XOR, 24-dimensional vectors containing either the spurious scalar or the causal feature pair padded with independent noise; for Colored MNIST, 784-dimensional pixel-wise color-difference or luminance images.