

Targeted Behavioral Repair with Contaminated Anchors

Kasra Fouladi

Iran University of Science and Technology (IUST)

k4sr405@gmail.com

Github: <https://github.com/kasrafouladi/StrikeForce-No-BPTT-via-FAP>

September 2026

Domain: Continual Learning, Model Repair, Imitation Learning

Keywords: elastic weight consolidation, knowledge distillation, catastrophic forgetting, targeted correction, contaminated anchors

1 Abstract

We study the problem of correcting a specific, identified behavioral failure in a trained model without retraining from scratch and without full access to the original training distribution. Given a competent base model θ_0 and a small corrective dataset D_{fix} exhibiting the failure, the standard remedy for avoiding collateral damage elsewhere is to regularize the update toward θ_0 using an anchor set D_{anc} sampled from the original distribution. We highlight and formalize a complication that this standard recipe does not address: the anchor set is not guaranteed to be failure-free. When the failure region has nonzero mass under the original distribution, D_{anc} may itself contain instances of the very behavior being corrected, so a preservation term applied indiscriminately can directly oppose the corrective signal. We combine parameter-space preservation (Elastic Weight Consolidation) with output-space preservation (Knowledge Distillation) into a single objective, characterize the repair-preservation trade-off this creates as a function of failure prevalence ρ and dataset sizes, and analyze the resulting multi-step dynamics locally. The analysis shows that, under standard local assumptions, repeated updates converge to a curvature-weighted repair displacement determined by the preservation terms, while anchor contamination perturbs that equilibrium continuously, with the empirical estimate indicating a non-negligible contamination rate. We describe a concrete instantiation used to repair a behavioral-cloning policy trained on the StrikeForce environment.

2 Introduction

2.1 Problem

A model θ_0 that is already competent on its original task can still exhibit a specific, identifiable failure mode on a subset of the input space. Retraining from scratch is often infeasible: the original training set may no longer be fully available, or reacquiring it is costly relative to simply collecting examples of the failure itself.

2.2 Why Naive Fine-Tuning Fails

Fine-tuning θ_0 directly on failure examples D_{fix} corrects the targeted behavior but risks catastrophic forgetting [1]: nothing in the objective discourages the update from disrupting behavior the model already had right. The standard fix is to add a preservation term computed on an anchor set drawn from the original distribution, whether at the parameter level (EWC) or the output level (KD) [1, 2, 3].

2.3 The Contaminated-Anchor Complication

This standard recipe implicitly assumes the anchor set represents only desirable behavior. That assumption is not free: if the failure region $\mathcal{F}$ has nonzero probability $\rho = P_{\text{orig}}(x \in \mathcal{F})$ under the original distribution, an anchor set sampled from that same distribution will, in expectation, contain a ρ -fraction of failure examples. On those examples, the preservation term explicitly pulls the model back toward the very behavior the corrective term is trying to remove. We refer to this as the **contaminated-anchor** setting and treat it as a first-class part of the problem rather than an edge case to be assumed away.

2.4 Our Approach

We combine EWC (parameter-space preservation, weighted by Fisher-style sensitivity) and KD (output-space preservation through temperature-scaled softmax agreement with θ_0) into a single joint objective (Section 4). We then analyze the resulting multi-step repair dynamics and the effect of anchor contamination on the local preservation geometry (Section 6), followed by the practical repair-preservation trade-off in terms of ρ , $|D_{\text{fix}}|$, $|D_{\text{anc}}|$, and the regularization strengths λ_1, λ_2 (Section 7). We do not claim the contamination problem is solved outright; rather, we make its effect explicit so that the method’s behavior under it is predictable rather than silently absorbed into noise.

2.5 Summary of Contributions

- A formalization of the contaminated-anchor setting for behavioral repair, distinguishing it from standard continual-learning preservation, which typically assumes clean task boundaries.
- A combined EWC+KD objective (Section 4) with an explicit account of how the corrective and preservation gradients interact on contaminated anchor points (Section 5).
- A local multi-step analysis of the repair dynamics, showing that the joint update converges, under standard local assumptions, to a curvature-weighted repair displacement determined by the preservation geometry (Section 6).
- A characterization of how anchor contamination perturbs that preservation geometry and therefore the resulting repair point as a function of failure prevalence ρ (Section 6).
- A concrete implementation and repair pipeline applied to a behavioral-cloning policy for the StrikeForce environment (Section 8), with pre-repair (θ_0), post-repair (θ^*), and naive fine-tuning (θ_{FT}) checkpoints evaluated over $n = 64$ paired episodes on base, derived, and composite engagement metrics (Section 11).

3 Problem Setting

Let $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^K$ be a trained model with parameters θ . We assume an initial model θ_0 is already competent on the original task distribution P_{orig} , but exhibits a specific behavioral failure on a subset $\mathcal{F} \subseteq \mathcal{X}$. The goal is not to retrain from scratch, but to modify θ_0 so that the undesired behavior is corrected while preserving the model’s behavior outside $\mathcal{F}$.

We assume access to:

- a **corrective dataset** $D_{\text{fix}} = \{(x_i, y_i)\}_{i=1}^{n_f}$, containing examples of the undesired behavior with corrected labels;
- an **anchor dataset** $D_{\text{anc}} = \{x_j\}_{j=1}^{n_a}$, sampled from P_{orig} ;
- the original parameters θ_0 and the original model’s outputs on the anchor examples.

The original training set itself is assumed unavailable or too costly to reacquire, a setting also studied in continual-learning work such as Learning without Forgetting [3]. The operational criterion used to construct D_{fix} is described in Section 8.3.

Definition 1 (Contaminated-Anchor Assumption). *Let $\rho = P_{\text{orig}}(x \in \mathcal{F}) > 0$ denote the failure prevalence under the original distribution. Since D_{anc} is sampled from P_{orig} , ρ is equivalently the expected failure frequency within D_{anc} ; it is not the failure frequency within the corrective set D_{fix} , which is failure-dominated by construction: every frame in D_{fix} is an expert action recorded during an intervention.*

$$P_{\text{orig}} = (1 - \rho) P_{\text{good}} + \rho P_{\text{fail}},$$

and is therefore not assumed failure-free: $D_{\text{anc}} \cap \mathcal{F} \neq \emptyset$ in general.

This creates the central tension of the problem: a preservation mechanism applied indiscriminately to all anchor samples can oppose the desired repair.

4 Combined Objective

We seek $\theta^* \approx \theta_0$ in the sense of preserving general behavior, while improving performance on $\mathcal{F}$. The update $\Delta\theta = \theta^* - \theta_0$ should be targeted rather than globally disruptive.

4.1 Corrective Term

$$L_{\text{fix}}(\theta) = \frac{1}{n_f} \sum_{(x_i, y_i) \in D_{\text{fix}}} \ell(f_\theta(x_i), y_i).$$

D_{fix} is not intended to represent the full original distribution; it supplies a targeted signal for the specific missing or incorrect behavior.

Choice of ℓ and the detached focal weight. We use a focal-style per-example loss

$$\ell(f_\theta(x), y) = -(1 - p_t(x; \theta))^\gamma \log p_t(x; \theta),$$

where $p_t(x; \theta)$ is the model’s softmax probability on the target class and $\gamma \geq 0$ is the focal parameter. Throughout the analysis we treat the modulating factor $w(x; \theta) = (1 - p_t(x; \theta))^\gamma$ as a *detached* per-example weight: it is evaluated at the current parameters and is treated as a constant with respect to differentiation. This is a standard implementation choice for focal-style losses and is used here so

that the per-example loss reduces to a weighted cross-entropy with a parameter-dependent weight. Under this convention the per-example gradient with respect to the model logit is

$$\frac{\partial \ell}{\partial z} = w(x; \theta) (p_t(x; \theta) - 1),$$

and the corresponding per-example second derivative is

$$\frac{\partial^2 \ell}{\partial z^2} = w(x; \theta) p_t(x; \theta) (1 - p_t(x; \theta)) \geq 0.$$

This non-negativity is what makes the corrective term locally convex in the model’s own output, a property we use in Section 6.

4.2 Parameter-Space Preservation (EWC)

For each anchor example x_j , let $\ell_{\text{anc}}(x_j; \theta_0)$ denote the per-example loss used to characterize the original model’s behavior on that example. In the current implementation, the diagonal Fisher-style coefficient for parameter θ_i is estimated by accumulating the squared per-sample gradients and averaging over the anchor set:

$$\hat{F}_i = \frac{1}{n_a} \sum_{j=1}^{n_a} \left(\frac{\partial \ell_{\text{anc}}(x_j; \theta_0)}{\partial \theta_i} \right)^2.$$

Thus $\hat{F}_i$ is the mean squared sensitivity of parameter θ_i to the anchor-set loss. When the per-example loss is a negative log-likelihood under the usual Fisher construction, this is an empirical diagonal Fisher approximation; more generally, it should be interpreted as a squared-gradient sensitivity statistic. We use $\hat{F}_i$ in the EWC penalty

$$L_{\text{EWC}}(\theta) = \frac{1}{2} \sum_i \hat{F}_i (\theta_i - \theta_{0,i})^2.$$

Parameters with larger $\hat{F}_i$ therefore incur a larger cost for moving away from $\theta_{0,i}$ [1].

4.3 Output-Space Preservation (KD)

For anchor input x_j , let $z_0(x_j) = f_{\theta_0}(x_j)$ and $z_\theta(x_j) = f_\theta(x_j)$. With temperature T ,

$$p_0^T(x_j) = \text{softmax}\left(\frac{z_0(x_j)}{T}\right), \quad p_\theta^T(x_j) = \text{softmax}\left(\frac{z_\theta(x_j)}{T}\right),$$

and we use a forward-KL output-preservation objective,

$$L_{\text{KD}}(\theta) = \frac{1}{n_a} \sum_{j=1}^{n_a} \text{KL}(p_\theta^T(x_j) \| p_0^T(x_j)) = \frac{1}{n_a} \sum_{j=1}^{n_a} \sum_{k=1}^K p_{\theta,k}^T(x_j) \log \frac{p_{\theta,k}^T(x_j)}{p_{0,k}^T(x_j)}.$$

This is a KL-based behavioral anchoring mechanism related to the distillation literature [2, 4]; the present formulation specifically uses the positive forward KL implemented by the repair procedure.

4.4 Combined Objective

$$L_{\text{total}}(\theta) = L_{\text{fix}}(\theta) + \lambda_1 L_{\text{EWC}}(\theta) + \lambda_2 L_{\text{KD}}(\theta), \quad \lambda_1, \lambda_2 \geq 0.$$

5 The Conflict Induced by Anchor Contamination

If every anchor example represented desirable behavior, $\nabla_{\theta} L_{\text{EWC}}$ and $\nabla_{\theta} L_{\text{KD}}$ would generally align with the goal of behavioral preservation without opposing repair. Under Definition 1, this no longer holds unconditionally.

For $x \in D_{\text{anc}} \cap \mathcal{F}$, the reference output $f_{\theta_0}(x)$ is precisely the behavior the corrective term seeks to change. The KD gradient at such a point pulls $p_{\theta}^T(x)$ toward $p_0^T(x)$, i.e., back toward the failure behavior, directly opposing

$$\nabla_{\theta} L_{\text{fix}} \quad \text{vs.} \quad \nabla_{\theta} L_{\text{KD}}.$$

The method does not attempt to preserve θ_0 's behavior everywhere; it seeks an update in which the corrective signal is strong enough to overcome preservation pressure specifically on $\mathcal{F}$, while preservation remains effective on the (typically much larger) non-failure portion of the anchor set.

Remark 1. *This is an asymmetry, not a paradox: D_{fix} is concentrated on $\mathcal{F}$ by construction, while the contaminated fraction of D_{anc} is only $\rho \cdot n_a$ examples in expectation. The repair signal on $\mathcal{F}$ can dominate provided $|D_{\text{fix}}|$ and its effective weighting are large relative to $\lambda_2 \rho n_a$. Because the two gradient fields are architecture-, data-, and loss-dependent, we do not claim a closed-form sufficient condition for dominance; the present paper treats this dependence as a qualitative trade-off to be evaluated empirically rather than as a universal analytical guarantee.*

6 Local Multi-Step Repair Dynamics

The key theoretical question is not whether the corrective gradient dominates on the first optimization step, but where the sequence of updates moves after many steps. Define the preservation objective

$$R(\theta) = \lambda_1 L_{\text{EWC}}(\theta) + \lambda_2 L_{\text{KD}}(\theta),$$

so that

$$L_{\text{total}}(\theta) = L_{\text{fix}}(\theta) + R(\theta).$$

The repair procedure follows

$$\theta_{t+1} = \theta_t - \eta [\nabla L_{\text{fix}}(\theta_t) + \nabla R(\theta_t)].$$

6.1 Local Curvature of the Corrective and Preservation Terms

We work with the second-order expansion of L_{total} around θ_0 . Let

$$H_F = \nabla^2 L_{\text{fix}}(\theta_0), \quad H_R = \nabla^2 R(\theta_0), \quad H_{\text{total}} = H_F + H_R,$$

denote the corrective, preservation, and total Hessians at the original parameters, and let

$$g_F = \nabla L_{\text{fix}}(\theta_0).$$

Because both preservation terms are minimized at θ_0 , we have $R(\theta_0) = 0$ and $\nabla R(\theta_0) = 0$, so g_F is the only first-order term in the expansion.

Corrective curvature. By Section 4.1, the detached focal weight makes the per-example second derivative in the target logit non-negative. For a multi-class softmax output $p(x; \theta) \in \Delta^{K-1}$, the per-example Hessian of $\tilde{L}_{\text{fix}}$ with respect to the logits is

$$C_F(x) = w(x; \theta_0) \left[\text{Diag}(p(x; \theta_0)) - p(x; \theta_0) p(x; \theta_0)^\top \right] \succeq 0,$$

which is the standard softmax Fisher matrix scaled by the detached focal weight $w(x; \theta_0) \geq 0$. Writing $J_F(x)$ for the Jacobian of the logits with respect to θ at θ_0 , the parameter-space Hessian of the corrective term is

$$H_F = \frac{1}{n_f} \sum_{i=1}^{n_f} J_F(x_i)^\top C_F(x_i) J_F(x_i) \succeq 0.$$

Thus, under the detached-focal convention, the corrective term is locally convex in the parameters up to the rank structure of the Jacobians.

Preservation curvature. For the EWC term,

$$H_{\text{EWC}} = \text{diag}(\hat{F}_1, \dots, \hat{F}_d),$$

so the mean-squared per-sample gradient estimate enters the local preservation geometry directly as diagonal curvature.

For the KL term, let J_j be the Jacobian of the model logits with respect to θ at θ_0 for anchor x_j , and define

$$C_j^T = \frac{1}{T^2} \left[\text{Diag}(p_0^T(x_j)) - p_0^T(x_j) p_0^T(x_j)^\top \right].$$

At θ_0 , the Hessian of the KL term with respect to parameters is

$$H_{\text{KD}} = \frac{1}{n_a} \sum_{j=1}^{n_a} J_j^\top C_j^T J_j \succeq 0.$$

Therefore,

$$H_R = \lambda_1 \text{diag}(\hat{F}) + \lambda_2 H_{\text{KD}} \succeq 0.$$

Total curvature. Combining the two, locally at θ_0 ,

$$\boxed{H_{\text{total}} = H_F + \lambda_1 \text{diag}(\hat{F}) + \lambda_2 H_{\text{KD}} + \gamma I \succ 0}$$

where $\gamma > 0$ is a small fixed damping constant. Absorbing γI into the definition of H_{total} ensures positive definiteness on the full parameter space and makes the equilibrium $\Delta^* = -H_{\text{total}}^{-1} g_F$ well defined without a separate case distinction.

The corrective term contributes curvature that locally at θ_0 reflects where the model's own output is still uncertain on the corrective set, the EWC component penalizes parameter-space displacement according to squared gradient sensitivity, and the KD component penalizes displacement that changes the model outputs on the anchors. We assume throughout the analysis that, with the damping γI included, H_{total} has eigenvalues in $[\mu, L]$ on the subspace $\mathcal{S}$ spanned by the optimization trajectory, for some $0 < \mu \leq L$. Since $\gamma > 0$, the lower bound $\mu \geq \gamma$ holds automatically and the equilibrium below is well defined.

6.2 Multi-Step Dynamics

Within a sufficiently small neighborhood of θ_0 , retaining the quadratic terms of both the corrective and preservation objectives gives the local surrogate

$$L_{\text{local}}(\Delta) = L_{\text{total}}(\theta_0) + g_F^\top \Delta + \frac{1}{2} \Delta^\top H_{\text{total}} \Delta, \quad \Delta = \theta - \theta_0.$$

The corresponding multi-step dynamics are

$$\Delta_{t+1} = (I - \eta H_{\text{total}}) \Delta_t - \eta g_F.$$

The corrective gradient g_F drives the model away from θ_0 , while the total curvature H_{total} grows the restoring force with displacement. Both the preservation terms and the local curvature of the corrective term contribute to this restoring effect; the focal-detached weight w sets the scale of the corrective contribution.

Proposition 1 (Local Equilibrium of the Repair Dynamics). *With H_{total} defined as in Section 6.1 (including the damping γI), let its eigenvalues on the subspace $\mathcal{S}$ containing the optimization trajectory lie in $[\mu, L]$, where $0 < \mu \leq L$, and suppose the local surrogate above is valid. If*

$$0 < \eta < \frac{2}{L},$$

then gradient descent on the local surrogate converges linearly to the unique stationary point

$$\Delta^* = -H_{\text{total}}^{-1} g_F$$

on $\mathcal{S}$. Moreover,

$$\|\Delta_t - \Delta^*\| \leq q^t \|\Delta_0 - \Delta^*\|, \quad q = \max_{h \in [\mu, L]} |1 - \eta h| < 1.$$

Proof. The stationary-point equation of the local surrogate is

$$g_F + H_{\text{total}} \Delta^* = 0,$$

which yields $\Delta^* = -H_{\text{total}}^{-1} g_F$. Subtracting the fixed-point equation from the update gives

$$\Delta_{t+1} - \Delta^* = (I - \eta H_{\text{total}})(\Delta_t - \Delta^*).$$

Because the eigenvalues of H_{total} lie in $[\mu, L]$ and $0 < \eta < 2/L$, the spectral radius of $I - \eta H_{\text{total}}$ is strictly smaller than one, giving the stated linear convergence bound. $\square$

This result formalizes the multi-step interpretation of the method. The repair does not simply follow the corrective gradient indefinitely. Instead, the corrective signal drives the model away from θ_0 , while the total curvature—comprising both the corrective curvature and the preservation curvature—increasingly resists displacement. Training settles when these forces balance:

$$\nabla L_{\text{fix}}(\theta^*) + \nabla R(\theta^*) = 0.$$

Under the local model, this balance is precisely $\Delta^* = -H_{\text{total}}^{-1} g_F$.

6.3 Interpretation as a Curvature-Weighted Repair

The total Hessian H_{total} is symmetric positive definite, so it admits the spectral decomposition

$$H_{\text{total}} = Q\Lambda Q^\top, \quad \Lambda = \text{diag}(h_1, \dots, h_d),$$

where the columns $v_1, \dots, v_d$ of Q are an orthonormal basis of eigenvectors and $h_i \geq 0$ are the associated total curvatures. With the damping γI included in H_{total} , we have $h_i \geq \gamma > 0$ for all i .

Working in the eigenbasis of H_{total} , let

$$\tilde{\Delta} = Q^\top \Delta, \quad \tilde{g}_F = Q^\top g_F,$$

so that $\tilde{\Delta}_i = \langle \Delta, v_i \rangle$ is the component of the parameter displacement along the i -th eigenvector of the total Hessian. The stationary condition $\tilde{g}_F + \Lambda \tilde{\Delta}^* = 0$ then decouples coordinate-wise,

$$\tilde{\Delta}_i^* = -\frac{(\tilde{g}_F)_i}{h_i}.$$

The equilibrium displacement therefore permits larger movement along eigenvectors with small total curvature h_i and suppresses movement along eigenvectors with large h_i . The resulting repair direction is not the raw corrective gradient direction; it is the corrective gradient filtered through the spectrum of the total curvature, i.e. through what the corrective term, the EWC term, and the KD term jointly consider important. In particular, the corrective curvature H_F can only *increase* the h_i of directions in which the model is uncertain on D_{fix} , thereby further limiting the equilibrium displacement along those directions.

The same spectral picture controls both the achievable repair gain and the preservation cost. Under the local surrogate,

$$L_{\text{fix}}(\theta_0 + \Delta^*) - L_{\text{fix}}(\theta_0) \approx -g_F^\top H_{\text{total}}^{-1} g_F = -\sum_{i=1}^d \frac{(\tilde{g}_F)_i^2}{h_i},$$

whereas

$$R(\theta_0 + \Delta^*) \approx \frac{1}{2} g_F^\top H_{\text{total}}^{-1} H_R H_{\text{total}}^{-1} g_F.$$

If H_R and H_F are simultaneously diagonalized by the same basis (a simplification that holds exactly only when they commute), the preservation cost reduces to

$$R(\theta_0 + \Delta^*) \approx \frac{1}{2} \sum_{i=1}^d \frac{(\tilde{g}_F)_i^2}{h_i} \cdot \frac{(h_R)_i}{h_i},$$

so that the ratio $(h_R)_i/h_i$ controls how much of the corrective displacement is “paid for” in preservation cost along each direction. The quantity

$$S = g_F^\top H_{\text{total}}^{-1} g_F$$

therefore summarizes the achievable first-order repair gain under the local quadratic model, with the preservation cost bounded by $S/2$ in the commuting case and by $\frac{1}{2} h_{\text{max}}(H_R) \|H_{\text{total}}^{-1} g_F\|^2$ in general. The value of S is larger when the corrective gradient has substantial components along eigenvectors of low total curvature, which is precisely the regime in which the repair can be achieved with limited collateral movement. Conversely, when the corrective gradient is concentrated along high-curvature eigenvectors, S is small: the combined curvature suppresses the displacement and the achievable repair gain is correspondingly limited.

6.4 Behavior-Level Targetedness

Parameter displacement alone does not imply behavioral locality. To connect the parameter-space result to the desired behavioral property, let $z_G(\theta)$ and $z_F(\theta)$ denote the stacked model outputs on representative good and failure inputs, respectively, with Jacobians at θ_0

$$J_G = \left. \frac{\partial z_G}{\partial \theta} \right|_{\theta_0}, \quad J_F = \left. \frac{\partial z_F}{\partial \theta} \right|_{\theta_0}.$$

At the local equilibrium,

$$z_G(\theta^*) - z_G(\theta_0) \approx -J_G H_{\text{total}}^{-1} g_F,$$

while

$$z_F(\theta^*) - z_F(\theta_0) \approx -J_F H_{\text{total}}^{-1} g_F.$$

Therefore, a sufficient local condition for a targeted repair is that the total-curvature-weighted corrective direction has a substantially larger effect on the failure outputs than on the good outputs, for example

$$\|J_F H_{\text{total}}^{-1} g_F\| \gg \|J_G H_{\text{total}}^{-1} g_F\|.$$

A stronger structural version is the existence of a repair direction d such that

$$g_F^\top d < 0, \quad \|J_G d\| \approx 0, \quad \|J_F d\| > 0.$$

Such a direction improves the corrective objective while producing little first-order output change on good behavior. The combined EWC+KD+corrective geometry favors precisely those parameter directions that can produce the required repair with lower total curvature. This does not constitute a global guarantee of behavioral equivalence; it is a local sufficient condition linking the parameter-space equilibrium to the desired behavior-level locality.

6.5 Effect of Anchor Contamination

Under the contaminated-anchor model,

$$P_{\text{anc}} = (1 - \rho)P_{\text{good}} + \rho P_{\text{fail}},$$

the population version of the preservation Hessian decomposes accordingly:

$$H_R(\rho) = (1 - \rho)H_{R,\text{good}} + \rho H_{R,\text{fail}} + H_{\text{shared}},$$

where H_{shared} denotes any ρ -independent contribution such as a separately fixed preservation component. Equivalently, after absorbing the common term into $H_R(0)$,

$$H_R(\rho) = H_R(0) + \rho(H_{R,\text{fail}} - H_{R,\text{good}}).$$

The corrective Hessian H_F does not depend on ρ , since D_{fix} is failure-dominated by construction and is not resampled from P_{orig} . Thus the total Hessian inherits the contamination dependence through H_R alone:

$$H_{\text{total}}(\rho) = H_F + H_R(\rho) = H_{\text{total}}(0) + \rho(H_{R,\text{fail}} - H_{R,\text{good}}).$$

Let

$$E_\rho = \rho(H_{R,\text{fail}} - H_{R,\text{good}}).$$

If $H_{\text{total}}(0) \succeq \mu I$ and

$$\|E_\rho\| < \mu,$$

then $H_{\text{total}}(\rho)$ remains positive definite and standard matrix perturbation gives

$$\|H_{\text{total}}(\rho)^{-1} - H_{\text{total}}(0)^{-1}\| \leq \frac{\|E_\rho\|}{\mu(\mu - \|E_\rho\|)}.$$

Consequently, with the same local corrective gradient g_F ,

$$\|\Delta^*(\rho) - \Delta^*(0)\| \leq \frac{\|E_\rho\|}{\mu(\mu - \|E_\rho\|)} \|g_F\| = O(\rho)$$

for sufficiently small ρ . Hence rare failure cases in the anchor set perturb the local repair equilibrium continuously rather than changing it discontinuously. The contamination remains harmful because its curvature points toward preserving the old failure behavior, but when ρ is small its aggregate influence is correspondingly limited under the stated conditioning assumptions. When $\ker(H_R) \subseteq \text{range}(H_F)$, the corrective curvature H_F raises the lower bound μ in the perturbation bound and reduces the sensitivity of the equilibrium to anchor contamination. Outside this case, H_F does not change μ and the bound is governed by the preservation geometry alone.

The bound is stated under the perturbative condition $\|E_\rho\| < \mu$ and is a first-order statement in ρ ; it is not claimed to hold at the empirically estimated $\hat{\rho} \approx 0.185$ (Section 8.3), where the condition is correspondingly harder to satisfy. The result should therefore be read as characterizing the local sensitivity of the repair equilibrium to contamination in the small- ρ regime, not as a guarantee that the shift is negligible at the operating point of the experiments.

6.6 Theoretical Interpretation

The resulting picture is therefore a multi-step balance rather than a one-step dominance argument:

$$\underbrace{\nabla L_{\text{fix}}}_{\text{repair drive}} + \underbrace{\lambda_1 \nabla L_{\text{EWC}} + \lambda_2 \nabla L_{\text{KD}}}_{\text{restoring preservation force}} = 0 \quad \text{at equilibrium,}$$

and the local geometry of this balance is governed by the total Hessian

$$H_{\text{total}} = H_F + \lambda_1 \text{diag}(\hat{F}) + \lambda_2 H_{\text{KD}}.$$

The corrective curvature H_F reflects where the model is still uncertain on the corrective set; the EWC coefficients measured as mean squared per-sample sensitivities determine which parameter directions are expensive to change; KD determines which anchor-output changes are expensive. When the corrective signal admits a low-total-curvature direction, the equilibrium can produce a substantial correction with limited collateral movement. Anchor contamination perturbs H_R and therefore H_{total} in proportion to its prevalence, so the main failure mode of contamination is a shift of the equilibrium toward preserving part of the behavior that the repair objective seeks to eliminate.

The analysis is local and does not claim that neural-network optimization is globally convex, that the empirical Fisher equals the exact Fisher in every implementation, or that the resulting equilibrium necessarily gives perfect failure correction and zero general-performance loss.

7 Repair–Preservation Trade-off

As ρ decreases, most anchor samples represent desirable behavior and the preservation terms receive a strong aggregate signal from non-failure examples, reducing the conflict of Section 5. In the local analysis of Section 6, this corresponds to a smaller perturbation of the preservation Hessian $H_R(\rho)$. The corrective dataset provides a concentrated signal $D_{\text{fix}} \sim P(x \mid x \in \mathcal{F})$ regardless of ρ . The effective optimization is therefore governed jointly by $|D_{\text{fix}}|$, $|D_{\text{anc}}|\rho$, the preservation geometry, and the regularization strengths λ_1, λ_2 , rather than by either dataset size alone.

We do not assume a fixed value of λ_1, λ_2 is universally optimal; the method should be evaluated as a function of $|D_{\text{fix}}|, |D_{\text{anc}}|, \rho, \lambda_1, \lambda_2$.

Target Property

The desired solution satisfies, qualitatively,

$$\text{Behavior}_{\text{fail}}(\theta^*) \succ \text{Behavior}_{\text{fail}}(\theta_0), \quad \text{Performance}_{\text{good}}(\theta^*) \approx \text{Performance}_{\text{good}}(\theta_0),$$

where $\text{Behavior}_{\text{fail}}(\theta)$ denotes the operational failure behavior of θ on the intervention-defined failure region of Section 8.3, and $\succ$ means strictly better on that axis. The objective is a *localized* behavioral correction with limited collateral change, not merely a lower training loss.

8 Implementation

We instantiate the combined objective for a behavioral-cloning policy [9] trained in the *StrikeForce* 2D partially observable battle-arena testbed [10], using the decoupled Future Action Prediction (FAP)/Transformer-aggregator architecture described in [11]. We therefore apply the repair directly to the converged checkpoint θ_0 produced by that training procedure, without independently retraining the underlying policy architecture for the present study.

8.1 Precomputed Anchor Statistics

Before repair, we precompute and pin, for each anchor episode: θ_0 itself, an estimate of the Fisher diagonal $\{F_i\}$, and the reference future-prediction and policy logits $\{z_0(x_j)\}$ under θ_0 . Pinning the logits rather than recomputing $f_{\theta_0}(x_j)$ at every step avoids carrying a second live copy of the model during repair and fixes the KD reference exactly.

8.2 Per-Step Composition

At each optimization step we compute three gradient contributions and sum them before the optimizer step: (i) the corrective gradient from a shuffled minibatch of D_{fix} , using the same time-discounted focal loss as the original training procedure; (ii) the EWC gradient, evaluated over all parameters against the pinned (θ_0, F) ; (iii) the KD gradient, evaluated as a KL-type divergence between current and pinned softmax outputs at both the future-prediction heads and the main policy head.

Remark 2. *In the current implementation, the corrective term is minibatched (shuffled subsets of D_{fix} each step) while the KD term is evaluated over the entire anchor set at every step. This gives the anchor term zero gradient stochasticity relative to the corrective term and dominates per-step compute. The asymmetry is intentional: the anchor term is kept deterministic across*

optimization steps so that the preservation signal is applied consistently against the fixed reference outputs. This does not change the joint-objective gradient identity of Section 4, and full-batch KD is an implementation choice rather than a requirement of the method; minibatched KD would be a valid computational alternative.

8.3 Anchor and Corrective Set Provenance

The corrective set D_{fix} is constructed from *intervention episodes* collected during unmasked inference. Concretely, while the policy is deployed in inference mode, it occasionally enters a state in which it produces an unpleasant, illogical, or otherwise unacceptable action. A human expert intervenes at that point, takes control, and executes the intended action until the policy has escaped the problematic situation, after which inference resumes. Each intervention episode therefore supplies a short trajectory of states paired with the expert’s actions, i.e. the corrected labels for the failure region. During repair, the model is trained only on these expert actions. The corrective set is thus concentrated on $\mathcal{F}$ by construction, and its failure density is close to one, in contrast to the anchor set. Operationally, the failure region $\mathcal{F}$ in this study corresponds to states in which the policy disengages from combat when engagement is warranted; the expert interventions recorded in D_{fix} are precisely the actions that restore engagement. The composite metrics of Section 11.3 are designed to measure this axis directly.

The anchor set D_{anc} , by contrast, is sampled from the original distribution via the stage-1 validation episodes. It is not constructed to exclude failure cases: the behavioral failure is identifiable but is not assumed to be cleanly separable from the original distribution. Instead, the intended setting is that such cases are relatively rare, so most anchors still represent desirable behavior while a minority may overlap with $\mathcal{F}$. This is precisely the contamination scenario formalized in Definition 1.

Here ρ denotes the empirical failure frequency *within* D_{anc} , not within D_{fix} : the two sets have very different failure densities by construction. Every frame in D_{fix} is an expert action recorded during an intervention, so the failure density in D_{fix} is close to one. The anchor set, by contrast, contains only the ρ -fraction of frames that coincide with the failure region.

We estimate ρ empirically from the rate at which the policy required expert intervention during normal operation, using the intervention episodes that also supply D_{fix} ; this rate serves as a proxy for $P_{\text{orig}}(x \in \mathcal{F})$ under the assumption that the deployment episodes are representative of the original distribution. Across 28 episodes of 1024 frames each (28,672 frames total), 5300 frames required intervention, giving

$$\hat{\rho} = \frac{5300}{28672} \approx 0.185.$$

This estimate is non-negligible: the contaminated fraction of D_{anc} is a substantial minority rather than a vanishing edge case. The contaminated-anchor setting formalized in this paper is therefore a practical concern rather than a theoretical curiosity, and the conflict between the corrective and preservation gradients analyzed in Section 5 is expected to be observable in the experiments rather than masked by a vanishing ρ . When episode-level labels or behavioral inspection are available, the observed failure frequency in D_{anc} can be reported directly as an empirical estimate of the contamination rate; the method itself does not require explicit filtering of those examples.

8.4 Hyperparameters

We use $T = 1$ in the current implementation, i.e., ordinary softmax rather than temperature-softened distillation targets. The regularization strengths of the combined objective are set to

$$\lambda_1 = \lambda_2 = 5,$$

i.e., the EWC and KD preservation terms are weighted equally in Section 4. This is the configuration used for the θ^* checkpoint reported in Section 11; the values were fixed for the repair runs and were not separately tuned per metric or per failure mode. As stated in Section 7, the trade-off between repair and preservation is governed jointly by $(|D_{\text{fix}}|, |D_{\text{anc}}|, \rho, \lambda_1, \lambda_2)$, so this choice is a single operating point on that trade-off rather than a claim of optimality. The naive fine-tuning baseline θ_{FT} corresponds to the same pipeline with $\lambda_1 = \lambda_2 = 0$.

9 Related Work

Elastic Weight Consolidation [1] penalizes movement of parameters deemed important to previous tasks, estimated via the Fisher information matrix; we use the same diagonal-Fisher penalty but apply it against a single identified failure region rather than a sequence of task boundaries. Knowledge distillation [2] was originally proposed for transferring information from a teacher to a student; here θ_0 and θ^* share an architecture, and KD instead serves as an output-space behavioral anchor. Learning without Forgetting [3] is also closely related to our data-availability constraint because it studies learning new capabilities when the original training data are unavailable, although its objective is preservation of existing capabilities while adding new ones rather than correcting a known failure.

More broadly, our setting belongs to the family of post-hoc model-editing approaches that seek to alter a pretrained model while limiting unintended changes. MEND [5] learns to transform gradients for fast, localized edits, ROME [6] directly modifies selected model parameters to rewrite factual associations, and task arithmetic [7] steers pretrained-model behavior through parameter-space task vectors. These approaches demonstrate the broader feasibility of targeted behavioral modification, but they address different editing mechanisms and problem formulations from the EWC+KD repair procedure studied here. In particular, our focus is not on editing factual associations in language models or learning a separate editor; it is on repairing a behavioral failure in an existing policy using a small corrective set plus limited anchors while explicitly accounting for anchor contamination.

The closest methodological precedent is [4], who combine a task-learning loss with EWC and KD to balance new-task learning against retention in a continual-classification setting. Our problem differs in framing: rather than sequential task learning, we address targeted correction of a single identified behavioral failure under incomplete access to the original training data, and we treat the possibility that the anchor set itself contains the failure ($D_{\text{anc}} \cap \mathcal{F} \neq \emptyset$) as a first-class part of the problem rather than an incidental detail. Recent analysis of model editing also shows that repeated edits can induce forgetting and downstream performance degradation [8], reinforcing the importance of evaluating collateral behavioral change rather than reporting only edit success.

10 Scope Statement

Explicitly not claimed:

- That D_{fix} or $\mathcal{F}$ admits a simple task-independent mathematical definition; failure is identified operationally from unmasked inference behavior, with task-specific or human judgment where needed.
- That preservation on D_{anc} guarantees global behavioral equivalence $f_{\theta^*}(x) \approx f_{\theta_0}(x)$ for all $x \in \mathcal{X}$; it is an empirical, regularized approximation over representative samples.
- That a specific (λ_1, λ_2) setting is universally optimal; the trade-off is a function of ρ , $|D_{\text{fix}}|$, and $|D_{\text{anc}}|$ and should be tuned per setting.
- That the local equilibrium analysis of Section 6 is a global convergence theorem for a non-convex neural network; its convergence result applies to the stated local surrogate and to the subspace where the preservation Hessian is positive definite.
- That the empirical diagonal Fisher estimate used here is the exact Fisher matrix in every loss construction; outside the standard likelihood setting, it is explicitly a squared-gradient sensitivity statistic.
- That the current implementation’s full-batch anchor evaluation is necessary to the method; it is a design choice of the present pipeline (Section 8).
- That the sequential learning setting of [4] and the single-failure-correction setting studied here are the same problem; we adopt their combined-loss mechanism but not their task-boundary framing.
- That the base, derived, and composite metrics are computed from independent measurement runs. All three metric groups are aggregated from the same $n = 64$ paired evaluation episodes described in Section 11.2; the derived and composite groups are deterministic transformations of the base group and are not independent measurements.
- That the figures in Section 11 establish statistical significance on their own. They are descriptive visualizations of the tables and the paired test summaries; inferential claims are supported only by the tables and their associated corrected q -values.

11 Experimental Design

11.1 Checkpoints Under Comparison

Three checkpoints are compared. All three share the same architecture, environment, and inference-time action mask; they differ only in the parameters produced by their respective optimization procedures.

- θ_0 (**P**) — **Pre-repair**: the converged policy checkpoint before any behavioral repair, produced by the Future Action Prediction / Transformer-aggregator training procedure of [11].
- θ_{FT} (**F**) — **Naive fine-tuning**: obtained by optimizing θ_0 on D_{fix} alone, i.e. $\lambda_1 = \lambda_2 = 0$ in Section 4. This baseline isolates the effect of the corrective signal without any explicit preservation mechanism.
- θ^* (**B**) — **EWC+KD repair**: the checkpoint produced by the combined objective with $\lambda_1 = \lambda_2 = 5$ (Section 8.4), implemented as described in Section 8.

Throughout this section we use the short names **P**, **F**, and **B** for θ_0 , θ_{FT} , and θ^* respectively.

11.2 Evaluation Protocol

Evaluation uses a fixed set of $n = 64$ paired episodes, with the heuristic inference-time action mask (`bc_inference`) applied uniformly to all three checkpoints. The same random seed, map, and initial condition are used within each paired episode, so that differences between checkpoints within a pair are attributable to the policy rather than to environment stochasticity. All results are reported as paired comparisons with $n = 64$ and are corrected for multiple comparisons using Benjamini–Hochberg FDR control over all $(3 \text{ pairs}) \times (14 \text{ metrics})$ tests jointly. The analysis is implemented in the accompanying `compare_models.py` script, which also generates all LaTeX tables and figures referenced below.

Why a paired design and not aggregate averages. The pre-repair policy exhibits high per-episode variance in survival time (sample standard deviation of roughly 6.7×10^3 against a mean of 1.2×10^4 , i.e. a coefficient of variation above 0.5). Unpaired comparisons at this variance level cannot reliably distinguish policy differences from episode-to-episode environment randomness. Pairing on identical initial conditions removes this shared variance and allows the paired t -test or Wilcoxon signed-rank test to detect policy-level effects that would otherwise be masked at this sample size.

11.3 Metrics

We use three groups of metrics. All metrics are oriented so that larger values indicate a better policy; none of the metrics is an accumulator that grows with episode length.

Base metrics. The five quantities extracted from each episode log are ¹

$$N_{\text{kills}} \text{ (kills), } T_{\text{alive}} \text{ (frames alive), } A_{\text{HP}} \text{ (area under HP curve),}$$

$$A_{\text{stamina}} \text{ (area under stamina curve), } A_{\text{damage}} \text{ (area under damage-per-bullet curve).}$$

These are the raw quantities on which the derived metrics below are built. They are reported for completeness but are not used as primary evidence: A_{HP} , A_{stamina} , and A_{damage} are accumulators and therefore correlate with T_{alive} ; N_{kills} and T_{alive} are the volume measures on which the targeted failure is defined.

Derived metrics. Six normalized quantities are computed from the base metrics:

$$S_1 = \frac{N_{\text{kills}}}{T_{\text{alive}}}, \quad S_2 = \frac{N_{\text{kills}}}{A_{\text{stamina}}}, \quad S_3 = \frac{N_{\text{kills}}^2}{T_{\text{alive}} A_{\text{damage}}},$$

$$r_{\text{HP}} = \frac{A_{\text{HP}}}{T_{\text{alive}}}, \quad r_{\text{stam}} = \frac{A_{\text{stamina}}}{T_{\text{alive}}}, \quad r_{\text{dmg}} = \frac{A_{\text{damage}}}{T_{\text{alive}}}.$$

S_1 is the kill rate, S_2 is kills per unit of accumulated stamina (an engagement-efficiency measure), and S_3 is a damage-normalized kill rate. r_{HP} , r_{stam} , and r_{dmg} are the per-frame HP, stamina, and damage rates; larger values indicate a policy that maintains a higher combat capacity per unit time. All six are higher-is-better.

¹<https://github.com/bistoyek21-ric/StrikeForce/blob/main/StrikeForce-client/StrikeForce-Lab-Protocol.md>

Composite metrics. Three kill-weighted resource metrics are defined as

$$M_{\text{HP}} = \frac{N_{\text{kills}} A_{\text{HP}}}{T_{\text{alive}}^2}, \quad M_{\text{stam}} = \frac{N_{\text{kills}} A_{\text{stamina}}}{T_{\text{alive}}^2}, \quad M_{\text{dmg}} = \frac{N_{\text{kills}} A_{\text{damage}}}{T_{\text{alive}}^2}.$$

Each metric combines the kill count with a resource accumulator and normalizes by squared episode duration, which penalizes duration more strongly than a single-time normalization: when a resource accumulator grows approximately linearly with time, the metric behaves roughly like kills per unit time multiplied by the average resource level per unit time. It therefore measures a duration-normalized engagement–resource trade-off, not total accumulated combat output. A policy that survives for a long time while producing few kills is penalized, while a policy that produces kills and maintains resource capacity can score well. This interpretation is useful for comparing policies under the same episode protocol, but the metric is a designed proxy rather than a universally valid measure of policy quality. We consequently report it alongside the base and derived metrics rather than treating it as standalone proof of behavioral repair.

Why the evaluation does not use an argmax-based failure rate. One candidate metric is the per-frame failure rate, defined as the fraction of decisions on which the raw argmax of $f_{\theta}(s)$ does not match an operationally defined task-correct action. We do not use this metric, for two reasons. First, it is not a genuine measurement of policy behavior: the argmax is a single point in the simplex, and small logit perturbations near the decision boundary can flip a decision without producing any behavioral change, so the metric can move by ϵ -perturbation rather than by genuine repair. Second, the metric is self-referential: it is aggregated over the state distribution $d_{\pi}(s)$ induced by the policy itself, so a policy can improve its apparent failure rate by shifting d_{π} toward low-risk states without changing behavior on $\mathcal{F}$. Both issues are absent from the rate-based and composite metrics above, which are aggregated over a fixed episode budget rather than over a policy-dependent trajectory length. The operational definition of failure on which D_{fix} is constructed remains in Section 8.3 for provenance; the empirical evaluation of the repair is based on the rate-based and composite metrics.

11.4 Results

The evaluation contains 64 valid paired episodes for each checkpoint (P , B , and F), with no malformed data blocks. Composite metrics are computed per episode using a squared-time denominator (T_{alive}^2).

Metric	P (θ_0)	B (θ^*)	F (θ_{FT})
T_{alive}	11908 $\pm$ 6714	10991 $\pm$ 5920	9630 $\pm$ 4554
N_{kills}	184.1 $\pm$ 100.0	188.5 $\pm$ 88.9	167.8 $\pm$ 78.5
A_{HP}	$(1.032 \pm 0.680) \times 10^8$	$(0.989 \pm 0.594) \times 10^8$	$(0.804 \pm 0.502) \times 10^8$
A_{stamina}	$(1.214 \pm 0.456) \times 10^8$	$(1.150 \pm 0.410) \times 10^8$	$(1.050 \pm 0.351) \times 10^8$
A_{damage}	$(2.277 \pm 1.739) \times 10^7$	$(2.374 \pm 1.572) \times 10^7$	$(2.050 \pm 1.340) \times 10^7$

Table 1: Base metrics, mean $\pm$ standard deviation across 64 episodes.

Metric	P	B	F
$S_1 = N_{\text{kills}}/T_{\text{alive}}$	0.015658	0.017330	0.017615
$S_2 = N_{\text{kills}}/A_{\text{stamina}}$	1.45085×10^{-6}	1.58657×10^{-6}	1.55798×10^{-6}
$S_3 = N_{\text{kills}}^2/(T_{\text{alive}}A_{\text{damage}})$	1.46705×10^{-7}	1.53449×10^{-7}	1.58177×10^{-7}
$r_{\text{HP}} = A_{\text{HP}}/T_{\text{alive}}$	8540.28	8876.13	8243.26
$r_{\text{stam}} = A_{\text{stamina}}/T_{\text{alive}}$	11021.9	11089.0	11451.8
$r_{\text{dmg}} = A_{\text{damage}}/T_{\text{alive}}$	1800.78	2075.10	2050.61

Table 2: Derived metrics. Entries are means of the episode-level metrics; bold marks the largest mean in each row.

Metric	P	B	F
$M_{\text{HP}} = N_{\text{kills}}A_{\text{HP}}/T_{\text{alive}}^2$	133.708	153.412	145.013
$M_{\text{stam}} = N_{\text{kills}}A_{\text{stamina}}/T_{\text{alive}}^2$	173.693	193.041	202.676
$M_{\text{dmg}} = N_{\text{kills}}A_{\text{damage}}/T_{\text{alive}}^2$	28.7439	36.5019	36.6080
Mean relative composite score	0.8379	0.9832	0.9818

Table 3: Composite metrics, computed with T_{alive}^2 in the denominator. The relative score is the mean of each model’s three composite metrics normalized by the across-model maximum for each metric.

11.5 Paired Statistical Analysis

All pairwise tests use the 64 matched episodes. The analysis reports paired t -tests or Wilcoxon signed-rank tests as appropriate, and Benjamini–Hochberg correction across the 42 tests (three pairs by 14 metrics). The table below highlights the principal corrected results; the complete test summaries are in Appendix Tables 5–7.

Pair	Metric (difference is second minus first)	Δ	q_{FDR}	d_z
P vs B	S_1 (B–P)	+0.001672	< 0.001	+0.62
P vs B	S_2 (B–P)	$+1.357 \times 10^{-7}$	0.0051	+0.44
P vs B	r_{dmg} (B–P)	+274.32	0.0123	+0.40
P vs B	M_{HP} (B–P)	+19.70	0.0015	+0.51
P vs B	M_{stam} (B–P)	+19.35	0.0175	+0.38
P vs B	M_{dmg} (B–P)	+7.758	< 0.001	+0.56
P vs F	S_1 (F–P)	+0.001957	< 0.001	+0.67
P vs F	r_{dmg} (F–P)	+249.83	0.0334	+0.34
P vs F	M_{stam} (F–P)	+28.98	0.0019	+0.60
P vs F	M_{dmg} (F–P)	+7.864	0.0051	+0.51
B vs F	r_{HP} (F–B)	−632.87	0.0442	−0.30
B vs F	A_{HP} (F–B)	-1.847×10^7	0.0177	−0.24

Table 4: Selected paired comparisons. Positive Δ means the second checkpoint has a larger value; the interpretation as an improvement depends on the metric.

Figure 1 visualizes the same corrected results across all fourteen metrics and all three pairwise comparisons, with 95% confidence intervals on the paired mean difference and a filled/open marker convention for corrected significance. The summary rows of Table 4 highlight the principal effects;

the figure shows the distribution of effect sizes and the location of the Benjamini–Hochberg threshold across the full set of 42 tests.

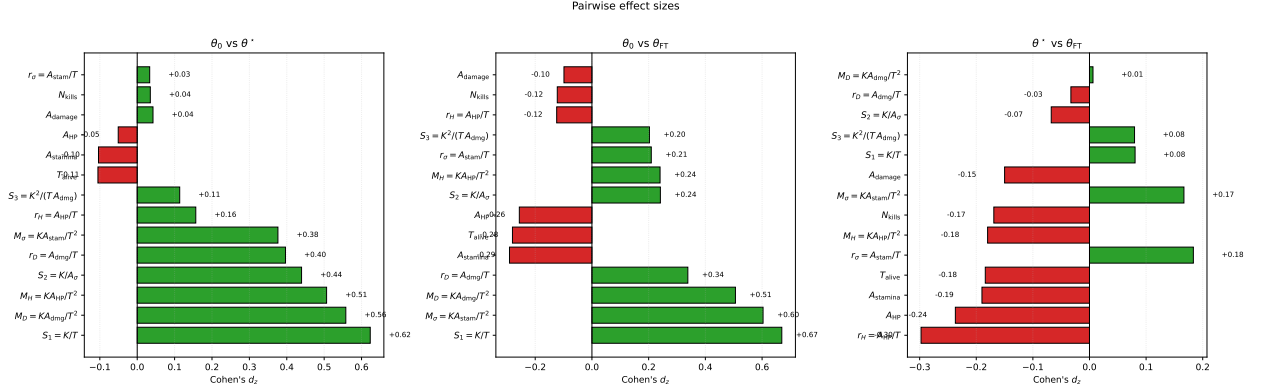


Figure 1: Paired effect sizes with 95% confidence intervals on the paired mean difference $\Delta = M_2 - M_1$, for all fourteen metrics across the three pairwise comparisons (P vs B, P vs F, B vs F). Filled markers indicate $q_{\text{FDR}} < 0.05$ after Benjamini–Hochberg correction across the 42 pair-by-metric tests of Section 11.5; open markers indicate comparisons that are not significant after correction.

11.6 Interpretation of Results

The EWC+KD checkpoint improves several engagement-oriented metrics over the pre-repair checkpoint. Compared with P, B has significantly higher kill rate S_1 ($q < 0.001$, $d_z = 0.62$), kills per accumulated stamina S_2 ($q = 0.0051$, $d_z = 0.44$), damage rate r_{dmg} ($q = 0.0123$, $d_z = 0.40$), and all three composite metrics: M_{HP} ($q = 0.0015$, $d_z = 0.51$), M_{stam} ($q = 0.0175$, $d_z = 0.38$), and M_{dmg} ($q < 0.001$, $d_z = 0.56$). The composite gains are approximately 14.7%, 11.1%, and 27.0%, respectively, relative to P. By contrast, the raw kill count, survival time, and accumulated resource totals do not show corrected significant differences between P and B. This pattern is consistent with improvement in normalized engagement and kill-weighted resource use, rather than a demonstrated increase in the total number of kills or episode duration.

Figure 2 summarizes the resulting profile across the six engagement-oriented metrics that show corrected significant differences in at least one pairwise comparison. Because each axis is min-max normalized across the three checkpoints, the figure is a descriptive profile rather than an inferential test; the corrected q -values reported above and in the appendix remain the basis for the statistical claims.

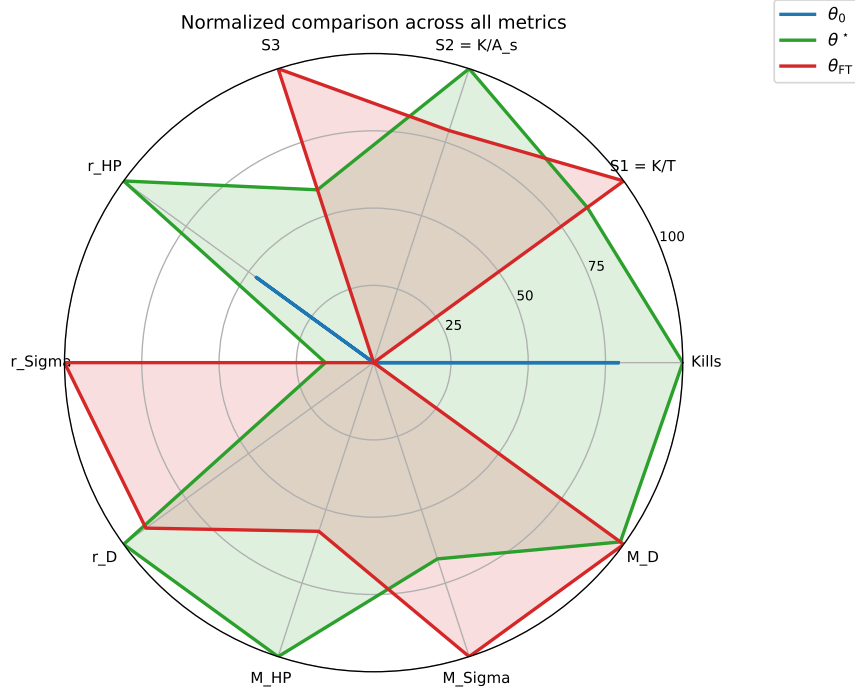


Figure 2: Normalized profile of the three checkpoints over the six engagement-oriented metrics that show corrected significant differences in at least one pairwise comparison (S_1 , S_2 , r_{dmg} , M_{HP} , M_{stam} , M_{dmg} ; Section 11.5). Each axis is min-max normalized across the three checkpoints, so that the best checkpoint on a given axis scores 1 and the worst scores 0. The figure is descriptive and is not used as an inferential test.

Naive fine-tuning also improves some engagement measures, but does not uniformly dominate the repaired checkpoint. Compared with P, F significantly improves S_1 ($q < 0.001$), r_{dmg} ($q = 0.0334$), M_{stam} ($q = 0.0019$), and M_{dmg} ($q = 0.0051$). The difference between B and F in S_1 is not significant ($q = 0.6506$); therefore the data do not establish that either repair procedure is better on kill rate. F has a slightly higher mean M_{dmg} (36.608 versus 36.502), but that difference is not significant ($q = 0.9606$). F’s mean M_{stam} is also higher, but the B–F difference is not significant ($q = 0.3017$).

Preservation evidence is mixed, not uniformly favorable to B. B has a higher mean HP rate than F (8876 versus 8243), and the paired B–F comparison is significant after correction ($q = 0.0442$). Total accumulated HP is also higher for B than F in the reported corrected comparison ($q = 0.0177$). However, these comparisons alone do not prove global behavioral preservation, and most other B–F comparisons are not significant. We therefore interpret the results as evidence that EWC+KD changes the repair–preservation trade-off, not as proof that it preserves every aspect of the original policy.

Pareto comparison and overall summary. Table 3 places the three checkpoints in the two-dimensional plane of mean raw kills and corrected composite damage M_{dmg} . Reading the two columns jointly, P is Pareto-dominated: it is lower than B on mean kills (167.81 vs. 188.47) and lower than F on M_{dmg} (28.74 vs. 36.61), so no reader can prefer P on either axis. Both B and F are Pareto-optimal: B has more mean kills (188.47 vs. 167.81), whereas F has a marginally larger

M_{dmg} (36.61 vs. 36.50). The normalized mean composite scores are 0.8379 for P, 0.9832 for B, and 0.9818 for F. These scores are descriptive summaries, not inferential tests; B and F are nearly tied on this aggregate. Figure 3 visualizes this comparison.

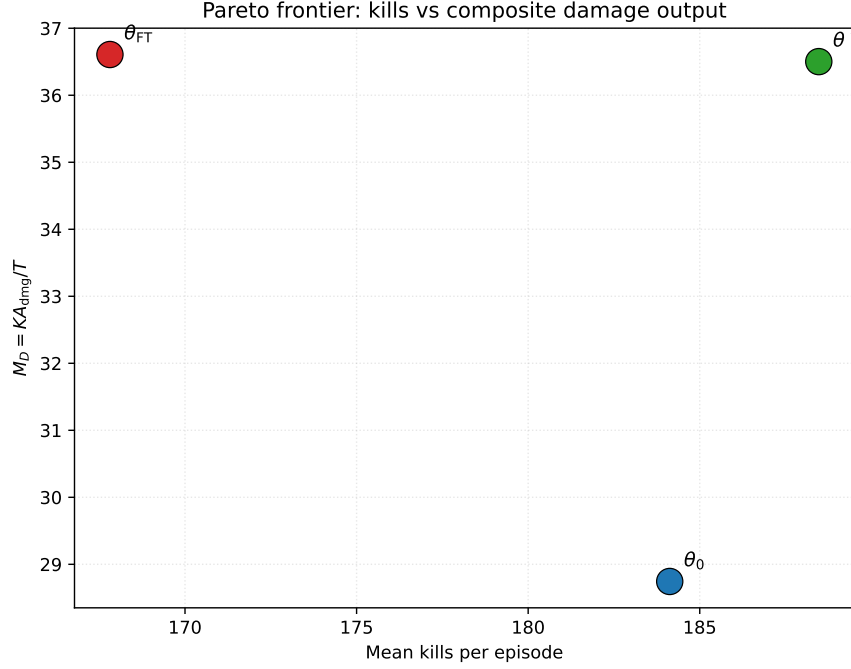


Figure 3: Pareto frontier in the plane of mean kills per episode and the corrected composite damage metric $M_{\text{dmg}} = N_{\text{kills}} A_{\text{damage}} / T_{\text{alive}}^2$. P is dominated by B; both B and F are Pareto-optimal, with B producing more mean kills and F a marginally larger M_{dmg} . This figure summarizes the two-dimensional comparison described in the surrounding text and is descriptive rather than inferential.

Limitations. The experiment evaluates three checkpoints on 64 paired episodes in one environment and uses episode-level proxy metrics rather than a direct behavioral-failure rate. The findings support claims about these measured metrics under this evaluation protocol; they do not establish universal repair guarantees or rule out regressions on unmeasured behaviors. The figures in this section visualize the same measurements that the tables and paired test summaries report; they are not independent measurement runs and do not carry additional statistical information beyond what the tables and their corrected q -values already provide.

A Full Paired Statistical Results

For each comparison, Δ is the second checkpoint minus the first. q_{FDR} denotes the Benjamini–Hochberg adjusted p -value across the 42 pair-by-metric tests. The wins column counts episode pairs favoring the first and second checkpoint, respectively.

Metric ($\Delta = B - P$)	Δ mean	q_{FDR}	d_z	Wins P:B
N_{kills}	+4.344	0.8128	0.04	34:29
T_{alive}	−916.844	0.5542	−0.11	37:27
A_{HP}	-4.291×10^6	0.7785	−0.05	35:29
A_{stamina}	-6.409×10^6	0.5542	−0.10	37:27
A_{damage}	+966423	0.8102	0.04	33:31
S_1	+0.00167182	< 0.001	0.62	17:47
S_2	$+1.35721 \times 10^{-7}$	0.0051	0.44	21:43
S_3	$+6.74418 \times 10^{-9}$	0.5310	0.11	24:40
r_{HP}	+335.846	0.3340	0.16	29:35
r_{stam}	+67.119	0.8128	0.03	35:29
r_{dmg}	+274.319	0.0123	0.40	27:37
M_{HP}	+19.7043	0.0015	0.51	21:43
M_{stam}	+19.3478	0.0175	0.38	24:40
M_{dmg}	+7.75799	< 0.001	0.56	20:44

Table 5: Full P–B paired comparison.

Metric ($\Delta = F - P$)	Δ mean	q_{FDR}	d_z	Wins P:F
N_{kills}	−16.313	0.6728	−0.12	33:31
T_{alive}	−2277.75	0.1628	−0.28	36:28
A_{HP}	-2.276×10^7	0.0891	−0.26	37:27
A_{stamina}	-1.637×10^7	0.0752	−0.29	38:26
A_{damage}	-2.269×10^6	0.7023	−0.10	32:32
S_1	+0.00195713	< 0.001	0.67	20:44
S_2	$+1.07132 \times 10^{-7}$	0.0752	0.24	24:40
S_3	$+1.14717 \times 10^{-8}$	0.2095	0.20	26:38
r_{HP}	−297.025	0.4499	−0.12	34:30
r_{stam}	+429.849	0.1981	0.21	28:36
r_{dmg}	+249.833	0.0334	0.34	27:37
M_{HP}	+11.3052	0.1533	0.24	24:40
M_{stam}	+28.9827	0.0019	0.60	22:42
M_{dmg}	+7.86402	0.0051	0.51	27:37

Table 6: Full P–F paired comparison.

Metric ($\Delta = F - B$)	Δ mean	q_{FDR}	d_z	Wins B:F
N_{kills}	−20.656	0.1867	−0.17	36:28
T_{alive}	−1360.91	0.1867	−0.18	39:25
A_{HP}	-1.847×10^7	0.0177	−0.24	43:21
A_{stamina}	-9.964×10^6	0.2439	−0.19	39:25
A_{damage}	-3.236×10^6	0.1981	−0.15	39:25
S_1	+0.000285309	0.6506	0.08	32:32
S_2	-2.85897×10^{-8}	0.5542	−0.07	35:29
S_3	$+4.72752 \times 10^{-9}$	0.6506	0.08	30:34
r_{HP}	−632.872	0.0442	−0.30	41:23
r_{stam}	+362.730	0.2574	0.18	26:38
r_{dmg}	−24.4861	0.8128	−0.03	34:30
M_{HP}	−8.39911	0.2602	−0.18	36:28
M_{stam}	+9.63497	0.3017	0.17	27:37
M_{dmg}	+0.106034	0.9606	0.01	36:28

Table 7: Full B–F paired comparison.

References

- [1] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. *Overcoming Catastrophic Forgetting in Neural Networks*. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017. doi:10.1073/pnas.1611835114.
- [2] G. Hinton, O. Vinyals, and J. Dean. *Distilling the Knowledge in a Neural Network*. arXiv:1503.02531, 2015.
- [3] Z. Li and D. Hoiem. *Learning Without Forgetting*. In *European Conference on Computer Vision (ECCV)*, pages 614–629, 2016. doi:10.1007/978-3-319-46493-0_37.
- [4] J. Zhang, X. Xu, G. Xie, and H. Liu. *Continuous Classification of Garbage Based on the Elastic Weight Consolidation and Knowledge Distillation*. CAAI Transactions on Intelligent Systems, 18(4):878–885, 2023. doi:10.11992/tis.202211023.
- [5] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. *Fast Model Editing at Scale*. In *International Conference on Learning Representations (ICLR)*, 2022.
- [6] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. *Locating and Editing Factual Associations in GPT*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022.
- [7] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. *Editing Models with Task Arithmetic*. In *International Conference on Learning Representations (ICLR)*, 2023.
- [8] A. Gupta, A. Rao, and G. Anumanchipalli. *Model Editing at Scale leads to Gradual and Catastrophic Forgetting*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15202–15232, 2024. doi:10.18653/v1/2024.findings-acl.902.
- [9] F. Torabi, G. Warnell, and P. Stone. *Behavioral Cloning from Observation*. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4950–4957, 2018. doi:10.24963/IJCAI.2018/687.
- [10] K. Fouladi. *StrikeForce: A 2D Partially Observable Battle-Arena Testbed for Imitation Learning and Multi-Agent Research*. ResearchGate, 2026. doi:10.13140/RG.2.2.24446.96326. <https://doi.org/10.13140/RG.2.2.24446.96326>
- [11] K. Fouladi. *No Back-Propagation Through Time via Future Action Prediction*. ResearchGate, 2026. doi:10.13140/RG.2.2.18971.43049. <https://doi.org/10.13140/RG.2.2.18971.43049>