ERROR DETECTABILITY IN MODEL MERGING: DIAGNOSIS AND CONSTRAINED REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging can preserve a classifier's decisions while degrading the confidence ranking used to detect its errors. We isolate this effect on expert–merge *agreement subsets*, where predictions, and hence correctness labels, coincide: on the five CLIP ViT-B/32 tasks with at least ten shared errors, expert macro AUROC on these subsets is 86.4%, versus 69.2% for TIES. A three-class linear construction shows that parameter averaging can reverse a perfect error ranking while preserving every prediction, through an example-dependent Jensen gap in the competing logits. We then introduce PED (Preserving Error Detectability), a local task-vector correction that optimizes error ranking under margin constraints on protected correct fitting examples and retains only candidates whose macro fixed-label fitting AUROC is at least that of the base. With 128 fitting and 128 selection labels per task, PED raises eight-task held-out AUROC from 84.7% to 88.2% and accuracy from 71.7% to 73.4%, exceeding LMN and ProbSurgery by 2.8 and 2.3 points in direct paired comparisons. On ViT-L/14, it improves AUROC by 3.9 points on TIES and by 4.2 points on task arithmetic. With correctness labels held fixed, PED recovers 72.1% of the expert–TIES ranking gap on shared ViT-B/32 decisions, so its gains include better ranking of the base model's own errors.

1 INTRODUCTION

Model merging combines task-specific fine-tunes of a shared backbone into a single model (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023). Evaluations typically focus on how much of the experts' accuracy the merged model retains. For selective prediction, routing, and escalation to human review, however, a second property matters: confidence must separate correct decisions from errors (Geifman & El-Yaniv, 2017; Corbière et al., 2019), as measured by correctness AUROC, the probability that a random correct prediction is more confident than a random error, with ties counted as one half. Retaining predictions does not guarantee that this ranking survives merging.

Comparing the correctness AUROC of an expert with that of the merged model conflates two changes: confidence scores may shift, and different examples may become errors. We separate the two by evaluating both models on their *agreement subset*, the examples on which their predictions, and hence their correctness labels, coincide. On the five CLIP ViT-B/32 tasks with at least ten shared errors, the experts reach 86.4% macro AUROC on these subsets versus 69.2% for TIES, and they lead on every task. These subsets cover 63.8% of the five tasks' held-out audit examples.

The effect already arises in linear models. Two three-class classifiers can agree on every prediction, and each can rank all correct examples above all errors, yet their parameter mean reverses this ordering (Figure 1a,b). The log-odds of averaged logits equal the average of the experts' log-odds plus a Jensen gap over the competing logits. When this example-dependent gap is sufficiently larger on an error than on a correct prediction, it overturns the experts' ordering without changing either prediction. For nonlinear networks, the decomposition includes a residual between the weight merge and a reference logit ensemble, separating the linear averaging mechanism from the effects of the nonlinear merge.

This diagnosis motivates PED (*Preserving Error Detectability*), a local correction in a task-vector subspace. PED minimizes a pairwise ranking loss subject to margin floors on protected correct fitting examples; a step is accepted only after a forward pass verifies these floors, and a separate eligibility test retains only candidates whose macro fixed-label fitting AUROC is at least that of the base. The

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.

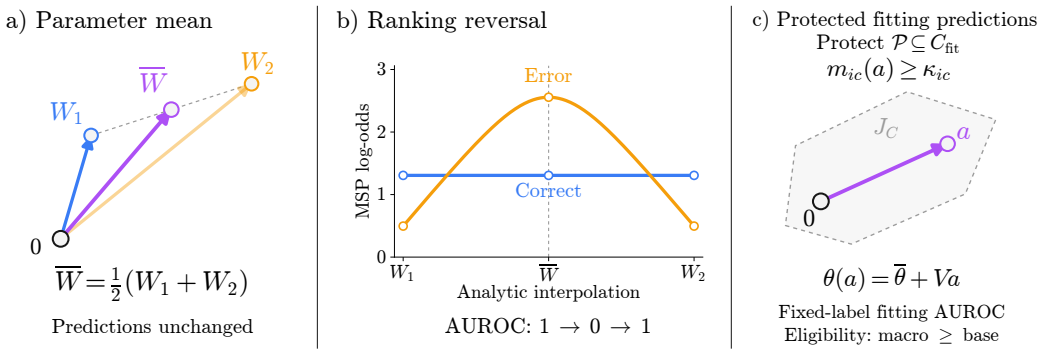


Figure 1: **Prediction agreement does not preserve error ranking.** (a,b) The linear construction of Proposition 1 retains every prediction but reverses the ranking at the parameter mean; curves show exact log-odds. (c) PED constrains the margins of a fixed set of protected correct fitting examples, verifies each proposed step with a forward pass, and deems a candidate eligible only if its macro fixed-label fitting AUROC is at least that of the base.

correction is folded into the merged weights and adds no module at inference. With 128 fitting and 128 selection labels per task, PED improves eight-task audit AUROC by 3.5 points and accuracy by 1.7 points, whereas LMN (Huang et al., 2026) and ProbSurgery (Wei et al., 2025) improve AUROC by 0.7 and 1.2 points with the same image budget.

Our contributions are threefold. **(i) Diagnosis:** agreement-subset evaluation compares error ranking on identical instances and correctness labels. **(ii) Mechanism:** an exact linear construction identifies the competing-logit Jensen gap as a mechanism by which averaging reverses ranking despite complete prediction agreement. **(iii) Repair:** PED combines local ranking optimization with explicit decision constraints, improving held-out AUROC for both merge methods and both backbones, as well as the fixed-label ranking of base decisions.

2 Related Work

Model merging and post-merge correction. Task arithmetic adds task vectors to a shared initialization (Ilharco et al., 2023), and TIES and DARE reduce interference among task vectors (Yadav et al., 2023; Yu et al., 2024). AdaMerging learns merging coefficients, while representation surgery and ProbSurgery correct the merged model's features (Yang et al., 2024b;a; Wei et al., 2025). Expert Merging aligns hidden states and logits with those of the experts (Zhang et al., 2026); feature-drift minimization, WUDI, and output-space projection choose merged weights that reduce feature or output discrepancies (Sun et al., 2025; Cheng et al., 2025; Evans et al., 2026). FeatCal uses a small calibration set and closed-form layerwise weight updates to reduce post-merge feature drift without an auxiliary inference module (Gu et al., 2026). PED also folds its correction into the weights, but optimizes fixed-label error ranking under constraints on correct fitting predictions. Our evaluation asks whether confidence still identifies errors on shared decisions.

Confidence, calibration, and selective prediction. Correctness AUROC, confidence prediction, and selective risk are established tools (Geifman & El-Yaniv, 2017; Corbière et al., 2019; Traub et al., 2024). Calibration differs from ranking: a strictly increasing map of a scalar confidence can change calibration but preserves AUROC, whereas multiclass temperature scaling can reorder examples. Several merging methods involve confidence or uncertainty: Model Soups relates weight averaging to ensembling (Wortsman et al., 2022), uncertainty-based gradient matching weights model updates by their uncertainty (Daheim et al., 2024), DisTaC conditions task vectors through distillation (Yoshida et al., 2026), and Hu et al. (2026) interpolate weights to improve the alignment–calibration trade-off of language models. LMN learns sample-specific temperatures with a lightweight meta-network to improve misclassification detection while preserving predictions (Huang et al., 2026). It is a direct score-only comparator. PED studies this ranking problem after model merging and changes the weights under constraints on protected correct fitting examples.

3 ERROR RANKING ON FIXED DECISIONS

Setting. Let θ_0 be a common initialization, θ_t the expert for task t, $\bar{\theta}$ a base merge, and $z_{\theta,c}(x)$ the logit of class c under parameters θ. Task identity is known at inference, and task heads are frozen. We restrict each task vector $\theta_t - \theta_0$ to each parameter group in a chosen set, normalize the nonzero restrictions, and collect them as the columns of $V \in \mathbb{R}^{P\times d}$, where P is the number of model parameters. Local corrections take the form

$$\theta(a) = \bar{\theta} + Va, \qquad a \in \mathbb{R}^d, \qquad a^0 = 0, \tag{1}$$

where the base displacement $\bar{\theta} - \theta_0$ need not lie in the column span of V.

Agreement-subset comparison. For an expert f and a merge g, the agreement subset $\mathcal{A}_{f,g}$ contains the examples (x, y) with $\arg\max f(x) = \arg\max g(x)$. On this subset, both models have the same correct and the same incorrect examples, so correctness AUROC restricted to $\mathcal{A}_{f,g}$ compares their confidence scores on identical positive and negative groups; we call the difference the *agreement-subset ranking gap*. The gap is a conditional comparison of scores rather than the causal effect of merging on the full population. It is defined only when both correctness groups are nonempty, and we report aggregate coverage for each task set.

Protection and fixed-label ranking. For each example in the labeled fitting set $D_{\mathrm{fit},t}$ of task t, fix the base prediction $h_i = \arg\max_c z_{\bar{\theta},c}(x_i)$, with ties broken by a fixed rule, and its correctness $q_i = \mathbf{1}\{h_i = y_i\}$. Write $C_t^{\mathrm{fit}} = \{i : q_i = 1\}$ and $E_t^{\mathrm{fit}} = \{i : q_i = 0\}$, and fix a protected set $P_t \subseteq C_t^{\mathrm{fit}}$ before optimization.

Let $\mathcal{H}_t \subseteq D_{\mathrm{fit},t}$ be the fixed ranking pool supplied to the optimizer, with $C_t = \mathcal{H}_t \cap C_t^{\mathrm{fit}}$ and $E_t = \mathcal{H}_t \cap E_t^{\mathrm{fit}}$; these groups define both the ranking pairs and the eligibility test. The anchors h_i and correctness labels q_i remain fixed throughout fitting, even if an initially incorrect prediction changes class. For every fitting example, define the anchored margins and log-odds

$$m_{ic}(a) = z_{\theta(a),h_i}(x_i) - z_{\theta(a),c}(x_i), \quad c \neq h_i, \tag{2}$$

$$s_i(a) = z_{\theta(a),h_i}(x_i) - \log \sum_{c\neq h_i} \exp z_{\theta(a),c}(x_i) = -\log \sum_{c\neq h_i} e^{-m_{ic}(a)}. \tag{3}$$

While h_i remains the prediction, s_i equals the logit of the maximum softmax probability (MSP). Once a prediction changes, s_i still scores the base class and need not rank examples as the candidate's MSP would. Hence $\widehat{A}_t$ below is a fixed-label metric on $\mathcal{H}_t$, distinct from own-prediction AUROC.

Correctness AUROC and generalized risk. For $c_t = |C_t| > 0$ and $e_t = |E_t| > 0$,

$$\widehat{A}_t(a) = \frac{1}{c_t e_t} \sum_{i\in C_t, j\in E_t} \left[\mathbf{1}\{s_i(a) > s_j(a)\} + \tfrac{1}{2}\mathbf{1}\{s_i(a) = s_j(a)\}\right]. \tag{4}$$

Tasks with an empty ranking group have undefined AUROC but may still contribute correct-example constraints; $\mathcal{T}_0$ denotes the set of nondegenerate tasks. Own-prediction audit metrics use each model's own predictions and correctness groups, whereas fixed-label evaluations, marked as such, retain the base correctness labels. For n scored examples with correctness labels, let E_k count the errors among the k most confident, with $E_0 = 0$. The trapezoidal area under the generalized risk E_k/n over coverage k/n (AUGRC), including the origin and averaged over permutations of tied scores, satisfies (Traub et al., 2024)

$$\widehat{\mathrm{AUGRC}} = \frac{1}{2n^2}\sum_{k=1}^{n}(E_{k-1} + E_k) = \frac{e^2}{2n^2} + \frac{ce}{n^2}(1 - \widehat{A}), \tag{5}$$

where c and $e = n - c$ count the correct and incorrect examples and $\widehat{A}$ uses the same scores and labels (Appendix A). At fixed labels, the two metrics thus encode the same ranking change; own-prediction labels can also change error prevalence. The identity holds per task, not for macro averages, and does not cover ordinary selective risk (E_k/k).

4 AVERAGING CAN REVERSE ERROR RANKING WITHOUT CHANGING A PREDICTION

Consider K experts whose logits $z_i^{(k)}$ on example i share a unique predicted class h_i, weights $\alpha_k \geq 0$ with $\sum_k \alpha_k = 1$, and averaged logits $\bar{z}_i = \sum_k \alpha_k z_i^{(k)}$. Every margin against h_i remains positive after averaging, so predictions are preserved. Write LSE for log-sum-exp, let the subscript $-h_i$ select the competing classes, and define the Jensen gap of the competing logits

$$\mathcal{J}_i = \sum_k \alpha_k \operatorname{LSE}(z_{i,-h_i}^{(k)}) - \operatorname{LSE}\left(\sum_k \alpha_k z_{i,-h_i}^{(k)}\right) \geq 0. \tag{6}$$

Substituting into Equation (3) yields the exact identity

$$\bar{s}_i = \sum_k \alpha_k s_i^{(k)} + \mathcal{J}_i, \qquad \bar{s}_i - \bar{s}_j = \sum_k \alpha_k (s_i^{(k)} - s_j^{(k)}) + \mathcal{J}_i - \mathcal{J}_j. \tag{7}$$

The first term averages expert log-odds; the second adds an example-dependent offset unconstrained by prediction agreement. Even if every expert ranks a correct example i above an error j, averaging reverses the pair exactly when

$$\mathcal{J}_j - \mathcal{J}_i > \sum_k \alpha_k (s_i^{(k)} - s_j^{(k)}). \tag{8}$$

Intuitively, $\mathcal{J}_i$ is large when the experts disagree about *which* competing class has the largest logit: averaging then flattens the competing logits, lowers their LSE, and inflates confidence in the shared prediction.

Proposition 1 (Complete reversal under parameter averaging)**.** *There exist two three-class linear classifiers that agree on every prediction on a two-example dataset and each attain correctness AUROC* 1*, whereas their parameter mean attains AUROC* 0 *on the same examples. Predictions, and hence accuracy, are constant along the entire convex interpolation.*

Proof. Take $x_C = (1, 0)^\top$ with $y_C = 1$, and $x_E = (0, 1)^\top$ with $y_E = 2$. For logits $z = Wx$, set

$$W_1 = \begin{pmatrix} 0 & 0 \\ -2 & -0.5 \\ -2 & -6 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} 0 & 0 \\ -2 & -6 \\ -2 & -0.5 \end{pmatrix}.$$

Every convex combination of W_1 and W_2 predicts class 1 on both examples, so x_C is correct and x_E is an error throughout. Each expert has $s_C = 2 - \log 2 \approx 1.307$ and $s_E = -\log(e^{-0.5} + e^{-6}) \approx 0.496$, hence AUROC 1. The parameter mean leaves s_C unchanged but gives $s_E = 3.25 - \log 2 \approx 2.557$, hence AUROC 0. All inequalities are strict. □

The experts assign the error different runner-up classes, so averaging lowers its competing log-sum-exp: $\mathcal{J}_C = 0$, whereas $\mathcal{J}_E \approx 2.061$ exceeds the experts' pair gap of about 0.811. The construction is robust to small perturbations and to repeated examples; it relies neither on ties nor on vanishing margins. At least three classes are necessary: in binary logit-affine models, the competing logit is a scalar, so $\mathcal{J}_i = 0$, and a pair ranked correctly by every expert cannot be reversed by convex averaging.

From logit averaging to weight merging. For linear classifiers, parameter and logit averaging coincide, so Proposition 1 shows that prediction agreement and perfect endpoint ranking alone imply no positive lower bound on the ranking of the parameter mean. For nonlinear backbones, fix a reference logit ensemble with weights α_k and score the experts, ensemble, and weight merge using the same class anchor h_i. Defining $r_i = s_i^{\text{merge}} - s_i^{\text{ens}}$ gives the exact decomposition $s_i^{\text{merge}} = \sum_k \alpha_k s_i^{(k)} + \mathcal{J}_i + r_i$. Here $\mathcal{J}_i$ describes averaging in the reference logit ensemble, and r_i records its discrepancy from the weight merge. Their relative contributions depend on the experts, ensemble weights, and merge rule (Appendix C.4).

5 LOCAL REPAIR WITH PROTECTED FITTING PREDICTIONS

PED optimizes the fixed-label error ranking over the coefficients a of Equation (1), subject to verified margin constraints on the protected sets P_t and a fitting-AUROC eligibility test on the ranking pools $\mathcal{H}_t$.

5.1 CORRECT-EXAMPLE MARGIN CONSTRAINTS

Assume that every example in P_t has a unique base prediction. For these protected examples only, set

$$\kappa_{ic} = \eta_m m_{ic}(0) > 0, \qquad m_{ic}(a) \geq \kappa_{ic}, \quad i \in P_t, \quad 0 < \eta_m < 1. \tag{9}$$

Let $J_C(a)$ stack the margin gradients of the protected examples across tasks, and let $J_{\text{all}}(a)$ stack those of the full fitting population; all derivatives are with respect to a.

Lemma 1 (Score differentials)**.** *For each fitting example, $\nabla s_i = \sum_{c \neq h_i} w_{ic} \nabla m_{ic}$, where $w_{ic} = e^{z_c} / \sum_{b \neq h_i} e^{z_b}$. Consequently, every differentiable function of these scores has zero directional derivative along* $\ker J_{\text{all}}$. *In contrast, a direction δ with $J_C \delta = 0$ fixes only the protected examples' scores to first order; unprotected scores may still change.*

Freezing all fitting margins would therefore freeze every fixed-label score to first order (Appendix A); PED instead imposes inequality constraints on a protected subset only. The floors bound each protected score from below by $-\log \sum_{c \neq h_i} e^{-\kappa_{ic}}$ (Remark 1). This leaves room for ranking changes but does not guarantee repair, because all examples share the same coefficient vector.

5.2 OBJECTIVE AND SURROGATE BOUND

For $t \in \mathcal{T}_0$, fix the robust scale $\sigma_t = \max\{\text{MAD}(s_i(0) : i \in \mathcal{H}_t), 10^{-3}\}$ and define

$$F_t(a) = \frac{1}{c_t e_t} \sum_{i \in C_t, j \in E_t} \log\left(1 + \exp\left[-\frac{s_i(a) - s_j(a)}{\beta \sigma_t}\right]\right), \tag{10}$$

$$\Phi(a) = \omega \log \sum_{t \in \mathcal{T}_0} \exp(F_t(a)/\omega), \tag{11}$$

where $\beta, \omega > 0$; if $\mathcal{T}_0$ is empty, PED returns the base. The pointwise logistic bound gives

$$1 - \widehat{A}_t(a) \leq F_t(a)/\log 2, \qquad \max_{t \in \mathcal{T}_0}(1 - \widehat{A}_t(a)) \leq \Phi(a)/\log 2. \tag{12}$$

Because surrogate descent need not improve AUROC (Appendix A), eligibility is checked on the fixed-label AUROC itself.

5.3 PROPOSAL, ACCEPTANCE, AND ELIGIBILITY

Let $v_{C,k} = m_C(a^k) - \kappa_C$ be the protected slacks, $g_k = \nabla \Phi(a^k)$, and $G = V^\top V + \epsilon I \succ 0$. With damping $\mu > 0$, trust-region radius r_k, and slack reserve $\xi = 0.1$, the proposal δ_k solves

$$\begin{aligned} \min_{\delta} \quad & g_k^\top \delta + \tfrac{\mu}{2} \delta^\top G \delta \\ \text{s.t.} \quad & J_{C,k}^{\text{prop}} \delta \geq -(1 - \xi) v_{C,k}, \qquad \delta^\top G \delta \leq r_k^2, \end{aligned} \tag{13}$$

where J_C^{prop} is a forward-difference Jacobian of the protected margins. The zero step is feasible at every accepted iterate. Rows that cannot become active within the radius may be screened from the proposal, but all protected examples remain in the actual forward check. A proposal with $g_k^\top \delta_k < 0$ is backtracked over $\alpha = 1, 1/2, \ldots$ until

$$m_C(a^k + \alpha \delta_k) \geq \kappa_C, \quad \arg\max z_{\theta(a^k + \alpha \delta_k)}(x_i) = h_i \quad \forall i \in P_t, \ \forall t, \tag{14}$$

$$\Phi(a^k + \alpha \delta_k) \leq \Phi(a^k) + c_A \alpha g_k^\top \delta_k, \quad 0 < c_A < 1. \tag{15}$$

Only iterates that satisfy

$$\frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \widehat{A}_t(a) \geq \frac{1}{|\mathcal{T}_0|} \sum_{t \in \mathcal{T}_0} \widehat{A}_t(0) \tag{16}$$

enter selection, and the base is always available. Eligibility compares each candidate with the base; it enforces neither per-task nor successive-iterate monotonicity of AUROC.

Proposition 2 (Preservation of protected fitting predictions)**.** *Assume fixed protected sets with unique base anchors, exact forward checks, fixed ranking groups, and truly negative directional derivatives for accepted nonzero Armijo steps. Then every eligible candidate preserves the predictions on each P_t and has macro fixed-label fitting AUROC on $\mathcal{T}_0$ at least that of the base. If $P_t = C_t^{\text{fit}}$ for every task, each task's fitting accuracy is at least its base accuracy. Every accepted nonzero step strictly decreases* Φ.

Proof sketch (full proof in Appendix A). Positive floors preserve the protected predictions. If $P_t = C_t^{\text{fit}}$, every initially correct prediction remains correct, which yields the accuracy comparison. Eligibility and Armijo acceptance yield the other conclusions. The forward checks evaluate the actual margins and predictions, independently of the approximate proposal Jacobian.

5.4 Local pair reachability

For a reversed fixed-label pair $p = (i, j)$ at a feasible iterate, let $d_p = (s_i - s_j)/\sigma_t < 0$, $b_p = \nabla d_p$, and

$$\mathcal{C}_C = \{\delta : J_C\delta \geq -v_C,\ \|\delta\|_G \leq r\}, \qquad R_p = \max_{\delta \in \mathcal{C}_C} b_p^\top \delta,$$

where $v_C = m_C(a) - \kappa_C$. Under exact derivatives, strict affine repair of the pair within this set is possible exactly when $R_p > -d_p$ (Proposition 3, Appendix D): a feasible primal step certifies repair, and a dual upper bound no larger than $-d_p$ rules it out within the affine model. The criterion is local to the specified constraints and radius; it does not establish nonlinear or global impossibility.

6 Experiments

The experiments address three questions: whether PED improves held-out performance over post-merge corrections, whether its ranking gains persist when correctness labels are held fixed, and how the results vary with the backbone, the merge base, and the adaptation budget.

6.1 Experimental setup

Tasks and models. We use SUN397 (Xiao et al., 2010), Cars (Krause et al., 2013), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), SVHN (Netzer et al., 2011), GTSRB (Stallkamp et al., 2011), MNIST (LeCun et al., 1998), and DTD (Cimpoi et al., 2014) with public CLIP task experts (Radford et al., 2021; Ilharco et al., 2023), known task identity, and fixed zero-shot text heads. The primary ViT-B/32 base is TIES (Yadav et al., 2023) with top-20% trimming and scale 1.0, chosen by accuracy from $\{0.3, 1.0\}$ on the base-selection split. Task arithmetic (Ilharco et al., 2023) with scale 0.3 provides a second base. We also evaluate both merge methods on ViT-L/14.

Data and model selection. For both backbones, each task has 64 base-selection examples, 936 audit examples, and three mutually disjoint fitting/selection pairs with 128 examples per role, using the same sample splits (Appendix B). The experts and audit examples are fixed across adaptation repeats. Supervised objectives use the fitting labels; label-free objectives use the same fitting images and expert outputs. Audit examples are never used for adaptation. On the selection set, candidates may lose at most 0.5 macro accuracy points relative to the base and at most 2.0 points on any task; among them, we select the candidate with the smallest macro pairwise loss. This operating-point selection is separate from the fitting-AUROC eligibility test.

Comparators and controls. LMN (Huang et al., 2026) learns a positive, input-dependent temperature for frozen logits. ProbSurgery (Wei et al., 2025) corrects merged representations using expert features and keeps an auxiliary module at inference. DisTaC (Yoshida et al., 2026) conditions task vectors before merging. The comparators share fitting images and selection rules with PED but keep their own objectives and access requirements. The internal controls CE-L, Brier-L, Rank-L, and SoftSafe-L share PED's coefficient space, which covers the last two visual residual blocks and final LayerNorm ($d = 24$). CE-L minimizes cross-entropy and Brier-L the Brier score (Brier, 1950); Rank-L minimizes the ranking loss with Adam and no preservation gate, and SoftSafe-L adds a squared-hinge margin penalty. Probe is an L_2-regularized correctness predictor on frozen features and base log-odds (Corbière et al., 2019).

Evaluation. Accuracy and correctness AUROC are computed per task and then macro-averaged. Own-prediction evaluation uses each model's own correctness groups, whereas fixed-label evaluation retains the base decisions. Confidence intervals use 1,000 paired bootstrap resamples of audit examples; direct method comparisons additionally match adaptation repeats. These intervals quantify audit-sampling uncertainty conditional on the evaluated experts and fitted models.

Table 1: **Post-merge correction on the TIES base (ViT-B/32).** Accuracy and AUROC are in percent; changes are in percentage points and are rounded separately from the absolute values. R is the number of completed adaptation repeats. Paired 95% intervals are for ΔAUROC relative to TIES. Superscript c marks internal controls.

Method	R	Acc	ΔAcc	AUROC	ΔAUROC	95% CI
TIES (Yadav et al., 2023)	–	71.7	–	84.7	–	–
+ LMN (Huang et al., 2026)	3	71.7	+0.0	85.4	+0.7	[+0.3, +1.1]
+ ProbSurgery (Wei et al., 2025)	3	72.6	+0.9	85.9	+1.2	[+0.7, +1.7]
+ CE-L^c	1	72.0	+0.3	84.6	−0.1	[−0.5, +0.2]
+ Brier-L^c (Brier, 1950)	1	71.7	+0.0	84.7	0.0	[0.0, 0.0]
+ Rank-L^c	3	71.9	+0.1	84.9	+0.2	[−0.1, +0.5]
+ SoftSafe-L^c	1	72.1	+0.4	85.2	+0.5	[+0.1, +0.9]
+ Probe^c (Corbière et al., 2019)	3	71.7	+0.0	84.2	−0.5	[−0.9, −0.1]
+ PED (ours)	3	**73.4**	**+1.7**	**88.2**	**+3.5**	[**+2.9**, **+4.1**]

Table 2: **Per-task own-prediction audit AUROC (%).** Values in parentheses are PED minus TIES in percentage points; all eight are positive. The macro row is computed from unrounded results and may differ slightly from the mean of the displayed rows.

Task	TIES (Yadav et al., 2023)	LMN (Huang et al., 2026)	ProbSurgery (Wei et al., 2025)	PED (ours)
SUN397	79.1	79.8	80.5	**83.2** (+4.1)
Cars	79.4	80.2	81.1	**83.9** (+4.5)
RESISC45	81.4	82.1	82.8	**86.1** (+4.7)
EuroSAT	82.5	83.0	83.6	**85.4** (+2.9)
SVHN	88.7	89.2	89.8	**91.5** (+2.8)
GTSRB	91.8	92.3	92.5	**94.0** (+2.2)
MNIST	97.8	97.9	98.1	**98.6** (+0.8)
DTD	76.7	78.4	79.0	**82.6** (+5.9)
Macro	84.7	85.4	85.9	**88.2** (**+3.5**)

6.2 Held-out performance

PED raises the AUROC of TIES from 84.7% to 88.2% and its accuracy from 71.7% to 73.4% (Table 1); the paired AUROC gain is 3.5 points (95% CI $[2.9, 4.1]$). LMN and ProbSurgery reach 85.4% and 85.9% AUROC. In direct paired comparisons (Table 9), PED exceeds LMN by 2.8 points ($[2.1, 3.5]$) and ProbSurgery by 2.3 points ($[1.6, 3.0]$).

No internal control gains more than 0.5 AUROC points, the gain attained by SoftSafe-L (Figure 2a). Rank-L, which optimizes the ranking loss in the same coefficient space without the preservation gate, gains 0.2 points, and CE-L and Brier-L change AUROC by -0.1 and 0.0 points. These comparisons evaluate the complete procedures: Rank-L and PED differ in both optimization and candidate acceptance. PED yields the largest ranking and accuracy gains among the evaluated corrections; the separate contributions of its optimizer, margin constraints, and eligibility test are not identified by this comparison.

Variation across tasks. PED has the highest AUROC point estimate on all eight tasks (Table 2). Its gain over TIES ranges from 0.8 points on MNIST, where the base already reaches 97.8%, to 5.9 points on DTD; RESISC45, Cars, and SUN397 improve by 4.7, 4.5, and 4.1 points. The macro gain thus reflects improvements across the benchmark rather than on a single task. Because predictions can change on audit examples, we next compare rankings under a common set of correctness labels.

6.3 Ranking the same base decisions

For task t, let $\mathcal{A}_t = \{x : h_t^{\text{exp}}(x) = h_b(x)\}$ be the audit examples on which the task expert (h_t^{exp}) and TIES (h_b) agree. We fix these instances and their base correctness labels $q_i = \mathbf{1}\{h_b(x_i) = y_i\}$

Table 3: **Fixed-label error ranking on the five eligible agreement tasks.** Anchored AUROC uses Equation (17) on identical instances with fixed TIES correctness labels. Changes and paired 95% intervals are relative to TIES; values are in percent or percentage points.

Scorer	Anchored AUROC (%)	Δ vs. TIES	Paired 95% CI
Task expert (Ilharco et al., 2023)	86.4	+17.2	[+15.1, +19.4]
TIES (Yadav et al., 2023)	69.2	0.0	–
+ LMN (Huang et al., 2026)	72.8	+3.6	[+1.8, +5.3]
+ ProbSurgery (Wei et al., 2025)	74.5	+5.3	[+3.2, +7.4]
+ PED (ours)	**81.6**	**+12.4**	**[+10.1, +14.6]**

Table 4: **Post-merge correction on CLIP ViT-L/14.** Eight-task macro accuracy and AUROC (%). Changes (percentage points) and paired 95% intervals for ΔAUROC are relative to the merge base of each block.

Method	Acc	ΔAcc	AUROC	ΔAUROC	95% CI
TIES (Yadav et al., 2023)	78.2	–	87.5	–	–
+ LMN (Huang et al., 2026)	78.2	+0.0	88.3	+0.8	[+0.4, +1.2]
+ ProbSurgery (Wei et al., 2025)	79.1	+0.9	88.9	+1.4	[+0.9, +1.9]
+ PED (ours)	**80.3**	**+2.1**	**91.4**	**+3.9**	**[+3.3, +4.5]**
Task arithmetic (Ilharco et al., 2023)	74.6	–	85.1	–	–
+ LMN (Huang et al., 2026)	74.6	+0.0	85.9	+0.8	[+0.4, +1.2]
+ ProbSurgery (Wei et al., 2025)	75.8	+1.2	86.6	+1.5	[+1.0, +2.0]
+ PED (ours)	**77.2**	**+2.6**	**89.3**	**+4.2**	**[+3.6, +4.8]**

across scorers. The primary macro average covers SUN397, Cars, RESISC45, SVHN, and DTD, each with at least ten shared errors and a nonempty correct group (Table 8). Agreement coverage is 63.8% on these five tasks and 71.2% on all eight. We score each model m by its log-odds for the base class,

$$u_m(x) = z_{m,h_b(x)}(x) - \log \sum_{c \neq h_b(x)} \exp z_{m,c}(x). \tag{17}$$

For LMN, z_m are the recalibrated logits; for ProbSurgery, they are computed from the corrected representation with the common task head. The anchor remains $h_b(x)$ even if a corrected model predicts another class.

The experts reach 86.4% anchored AUROC and TIES 69.2%; the paired TIES-minus-expert difference is -17.2 points ($[-19.4, -15.1]$), a substantial ranking loss on decisions that merging leaves unchanged. PED reaches 81.6% (Table 3), a gain of 12.4 points ($[10.1, 14.6]$) that recovers 72.1% of the expert–TIES gap, compared with 20.9% for LMN and 30.8% for ProbSurgery; the remaining gap is 4.8 points. Because instances and labels are shared across scorers, this gain shows that PED better separates the base model's correct and incorrect decisions; it does not rely on relabeling corrected predictions.

6.4 Backbones, merge bases, and adaptation budgets

On ViT-L/14, PED raises the AUROC of TIES from 87.5% to 91.4% and its accuracy from 78.2% to 80.3% (Table 4), an AUROC gain of 3.9 points ($[3.3, 4.5]$). On task arithmetic, the AUROC gain is 4.2 points ($[3.6, 4.8]$), and accuracy rises from 74.6% to 77.2%. LMN gains 0.8 AUROC points on each base, and ProbSurgery gains 1.4 and 1.5 points. The ranking improvements therefore extend to a larger CLIP backbone and to both merge methods.

Correction capacity and labels. For ViT-B/32 with 128 examples per role, enlarging the correction space from the last two blocks ($d = 24$) to the last four ($d = 40$), with final LayerNorm in both settings, raises the audit AUROC gain from 3.5 to 4.1 points (Table 10). Pooling the three fitting splits and, separately, the three selection splits yields 384 examples per role; the gains then become 4.8 points for $d = 24$ and 5.2 points for $d = 40$, the latter reaching 89.9% AUROC from the 84.7% TIES base. Within this range, the gains grow with both labels and capacity (Figure 2b).

Other merge bases and pipelines. PED improves the ViT-B/32 task-arithmetic base by 3.8 AUROC points ($[3.1, 4.5]$; Table 12). DisTaC followed by TIES reaches 73.1% accuracy and 85.8% AUROC, compared with 73.4% and 88.2% for TIES followed by PED (Table 7); this comparison is between complete pipelines with different adaptation and expert-access requirements.

7 Discussion and limitations

Accuracy and error detectability capture different consequences of merging. The agreement comparison exposes ranking losses even on retained decisions, and the linear construction shows how averaging alone can produce such losses. PED improves ranking both on each model's own predictions and on a common set of base decisions; the latter comparison attributes part of the gain to better separation of existing correct and incorrect predictions, independently of changes in their labels.

Our evaluation covers eight image-classification tasks, two CLIP backbones, and fixed expert checkpoints. Agreement-subset estimates are conditional on shared predictions, and three tasks have too few shared errors to enter the primary macro average. PED requires a small labeled set per task and known task identity at inference. Its preservation guarantee applies to the specified fitting examples, whereas held-out performance is measured empirically. Fixed-label repair is evaluated on the eligible agreement subsets; aggregate accuracy gains do not measure retention of every base decision. The intervals describe audit-sampling uncertainty conditional on the fitted models. The role of the nonlinear merge residual and behavior under distribution shift remain open questions.

8 Conclusion

Model merging can preserve every prediction and still reverse the confidence ranking of correct and incorrect decisions. Agreement-subset evaluation makes this distinction observable, and the competing-logit Jensen gap explains it exactly in linear models. PED improves error ranking through constrained local corrections, with consistent audit gains across tasks, merge methods, and backbones.

AI use statement

In this work, we used generative AI tools (large language models) solely for language polishing, including improving grammar, clarity, and readability. These tools were not used to generate substantive content or to perform any research tasks, including developing methods or theoretical frameworks, formulating claims or hypotheses, constructing proofs, designing experiments, generating or processing data, analyzing or interpreting results, writing code, searching the literature, translating text, or preparing figures. All AI-assisted revisions were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including all claims, results, and references.

References

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998. URL https://arxiv.org/abs/1703.00121.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10121–10143. PMLR, 2025. URL https://proceedings.mlr.press/v267/cheng25h.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. URL https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=D7KJmfEDQP.

Bethan Evans, Benjamin Etheridge, Stephen Roberts, and Jared Tanner. Model merging by output-space projection, 2026. URL https://arxiv.org/abs/2605.29101.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yanggan Gu, Shuo Cai, Zihao Wang, Wenjun Wang, Yuanyi Wang, Pengkai Wang, Sirui Huang, Su Lu, Jianmin Wu, and Hongxia Yang. FeatCal: Feature calibration for post-merging models. *arXiv preprint arXiv:2605.13030*, 2026. URL https://arxiv.org/abs/2605.13030.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. URL https://arxiv.org/abs/1709.00029.

Tiancheng Hu, Benjamin Minixhofer, and Nigel Collier. Navigating the alignment–calibration trade-off: A pareto-superior frontier via model merging. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42405–42422, 2026. URL https://aclanthology.org/2026.findings-acl.2104/.

Jincheng Huang, Jie Xu, Xiaoshuang Shi, Ping Hu, Lei Feng, and Xiaofeng Zhu. Revisiting confidence calibration for misclassification detection in VLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5393eaf0b723278738788520ec54e4b4-Abstract-Conference.html.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. URL https://www.cv-foundation.org/openaccess/content_iccv_workshops_2013/W19/html/Krause_3D_Object_Representations_2013_ICCV_paper.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://leon.bottou.org/papers/lecun-98h.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1453–1460, 2011. doi: 10.1109/IJCNN.2011.6033395. URL https://doi.org/10.1109/IJCNN.2011.6033395.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangliao Geng, and Boyang Li. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5745. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fbb4f2269267de23e0240e1c01242981-Abstract-Conference.html.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

Qi Wei, Shuo He, Enneng Yang, Tingcong Liu, Haobo Wang, Lei Feng, and Bo An. Representation surgery in model merging with probabilistic modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66015–66032, 2025. URL https://proceedings.mlr.press/v267/wei25c.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://dspace.mit.edu/entities/publication/092e56e9-7936-4ce3-91d4-43ae7ec24708.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a. URL https://arxiv.org/abs/2402.02705.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2310.02575.

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryotaro Shimizu, Ioannis Mitliagkas, and Hiroki Naganuma. DisTaC: Conditioning task vectors via distillation for robust model merging. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=W70w5JCzdq.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2311.03099.

Dengming Zhang, Xiaowen Ma, Zhenliang Ni, Zhenkai Wu, Han Shu, Xin Jiang, and Xinghao Chen. Expert merging: Model merging with unsupervised expert alignment and importance-guided layer chunking. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/24dbe8babeca9f372712bbf19dd3584e-Abstract-Conference.html.

A PROOFS

Lemma 1 (score differentials). Differentiating Equation (3) gives

$$\nabla s_i = \nabla z_{h_i} - \sum_{c \neq h_i} w_{ic} \nabla z_c = \sum_{c \neq h_i} w_{ic} \nabla m_{ic}.$$

Every score gradient therefore lies in $\mathrm{row}(J_{\mathrm{all}})$, and by the chain rule every differentiable function of the fixed-anchor scores has zero directional derivative along $\ker J_{\mathrm{all}}$. If $\mathrm{rank}\, J_{\mathrm{all}} = d$, this kernel is trivial; we do not assume this rank condition. In contrast, $J_C \delta = 0$ sets only the protected examples' score derivatives to zero. Unprotected scores can change, so the ranking objective need not be stationary on $\ker J_C$. All of these statements are first-order statements at the given iterate.

Remark 1 (Slack-limited score movement)**.** *For $i \in P_t$, the margin floors imply $s_i \geq -\log \sum_{c \neq h_i} e^{-\kappa_{ic}}$. A linearized feasible step δ satisfying $J_C \delta \geq -v_C$ obeys $\nabla s_i^\top \delta \geq -\sum_{c \neq h_i} w_{ic}(m_{ic} - \kappa_{ic})$. No such floor bound applies outside the protected set.*

Proof. The score $-\log \sum_c e^{-m_{ic}}$ is increasing in every margin, so substituting the floors yields the nonlinear lower bound. For a protected margin row, $(J_C)_{ic}\delta \geq -(m_{ic} - \kappa_{ic})$; multiplying by $w_{ic} \geq 0$ and summing yields the derivative bound. Because the coefficients are shared across examples, the per-example bounds need not be attainable simultaneously. □

Equation (5) **(trapezoidal AUGRC identity).** Consider a strict score order, and let D count the inversions in which an error is ranked above a correct example. The oracle order has $\sum_k E_k = e(e+1)/2$, and each adjacent swap that moves an error above a correct example increases this sum by one. Hence $\sum_k E_k = e(e+1)/2 + D$, with $D = ce(1 - \widehat{A})$. Because $E_0 = 0$ and $E_n = e$, trapezoidal integration of the generalized risk E_k/n over coverage k/n gives

$$\widehat{\mathrm{AUGRC}} = \frac{\sum_{k=1}^{n} E_k - e/2}{n^2} = \frac{e^2}{2n^2} + \frac{ce}{n^2}(1 - \widehat{A}).$$

Averaging uniformly over permutations of tied scores gives half credit to tied correct–error pairs and preserves this equality. The right-endpoint sum $n^{-2}\sum_k E_k$ instead contains the additional term $e/(2n^2)$. Ordinary risk–coverage integration uses E_k/k and yields a different metric. The macro generalized risk is the mean of the task-level areas and cannot be reconstructed from macro accuracy and macro AUROC alone. The corresponding population decomposition is due to Traub et al. (2024).

Equation (12) **and a step that lowers both the loss and AUROC.** For $u \in \mathbb{R}$ and $\beta > 0$, $\mathbf{1}\{u < 0\} + \frac{1}{2}\mathbf{1}\{u = 0\} \leq \log(1 + e^{-u/\beta})/\log 2$. Averaging over all pairs gives the first inequality; the smooth maximum dominates every task loss, which gives the second. For $\beta = \sigma = 1$, consider one incorrect score 0 and two correct scores $(0.01, 0.01)$: AUROC is 1, and the mean loss is about 0.688. Changing the correct scores to $(-0.01, 10)$ reduces the loss to about 0.349 but lowers AUROC to 0.5. Surrogate descent therefore cannot replace the eligibility check of Equation (16).

Proposition 2 (preservation of protected fitting predictions). Each protected anchor initially has margins above its positive floors. Equation (14) maintains these floors and checks the actual argmax, so the predictions on P_t remain correct at every accepted iterate. If $P_t = C_t^{\mathrm{fit}}$, every initially correct fitting example remains correct, and each initially incorrect example either stays incorrect or becomes correct; this proves the fitting-accuracy comparison with the base. If P_t is a proper subset, protection alone gives no such guarantee for the full fitting set.

Equation (16) supplies the macro fixed-label AUROC inequality for every candidate passed to selection, and the base satisfies it with equality. Armijo acceptance with a truly negative directional derivative gives strict surrogate descent at every accepted nonzero step. Because acceptance relies on exact checks, these conclusions hold even with approximate proposal Jacobians.

The ranking groups C_t, E_t and the anchors remain fixed on $\mathcal{H}_t$, so an initial error that becomes correct remains a negative in this ranking test. The eligibility guarantee therefore concerns neither the candidate's own correctness groups nor its own-prediction AUROC or AUGRC. Even at fixed labels, a macro-AUROC comparison alone does not ensure that macro AUGRC does not increase, because the task weights $c_t e_t / n_t^2$ differ.

Small-step acceptance. If the protected margins strictly exceed their floors, Φ has a locally Lipschitz gradient, and $g^\top \delta < 0$, then sufficiently small positive steps preserve these margins and satisfy the Armijo condition. If, in addition, the current iterate is eligible and every fixed-label pair gap is nonzero, a sufficiently small step preserves the sign of every pair and hence eligibility. Active floors, tied pair scores, and finite backtracking budgets preclude a general acceptance or convergence guarantee.

Proposition 3 (reachability duality; stated in Appendix D). For $v_C \geq 0$, $r > 0$, and $G \succ 0$, the set $\mathcal{C}_C$ contains zero and is compact, so its support value R_p is attained. The strict affine inequality $d_p + b_p^\top \delta > 0$ is therefore achievable exactly when $R_p > -d_p$. For feasible δ and any $\lambda \geq 0$,

$$b_p^\top \delta \leq b_p^\top \delta + \lambda^\top (J_C \delta + v_C) \leq \lambda^\top v_C + r \left\| G^{-1/2}(b_p + J_C^\top \lambda) \right\|_2 .$$

Strong duality holds under Slater's condition; in particular, $v_C > 0$ and $r > 0$ make zero strictly feasible (Boyd & Vandenberghe, 2004). The support criterion and the upper bound remain valid when some protected floors are active.

Nonlinear sufficient conditions. Define the G-metric Hessian norm at u as $\left\| G^{-1/2} \nabla^2 f(u) G^{-1/2} \right\|_{\mathrm{op}}$. Suppose that along $a + \rho\delta$, $\rho \in [0, 1]$, the pair Hessian norm is bounded by L_p and each protected margin Hessian norm by H_j. Then

$$d_p + b_p^\top \delta - \tfrac{1}{2} L_p \|\delta\|_G^2 > 0, \qquad m_{C,j} + (J_C)_j \delta - \tfrac{1}{2} H_j \|\delta\|_G^2 \geq \kappa_{C,j}$$

suffice for a nonlinear repair of the pair that retains the protected floors. To exclude every $\delta \in \mathcal{C}_C$ via $d_p + R_p + L_p r^2/2 \leq 0$, the pair Hessian bound must hold uniformly on $\{a + \rho\delta : \delta \in \mathcal{C}_C, \rho \in [0, 1]\}$. This excludes only the specified local set, not all nonlinear feasible steps or multistep trajectories. The empirical comparisons assume no certified Hessian bounds.

B Implementation and protocol

Table 5: Per-task allocation of the official test data for ViT-B/32 and ViT-L/14, using identical sample IDs for every role and repeat. The three fitting/selection pairs are disjoint from one another and from the base-selection and audit roles.

Role	Count	Use
Base selection	64	Choose the TIES scale by accuracy
Audit	936	Final model metrics
Fit, each of three repeats	128	Fit coefficients or the correctness probe
Selection, each repeat	128	Choose one operating point
Total allocated	1768	$64 + 936 + 3(128 + 128)$

The 64 base-selection examples are drawn from, and removed from, an initial pool of 1,000 evaluation examples; the remaining 936 form the audit set. The public experts were fine-tuned on the official training splits, and all adaptation and evaluation roles here use held-out test examples.

Table 6: Local-repair configuration shared by ViT-B/32 and ViT-L/14. Model-dependent quantities are computed separately for each backbone.

Component	Value
Margin floor fraction η_m	0.25
Surrogate temperature β, smooth-max ω	1, 1
Armijo constant c_A, proposal damping μ	10^{-4}, 0.1
Metric G	$V^\top V + 10^{-6} I$
Initial radius r_0	$0.16 \left\|\bar{\theta} - \theta_0\right\|_2$ on the selected layers
Slack reserve ξ	0.1
Budgets	12 outer steps, 20 backtracking trials, 3 radius halvings
Forward-difference step	10^{-2}
Gradient baselines	Adam, lr 5×10^{-3}, 40 steps, checkpoint every 4 steps
Probe	L_2 logistic regression; 10 regularization values in $[10^{-3}, 30]$

Selected layers and coefficient space. The main experiments use the visual encoder's last two residual blocks and final LayerNorm. With zero-based residual-block indexing, these are blocks $10, 11$ for ViT-B/32 and $22, 23$ for ViT-L/14. The ViT-B/32 capacity extension uses blocks $8, 9, 10, 11$ and the same final LayerNorm. Each block is one parameter group; the final LayerNorm's scale and bias form an additional group. For each of the eight experts, we restrict its task vector to each selected group and normalize the nonzero restriction to unit Euclidean norm, giving $d = 8(2+1) = 24$ in the main setting and $d = 8(4+1) = 40$ in the capacity extension. All other parameters stay at the merge base.

ViT-L/14 protocol. ViT-L/14 uses the same base-selection, audit, fitting, and selection samples as ViT-B/32, including all three adaptation repeats in Table 5. Its PED hyperparameters, optimization budgets, and selection rules are those in Table 6. For each L/14 merge base, we construct V from the L/14 task vectors and recompute $G = V^\top V + 10^{-6} I$ and $r_0 = 0.16 \left\|\bar{\theta} - \theta_0\right\|_2$ on its selected layers. The task score scales and protected margin floors are likewise recomputed from that L/14 base on each fitting split. Thus the backbone, selected block indices, and model-dependent quantities change, while the data protocol and numerical hyperparameters remain shared.

Algorithm inputs and acceptance. In every experiment, $P_t = C_t^{\text{fit}}$ protects all fitting examples that the merge base predicts correctly, and $\mathcal{H}_t = D_{\text{fit},t}$ uses the full fitting set for ranking and the fitting-AUROC eligibility test. We apply no additional τ threshold to either set. The parameter $\xi = 0.1$ reserves slack in the proposal constraints; it does not filter the protected or ranking examples. These sets, base anchors, and correctness labels remain fixed during optimization. With this choice of P_t, Proposition 2 preserves every initially correct fitting prediction and gives nondecreasing fitting accuracy for each task. Each outer step computes g_k by automatic differentiation and $J_{C,k}^{\text{prop}}$ by forward differences, screens inactive proposal rows, and solves Equation (13). Backtracking evaluates all protected examples in an actual forward pass and checks Equations (14)–(15); if the trial budget is exhausted, the radius is halved. Candidates that satisfy Equation (16), together with the base, enter selection. The selection rule permits at most a 0.5-point macro accuracy loss relative to the base and at most a 2.0-point loss on any task, and then minimizes the macro pairwise loss. The selected coefficients are folded into the merged weights.

LMN implementation and training. We retain the network and surrogate loss of Huang et al. (2026), following their Figure 5 and Appendix C.2. Separate fully connected layers project the logits, image embedding, and predicted-class text embedding; their concatenation is mapped to a positive sample-specific temperature through a fully connected layer and softplus. We train the meta network with SGD at learning rate 0.002, batch size 16, for 150 epochs on the fitting split, with the merged backbone and text head frozen. The original surrogate objective retains both its confidence term and full-probability constraint.

ProbSurgery implementation and training. We follow the published ProbSurgery implementation of Wei et al. (2025)[1], retaining its probabilistic correction network and representation-matching

[1] https://github.com/1998v7/ProbSurgery

objective. We train the correction module with Adam at learning rate 0.001, batch size 16, for $5{,}000$ iterations on the fitting images using expert features as targets. Training uses stochastic samples from the learned correction distribution; inference uses its mean. Both LMN and ProbSurgery use the shared selection split and operating-point rule above.

Candidate accounting. The local optimizer allows twelve outer steps, the gradient controls expose checkpoints, and the correctness probe uses a regularization grid. Under a common proposal budget, eligibility can change the number of candidates passed to selection. Equal budgets of fitting images and selection labels do not imply equal expert access or compute. Folding modifies existing weights without adding a correction module.

C Additional Experimental Results

C.1 Merging Baselines and Agreement Subsets

Table 7 compares complete pipelines, with the task-specific experts as a reference for performance before merging. Table 8 compares expert and TIES scores on identical instances and correctness groups; the three tasks with fewer than ten shared errors are excluded from the primary agreement macro.

Table 7: **Merging baselines and pre-merge conditioning.** Own-prediction macro accuracy and AUROC (%) over eight tasks.

Model / pipeline	Acc ↑	AUROC ↑
Zero-shot CLIP (Radford et al., 2021)	47.8	76.3
Task arithmetic (Ilharco et al., 2023)	67.9	82.8
TIES (Yadav et al., 2023)	71.7	84.7
DisTaC (Yoshida et al., 2026) → TIES (Yadav et al., 2023)	73.1	85.8
TIES (Yadav et al., 2023) → PED (ours)	**73.4**	**88.2**
Task experts (Ilharco et al., 2023)	90.9	90.8

Table 8: **Expert–TIES error ranking on agreement subsets.** Own-prediction accuracy and AUROC, and agreement-subset AUROC (%). Agreement columns evaluate identical instances and correctness groups. The primary agreement macro covers SUN397, Cars, RESISC45, SVHN, and DTD ($e_\cap \geq 10$). Each macro row averages every metric column over its task set, and $e_\cap$ is the total number of shared errors. Superscript s marks sparse-error estimates excluded from the primary macro; the eight-task agreement macro is descriptive only. Agreement coverage is 63.8% on the five eligible tasks and 71.2% on all eight.

	Expert (Ilharco et al., 2023)		TIES (Yadav et al., 2023)		Agreement AUROC		
Task	Acc	AUROC	Acc	AUROC	Expert	TIES	$e_\cap$
SUN397	78.1	84.6	63.1	79.1	84.2	71.7	81
Cars	79.2	83.7	58.5	79.4	84.2	71.2	63
RESISC45	95.4	93.0	69.0	81.4	92.1	59.9	13
EuroSAT	98.8	97.8	74.0	82.5	94.6^s	68.1^s	6
SVHN	98.0	89.9	87.5	88.7	89.0	76.9	12
GTSRB	99.3	97.8	77.7	91.8	96.4^s	80.2^s	2
MNIST	99.7	96.0	98.1	97.8	98.2^s	91.5^s	1
DTD	79.2	83.8	45.9	76.7	82.3	66.3	62
Macro (5 eligible)	86.0	87.0	64.8	81.1	86.4	69.2	231
Macro (all 8)	90.9	90.8	71.7	84.7	90.1	73.2	240

C.2 Direct paired comparisons

Matching adaptation repeats and audit examples gives the method-to-method AUROC differences in Table 9. Each interval is computed from paired resamples of the named pair rather than by subtracting the endpoints of two intervals against the base.

Table 9: **Direct paired AUROC differences on the TIES base.** Differences are PED minus the named comparator, in percentage points; the 95% intervals use matched adaptation repeats and audit instances. These comparisons differ from the changes relative to TIES in Table 1.

Comparison	AUROC difference (points)	Paired 95% CI
PED − LMN (Huang et al., 2026)	$+\mathbf{2.8}$	$[+2.1, +3.5]$
PED − ProbSurgery (Wei et al., 2025)	$+\mathbf{2.3}$	$[+1.6, +3.0]$

C.3 Correction capacity and labels

The larger coefficient space corrects the last four visual residual blocks (zero-based indices 8–11 in ViT-B/32) and final LayerNorm, giving $d = 40$. The main $d = 24$ space uses blocks $10, 11$ and final LayerNorm. The 384-example condition pools the fitting splits and, separately, the selection splits (Table 10). With this budget, $d = 40$ reaches 89.9% audit AUROC, a gain of 5.2 points over the TIES base of 84.7%.

Table 10: **Correction capacity and adaptation labels (ViT-B/32).** Fitting and selection use separate sets of the indicated size per task. Changes are in own-prediction audit AUROC percentage points relative to TIES (Yadav et al., 2023). Intervals are paired 95% audit-bootstrap intervals against the base; an em dash marks an unreported interval.

Correction space	d	Labels/role	Audit Δ	95% CI
Two blocks + final norm	24	128	$+3.5$	$[+2.9, +4.1]$
Four blocks + final norm	40	128	$+4.1$	$[+3.5, +4.8]$
Two blocks + final norm	24	384	$+4.8$	$[+4.2, +5.4]$
Four blocks + final norm	40	384	$+5.2$	—

C.4 Pairwise interpretation of the nonlinear residual

For a correct–error pair (i, j) scored with fixed anchors, define $\gamma = s_i^{\text{merge}} - s_j^{\text{merge}}$. Section 4 gives

$$\gamma = \sum_k \alpha_k \big(s_i^{(k)} - s_j^{(k)}\big) + \mathcal{J}_i - \mathcal{J}_j + r_i - r_j.$$

When $\gamma < 0$, subtracting the Jensen term from each score changes the pair gap to $\gamma + \mathcal{J}_j - \mathcal{J}_i$, which is strictly positive exactly when $\mathcal{J}_j - \mathcal{J}_i > -\gamma$. This is an algebraic diagnostic relative to the chosen logit ensemble. It leaves the merged weights unchanged and does not identify the effect of removing a mechanism from a nonlinear merging procedure.

C.5 Additional baseline metrics

Table 11 reports per-task metrics of zero-shot CLIP and task arithmetic, Table 12 compares PED gains on the two ViT-B/32 merge bases, and Figure 2 summarizes the audit gains of all corrections and of the capacity and label settings.

Table 11: Own-prediction audit metrics (%) of zero-shot CLIP and task arithmetic. Macro rows are computed from unrounded values.

	Zero-shot (Radford et al., 2021)		Task arithmetic (Ilharco et al., 2023)	
Task	Acc	AUROC	Acc	AUROC
SUN397	65.0	80.0	59.5	77.4
Cars	59.2	81.8	56.1	78.8
RESISC45	59.0	78.6	63.0	80.1
EuroSAT	43.5	74.0	73.9	77.2
SVHN	31.5	70.3	79.9	86.4
GTSRB	33.7	76.1	68.5	88.5
MNIST	48.3	75.1	96.5	96.7
DTD	42.6	74.3	45.6	77.4
Macro	47.8	76.3	67.9	82.8

Table 12: **PED audit gains on two merge bases (ViT-B/32).** Base AUROC is in percent; PED changes and paired 95% audit-bootstrap intervals are in percentage points.

Base	Base AUROC	PED ΔAUROC	95% CI
TIES (Yadav et al., 2023)	84.7	+3.5	[+2.9, +4.1]
Task arithmetic (Ilharco et al., 2023)	82.8	+3.8	[+3.1, +4.5]

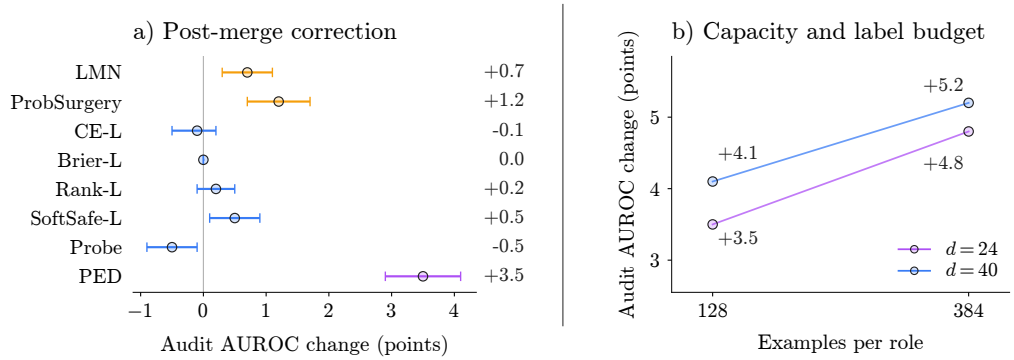


Figure 2: **Audit AUROC gains from correction and adaptation capacity.** (a) Changes relative to TIES for LMN (Huang et al., 2026), ProbSurgery (Wei et al., 2025), the internal controls, and PED; error bars show paired 95% audit-bootstrap intervals against the base. (b) PED gains for two coefficient dimensions and two fitting-label budgets, with selection using the same number of separate examples; the panel shows point estimates only, and lines connect the reported settings.

D REACHABILITY DIAGNOSTIC

Proposition 3 (Affine reachability and dual bound)**.** *Let* $\mathcal{C}_C = \{\delta : J_C\delta \geq -v_C,\ \|\delta\|_G \leq r\}$ *with* $v_C \geq 0$, $r > 0$, $G \succ 0$, *and* $R_p = \max_{\delta\in\mathcal{C}_C} b_p^\top \delta$. *With exact derivatives, strict affine repair of the fixed-label pair is possible if and only if* $R_p > -d_p$. *For every* $\lambda \geq 0$,

$$R_p \leq \lambda^\top v_C + r \left\| G^{-1/2}(b_p + J_C^\top \lambda) \right\|_2 . \tag{18}$$

The infimum over λ *equals* R_p *under Slater's condition, for which* $v_C > 0$ *suffices when* $r > 0$.

A feasible primal step with $b_p^\top \delta > -d_p$ certifies strict affine repairability. A valid dual upper bound no larger than $-d_p$ rules out strict affine repair within $\mathcal{C}_C$, whereas a large upper bound alone certifies nothing. Numerical comparisons require a stated solver tolerance.

With a forward-difference Jacobian, these statements concern the approximate affine model unless the derivative error is bounded. At a common iterate, radius, and Jacobian, $\mathcal{C}_C$ is an outer relaxation of the slack-reserved proposal set; this inclusion does not compare an exact-Jacobian diagnostic with an approximate proposal model. A multistep trajectory can leave the initial trust region, so a local diagnostic does not certify global impossibility.

The feasible set depends on the protected population, the retained margin rows, their floors, and the radius, so reachability labels for different protected sets are not interchangeable. The criterion describes the specified affine model; it is not an empirical preservation result.