NEVER OVERSHOOT: EXPERT-ANCHORED GRADIENT DESCENT FOR MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging combines task-specific experts fine-tuned from a common pretrained model into a single model, which often underperforms the experts on their respective tasks. We study refinement using a few labeled examples per task, already available for selecting or fitting the merge. We introduce expert-anchored gradient descent (XGD), which uses the original expert checkpoints to guide updates from these small buffers. XGD scales each coordinate's learning rate by its distance to the task expert and permits only movements toward that expert, without passing it. Replay gradients select locally favorable movements, and expert offsets are recomputed as refinement proceeds. On GLUE-8, CLIP-8 and CLIP-20 with 8 or 16 labels per task, both XGD and AdamW improve every evaluated merge in mean task performance, reusing the examples already used to select or fit it. Thus, few-shot refinement offers gains without additional labels. Across five buffer draws, XGD has the best mean among the compared refinements in five of six benchmark–budget settings. Starting directly from the pretrained model, XGD also outperforms many merge baselines, combining expert information without a separate merging algorithm. From the same merges at eight labels per task, XGD retains higher mean ImageNet accuracy than AdamW in every comparison and lower mean WikiText-2 masked-LM loss in all but one.

1 INTRODUCTION

A promising approach to building multi-task models is to combine separately fine-tuned variants of a shared pretrained model directly in parameter space, without extensive retraining, ensembling, or learning an explicit routing mechanism (Ilharco et al., 2023; Yadav et al., 2023). This makes merging attractive when task experts are already available and access to their original training data or resources for additional training is limited. However, the resulting model can still underperform the individual experts on their respective tasks. This leaves a second problem after merge construction: how to recover task performance using the limited supervision already available.

A small *labeled* replay buffer may still be available for each task even when the original training data is no longer accessible, for example to select merging hyperparameters or fit the combination of experts. Such buffers provide a lightweight source of task-specific supervision for refinement. The simplest strategy is to apply gradient descent to the replay data, starting from an initial merge or the pretrained model. Our experiments show that these labels remain useful after merge construction: both AdamW and XGD improve the selected merges by reusing the same examples, without additional supervision. Yet a gradient estimated from so few examples may be noisy or poorly aligned with the desired task behavior. The expert checkpoints contain information about parameter changes already found useful for their respective tasks. A refinement procedure can therefore use the replay signal to determine which movements toward those experts are useful at the current model.

We propose *expert-anchored gradient descent* (XGD), which uses task experts to set coordinate-wise learning rates during refinement. At each task update, XGD computes the task-loss gradient on the small labeled buffer and scales each coordinate by its current distance to the corresponding expert parameter. A directional gate permits only movements toward the expert, and a cap prevents passing its value. We recompute both the gradient and expert distances as refinement proceeds. The labeled examples thus provide the learning signal, while the checkpoints supply coordinate-wise step sizes,

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.

directions, and bounds. XGD can refine an existing merge or start directly from the pretrained model without a separate merging algorithm.

We evaluate XGD on GLUE-8, CLIP-8, and CLIP-20 with eight or sixteen labels per task. It improves mean task performance for every evaluated combination of eight merging methods, three benchmarks, and two label budgets, with larger gains generally occurring when the initial merge has a larger performance gap to the experts. The best mean among the compared refinements comes from a merge followed by XGD in five of six benchmark–budget settings. XGD's advantage over AdamW generally increases with the initial performance gap to the experts and is particularly pronounced from the pretrained model. From that initialization, it also outperforms many initial merges.

XGD also offers benefits for retaining pretrained capabilities outside the merged tasks. On CLIP-8, it improves both merged-task performance and zero-shot ImageNet accuracy for most initial merges. At eight labels per task, XGD achieves higher mean ImageNet accuracy than AdamW from every evaluated merge. On GLUE-8, it yields lower mean WikiText-2 masked-language-model loss than AdamW from all but one merge (Section 5.2). These retention advantages accompany larger parameter changes from the initial merge, indicating that update magnitude alone does not explain the retention differences.

Overall, we make the following contributions: (i) XGD, a refinement method that combines task-loss gradients with expert-derived coordinate-wise learning rates and movement constraints, starting from either a merge or the pretrained model; (ii) evidence across three benchmarks that labels already used to select or fit a merge remain useful for further improvement through either AdamW or XGD refinement; (iii) evidence that XGD generally outperforms matched AdamW refinement in task performance, with an advantage that tends to grow with the initial performance gap to the experts and is particularly pronounced from the pretrained model; (iv) evidence that, when refining merges, XGD generally retains more of the pretrained capabilities measured by our probes than matched AdamW refinement.

2 Background and Related Work

Task-vector merging combines the parameter changes made by experts fine-tuned from a common pretrained model. Task arithmetic adds a weighted sum of these changes, or *task vectors*, to the pretrained parameters (Ilharco et al., 2023). Interference between task vectors motivates methods that modify their components before combination: TIES trims small changes and resolves sign disagreements (Yadav et al., 2023), DARE randomly drops and rescales changes (Yu et al., 2024), and Breadcrumbs removes small and unusually large perturbations (Davari and Belilovsky, 2024). These methods use the structure of the expert checkpoints to determine which changes to retain in the merge.

Task data can also guide how the experts are combined. Fisher merging weights parameters by estimated Fisher information (Matena and Raffel, 2022). With labels, aTLAS learns coefficients for task-vector blocks (Zhang et al., 2024), and the learned variant of Localize-and-Stitch identifies sparse task-specific regions to combine (He et al., 2025). TATR uses gradients at the pretrained model to identify a region with limited interference (Sun et al., 2025). RegMean uses activation statistics to match linear-layer outputs (Jin et al., 2023). These approaches establish several ways to use limited task data when constructing a merge. For a broader overview of model merging, see Yang et al. (2026).

Several merging methods already assume access to a small labeled dataset for selecting hyperparameters or fitting the merge. We study how to exploit such limited supervision for refinement, starting from either an existing merge or the pretrained model directly. Ordinary fine-tuning updates the model using task-loss gradients; XGD additionally uses the expert checkpoints to guide those updates.

3 Expert-Anchored Gradient Descent (XGD)

3.1 Problem Setup

We are given T task experts, indexed by $\mathcal{T} = \{1, \ldots, T\}$, obtained by fine-tuning a common pretrained model. Let $\theta_0 \in \mathbb{R}^d$ and $\theta_i \in \mathbb{R}^d$ denote the pretrained and task-i expert parameters in

the shared parameter space to be merged, with $r \in \{1, \ldots, d\}$ indexing coordinates. The goal is to combine the experts' task abilities in a single parameter state θ.

For each task i, a small labeled replay buffer $\mathcal{D}_i^{\text{probe}}$ provides a differentiable loss $\mathcal{L}_i(\theta; \mathcal{D}_i^{\text{probe}})$. These buffers supply the data for refinement.

Refinement starts from an initial model $\theta^{(0)} \in \mathbb{R}^d$, which may be an existing merge or the pretrained model θ_0 itself. The expert checkpoints remain available throughout refinement. We seek to improve the initial model's performance across the tasks using the limited replay data.

3.2 INTRODUCING XGD

A naive way to refine a merged model or the pretrained initialization given a few labaled samples per task is to perform gradient descent on the replay losses. With few replay examples, however, these gradients may poorly represent the full task, and ordinary GD leaves the expert checkpoints unused. XGD instead moves each coordinate toward its expert value in proportion to the first-order predicted improvement from adopting that value, and never past it.

In every iteration, XGD goes through the tasks sequentially, and for each task i it computes the gradient of the replay loss and the offset to the expert, then makes a step aiming at improving performance of task i. Consider an update for task i at the current model parameters θ. Its replay gradient g_i and offset to the expert v_i[1] are

$$g_i = \nabla_\theta \mathcal{L}_i(\theta; \mathcal{D}_i^{\text{probe}}), \tag{1}$$

$$v_i = \theta_i - \theta. \tag{2}$$

For a learning rate $\eta \geq 0$, XGD updates each coordinate by

$$u_{i,r} = \begin{cases} \min\big(\eta|g_{i,r}|, 1\big)\, v_{i,r} & \text{if } g_{i,r}v_{i,r} < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The model then becomes $\theta \leftarrow \theta + u_i$.

The update follows from a first-order estimate of what each coordinate would gain from the expert's value. Let e_r be the r-th unit vector. Replacing coordinate r of the current model by the expert's value changes the replay loss, to first order, by

$$\mathcal{L}_i(\theta + v_{i,r}e_r) - \mathcal{L}_i(\theta) \approx g_{i,r}v_{i,r}. \tag{4}$$

XGD moves each coordinate toward the expert by an amount proportional to this predicted improvement. Where $g_{i,r}v_{i,r} \geq 0$, going towards the expert is predicted not to help and the coordinate does not move: this is the *gate*. Where $g_{i,r}v_{i,r} < 0$, the coordinate moves toward the expert by $\eta|g_{i,r}v_{i,r}|$, that is, η times the predicted improvement, but never by more than the distance $|v_{i,r}|$ itself: this is the *expert-distance cap*. Together they give Equation (3), since $\min(\eta|g_{i,r}|, 1)\, v_{i,r}$ is a step of length $\min(\eta|g_{i,r}v_{i,r}|, |v_{i,r}|)$ in the direction of $v_{i,r}$. Coordinates on which the expert's value promises a large improvement move far, coordinates on which it promises little barely move.

When the gate is active ($g_{i,r}v_{i,r} < 0$) and the cap is not reached ($\eta|g_{i,r}| < 1$), XGD takes the gradient step $-\eta g_{i,r}|v_{i,r}|$, with *per-coordinate learning rate* $\eta|v_{i,r}|$. The full update additionally blocks movements away from the expert and clips steps at its value. Unlike Adam's scaling from running gradient moments, XGD's scaling comes from the current expert distances, which are recomputed as the model changes and require no optimizer state. This also distinguishes XGD from projected gradient descent onto the coordinate-wise interval between the current model and the expert. On an admitted coordinate, projected descent takes a step of length $\eta|g_{i,r}|$ before clipping; XGD uses $\eta|g_{i,r}v_{i,r}|$, accounting for both the gradient and the distance to the expert. Section 5 evaluates both ingredients: the ablations support combining gating and distance scaling, with the benefit of scaling more consistent from merges than from pretrained.

The learning rate controls how many coordinates reach the cap, while the current expert distance bounds each coordinate step. We apply task updates sequentially and recompute the gradient and expert offset at the updated model before each step.

[1]The task vector $\tau_i = \theta_i - \theta_0$ is fixed, whereas the expert offset $v_i = \theta_i - \theta$ changes with the current model; they coincide when $\theta = \theta_0$.

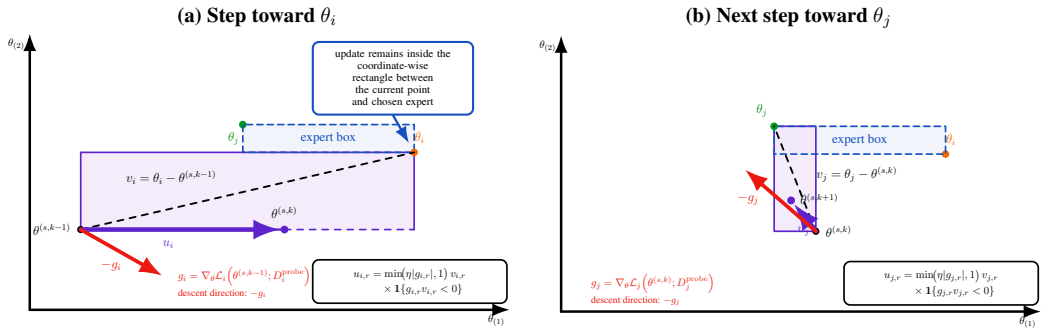


Figure 1: Sequential expert-anchored updates. (a) Starting from $\theta^{(s,k-1)}$, the task-i update u_i moves coordinate-wise toward θ_i and produces $\theta^{(s,k)}$. (b) The subsequent task-j update u_j moves from $\theta^{(s,k)}$ toward θ_j and produces $\theta^{(s,k+1)}$; the previous update is retained in light violet. Each violet rectangle is the coordinate-wise admissible region between the current iterate and the selected expert; the dashed blue rectangle is the interval hull spanned by θ_i and θ_j.

For the full procedure, a sweep visits all T tasks in an order π_s. Write $\theta^{(s,k)}$ for the model after step k of sweep s, starting with $\theta^{(s,0)} = \theta^{(s)}$ and ending with $\theta^{(s+1)} = \theta^{(s,T)}$. Figure 1 illustrates two steps; Algorithm 1 shows the pseudocode. Further implementation details are in Appendix A.

Algorithm 1 Sequential expert-anchored gradient descent (XGD)

Require: Task experts $\{\theta_i\}_{i=1}^T$; replay buffers $\{\mathcal{D}_i^{\text{probe}}\}_{i=1}^T$; initial model $\theta^{(0)}$ (any weight-space merge of the experts, or the pretrained model θ_0 itself); sweeps S; learning rate $\eta \geq 0$.

for $s = 0, \ldots, S-1$ **do**
  $\theta^{(s,0)} \leftarrow \theta^{(s)}$.
  Sample a random permutation π_s of $\{1, \ldots, T\}$.
  for $k = 1, \ldots, T$ **do**
    $i \leftarrow \pi_s(k)$.
    $g_i^{(s,k)} \leftarrow \nabla_\theta \mathcal{L}_i(\theta^{(s,k-1)}; \mathcal{D}_i^{\text{probe}})$ using task i's head.
    $v_i^{(s,k)} \leftarrow \theta_i - \theta^{(s,k-1)}$.
    for each mergeable coordinate r **do**
      $u_{i,r}^{(s,k)} \leftarrow \min\left(\eta\,|g_{i,r}^{(s,k)}|,\, 1\right) v_{i,r}^{(s,k)} \cdot \mathbf{1}\{g_{i,r}^{(s,k)} v_{i,r}^{(s,k)} < 0\}$. {gated, capped step toward θ_i; Eq. (3)}
    end for
    $\theta^{(s,k)} \leftarrow \theta^{(s,k-1)} + u_i^{(s,k)}$.
  end for
  $\theta^{(s+1)} \leftarrow \theta^{(s,T)}$.
end for

return $\theta^{(S)}$.

The coordinate interval constraint ensures that refinement cannot move farther from the interval hull spanned by the experts at any nonnegative learning rate.

Theorem 1 (Distance monotonicity and boundedness)**.** *Let $I_r = [\min_j \theta_{j,r}, \max_j \theta_{j,r}]$ and $B_{\text{exp}} = \prod_{r=1}^d I_r$ be the interval hull spanned by the experts, and let $d_p(\cdot, B_{\text{exp}})$ denote ℓ^p distance to it. For any initialization and learning rate $\eta \geq 0$, every update of Algorithm 1 satisfies*

$$d_p(\theta^{(s,k)}, B_{\text{exp}}) \leq d_p(\theta^{(s,k-1)}, B_{\text{exp}}), \qquad 1 \leq p \leq \infty.$$

Moreover, all iterates remain in the interval hull spanned by the initialization and the experts, and are therefore bounded.

Proof. By Equation (3), a task-i update from y to z satisfies $z_r = (1-\alpha_r) y_r + \alpha_r \theta_{i,r}$ with $\alpha_r \in [0, 1]$. Let q_r be the closest point in I_r to y_r. Since $q_r, \theta_{i,r} \in I_r$, their convex combination $w_r = (1-\alpha_r) q_r + \alpha_r \theta_{i,r}$ also belongs to I_r. Thus,

$$\text{dist}(z_r, I_r) \leq |z_r - w_r| = (1-\alpha_r)\,\text{dist}(y_r, I_r) \leq \text{dist}(y_r, I_r).$$

Taking ℓ^p norms of these coordinate distances proves the distance inequality, including the maximum norm for $p = \infty$. Each updated coordinate also stays between its current and expert values, so the interval hull spanned by the initialization and all experts is invariant. Induction gives boundedness of all iterates. □

The interval hull may be quite large and may contain regions of poor quality. Hence, the theorem should not be interpreted as a guarantee of performance improvement or retention of pretrained capabilities. It guarantees non-increasing distance to the interval hull spanned by the experts and bounded iterates.

4 Experimental Setup

Settings. We evaluate RoBERTa-base on GLUE-8 (Wang et al., 2018; Liu et al., 2019) and CLIP ViT-B/32 on eight- and twenty-task vision benchmarks (Radford et al., 2021; Ilharco et al., 2023; Wang et al., 2024), with the published task experts. Each task keeps its own head (the expert's classifier on GLUE, the frozen zero-shot text embeddings on CLIP); only the shared encoder is merged and refined. Task lists and checkpoint details are in Appendix B.

Protocol. Methods share a budget of $B \in \{8, 16\}$ labeled training examples per task, disjoint from evaluation. Every refinement runs for a fixed horizon of $S = 50$ sweeps and selects only its learning rate (Appendix B.1); the selected rate is then refit on all B examples from the same initialization. The selection objective is the held-out task metric averaged over the four rotations of a four-fold split of the budget, so at $B = 8$ each candidate is built on six examples per task and scored on the other two (Appendix B.1). Merge hyperparameters are selected on the same budget by the same rule. All adaptation comparisons report mean ± sample standard deviation over five buffer draws (seeds 103–107; see Appendix B). Refinement reuses the examples used to select or fit the merge; it receives no additional labels.

Comparisons. We evaluate merges before and after XGD refinement. Initializations include task arithmetic, TIES, DARE-TIES, Breadcrumbs, Fisher merging, aTLAS, TATR, and Localize-and-Stitch; additional baseline results and benchmark-specific coverage are in Appendix C. From the pretrained model, we compare XGD with ordinary GD and AdamW, which update parameters without expert guidance, and two ablations of XGD: *ungated distance-scaled refinement* removes the directional gate while retaining the expert-distance scaling and step cap; *gate-only refinement* removes the distance scaling while retaining the gate and step cap. All use the same buffers, sequential schedule, and selection procedure. These comparisons distinguish the effects of XGD's ingredients from the gains of ordinary fine-tuning. From every merge we also refine with AdamW under the same budget and selection, and from TIES and task arithmetic with all four replay families, so that composition gains can be attributed. Implementation details are in Appendices A and B.

Metrics. We report the mean of each task's standard evaluation metric, averaged across the five buffer draws. The metrics are top-1 accuracy for CLIP and the GLUE task metrics. Pretrained and single-task expert scores provide references. Appendix B defines the metrics, and Appendix E gives individual-task results.

5 Results

We first evaluate XGD on the merged tasks, starting from either existing merges or the pretrained model directly, then examine how refinement affects capabilities outside those tasks.

5.1 Performance on the Merged Tasks

Table 1 compares refinement from existing merges and from the pretrained model on three encoder benchmarks at both label budgets.

Table 1: Merged-task performance at $B = 8$ and $B = 16$ labels per task. Mean task metric $\pm$ sample std over five buffer draws. Each +XGD row refines the merge immediately above it; θ_0+ rows start from the pretrained model. Bold marks the highest mean among the displayed models in each column.

	$B = 8$			$B = 16$		
Model	**CLIP-8**	**GLUE-8**	**CLIP-20**	**CLIP-8**	**GLUE-8**	**CLIP-20**
Pretrained θ_0	0.478	0.354	0.550	0.478	0.354	0.550
Task arithmetic	$0.684_{\pm 0.010}$	$0.539_{\pm 0.028}$	0.604	$0.684_{\pm 0.010}$	0.552	$0.601_{\pm 0.003}$
+XGD	$0.725_{\pm 0.010}$	$0.652_{\pm 0.059}$	$0.637_{\pm 0.007}$	$0.740_{\pm 0.009}$	$0.680_{\pm 0.009}$	$0.654_{\pm 0.005}$
Breadcrumbs	$0.701_{\pm 0.011}$	$0.471_{\pm 0.014}$	$0.600_{\pm 0.007}$	0.705	$0.470_{\pm 0.017}$	0.603
+XGD	$0.731_{\pm 0.011}$	$0.665_{\pm 0.016}$	$0.649_{\pm 0.006}$	$0.750_{\pm 0.008}$	$0.662_{\pm 0.019}$	$0.659_{\pm 0.004}$
DARE-TIES	0.708	$0.552_{\pm 0.015}$	0.520	0.708	$0.546_{\pm 0.015}$	0.520
+XGD	$0.755_{\pm 0.006}$	$0.666_{\pm 0.014}$	$0.636_{\pm 0.010}$	$\mathbf{0.776}_{\pm 0.003}$	$0.678_{\pm 0.016}$	$0.668_{\pm 0.008}$
TIES	$0.737_{\pm 0.005}$	0.549	0.619	$0.733_{\pm 0.005}$	0.549	$0.615_{\pm 0.010}$
+XGD	$\mathbf{0.763}_{\pm 0.008}$	$0.655_{\pm 0.010}$	$\mathbf{0.665}_{\pm 0.008}$	$0.773_{\pm 0.006}$	$0.682_{\pm 0.003}$	$\mathbf{0.676}_{\pm 0.006}$
aTLAS	$0.717_{\pm 0.008}$	$0.600_{\pm 0.036}$	$0.645_{\pm 0.009}$	$0.721_{\pm 0.014}$	$0.612_{\pm 0.019}$	$0.655_{\pm 0.006}$
+XGD	$0.728_{\pm 0.006}$	$0.617_{\pm 0.034}$	$0.659_{\pm 0.008}$	$0.744_{\pm 0.011}$	$0.649_{\pm 0.015}$	$0.674_{\pm 0.004}$
TATR	$0.719_{\pm 0.002}$	$0.549_{\pm 0.014}$	$0.605_{\pm 0.006}$	$0.717_{\pm 0.006}$	$0.563_{\pm 0.014}$	$0.607_{\pm 0.001}$
+XGD	$0.744_{\pm 0.004}$	$\mathbf{0.678}_{\pm 0.011}$	$0.649_{\pm 0.004}$	$0.757_{\pm 0.014}$	$\mathbf{0.684}_{\pm 0.007}$	$0.667_{\pm 0.003}$
Fisher (replay)	$0.642_{\pm 0.031}$	$0.505_{\pm 0.035}$	$0.607_{\pm 0.009}$	$0.656_{\pm 0.024}$	$0.509_{\pm 0.018}$	$0.603_{\pm 0.008}$
+XGD	$0.698_{\pm 0.017}$	$0.608_{\pm 0.059}$	$0.641_{\pm 0.012}$	$0.718_{\pm 0.017}$	$0.625_{\pm 0.035}$	$0.651_{\pm 0.005}$
Localize-and-Stitch	$0.580_{\pm 0.061}$	$0.459_{\pm 0.105}$	$0.584_{\pm 0.013}$	$0.640_{\pm 0.031}$	$0.528_{\pm 0.044}$	$0.588_{\pm 0.008}$
+XGD	$0.632_{\pm 0.038}$	$0.570_{\pm 0.106}$	$0.617_{\pm 0.007}$	$0.687_{\pm 0.029}$	$0.646_{\pm 0.013}$	$0.621_{\pm 0.006}$
θ_0+AdamW	$0.540_{\pm 0.007}$	$0.453_{\pm 0.028}$	$0.582_{\pm 0.005}$	$0.566_{\pm 0.007}$	$0.462_{\pm 0.034}$	$0.598_{\pm 0.004}$
θ_0+GD	$0.497_{\pm 0.014}$	$0.404_{\pm 0.071}$	$0.564_{\pm 0.006}$	$0.532_{\pm 0.018}$	$0.431_{\pm 0.062}$	$0.571_{\pm 0.011}$
θ_0+ungated	$0.522_{\pm 0.013}$	$0.438_{\pm 0.024}$	$0.564_{\pm 0.009}$	$0.541_{\pm 0.012}$	$0.443_{\pm 0.036}$	$0.576_{\pm 0.007}$
θ_0+gate-only	$0.615_{\pm 0.076}$	$0.505_{\pm 0.086}$	$0.556_{\pm 0.007}$	$0.645_{\pm 0.028}$	$0.572_{\pm 0.013}$	$0.615_{\pm 0.029}$
θ_0+XGD	$0.643_{\pm 0.031}$	$0.535_{\pm 0.023}$	$0.623_{\pm 0.011}$	$0.703_{\pm 0.014}$	$0.561_{\pm 0.019}$	$0.647_{\pm 0.006}$

Table 2: XGD against AdamW from the same initialization. XGD minus AdamW in mean task metric, reported as mean $\pm$ sample std of the differences over five paired buffer draws. Both methods use the same budget, selection procedure, and horizon. Positive values favor XGD; the last row counts merges with a positive mean difference.

	$B = 8$			$B = 16$		
Model	**CLIP-8**	**GLUE-8**	**CLIP-20**	**CLIP-8**	**GLUE-8**	**CLIP-20**
Pretrained θ_0	$+0.102_{\pm 0.033}$	$+0.083_{\pm 0.012}$	$+0.041_{\pm 0.014}$	$+0.136_{\pm 0.013}$	$+0.098_{\pm 0.051}$	$+0.049_{\pm 0.005}$
Task arithmetic	$+0.008_{\pm 0.008}$	$+0.012_{\pm 0.009}$	$+0.011_{\pm 0.003}$	$+0.007_{\pm 0.004}$	$+0.001_{\pm 0.002}$	$+0.009_{\pm 0.006}$
Breadcrumbs	$+0.002_{\pm 0.005}$	$+0.015_{\pm 0.009}$	$+0.010_{\pm 0.007}$	$+0.002_{\pm 0.007}$	$+0.006_{\pm 0.030}$	$+0.005_{\pm 0.005}$
DARE-TIES	$+0.010_{\pm 0.005}$	$+0.022_{\pm 0.025}$	$+0.038_{\pm 0.014}$	$+0.007_{\pm 0.004}$	$+0.006_{\pm 0.016}$	$+0.032_{\pm 0.023}$
TIES	$+0.004_{\pm 0.004}$	$+0.028_{\pm 0.018}$	$+0.007_{\pm 0.005}$	$+0.001_{\pm 0.005}$	$+0.040_{\pm 0.015}$	$+0.006_{\pm 0.005}$
aTLAS	$+0.004_{\pm 0.002}$	$+0.005_{\pm 0.011}$	$+0.011_{\pm 0.004}$	$+0.001_{\pm 0.003}$	$+0.012_{\pm 0.011}$	$+0.008_{\pm 0.003}$
TATR	$-0.003_{\pm 0.001}$	$+0.010_{\pm 0.012}$	$+0.006_{\pm 0.004}$	$-0.001_{\pm 0.007}$	$-0.004_{\pm 0.008}$	$+0.008_{\pm 0.003}$
Fisher (replay)	$+0.017_{\pm 0.013}$	$+0.024_{\pm 0.038}$	$+0.008_{\pm 0.004}$	$+0.012_{\pm 0.007}$	$+0.027_{\pm 0.015}$	$+0.010_{\pm 0.004}$
Localize-and-Stitch	$+0.029_{\pm 0.034}$	$+0.025_{\pm 0.008}$	$+0.022_{\pm 0.011}$	$+0.010_{\pm 0.007}$	$+0.005_{\pm 0.015}$	$+0.009_{\pm 0.002}$
XGD ahead (merges)	7 of 8	8 of 8	8 of 8	7 of 8	7 of 8	8 of 8

Refining a merge. Both XGD and AdamW improve mean task performance over the initial merge in all 24 benchmark–initialization combinations at each budget (Tables 1 and 2). The same labels therefore remain useful for refinement after guiding merge selection or fitting, including for aTLAS, TATR, and Localize-and-Stitch, which use labels during construction. Gains tend to be larger when the initialization has a larger gap to the task experts. For example, Breadcrumbs on GLUE-8 improves from 0.471 to 0.665 at $B = 8$, whereas the already stronger aTLAS improves from 0.600 to 0.617. The highest mean among all evaluated models comes from an XGD composition in five of six benchmark–budget settings; TATR+AdamW leads on GLUE-8 at $B = 16$.

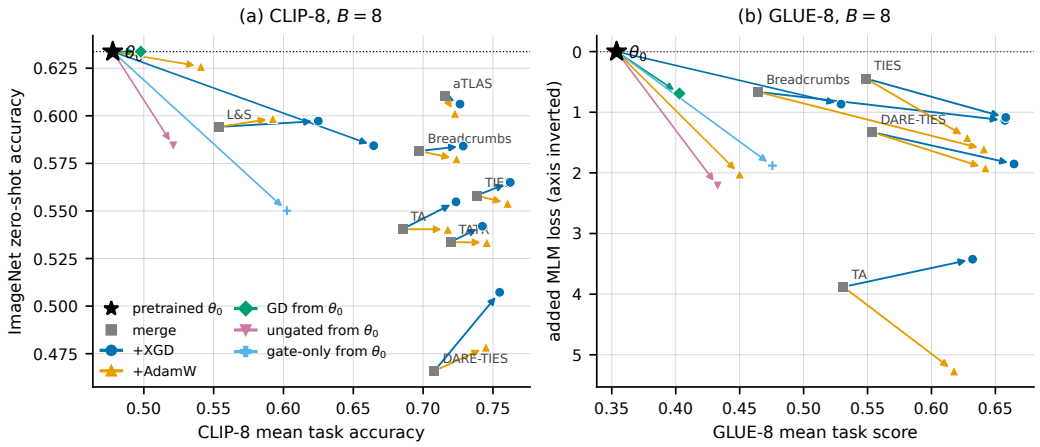


Figure 2: **Merged-task score against a held-out retention probe at $B = 8$.** (a) CLIP-8 mean accuracy against zero-shot ImageNet accuracy; (b) GLUE-8 mean score against the masked-LM loss added over the pretrained model, read through the original LM head on WikiText-2, with the axis inverted so that up is better in both panels. One point per model, mean over five buffer draws. Panel (b) omits Breadcrumbs and DARE-TIES for readability; Table 7 reports all eight methods. Grey squares are initial merges; solid blue arrows connect each merge to its XGD refinement (circles), and dashed orange arrows connect it to its AdamW refinement (triangles). The from-pretrained arms are arrows out of θ_0 (black star); the dotted line is the pretrained model's probe score. Full values and parameter distances are in Tables 5 and 7; Table 5 also reports ImageNet-V2, and Table 6 gives the vision results at $B = 16$.

These gains suggest that selecting or fitting a merge does not exhaust the value of its labels. A possible explanation is that refinement updates the current parameters directly, allowing corrections beyond the original merge construction. The AdamW results establish this benefit without expert anchoring; XGD adds expert guidance to that refinement step.

Compared with AdamW from the same merge, XGD has the higher mean in 23 of 24 comparisons at $B = 8$ and 22 at $B = 16$ (Table 2). The exceptions favor TATR+AdamW on CLIP-8 at both budgets and GLUE-8 at $B = 16$. Several margins are small relative to paired-draw variation. Further controls from TIES and task arithmetic favor XGD in all six benchmark–initialization means, though its advantage over ungated refinement is sometimes negligible (Table 4).

Across initializations, XGD's advantage over AdamW generally grows with the initial performance gap to the task experts and is most pronounced from the pretrained model (Table 2). This pattern is consistent with the motivation for expert anchoring, although its strength varies across benchmarks and label budgets.

Refining from the pretrained model. XGD exceeds GD, ungated refinement, and AdamW in mean task score on every benchmark at both budgets, and exceeds AdamW on every paired draw. It also leads gate-only refinement in five of six means; gate-only leads on GLUE-8 at $B = 16$ (0.572 versus 0.561). XGD wins every draw against gate-only on CLIP-8 at $B = 16$ and CLIP-20 at both budgets. At $B = 8$, however, the CLIP-8 and GLUE-8 mean advantages are largely driven by one poor gate-only draw each.

The ablations distinguish two uses of the experts. Ungated refinement retains distance-based step sizes and magnitude caps but permits movement away from the experts; its lower scores support the gate. Gate-only retains the direction constraint and cap. Although its comparison with XGD varies from pretrained, the merge controls more consistently favor scaling: XGD exceeds gate-only in all six TIES/task-arithmetic means by 0.014–0.038 (Table 4), though not on every draw. Together, these results support combining the gate and distance scaling, with benefits that depend on initialization.

Without a separate merging algorithm, pretrained-start XGD outperforms 28 of the 48 initial-merge cells, including every CLIP-20 merge except aTLAS at both budgets. On GLUE-8 it exceeds three

Table 3: Distances and retention before and after refinement at $B = 8$. $d_0 = \|\theta - \theta_0\|_2$ is distance to the pretrained model; $d_{\text{init}} = \|\theta - \theta^{(0)}\|_2$ is distance from the initialization listed in each row. Initial denotes the unrefined model. Distances are means over five buffer draws; retention scores are mean ± sample std. Higher ImageNet accuracy and lower masked-LM loss increase over the pretrained model (ΔMLM) indicate better retention. Compare distances within each benchmark. Full probe results are in Tables 5 and 7.

	d_0			d_{init}		Retention	
Initial model	Initial	XGD	AdamW	XGD	AdamW	XGD	AdamW
CLIP-8: ImageNet accuracy ↑							
Pretrained θ_0	0.000	0.730	0.186	0.730	0.186	$0.596_{\pm 0.020}$	$0.626_{\pm 0.007}$
Task arithmetic	1.75	1.70	1.76	0.141	0.129	$0.555_{\pm 0.037}$	$0.534_{\pm 0.048}$
Breadcrumbs	2.08	2.01	2.08	0.149	0.115	$0.580_{\pm 0.017}$	$0.572_{\pm 0.022}$
DARE-TIES	5.25	5.06	5.25	0.295	0.154	$0.508_{\pm 0.010}$	$0.483_{\pm 0.012}$
TIES	3.53	3.44	3.54	0.162	0.160	$0.574_{\pm 0.015}$	$0.562_{\pm 0.015}$
aTLAS	1.51	1.48	1.52	0.096	0.095	$0.609_{\pm 0.005}$	$0.604_{\pm 0.004}$
TATR	2.16	2.09	2.17	0.148	0.129	$0.547_{\pm 0.013}$	$0.540_{\pm 0.013}$
Fisher (replay)	1.78	1.69	1.80	0.205	0.171	$0.604_{\pm 0.005}$	$0.594_{\pm 0.007}$
Localize-and-Stitch	0.62	0.69	0.65	0.147	0.120	$0.601_{\pm 0.033}$	$0.599_{\pm 0.040}$
GLUE-8: ΔMLM loss ↓							
Pretrained θ_0	0.000	2.615	2.328	2.615	2.328	$0.95_{\pm 0.14}$	$1.56_{\pm 1.24}$
Task arithmetic	114.8	113.5	114.9	4.254	1.760	$3.30_{\pm 0.44}$	$5.16_{\pm 1.07}$
Breadcrumbs	98.7	98.3	98.7	1.631	1.274	$1.24_{\pm 0.29}$	$1.70_{\pm 0.70}$
DARE-TIES	164.0	162.7	164.0	4.579	0.922	$1.77_{\pm 0.48}$	$2.10_{\pm 1.00}$
TIES	90.0	89.5	90.0	2.667	0.848	$1.14_{\pm 0.11}$	$1.17_{\pm 0.38}$
aTLAS	50.7	50.1	50.7	2.280	0.548	$1.41_{\pm 0.18}$	$1.21_{\pm 0.21}$
TATR	126.7	126.2	126.7	1.716	1.442	$2.24_{\pm 0.56}$	$3.24_{\pm 0.71}$
Fisher (replay)	19.7	19.2	19.8	1.742	1.591	$1.20_{\pm 0.43}$	$1.53_{\pm 0.82}$
Localize-and-Stitch	20.8	19.3	20.9	2.966	1.348	$1.35_{\pm 0.49}$	$3.66_{\pm 2.95}$

of eight merges at $B = 8$ and six at $B = 16$; on CLIP-8 it exceeds two and three, respectively. It remains below aTLAS and the best XGD composition in each setting. Appendix C gives further comparisons, and Appendix E reports individual-task results.

5.2 Retention of Pretrained Capabilities

We probe capabilities outside the merged tasks using zero-shot ImageNet-1k accuracy for CLIP and WikiText-2 masked-language-model loss through RoBERTa's original pretrained LM head (Merity et al., 2016). ImageNet is absent from the experts' training sets; WikiText-2 samples the Wikipedia domain included in RoBERTa's pretraining, testing its original objective without fitting a new head (Liu et al., 2019). We use the pretrained model as reference and report $B = 8$ here; Appendix D gives probe details and additional vision results at $B = 16$.

Vision. On CLIP-8, XGD raises both mean merged-task and ImageNet accuracy for seven of eight initial merges; aTLAS gains only on the merged tasks (Figure 2a). XGD retains higher mean ImageNet accuracy than AdamW from every merge, while AdamW reduces it from five of eight initializations.

Table 3 shows that XGD moves farther from the initial merge than AdamW in all eight comparisons but ends closer to θ_0 in seven. For DARE-TIES, XGD moves 0.295 from the merge versus 0.154 for AdamW, yet ends at distance 5.06 from θ_0 versus 5.25, with ImageNet accuracy 0.508 versus 0.483. Localize-and-Stitch is the exception: XGD ends farther from θ_0 while retaining slightly higher ImageNet accuracy. Parameter distance alone therefore does not determine retention.

From the pretrained model, XGD makes larger task gains than AdamW but loses more ImageNet accuracy and moves farther from θ_0. Its merged-task and ImageNet means are 0.643 and 0.596, compared with 0.540 and 0.626 for AdamW. Against gate-only refinement, it improves both mean task performance and ImageNet accuracy at $B = 8$; at $B = 16$, it improves task performance while

retaining slightly lower ImageNet accuracy (Table 6). The retention advantage from merges therefore does not extend uniformly to pretrained-start refinement.

Text. On GLUE-8, XGD retains lower mean masked-LM loss than AdamW from seven of eight merges, with aTLAS the exception (Table 7; Figure 2b shows six methods): its added loss is 1.41 for XGD versus 1.21 for AdamW. This is a relative advantage: XGD still raises loss from six initial merges, improving it only from task arithmetic and Localize-and-Stitch. It ends closer to θ_0 despite moving farther from the initial merge in all eight comparisons (Table 3). From pretrained, XGD achieves both higher mean task scores and lower mean masked-LM loss than AdamW: it gains 0.181 on the merged tasks and adds 0.95 to MLM loss, compared with 0.099 and 1.56 for AdamW. Retention differences vary across draws, however: AdamW retains more from pretrained on two of five draws, and the TIES means are nearly tied (Table 7).

Role of expert anchoring. Expert anchoring offers a possible explanation for these retention patterns: the gate allows only locally favorable movements toward an expert, while the expert distance scales and caps the step. Coordinates already close to an expert can therefore move only a little. AdamW has no corresponding expert-based constraint, so a smaller update may leave it near the initial merge. If merging amplifies parameter changes beyond the range of expert values, moving toward an expert may partially reverse that amplification. Where those changes point away from θ_0, this could explain why XGD ends closer to the pretrained model despite moving farther from the merge than AdamW.

6 Conclusion

A few labeled examples remain useful after they have already guided merge construction or selection. Reusing those same examples for either AdamW or XGD refinement improves every evaluated merge in mean task performance, without acquiring additional labels. This supports refinement as a complement to merging in its own right. XGD additionally uses expert checkpoints to guide the updates: replay gradients identify locally favorable movements toward each expert, while expert distances scale and bound those movements.

XGD gains tend to be larger for initializations with larger gaps to the task experts, consistent with the motivation for expert anchoring. Starting directly from the pretrained model, XGD combines expert information without a separate merging algorithm and is competitive with many merge baselines. Both starting points benefit from expert guidance: XGD generally achieves higher mean task scores than AdamW from the same initialization, with larger differences from the pretrained model.

Expert-guided refinement also offers favorable retention of capabilities beyond the merged tasks. At eight labels per task, XGD achieves higher mean zero-shot ImageNet accuracy than AdamW from every evaluated merge on CLIP-8, and lower mean WikiText-2 masked-language-model loss from all but one merge on GLUE-8 (Section 5.2). On CLIP-8, it improves both merged-task and ImageNet accuracy for most initial merges. These findings suggest that expert anchoring can help preserve pretrained capabilities while adapting to the merged tasks.

Limitations. Our study covers two encoder architectures with published task experts and five replay-buffer draws per cell; shared-output architectures and other expert-training seeds remain untested. The set of initializations is a sample of the merging literature rather than an exhaustive one. Improvements in mean score can also accompany worse scores on individual tasks. Finally, retention is measured with two probes, zero-shot ImageNet accuracy for CLIP-8 and masked-LM loss for GLUE-8, which cover only part of the pretrained models' capabilities.

Reproducibility statement

Code, job files and result files are available at `https://anonymous.4open.science/r/XGD-6118`. The README gives the exact commands to rerun each experiment.

REFERENCES

MohammadReza Davari and Eugene Belilovsky. Model Breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, 2024.

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. Localize-and-Stitch: Efficient model merging via sparse task arithmetic. *Transactions on Machine Learning Research*, 2025.

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.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations*, 2023.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Michael S. Matena and Colin A. Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

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 *International Conference on Machine Learning*, 2021.

Wenju Sun, Qingyong Li, Wen Wang, Yangli-ao Geng, and Boyang Li. Task Arithmetic in Trust Region: A training-free model merging approach to navigate knowledge conflicts. *arXiv preprint arXiv:2501.15065*, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, 2018.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *International Conference on Machine Learning*, 2024.

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*, 2023.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8), Article 216, 2026. https://doi.org/10.1145/3787849.

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 *International Conference on Machine Learning*, 2024.

Frederic Z. Zhang, Paul Albert, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Knowledge composition using task vectors with learned anisotropic scaling. In *Advances in Neural Information Processing Systems*, 2024.

A METHOD DETAILS

Replay ablations and implementation. At the same within-sweep state and with $v_i = \theta_i - \theta$, the two ablations apply

$$u_{i,r}^{\mathrm{GD}} = -\eta g_{i,r}, \qquad u_{i,r}^{\mathrm{ungated}} = \mathrm{clip}(-\eta g_{i,r}|v_{i,r}|, -|v_{i,r}|, |v_{i,r}|)\,.$$

The ungated arm keeps the anchor-derived per-coordinate learning rate $\eta|v_{i,r}|$ and the symmetric magnitude cap, but allows steps away from the expert. It therefore does not preserve the current-to-expert interval and is not covered by Theorem 1. The gate-only arm, $u_{i,r}^{\text{gate-only}} = \mathrm{clip}(-\eta\, m_{i,r}\, g_{i,r}, -|v_{i,r}|, |v_{i,r}|)$, keeps the gate and the cap but steps along the raw gradient; it is covered by the theorem, and its learning-rate grid is $\{10^{-5}, \ldots, 1\}$ in decades, extended like every other arm. All three procedures use sequential task updates, constant learning rates, and no momentum or weight decay. The AdamW baseline applies the decoupled-weight-decay Adam update of Loshchilov and Hutter (2019) to the same plain replay gradient $g_{i,r}$, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, bias correction, zero weight decay, and a single moment state carried across every task step and sweep of a run; it uses no gate, scaling, or cap, and starts each selection candidate and the refit from zero moments. It is the standard few-shot fine-tuning recipe under XGD's schedule and selection rule, and it is the only arm whose per-coordinate step size is normalized independently of the gradient scale. The reported XGD runs use a zero gate threshold, no RMS normalization, and a task order reshuffled each sweep with seed 0. Each update accumulates a gradient over the current task buffer with dropout disabled. GLUE uses classification cross-entropy and unweighted STS-B mean-squared error; CLIP uses cross-entropy.

B EXTENDED EXPERIMENTAL PROTOCOL

This appendix specifies replay sampling, selection, implementation details, and the scope of the reported comparisons.

Benchmark settings. GLUE uses RoBERTa-base with the task experts specified in `configs/glue8.yaml`; the shared encoder is merged and each expert's supervised head is frozen. Evaluation uses the GLUE validation splits, with *matched validation only* for MNLI, and text truncated to 128 tokens. CLIP uses `openai/clip-vit-base-patch32` with the published task experts and class mappings configured in `src/apr/vision.py`; the vision tower is merged, while the pretrained visual projection, text-derived class embeddings, and logit scale remain fixed. The eight tasks are SUN397, Cars, RESISC45, EuroSAT, SVHN, GTSRB, MNIST, and DTD. CLIP-20 adds the twelve tasks listed in the held-out probe description below. All checkpoints and preprocessing settings are also recorded in the experiment configurations and result files.

Construction and compute. The evaluated aTLAS variant fits per-task, per-parameter-block coefficients initialized at zero, using 300 AdamW steps with learning rate 10^{-3}, weight decay 0.1, and batch size 16 on the full budget. Localize-and-Stitch selects learned-mask settings on the construction folds and refits its masks on all B examples; its recorded grid uses 300 training steps, penalties $\{0, 10^{-7}, 10^{-6}, 10^{-5}\}$, sparsities $\{0.01, 0.05, 0.1\}$, and scales $\{0.1, 0.2, 0.3, 0.5, 1\}$. Fisher construction estimates expected diagonal information and selects model coefficients through a 50-point simplex search.

Replay data. For each task, we first draw the full budget $B \in \{8, 16\}$ from its training split, then shuffle its indices and partition them into four disjoint folds. The full GLUE and CLIP classification budgets are class-balanced where possible: when $C \le B$, each class contributes $\lfloor B/C \rfloor$ examples and the remaining slots are sampled from the unused examples. When $C > B$, we sample B distinct classes and one example per class. The subsequent fold split is not stratified, so an individual fold need not be balanced. Sampling and partitioning are deterministic given the buffer seed. The same draw is reused across methods; none of the evaluation examples enters the adaptation budget.

Protocol details. Each learning rate is evaluated across four folds, with every run starting from the same initialization. We select the rate with the highest mean held-out task metric and restart

refinement on all B examples. Ties favor smaller learning rates or merge scales. Undefined GLUE selection metrics are assigned zero.

The encoder learning-rate grids are explicitly $\{1, 2, 4, 8, 16, 32\}$ for XGD, $\{0.5, 1, 2, 4, 8, 16\}$ for ungated replay, and $\{10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$ for GD. The AdamW baseline uses $\{10^{-6}, 2\times10^{-6}, \ldots, 2.56\times10^{-4}\}$.

B.1 Selection and refitting

The main task comparisons use metric-based, fixed-horizon select-then-refit on five buffer seeds 103–107. Every hyperparameter, those of the merges included, is selected by averaging the held-out task metric over the four rotations of a four-fold split of the buffer.

Choice of the horizon. We fix $S = 50$ sweeps for every refinement method and select only the learning rate, which gives all methods the same sweep budget and avoids choosing a second hyperparameter from validation folds of a few examples per task.

Per-task metrics. GLUE-8 comprises CoLA, SST-2, MRPC, STS-B, QQP, MNLI, QNLI, and RTE. CoLA is scored by Matthews correlation, STS-B by the mean of Pearson and Spearman correlations, MRPC and QQP by the mean of accuracy and F1, MNLI by matched-validation accuracy, and the remaining tasks by accuracy. Every CLIP task is scored by top-1 accuracy. We compute the benchmark mean within each draw before summarizing across draws. GLUE's pretrained reference uses the experts' frozen supervised heads, so it is not a zero-shot classifier. Raw minima across tasks are omitted because task difficulty and metric scales differ.

Baselines. The active comparisons are those in Tables 1, 2 and 4. Task arithmetic, TIES, DARE-TIES, and Breadcrumbs construct candidates from checkpoints; their hyperparameters are selected using the label budget. Expected-Fisher merging constructs candidates from the budget's unlabeled inputs and uses labels for selection. aTLAS, TATR and Localize-and-Stitch use labels in construction or reweighting. GD, ungated and gate-only refinement and the AdamW baseline are evaluated from the pretrained model; every composition applies XGD, and as a control AdamW, to the corresponding selected initialization, and TIES and task arithmetic are refined with all four replay families. Initialization selection and refinement deliberately draw on the same label budget: refinement reuses the labels that were already available for constructing and selecting the merge, so no method receives additional supervision beyond the shared budget.

C Additional Merged-Task Results

Replay comparisons from the pretrained model. XGD exceeds ungated replay in all six benchmark–budget means, supporting the gate's contribution beyond distance scaling and the cap. It also exceeds AdamW on every paired draw; the mean gap ranges from 0.041 to 0.136. AdamW has a higher mean than GD and ungated refinement in all six settings. Gate-only provides a stronger ablation: it leads XGD on GLUE-8 at $B = 16$, while XGD wins every draw on CLIP-8 at $B = 16$ and CLIP-20 at both budgets. On CLIP-8 and GLUE-8 at $B = 8$, the mean advantage of XGD coexists with gate-only wins on several draws. Thus, neither mean performance nor lower seed variation uniformly favors distance scaling.

Comparison with merging baselines. XGD from θ_0 outperforms 28 of the 48 initial-merge cells in Table 1, including 14 of the 24 checkpoint-only cells. On CLIP-20 it beats every merge except aTLAS at both budgets. On GLUE-8 it beats three of eight at $B = 8$ and six at $B = 16$. On CLIP-8 it trails all four checkpoint-only merges at $B = 8$ (0.643 versus 0.684–0.737), while at $B = 16$ it exceeds task arithmetic but trails the other three; it exceeds Fisher and Localize-and-Stitch at both budgets. aTLAS is the only initialization it never reaches. The best XGD composition gains 0.120 and 0.073 over the pretrained start on CLIP-8, 0.142 and 0.124 on GLUE-8, and 0.042 and 0.030 on CLIP-20, at $B = 8$ and $B = 16$, respectively.

Table 2 reports paired five-seed differences for every initialization. XGD has the higher mean from seven, eight and eight of the eight merges on CLIP-8, GLUE-8 and CLIP-20 at $B = 8$, and from

Table 4: Refining TIES and task arithmetic with each replay family at $B = 8$. Mean task metric $\pm$ sample std over five buffer draws. Merge is the unrefined initialization; bold marks the best refinement in each row.

Initialization	Merge	+XGD	+ungated	+gate-only	+GD	+AdamW
CLIP-8, TIES	$0.737_{\pm 0.005}$	$\mathbf{0.763}_{\pm 0.008}$	$0.758_{\pm 0.012}$	$0.742_{\pm 0.013}$	$0.741_{\pm 0.009}$	$0.759_{\pm 0.007}$
CLIP-8, Task arith.	$0.684_{\pm 0.010}$	$\mathbf{0.725}_{\pm 0.010}$	$0.713_{\pm 0.005}$	$0.711_{\pm 0.014}$	$0.691_{\pm 0.013}$	$0.717_{\pm 0.003}$
GLUE-8, TIES	0.549	$\mathbf{0.655}_{\pm 0.010}$	$0.633_{\pm 0.020}$	$0.631_{\pm 0.035}$	$0.620_{\pm 0.012}$	$0.627_{\pm 0.015}$
GLUE-8, Task arith.	$0.539_{\pm 0.028}$	$\mathbf{0.652}_{\pm 0.059}$	$0.651_{\pm 0.032}$	$0.614_{\pm 0.085}$	$0.617_{\pm 0.070}$	$0.640_{\pm 0.067}$
CLIP-20, TIES	0.619	$\mathbf{0.665}_{\pm 0.008}$	$0.658_{\pm 0.008}$	$0.651_{\pm 0.015}$	$0.636_{\pm 0.015}$	$0.658_{\pm 0.007}$
CLIP-20, Task arith.	0.604	$\mathbf{0.637}_{\pm 0.007}$	$0.616_{\pm 0.008}$	$0.609_{\pm 0.010}$	0.605	$0.626_{\pm 0.005}$

seven, seven and eight at $B = 16$. TATR favors AdamW on CLIP-8 at both budgets and on GLUE-8 at $B = 16$. Small positive means, such as task arithmetic on GLUE-8 at $B = 16$ (0.001), should not be read as statistically established improvements.

Twenty tasks. At $B = 16$, the best checkpoint-only CLIP-20 merge recovers about a fifth of the gap from the pretrained mean (0.550) to the expert reference (0.896), compared with roughly three fifths for TIES on CLIP-8. Task-arithmetic selection favors smaller coefficients on CLIP-20 than on CLIP-8.

Refinement controls from a merge. Table 4 refines TIES and task arithmetic at $B = 8$ with all replay families under four-fold selection, $S = 50$, and five draws. XGD has the highest mean in all six rows, although task arithmetic on GLUE-8 is almost tied with ungated refinement (0.652 versus 0.651) and has substantial seed variation. XGD exceeds gate-only in all six means, by 0.014–0.038, with paired wins on three to five of the five seeds. Plain GD has the lowest refinement mean in five of six rows; gate-only is slightly lower from task arithmetic on GLUE-8. These controls support an average benefit from combining the gate and distance scaling when refining these merges.

D Retention Probes and Extended Results

This appendix specifies the vision and text retention probes. All adapted models use the same five buffer draws and four-fold selection protocol as the merged-task results.

ImageNet zero-shot probe. For Figure 2 and Tables 5 and 6, the ImageNet-1k validation set and ImageNet-V2 are taken from the clip-benchmark webdataset release, with its 1,000 class names and the eighty OpenAI prompt templates; class embeddings are the template-averaged text features of CLIP ViT-B/32, and the pretrained model's 0.634 matches the published zero-shot accuracy of that backbone.

Masked-LM probe (GLUE-8). Each state's encoder is loaded under RoBERTa-base's original pretrained LM head and scored by masked-LM loss on WikiText-2 validation text; nothing is refit, so the probe reads a pretrained capability through the pretrained head, as the ImageNet probe does for CLIP. We concatenate the nonempty validation entries and use the first 256 complete blocks of 128 tokens (32,768 input tokens). Corruption is shared across states with seed 0: each token is selected with probability 0.15, then replaced by the mask token with probability 0.8, by a random vocabulary token with probability 0.1, or left unchanged otherwise, and loss is aggregated over the selected tokens. The probe was run at $B = 8$ only. Here, held-out refers to our merging and refinement protocol: RoBERTa was pretrained partly on Wikipedia (Liu et al., 2019), and we have not established that these articles were absent from its pretraining data. The resulting retention measure is specific to masked-token prediction on this corpus.

The vision track scales to the twenty-task benchmark of Wang et al. (2024) using the published fine-tuned ViT-B/32 encoders for all twenty tasks (the class-name orderings were reconstructed for three tasks, with zero-shot and expert scores respectively: PCAM $0.556 \to 0.887$, FER2013 $0.376 \to 0.711$, RenderedSST2 $0.586 \to 0.713$).

Table 5: ImageNet retention on CLIP-8 at $B = 8$. Zero-shot ImageNet-1k and ImageNet-V2 accuracy, merged-task mean (CLIP-8), and parameter distance to θ_0 (Dist.). Scores are mean $\pm$ sample std over five buffer draws (seeds 103–107) with four-fold selection; distances are means. Bold marks the highest ImageNet accuracy among adapted models.

Model	Dist.	CLIP-8	ImageNet	ImageNet-V2
Pretrained θ_0	0.00	0.478	**0.634**	0.558
θ_0+XGD	0.73	$0.643_{\pm 0.031}$	$0.596_{\pm 0.020}$	$0.519_{\pm 0.021}$
θ_0+AdamW	0.19	$0.540_{\pm 0.007}$	$0.626_{\pm 0.007}$	$0.552_{\pm 0.007}$
θ_0+GD	0.01	$0.497_{\pm 0.014}$	$\mathbf{0.634}_{\pm 0.001}$	$0.558_{\pm 0.002}$
θ_0+ungated	0.33	$0.522_{\pm 0.013}$	$0.591_{\pm 0.012}$	$0.516_{\pm 0.012}$
θ_0+gate-only	1.12	$0.615_{\pm 0.076}$	$0.560_{\pm 0.030}$	$0.483_{\pm 0.028}$
Task arithmetic	1.75	$0.684_{\pm 0.010}$	$0.533_{\pm 0.059}$	$0.458_{\pm 0.054}$
+XGD	1.70	$0.725_{\pm 0.010}$	$0.555_{\pm 0.037}$	$0.477_{\pm 0.037}$
+AdamW	1.76	$0.717_{\pm 0.003}$	$0.534_{\pm 0.048}$	$0.459_{\pm 0.045}$
Breadcrumbs	2.08	$0.701_{\pm 0.011}$	$0.575_{\pm 0.022}$	$0.496_{\pm 0.021}$
+XGD	2.01	$0.731_{\pm 0.011}$	$0.580_{\pm 0.017}$	$0.500_{\pm 0.017}$
+AdamW	2.08	$0.729_{\pm 0.013}$	$0.572_{\pm 0.022}$	$0.493_{\pm 0.021}$
DARE-TIES	5.25	0.708	0.466	0.396
+XGD	5.06	$0.755_{\pm 0.006}$	$0.508_{\pm 0.010}$	$0.434_{\pm 0.007}$
+AdamW	5.25	$0.745_{\pm 0.005}$	$0.483_{\pm 0.012}$	$0.412_{\pm 0.009}$
TIES	3.53	$0.737_{\pm 0.005}$	$0.566_{\pm 0.019}$	$0.490_{\pm 0.018}$
+XGD	3.44	$0.763_{\pm 0.008}$	$0.574_{\pm 0.015}$	$0.495_{\pm 0.016}$
+AdamW	3.54	$0.759_{\pm 0.007}$	$0.562_{\pm 0.015}$	$0.485_{\pm 0.014}$
aTLAS	1.51	$0.717_{\pm 0.008}$	$0.612_{\pm 0.003}$	$0.537_{\pm 0.004}$
+XGD	1.48	$0.728_{\pm 0.006}$	$0.609_{\pm 0.005}$	$0.534_{\pm 0.005}$
+AdamW	1.52	$0.724_{\pm 0.008}$	$0.604_{\pm 0.004}$	$0.530_{\pm 0.004}$
TATR	2.16	$0.719_{\pm 0.002}$	$0.541_{\pm 0.015}$	$0.463_{\pm 0.015}$
+XGD	2.09	$0.744_{\pm 0.004}$	$0.547_{\pm 0.013}$	$0.468_{\pm 0.014}$
+AdamW	2.17	$0.747_{\pm 0.003}$	$0.540_{\pm 0.013}$	$0.461_{\pm 0.012}$
Fisher (replay)	1.78	$0.642_{\pm 0.031}$	$0.603_{\pm 0.008}$	$0.526_{\pm 0.009}$
+XGD	1.69	$0.698_{\pm 0.017}$	$0.604_{\pm 0.005}$	$0.526_{\pm 0.005}$
+AdamW	1.80	$0.681_{\pm 0.028}$	$0.594_{\pm 0.007}$	$0.517_{\pm 0.006}$
Localize-and-Stitch	0.62	$0.580_{\pm 0.061}$	$0.597_{\pm 0.051}$	$0.522_{\pm 0.050}$
+XGD	0.69	$0.632_{\pm 0.038}$	$0.601_{\pm 0.033}$	$0.524_{\pm 0.032}$
+AdamW	0.65	$0.602_{\pm 0.057}$	$0.599_{\pm 0.040}$	$0.524_{\pm 0.039}$

Table 6: ImageNet retention on CLIP-8 at $B = 16$. Zero-shot ImageNet-1k and ImageNet-V2 accuracy, merged-task mean (CLIP-8), and parameter distance to θ_0 (Dist.). Scores are mean $\pm$ sample std over five buffer draws (seeds 103–107) with four-fold selection; distances are means. Bold marks the highest ImageNet accuracy among adapted models.

Model	Dist.	CLIP-8	ImageNet	ImageNet-V2
Pretrained θ_0	0.00	0.478	0.634	0.558
θ_0+XGD	0.81	$0.703_{\pm 0.014}$	$0.596_{\pm 0.012}$	$0.521_{\pm 0.012}$
θ_0+AdamW	0.24	$0.566_{\pm 0.007}$	$\mathbf{0.622}_{\pm 0.008}$	$0.547_{\pm 0.008}$
θ_0+GD	0.06	$0.532_{\pm 0.018}$	$0.620_{\pm 0.017}$	$0.545_{\pm 0.018}$
θ_0+ungated	0.44	$0.541_{\pm 0.012}$	$0.572_{\pm 0.027}$	$0.499_{\pm 0.025}$
θ_0+gate-only	0.55	$0.645_{\pm 0.028}$	$0.600_{\pm 0.023}$	$0.524_{\pm 0.023}$
Task arithmetic	1.75	$0.684_{\pm 0.010}$	$0.533_{\pm 0.059}$	$0.458_{\pm 0.054}$
+XGD	1.69	$0.740_{\pm 0.009}$	$0.559_{\pm 0.037}$	$0.481_{\pm 0.034}$
+AdamW	1.76	$0.734_{\pm 0.006}$	$0.538_{\pm 0.049}$	$0.462_{\pm 0.043}$
Breadcrumbs	2.16	0.705	0.565	0.486
+XGD	2.07	$0.750_{\pm 0.008}$	$0.577_{\pm 0.001}$	$0.497_{\pm 0.002}$
+AdamW	2.18	$0.748_{\pm 0.009}$	$0.559_{\pm 0.007}$	$0.481_{\pm 0.005}$
DARE-TIES	5.25	0.708	0.466	0.396
+XGD	4.95	$0.776_{\pm 0.003}$	$0.526_{\pm 0.013}$	$0.450_{\pm 0.012}$
+AdamW	5.25	$0.769_{\pm 0.002}$	$0.494_{\pm 0.005}$	$0.421_{\pm 0.004}$
TIES	3.25	$0.733_{\pm 0.005}$	$0.559_{\pm 0.022}$	$0.485_{\pm 0.021}$
+XGD	3.09	$0.773_{\pm 0.006}$	$0.577_{\pm 0.014}$	$0.499_{\pm 0.016}$
+AdamW	3.26	$0.772_{\pm 0.006}$	$0.558_{\pm 0.015}$	$0.482_{\pm 0.016}$
aTLAS	1.51	$0.721_{\pm 0.014}$	$0.612_{\pm 0.005}$	$0.537_{\pm 0.003}$
+XGD	1.48	$0.744_{\pm 0.011}$	$0.611_{\pm 0.005}$	$0.535_{\pm 0.005}$
+AdamW	1.52	$0.743_{\pm 0.012}$	$0.604_{\pm 0.006}$	$0.529_{\pm 0.005}$
TATR	2.02	$0.717_{\pm 0.006}$	$0.555_{\pm 0.030}$	$0.477_{\pm 0.030}$
+XGD	1.92	$0.757_{\pm 0.014}$	$0.564_{\pm 0.021}$	$0.488_{\pm 0.022}$
+AdamW	2.03	$0.758_{\pm 0.010}$	$0.554_{\pm 0.023}$	$0.477_{\pm 0.024}$
Fisher (replay)	1.75	$0.656_{\pm 0.024}$	$0.600_{\pm 0.030}$	$0.526_{\pm 0.029}$
+XGD	1.64	$0.718_{\pm 0.017}$	$0.606_{\pm 0.013}$	$0.530_{\pm 0.012}$
+AdamW	1.78	$0.705_{\pm 0.023}$	$0.590_{\pm 0.022}$	$0.514_{\pm 0.023}$
Localize-and-Stitch	0.95	$0.640_{\pm 0.031}$	$0.564_{\pm 0.047}$	$0.489_{\pm 0.046}$
+XGD	0.92	$0.687_{\pm 0.029}$	$0.593_{\pm 0.020}$	$0.517_{\pm 0.021}$
+AdamW	0.97	$0.677_{\pm 0.030}$	$0.587_{\pm 0.021}$	$0.511_{\pm 0.020}$

Table 7: Masked-language-model retention on GLUE-8 at $B = 8$. ΔMLM is the WikiText-2 validation loss increase over the pretrained model's 2.51 (lower is better). GLUE-8 is the merged-task mean; Dist. is parameter distance to θ_0. Scores are mean $\pm$ sample std over five buffer draws; distances are means.

Model	Dist.	GLUE-8	ΔMLM
Pretrained θ_0	0.0	0.354	0.00
θ_0+XGD	2.6	$0.535_{\pm 0.023}$	$0.95_{\pm 0.14}$
θ_0+AdamW	2.3	$0.453_{\pm 0.028}$	$1.56_{\pm 1.24}$
θ_0+GD	0.2	$0.404_{\pm 0.071}$	$0.78_{\pm 1.07}$
θ_0+ungated	1.8	$0.438_{\pm 0.024}$	$2.30_{\pm 0.34}$
θ_0+gate-only	3.4	$0.505_{\pm 0.086}$	$1.97_{\pm 1.22}$
Task arithmetic	114.8	$0.539_{\pm 0.028}$	$3.33_{\pm 1.84}$
+XGD	113.5	$0.652_{\pm 0.059}$	$3.30_{\pm 0.44}$
+AdamW	114.9	$0.640_{\pm 0.067}$	$5.16_{\pm 1.07}$
Breadcrumbs	98.7	$0.471_{\pm 0.014}$	$0.83_{\pm 0.36}$
+XGD	98.3	$0.665_{\pm 0.016}$	$1.24_{\pm 0.29}$
+AdamW	98.7	$0.650_{\pm 0.024}$	$1.70_{\pm 0.70}$
DARE-TIES	164.0	$0.552_{\pm 0.015}$	$1.28_{\pm 0.43}$
+XGD	162.7	$0.666_{\pm 0.014}$	$1.77_{\pm 0.48}$
+AdamW	164.0	$0.644_{\pm 0.013}$	$2.10_{\pm 1.00}$
TIES	90.0	0.549	0.44
+XGD	89.5	$0.655_{\pm 0.010}$	$1.14_{\pm 0.11}$
+AdamW	90.0	$0.627_{\pm 0.015}$	$1.17_{\pm 0.38}$
aTLAS	50.7	$0.600_{\pm 0.036}$	$1.02_{\pm 0.28}$
+XGD	50.1	$0.617_{\pm 0.034}$	$1.41_{\pm 0.18}$
+AdamW	50.7	$0.612_{\pm 0.039}$	$1.21_{\pm 0.21}$
TATR	126.7	$0.549_{\pm 0.014}$	$1.44_{\pm 0.54}$
+XGD	126.2	$0.678_{\pm 0.011}$	$2.24_{\pm 0.56}$
+AdamW	126.7	$0.668_{\pm 0.011}$	$3.24_{\pm 0.71}$
Fisher (replay)	19.7	$0.505_{\pm 0.035}$	$0.77_{\pm 0.34}$
+XGD	19.2	$0.608_{\pm 0.059}$	$1.20_{\pm 0.43}$
+AdamW	19.8	$0.583_{\pm 0.078}$	$1.53_{\pm 0.82}$
Localize-and-Stitch	20.8	$0.459_{\pm 0.105}$	$2.53_{\pm 2.46}$
+XGD	19.3	$0.570_{\pm 0.106}$	$1.35_{\pm 0.49}$
+AdamW	20.9	$0.545_{\pm 0.108}$	$3.66_{\pm 2.95}$

E PER-TASK RESULTS ON THE MERGED TASKS

These tables expand the aggregate results in Table 1 into fixed task columns. Each cell is the mean $\pm$ sample standard deviation over the same three buffer draws (seeds 103–105); at $B = 8$ the states are the four-fold-selected ones, at $B = 16$ the single-split ones, as in the main table. Pretrained and single-task expert references are fixed checkpoints. Each +XGD row refines the initialization immediately above it. Comparisons should be made within a task column: raw scores on different metrics are not directly comparable. Averaging the unrounded task scores recovers the reported suite mean; the standard deviation of that mean is computed across draws, not by averaging per-task standard deviations.

GLUE uses Matthews correlation for CoLA; mean Pearson/Spearman correlation for STS-B; mean accuracy/F1 for MRPC and QQP; and accuracy for the remaining tasks (matched validation for MNLI). CLIP uses top-1 accuracy. GradFix and DOGE/APGD at $B = 8$ were run on the first draw only.

Table 8: CLIP-8 per-task results at $B = 8$. Mean $\pm$ sample standard deviation over three buffer draws (GradFix and DOGE/APGD: one draw). Task metrics and reference rows are defined in Appendix E.

Model	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD
Pretrained θ_0	0.632	0.597	0.582	0.450	0.315	0.325	0.483	0.439
Single-task experts	0.749	0.783	0.949	0.991	0.972	0.989	0.996	0.794
Task arithmetic	$0.619_{\pm 0.042}$	$0.591_{\pm 0.033}$	$0.667_{\pm 0.034}$	$0.766_{\pm 0.029}$	$0.750_{\pm 0.025}$	$0.647_{\pm 0.033}$	$0.940_{\pm 0.018}$	$0.503_{\pm 0.026}$
+XGD	$0.630_{\pm 0.025}$	$0.611_{\pm 0.021}$	$0.698_{\pm 0.026}$	$0.828_{\pm 0.003}$	$0.808_{\pm 0.018}$	$0.743_{\pm 0.051}$	$0.938_{\pm 0.025}$	$0.534_{\pm 0.014}$
Breadcrumbs	$0.647_{\pm 0.010}$	$0.618_{\pm 0.009}$	$0.695_{\pm 0.003}$	$0.783_{\pm 0.013}$	$0.738_{\pm 0.052}$	$0.643_{\pm 0.043}$	$0.932_{\pm 0.029}$	$0.522_{\pm 0.002}$
+XGD	$0.647_{\pm 0.006}$	$0.626_{\pm 0.009}$	$0.720_{\pm 0.006}$	$0.824_{\pm 0.029}$	$0.801_{\pm 0.036}$	$0.730_{\pm 0.047}$	$0.937_{\pm 0.017}$	$0.546_{\pm 0.005}$
DARE-TIES	$0.595_{\pm 0.000}$	$0.577_{\pm 0.000}$	$0.663_{\pm 0.000}$	$0.713_{\pm 0.000}$	$0.862_{\pm 0.000}$	$0.792_{\pm 0.000}$	$0.977_{\pm 0.000}$	$0.486_{\pm 0.000}$
+XGD	$0.632_{\pm 0.006}$	$0.578_{\pm 0.052}$	$0.733_{\pm 0.008}$	$0.830_{\pm 0.007}$	$0.889_{\pm 0.010}$	$0.875_{\pm 0.012}$	$0.975_{\pm 0.004}$	$0.526_{\pm 0.006}$
TIES	$0.661_{\pm 0.010}$	$0.628_{\pm 0.014}$	$0.723_{\pm 0.016}$	$0.754_{\pm 0.027}$	$0.854_{\pm 0.008}$	$0.791_{\pm 0.015}$	$0.967_{\pm 0.003}$	$0.520_{\pm 0.014}$
+XGD	$0.666_{\pm 0.009}$	$0.638_{\pm 0.013}$	$0.748_{\pm 0.017}$	$0.820_{\pm 0.015}$	$0.874_{\pm 0.008}$	$0.845_{\pm 0.021}$	$0.966_{\pm 0.004}$	$0.542_{\pm 0.010}$
aTLAS	$0.667_{\pm 0.006}$	$0.652_{\pm 0.011}$	$0.727_{\pm 0.002}$	$0.801_{\pm 0.033}$	$0.665_{\pm 0.042}$	$0.747_{\pm 0.070}$	$0.866_{\pm 0.040}$	$0.599_{\pm 0.009}$
+XGD	$0.663_{\pm 0.005}$	$0.644_{\pm 0.013}$	$0.732_{\pm 0.006}$	$0.792_{\pm 0.037}$	$0.719_{\pm 0.028}$	$0.783_{\pm 0.066}$	$0.874_{\pm 0.044}$	$0.604_{\pm 0.009}$
TATR	$0.628_{\pm 0.001}$	$0.607_{\pm 0.000}$	$0.697_{\pm 0.001}$	$0.822_{\pm 0.001}$	$0.799_{\pm 0.000}$	$0.720_{\pm 0.001}$	$0.967_{\pm 0.000}$	$0.518_{\pm 0.002}$
+XGD	$0.631_{\pm 0.003}$	$0.609_{\pm 0.019}$	$0.716_{\pm 0.006}$	$0.847_{\pm 0.010}$	$0.843_{\pm 0.014}$	$0.793_{\pm 0.030}$	$0.963_{\pm 0.004}$	$0.538_{\pm 0.006}$
Fisher (replay)	$0.633_{\pm 0.039}$	$0.586_{\pm 0.050}$	$0.638_{\pm 0.104}$	$0.331_{\pm 0.071}$	$0.916_{\pm 0.044}$	$0.597_{\pm 0.241}$	$0.784_{\pm 0.036}$	$0.630_{\pm 0.082}$
+XGD	$0.635_{\pm 0.038}$	$0.604_{\pm 0.051}$	$0.694_{\pm 0.073}$	$0.613_{\pm 0.044}$	$0.913_{\pm 0.036}$	$0.668_{\pm 0.196}$	$0.847_{\pm 0.049}$	$0.630_{\pm 0.072}$
Localize-and-Stitch	$0.616_{\pm 0.035}$	$0.566_{\pm 0.061}$	$0.602_{\pm 0.009}$	$0.610_{\pm 0.074}$	$0.511_{\pm 0.187}$	$0.422_{\pm 0.099}$	$0.659_{\pm 0.138}$	$0.443_{\pm 0.001}$
+XGD	$0.620_{\pm 0.020}$	$0.564_{\pm 0.026}$	$0.628_{\pm 0.012}$	$0.691_{\pm 0.061}$	$0.676_{\pm 0.067}$	$0.549_{\pm 0.075}$	$0.796_{\pm 0.049}$	$0.477_{\pm 0.020}$
GradFix	0.638	0.569	0.625	0.590	0.484	0.463	0.692	0.469
DOGE/APGD	0.406	0.438	0.703	0.816	0.898	0.885	0.988	0.495
θ_0+GD	$0.634_{\pm 0.001}$	$0.592_{\pm 0.004}$	$0.587_{\pm 0.001}$	$0.557_{\pm 0.073}$	$0.354_{\pm 0.037}$	$0.325_{\pm 0.006}$	$0.499_{\pm 0.008}$	$0.435_{\pm 0.003}$
θ_0+ungated	$0.598_{\pm 0.008}$	$0.531_{\pm 0.024}$	$0.557_{\pm 0.031}$	$0.626_{\pm 0.018}$	$0.431_{\pm 0.015}$	$0.379_{\pm 0.043}$	$0.628_{\pm 0.027}$	$0.418_{\pm 0.012}$
θ_0+XGD	$0.611_{\pm 0.001}$	$0.495_{\pm 0.056}$	$0.617_{\pm 0.031}$	$0.754_{\pm 0.038}$	$0.818_{\pm 0.018}$	$0.631_{\pm 0.090}$	$0.900_{\pm 0.022}$	$0.491_{\pm 0.021}$

Table 9: CLIP-8 per-task results at $B = 16$. Mean $\pm$ sample standard deviation over three buffer draws. Task metrics and reference rows are defined in Appendix E.

Model	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD
Pretrained θ_0	0.632	0.597	0.582	0.450	0.315	0.325	0.483	0.439
Single-task experts	0.749	0.783	0.949	0.991	0.972	0.989	0.996	0.794
Task arithmetic	$0.619_{\pm 0.042}$	$0.591_{\pm 0.033}$	$0.667_{\pm 0.034}$	$0.766_{\pm 0.029}$	$0.750_{\pm 0.025}$	$0.647_{\pm 0.033}$	$0.940_{\pm 0.018}$	$0.503_{\pm 0.026}$
+XGD	$0.638_{\pm 0.021}$	$0.615_{\pm 0.006}$	$0.719_{\pm 0.023}$	$0.886_{\pm 0.006}$	$0.808_{\pm 0.021}$	$0.771_{\pm 0.054}$	$0.948_{\pm 0.018}$	$0.538_{\pm 0.015}$
Breadcrumbs	$0.641_{\pm 0.000}$	$0.613_{\pm 0.000}$	$0.693_{\pm 0.000}$	$0.791_{\pm 0.000}$	$0.769_{\pm 0.000}$	$0.668_{\pm 0.000}$	$0.949_{\pm 0.000}$	$0.521_{\pm 0.000}$
+XGD	$0.651_{\pm 0.002}$	$0.611_{\pm 0.027}$	$0.745_{\pm 0.026}$	$0.899_{\pm 0.007}$	$0.826_{\pm 0.009}$	$0.804_{\pm 0.041}$	$0.951_{\pm 0.010}$	$0.560_{\pm 0.013}$
DARE-TIES	$0.595_{\pm 0.000}$	$0.577_{\pm 0.000}$	$0.663_{\pm 0.000}$	$0.713_{\pm 0.000}$	$0.862_{\pm 0.000}$	$0.792_{\pm 0.000}$	$0.977_{\pm 0.000}$	$0.486_{\pm 0.000}$
+XGD	$0.635_{\pm 0.004}$	$0.612_{\pm 0.009}$	$0.746_{\pm 0.004}$	$0.877_{\pm 0.015}$	$0.882_{\pm 0.003}$	$0.881_{\pm 0.014}$	$0.977_{\pm 0.003}$	$0.521_{\pm 0.003}$
TIES	$0.655_{\pm 0.010}$	$0.620_{\pm 0.014}$	$0.714_{\pm 0.015}$	$0.769_{\pm 0.027}$	$0.849_{\pm 0.008}$	$0.783_{\pm 0.015}$	$0.966_{\pm 0.003}$	$0.512_{\pm 0.014}$
+XGD	$0.667_{\pm 0.006}$	$0.630_{\pm 0.011}$	$0.765_{\pm 0.016}$	$0.885_{\pm 0.023}$	$0.869_{\pm 0.007}$	$0.858_{\pm 0.018}$	$0.967_{\pm 0.004}$	$0.549_{\pm 0.011}$
aTLAS	$0.672_{\pm 0.004}$	$0.660_{\pm 0.006}$	$0.749_{\pm 0.041}$	$0.837_{\pm 0.028}$	$0.586_{\pm 0.071}$	$0.702_{\pm 0.081}$	$0.917_{\pm 0.015}$	$0.596_{\pm 0.004}$
+XGD	$0.666_{\pm 0.007}$	$0.636_{\pm 0.036}$	$0.751_{\pm 0.049}$	$0.880_{\pm 0.035}$	$0.719_{\pm 0.065}$	$0.771_{\pm 0.060}$	$0.928_{\pm 0.023}$	$0.591_{\pm 0.027}$
TATR	$0.634_{\pm 0.010}$	$0.613_{\pm 0.008}$	$0.705_{\pm 0.014}$	$0.823_{\pm 0.002}$	$0.784_{\pm 0.026}$	$0.708_{\pm 0.022}$	$0.962_{\pm 0.010}$	$0.525_{\pm 0.012}$
+XGD	$0.643_{\pm 0.008}$	$0.624_{\pm 0.011}$	$0.736_{\pm 0.010}$	$0.898_{\pm 0.014}$	$0.831_{\pm 0.007}$	$0.819_{\pm 0.027}$	$0.964_{\pm 0.003}$	$0.553_{\pm 0.006}$
Fisher (replay)	$0.685_{\pm 0.035}$	$0.655_{\pm 0.061}$	$0.788_{\pm 0.064}$	$0.652_{\pm 0.184}$	$0.778_{\pm 0.099}$	$0.409_{\pm 0.092}$	$0.771_{\pm 0.021}$	$0.526_{\pm 0.061}$
+XGD	$0.651_{\pm 0.033}$	$0.639_{\pm 0.065}$	$0.755_{\pm 0.092}$	$0.835_{\pm 0.067}$	$0.803_{\pm 0.026}$	$0.613_{\pm 0.105}$	$0.883_{\pm 0.040}$	$0.546_{\pm 0.042}$
Localize-and-Stitch	$0.619_{\pm 0.005}$	$0.584_{\pm 0.010}$	$0.642_{\pm 0.015}$	$0.767_{\pm 0.007}$	$0.732_{\pm 0.027}$	$0.583_{\pm 0.045}$	$0.882_{\pm 0.019}$	$0.480_{\pm 0.009}$
+XGD	$0.630_{\pm 0.006}$	$0.601_{\pm 0.014}$	$0.670_{\pm 0.018}$	$0.830_{\pm 0.010}$	$0.761_{\pm 0.024}$	$0.665_{\pm 0.013}$	$0.894_{\pm 0.009}$	$0.506_{\pm 0.010}$
GradFix	$0.602_{\pm 0.044}$	$0.515_{\pm 0.051}$	$0.587_{\pm 0.056}$	$0.632_{\pm 0.114}$	$0.563_{\pm 0.043}$	$0.441_{\pm 0.009}$	$0.769_{\pm 0.068}$	$0.457_{\pm 0.028}$
DOGE/APGD	$0.402_{\pm 0.008}$	$0.415_{\pm 0.040}$	$0.704_{\pm 0.002}$	$0.812_{\pm 0.006}$	$0.899_{\pm 0.003}$	$0.891_{\pm 0.010}$	$0.988_{\pm 0.000}$	$0.490_{\pm 0.008}$
θ_0+GD	$0.631_{\pm 0.001}$	$0.585_{\pm 0.011}$	$0.601_{\pm 0.017}$	$0.619_{\pm 0.124}$	$0.406_{\pm 0.079}$	$0.354_{\pm 0.027}$	$0.558_{\pm 0.055}$	$0.443_{\pm 0.004}$
θ_0+ungated	$0.603_{\pm 0.013}$	$0.533_{\pm 0.024}$	$0.598_{\pm 0.008}$	$0.703_{\pm 0.025}$	$0.471_{\pm 0.011}$	$0.402_{\pm 0.010}$	$0.653_{\pm 0.033}$	$0.424_{\pm 0.009}$
θ_0+XGD	$0.627_{\pm 0.010}$	$0.568_{\pm 0.022}$	$0.662_{\pm 0.014}$	$0.830_{\pm 0.035}$	$0.736_{\pm 0.151}$	$0.630_{\pm 0.115}$	$0.892_{\pm 0.071}$	$0.502_{\pm 0.020}$

Table 10: GLUE-8 per-task results at $B = 8$. Mean $\pm$ sample standard deviation over three buffer draws (GradFix and DOGE/APGD: one draw). Task metrics and reference rows are defined in Appendix E.

Model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
Pretrained θ_0	0.000	0.509	0.748	-0.047	0.316	0.336	0.495	0.473
Single-task experts	0.623	0.936	0.916	0.909	0.901	0.873	0.924	0.798
Task arithmetic	$0.120_{\pm 0.017}$	$0.814_{\pm 0.054}$	$0.727_{\pm 0.015}$	$0.367_{\pm 0.280}$	$0.503_{\pm 0.229}$	$0.657_{\pm 0.071}$	$0.597_{\pm 0.087}$	$0.463_{\pm 0.023}$
+XGD	$0.070_{\pm 0.047}$	$0.867_{\pm 0.030}$	$0.765_{\pm 0.021}$	$0.582_{\pm 0.393}$	$0.767_{\pm 0.063}$	$0.734_{\pm 0.044}$	$0.731_{\pm 0.117}$	$0.543_{\pm 0.053}$
Breadcrumbs	$0.129_{\pm 0.040}$	$0.882_{\pm 0.011}$	$0.412_{\pm 0.085}$	$0.340_{\pm 0.072}$	$0.329_{\pm 0.008}$	$0.627_{\pm 0.047}$	$0.496_{\pm 0.000}$	$0.501_{\pm 0.009}$
+XGD	$0.095_{\pm 0.054}$	$0.863_{\pm 0.028}$	$0.752_{\pm 0.015}$	$0.807_{\pm 0.023}$	$0.750_{\pm 0.070}$	$0.684_{\pm 0.081}$	$0.774_{\pm 0.020}$	$0.534_{\pm 0.054}$
DARE-TIES	$0.256_{\pm 0.021}$	$0.885_{\pm 0.003}$	$0.706_{\pm 0.028}$	$0.638_{\pm 0.074}$	$0.359_{\pm 0.012}$	$0.557_{\pm 0.024}$	$0.583_{\pm 0.013}$	$0.445_{\pm 0.010}$
+XGD	$0.181_{\pm 0.079}$	$0.878_{\pm 0.032}$	$0.752_{\pm 0.012}$	$0.818_{\pm 0.038}$	$0.761_{\pm 0.029}$	$0.659_{\pm 0.064}$	$0.780_{\pm 0.046}$	$0.489_{\pm 0.023}$
TIES	$0.234_{\pm 0.000}$	$0.788_{\pm 0.000}$	$0.757_{\pm 0.000}$	$0.659_{\pm 0.000}$	$0.546_{\pm 0.000}$	$0.413_{\pm 0.000}$	$0.519_{\pm 0.000}$	$0.473_{\pm 0.000}$
+XGD	$0.346_{\pm 0.010}$	$0.869_{\pm 0.035}$	$0.779_{\pm 0.019}$	$0.820_{\pm 0.010}$	$0.703_{\pm 0.012}$	$0.562_{\pm 0.031}$	$0.681_{\pm 0.058}$	$0.505_{\pm 0.027}$
aTLAS	$0.291_{\pm 0.053}$	$0.821_{\pm 0.081}$	$0.803_{\pm 0.010}$	$0.798_{\pm 0.038}$	$0.621_{\pm 0.014}$	$0.492_{\pm 0.091}$	$0.466_{\pm 0.074}$	$0.446_{\pm 0.015}$
+XGD	$0.264_{\pm 0.090}$	$0.849_{\pm 0.029}$	$0.797_{\pm 0.023}$	$0.808_{\pm 0.025}$	$0.661_{\pm 0.026}$	$0.541_{\pm 0.088}$	$0.537_{\pm 0.116}$	$0.487_{\pm 0.014}$
TATR	$0.212_{\pm 0.028}$	$0.860_{\pm 0.005}$	$0.595_{\pm 0.121}$	$0.596_{\pm 0.046}$	$0.318_{\pm 0.002}$	$0.677_{\pm 0.026}$	$0.567_{\pm 0.020}$	$0.542_{\pm 0.032}$
+XGD	$0.126_{\pm 0.028}$	$0.869_{\pm 0.031}$	$0.757_{\pm 0.039}$	$0.823_{\pm 0.010}$	$0.767_{\pm 0.060}$	$0.715_{\pm 0.042}$	$0.795_{\pm 0.038}$	$0.546_{\pm 0.048}$
Fisher (replay)	$0.004_{\pm 0.007}$	$0.681_{\pm 0.212}$	$0.424_{\pm 0.207}$	$0.774_{\pm 0.089}$	$0.312_{\pm 0.045}$	$0.507_{\pm 0.109}$	$0.574_{\pm 0.072}$	$0.664_{\pm 0.115}$
+XGD	$-0.012_{\pm 0.060}$	$0.769_{\pm 0.185}$	$0.548_{\pm 0.272}$	$0.816_{\pm 0.061}$	$0.665_{\pm 0.087}$	$0.658_{\pm 0.085}$	$0.688_{\pm 0.076}$	$0.680_{\pm 0.042}$
Localize-and-Stitch	$0.049_{\pm 0.041}$	$0.693_{\pm 0.182}$	$0.561_{\pm 0.349}$	$0.392_{\pm 0.398}$	$0.316_{\pm 0.003}$	$0.543_{\pm 0.172}$	$0.610_{\pm 0.122}$	$0.455_{\pm 0.045}$
+XGD	$0.066_{\pm 0.009}$	$0.740_{\pm 0.183}$	$0.695_{\pm 0.108}$	$0.643_{\pm 0.297}$	$0.597_{\pm 0.154}$	$0.566_{\pm 0.183}$	$0.669_{\pm 0.118}$	$0.557_{\pm 0.058}$
GradFix	0.000	0.509	0.748	0.067	0.494	0.327	0.519	0.473
DOGE/APGD	0.000	0.509	0.748	0.005	0.316	0.354	0.505	0.473
θ_0+GD	$0.030_{\pm 0.053}$	$0.606_{\pm 0.168}$	$0.745_{\pm 0.006}$	$0.144_{\pm 0.330}$	$0.396_{\pm 0.138}$	$0.338_{\pm 0.002}$	$0.492_{\pm 0.005}$	$0.473_{\pm 0.000}$
θ_0+ungated	$0.036_{\pm 0.039}$	$0.670_{\pm 0.028}$	$0.700_{\pm 0.035}$	$0.216_{\pm 0.101}$	$0.508_{\pm 0.068}$	$0.344_{\pm 0.007}$	$0.502_{\pm 0.042}$	$0.486_{\pm 0.020}$
θ_0+XGD	$0.111_{\pm 0.064}$	$0.763_{\pm 0.094}$	$0.720_{\pm 0.074}$	$0.638_{\pm 0.184}$	$0.571_{\pm 0.066}$	$0.401_{\pm 0.032}$	$0.567_{\pm 0.051}$	$0.465_{\pm 0.046}$

Table 11: GLUE-8 per-task results at $B = 16$. Mean $\pm$ sample standard deviation over three buffer draws. Task metrics and reference rows are defined in Appendix E.

Model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
Pretrained θ_0	0.000	0.509	0.748	−0.047	0.316	0.336	0.495	0.473
Single-task experts	0.623	0.936	0.916	0.909	0.901	0.873	0.924	0.798
Task arithmetic	$0.130_{\pm 0.000}$	$0.845_{\pm 0.000}$	$0.719_{\pm 0.000}$	$0.529_{\pm 0.000}$	$0.371_{\pm 0.000}$	$0.698_{\pm 0.000}$	$0.648_{\pm 0.000}$	$0.477_{\pm 0.000}$
+XGD	$0.052_{\pm 0.042}$	$0.883_{\pm 0.016}$	$0.729_{\pm 0.039}$	$0.817_{\pm 0.016}$	$0.811_{\pm 0.029}$	$0.740_{\pm 0.022}$	$0.813_{\pm 0.016}$	$0.560_{\pm 0.026}$
Breadcrumbs	$0.132_{\pm 0.017}$	$0.871_{\pm 0.041}$	$0.467_{\pm 0.136}$	$0.360_{\pm 0.109}$	$0.330_{\pm 0.010}$	$0.614_{\pm 0.089}$	$0.496_{\pm 0.000}$	$0.493_{\pm 0.027}$
+XGD	$0.114_{\pm 0.037}$	$0.874_{\pm 0.013}$	$0.713_{\pm 0.050}$	$0.832_{\pm 0.012}$	$0.771_{\pm 0.027}$	$0.731_{\pm 0.023}$	$0.765_{\pm 0.095}$	$0.538_{\pm 0.044}$
DARE-TIES	$0.256_{\pm 0.021}$	$0.885_{\pm 0.003}$	$0.706_{\pm 0.028}$	$0.638_{\pm 0.074}$	$0.359_{\pm 0.012}$	$0.557_{\pm 0.024}$	$0.583_{\pm 0.013}$	$0.445_{\pm 0.010}$
+XGD	$0.181_{\pm 0.013}$	$0.880_{\pm 0.007}$	$0.756_{\pm 0.030}$	$0.834_{\pm 0.007}$	$0.788_{\pm 0.009}$	$0.715_{\pm 0.001}$	$0.822_{\pm 0.019}$	$0.492_{\pm 0.028}$
TIES	$0.234_{\pm 0.000}$	$0.788_{\pm 0.000}$	$0.757_{\pm 0.000}$	$0.659_{\pm 0.000}$	$0.546_{\pm 0.000}$	$0.413_{\pm 0.000}$	$0.519_{\pm 0.000}$	$0.473_{\pm 0.000}$
+XGD	$0.290_{\pm 0.042}$	$0.878_{\pm 0.010}$	$0.750_{\pm 0.058}$	$0.838_{\pm 0.005}$	$0.725_{\pm 0.029}$	$0.633_{\pm 0.031}$	$0.754_{\pm 0.051}$	$0.520_{\pm 0.000}$
aTLAS	$0.301_{\pm 0.046}$	$0.833_{\pm 0.031}$	$0.807_{\pm 0.001}$	$0.839_{\pm 0.011}$	$0.631_{\pm 0.017}$	$0.516_{\pm 0.096}$	$0.490_{\pm 0.062}$	$0.454_{\pm 0.042}$
+XGD	$0.200_{\pm 0.048}$	$0.845_{\pm 0.038}$	$0.759_{\pm 0.070}$	$0.828_{\pm 0.006}$	$0.700_{\pm 0.018}$	$0.599_{\pm 0.067}$	$0.632_{\pm 0.133}$	$0.481_{\pm 0.021}$
TATR	$0.217_{\pm 0.025}$	$0.863_{\pm 0.004}$	$0.595_{\pm 0.095}$	$0.671_{\pm 0.013}$	$0.316_{\pm 0.000}$	$0.679_{\pm 0.023}$	$0.638_{\pm 0.018}$	$0.568_{\pm 0.011}$
+XGD	$0.103_{\pm 0.026}$	$0.882_{\pm 0.003}$	$0.711_{\pm 0.044}$	$0.836_{\pm 0.017}$	$0.814_{\pm 0.018}$	$0.729_{\pm 0.023}$	$0.805_{\pm 0.033}$	$0.597_{\pm 0.025}$
Fisher (replay)	$0.021_{\pm 0.032}$	$0.665_{\pm 0.182}$	$0.541_{\pm 0.273}$	$0.711_{\pm 0.088}$	$0.326_{\pm 0.040}$	$0.524_{\pm 0.236}$	$0.569_{\pm 0.087}$	$0.716_{\pm 0.049}$
+XGD	$0.042_{\pm 0.089}$	$0.808_{\pm 0.111}$	$0.610_{\pm 0.170}$	$0.794_{\pm 0.058}$	$0.707_{\pm 0.058}$	$0.673_{\pm 0.108}$	$0.667_{\pm 0.032}$	$0.680_{\pm 0.061}$
Localize-and-Stitch	$0.049_{\pm 0.021}$	$0.706_{\pm 0.160}$	$0.761_{\pm 0.006}$	$0.316_{\pm 0.539}$	$0.564_{\pm 0.209}$	$0.593_{\pm 0.125}$	$0.580_{\pm 0.179}$	$0.474_{\pm 0.037}$
+XGD	$0.050_{\pm 0.017}$	$0.832_{\pm 0.069}$	$0.699_{\pm 0.148}$	$0.579_{\pm 0.424}$	$0.675_{\pm 0.174}$	$0.620_{\pm 0.184}$	$0.674_{\pm 0.159}$	$0.538_{\pm 0.030}$
GradFix	$0.000_{\pm 0.000}$	$0.532_{\pm 0.039}$	$0.632_{\pm 0.180}$	$0.119_{\pm 0.108}$	$0.408_{\pm 0.078}$	$0.340_{\pm 0.022}$	$0.503_{\pm 0.008}$	$0.474_{\pm 0.002}$
DOGE/APGD	$0.000_{\pm 0.000}$	$0.509_{\pm 0.000}$	$0.748_{\pm 0.000}$	$0.002_{\pm 0.028}$	$0.316_{\pm 0.000}$	$0.354_{\pm 0.000}$	$0.505_{\pm 0.000}$	$0.473_{\pm 0.000}$
θ_0+GD	$0.026_{\pm 0.046}$	$0.570_{\pm 0.102}$	$0.705_{\pm 0.083}$	$0.117_{\pm 0.174}$	$0.385_{\pm 0.120}$	$0.354_{\pm 0.022}$	$0.489_{\pm 0.006}$	$0.511_{\pm 0.034}$
θ_0+ungated	$0.066_{\pm 0.055}$	$0.735_{\pm 0.039}$	$0.608_{\pm 0.059}$	$0.291_{\pm 0.085}$	$0.564_{\pm 0.024}$	$0.365_{\pm 0.013}$	$0.514_{\pm 0.028}$	$0.501_{\pm 0.021}$
θ_0+XGD	$0.062_{\pm 0.051}$	$0.793_{\pm 0.044}$	$0.690_{\pm 0.079}$	$0.799_{\pm 0.023}$	$0.599_{\pm 0.087}$	$0.492_{\pm 0.050}$	$0.605_{\pm 0.090}$	$0.477_{\pm 0.030}$

Table 12: CLIP-20 per-task results at $B = 8$. Tasks 1–10 of 20. Mean $\pm$ sample standard deviation over three buffer draws (GradFix and DOGE/APGD: one draw). Task metrics and reference rows are defined in Appendix E.

Model	SUN397	Cars	RESISC45	EuroSAT	SVHN
Pretrained θ_0	0.632	0.597	0.582	0.450	0.315
Single-task experts	0.749	0.783	0.949	0.991	0.972
Task arithmetic	$0.642_{\pm 0.000}$	$0.597_{\pm 0.000}$	$0.628_{\pm 0.000}$	$0.609_{\pm 0.000}$	$0.464_{\pm 0.000}$
+XGD	$0.633_{\pm 0.006}$	$0.592_{\pm 0.011}$	$0.640_{\pm 0.019}$	$0.688_{\pm 0.041}$	$0.606_{\pm 0.016}$
Breadcrumbs	$0.613_{\pm 0.017}$	$0.520_{\pm 0.036}$	$0.586_{\pm 0.032}$	$0.585_{\pm 0.033}$	$0.492_{\pm 0.015}$
+XGD	$0.627_{\pm 0.002}$	$0.578_{\pm 0.006}$	$0.646_{\pm 0.016}$	$0.706_{\pm 0.028}$	$0.658_{\pm 0.019}$
DARE-TIES	$0.465_{\pm 0.000}$	$0.318_{\pm 0.000}$	$0.426_{\pm 0.000}$	$0.367_{\pm 0.000}$	$0.508_{\pm 0.000}$
+XGD	$0.585_{\pm 0.014}$	$0.483_{\pm 0.006}$	$0.593_{\pm 0.039}$	$0.663_{\pm 0.067}$	$0.721_{\pm 0.024}$
TIES	$0.627_{\pm 0.000}$	$0.511_{\pm 0.000}$	$0.596_{\pm 0.000}$	$0.540_{\pm 0.000}$	$0.573_{\pm 0.000}$
+XGD	$0.642_{\pm 0.001}$	$0.570_{\pm 0.004}$	$0.652_{\pm 0.011}$	$0.683_{\pm 0.042}$	$0.700_{\pm 0.008}$
aTLAS	$0.628_{\pm 0.015}$	$0.601_{\pm 0.010}$	$0.689_{\pm 0.020}$	$0.746_{\pm 0.075}$	$0.563_{\pm 0.075}$
+XGD	$0.630_{\pm 0.012}$	$0.609_{\pm 0.012}$	$0.690_{\pm 0.019}$	$0.746_{\pm 0.068}$	$0.597_{\pm 0.062}$
TATR	$0.619_{\pm 0.016}$	$0.518_{\pm 0.035}$	$0.614_{\pm 0.020}$	$0.573_{\pm 0.009}$	$0.485_{\pm 0.004}$
+XGD	$0.624_{\pm 0.007}$	$0.578_{\pm 0.020}$	$0.637_{\pm 0.011}$	$0.694_{\pm 0.045}$	$0.688_{\pm 0.025}$
Fisher (replay)	$0.648_{\pm 0.041}$	$0.566_{\pm 0.029}$	$0.657_{\pm 0.106}$	$0.505_{\pm 0.059}$	$0.626_{\pm 0.133}$
+XGD	$0.644_{\pm 0.035}$	$0.574_{\pm 0.015}$	$0.661_{\pm 0.036}$	$0.654_{\pm 0.043}$	$0.677_{\pm 0.098}$
Localize-and-Stitch	$0.634_{\pm 0.002}$	$0.581_{\pm 0.024}$	$0.614_{\pm 0.016}$	$0.605_{\pm 0.048}$	$0.470_{\pm 0.040}$
+XGD	$0.629_{\pm 0.003}$	$0.579_{\pm 0.009}$	$0.623_{\pm 0.017}$	$0.677_{\pm 0.036}$	$0.593_{\pm 0.036}$
GradFix	0.628	0.564	0.576	0.464	0.388
DOGE/APGD	0.279	0.160	0.387	0.605	0.771
θ_0+GD	$0.632_{\pm 0.001}$	$0.591_{\pm 0.006}$	$0.588_{\pm 0.007}$	$0.566_{\pm 0.086}$	$0.355_{\pm 0.042}$
θ_0+ungated	$0.590_{\pm 0.018}$	$0.519_{\pm 0.013}$	$0.550_{\pm 0.033}$	$0.584_{\pm 0.077}$	$0.442_{\pm 0.032}$
θ_0+XGD	$0.620_{\pm 0.006}$	$0.569_{\pm 0.016}$	$0.610_{\pm 0.017}$	$0.680_{\pm 0.014}$	$0.664_{\pm 0.104}$

Model	GTSRB	MNIST	DTD	CIFAR100	STL10
Pretrained θ_0	0.325	0.483	0.439	0.617	0.971
Single-task experts	0.989	0.996	0.794	0.882	0.976
Task arithmetic	$0.431_{\pm 0.000}$	$0.718_{\pm 0.000}$	$0.471_{\pm 0.000}$	$0.694_{\pm 0.000}$	$0.970_{\pm 0.000}$
+XGD	$0.515_{\pm 0.013}$	$0.770_{\pm 0.042}$	$0.491_{\pm 0.008}$	$0.664_{\pm 0.009}$	$0.969_{\pm 0.001}$
Breadcrumbs	$0.499_{\pm 0.008}$	$0.826_{\pm 0.032}$	$0.472_{\pm 0.009}$	$0.631_{\pm 0.026}$	$0.946_{\pm 0.009}$
+XGD	$0.583_{\pm 0.005}$	$0.830_{\pm 0.045}$	$0.490_{\pm 0.001}$	$0.660_{\pm 0.002}$	$0.964_{\pm 0.002}$
DARE-TIES	$0.559_{\pm 0.000}$	$0.899_{\pm 0.000}$	$0.348_{\pm 0.000}$	$0.489_{\pm 0.000}$	$0.874_{\pm 0.000}$
+XGD	$0.696_{\pm 0.033}$	$0.918_{\pm 0.009}$	$0.439_{\pm 0.005}$	$0.608_{\pm 0.010}$	$0.933_{\pm 0.009}$
TIES	$0.596_{\pm 0.000}$	$0.884_{\pm 0.000}$	$0.471_{\pm 0.000}$	$0.652_{\pm 0.000}$	$0.947_{\pm 0.000}$
+XGD	$0.666_{\pm 0.029}$	$0.884_{\pm 0.016}$	$0.491_{\pm 0.003}$	$0.678_{\pm 0.001}$	$0.959_{\pm 0.000}$
aTLAS	$0.620_{\pm 0.073}$	$0.746_{\pm 0.021}$	$0.561_{\pm 0.018}$	$0.603_{\pm 0.040}$	$0.966_{\pm 0.001}$
+XGD	$0.647_{\pm 0.060}$	$0.791_{\pm 0.038}$	$0.566_{\pm 0.016}$	$0.608_{\pm 0.041}$	$0.966_{\pm 0.001}$
TATR	$0.500_{\pm 0.007}$	$0.806_{\pm 0.027}$	$0.468_{\pm 0.022}$	$0.618_{\pm 0.022}$	$0.945_{\pm 0.010}$
+XGD	$0.619_{\pm 0.054}$	$0.818_{\pm 0.052}$	$0.490_{\pm 0.002}$	$0.635_{\pm 0.017}$	$0.959_{\pm 0.006}$
Fisher (replay)	$0.362_{\pm 0.026}$	$0.677_{\pm 0.058}$	$0.469_{\pm 0.016}$	$0.719_{\pm 0.030}$	$0.963_{\pm 0.005}$
+XGD	$0.472_{\pm 0.034}$	$0.754_{\pm 0.050}$	$0.490_{\pm 0.008}$	$0.691_{\pm 0.042}$	$0.966_{\pm 0.001}$
Localize-and-Stitch	$0.390_{\pm 0.025}$	$0.675_{\pm 0.060}$	$0.443_{\pm 0.016}$	$0.677_{\pm 0.003}$	$0.969_{\pm 0.002}$
+XGD	$0.485_{\pm 0.035}$	$0.735_{\pm 0.048}$	$0.463_{\pm 0.012}$	$0.653_{\pm 0.012}$	$0.967_{\pm 0.002}$
GradFix	0.465	0.652	0.457	0.644	0.958
DOGE/APGD	0.731	0.967	0.330	0.409	0.861
θ_0+GD	$0.324_{\pm 0.002}$	$0.502_{\pm 0.013}$	$0.437_{\pm 0.001}$	$0.649_{\pm 0.024}$	$0.971_{\pm 0.002}$
θ_0+ungated	$0.361_{\pm 0.039}$	$0.589_{\pm 0.017}$	$0.415_{\pm 0.027}$	$0.576_{\pm 0.060}$	$0.962_{\pm 0.005}$
θ_0+XGD	$0.502_{\pm 0.050}$	$0.780_{\pm 0.072}$	$0.460_{\pm 0.011}$	$0.621_{\pm 0.032}$	$0.962_{\pm 0.005}$

Table 13: CLIP-20 per-task results at $B = 8$. Tasks 11–20 of 20. Mean ± sample standard deviation over three buffer draws (GradFix and DOGE/APGD: one draw). Task metrics and reference rows are defined in Appendix E.

Model	Flowers102	Pets	PCAM	FER2013	EMNIST
Pretrained θ_0	0.664	0.833	0.556	0.376	0.109
Single-task experts	0.886	0.925	0.887	0.711	0.942
Task arithmetic	$0.665_{\pm 0.000}$	$0.841_{\pm 0.000}$	$0.571_{\pm 0.000}$	$0.429_{\pm 0.000}$	$0.165_{\pm 0.000}$
+XGD	$0.666_{\pm 0.006}$	$0.834_{\pm 0.004}$	$0.694_{\pm 0.037}$	$0.485_{\pm 0.028}$	$0.190_{\pm 0.019}$
Breadcrumbs	$0.593_{\pm 0.038}$	$0.816_{\pm 0.012}$	$0.637_{\pm 0.026}$	$0.465_{\pm 0.002}$	$0.191_{\pm 0.002}$
+XGD	$0.633_{\pm 0.009}$	$0.832_{\pm 0.003}$	$0.734_{\pm 0.030}$	$0.517_{\pm 0.018}$	$0.219_{\pm 0.015}$
DARE-TIES	$0.357_{\pm 0.000}$	$0.652_{\pm 0.000}$	$0.647_{\pm 0.000}$	$0.464_{\pm 0.000}$	$0.202_{\pm 0.000}$
+XGD	$0.483_{\pm 0.034}$	$0.743_{\pm 0.019}$	$0.751_{\pm 0.020}$	$0.530_{\pm 0.014}$	$0.264_{\pm 0.007}$
TIES	$0.536_{\pm 0.000}$	$0.794_{\pm 0.000}$	$0.743_{\pm 0.000}$	$0.475_{\pm 0.000}$	$0.213_{\pm 0.000}$
+XGD	$0.581_{\pm 0.010}$	$0.811_{\pm 0.007}$	$0.759_{\pm 0.028}$	$0.530_{\pm 0.019}$	$0.240_{\pm 0.006}$
aTLAS	$0.709_{\pm 0.012}$	$0.840_{\pm 0.013}$	$0.572_{\pm 0.056}$	$0.525_{\pm 0.034}$	$0.257_{\pm 0.031}$
+XGD	$0.705_{\pm 0.005}$	$0.844_{\pm 0.012}$	$0.634_{\pm 0.071}$	$0.535_{\pm 0.031}$	$0.246_{\pm 0.029}$
TATR	$0.604_{\pm 0.036}$	$0.818_{\pm 0.009}$	$0.655_{\pm 0.046}$	$0.534_{\pm 0.009}$	$0.195_{\pm 0.004}$
+XGD	$0.635_{\pm 0.018}$	$0.829_{\pm 0.010}$	$0.733_{\pm 0.026}$	$0.525_{\pm 0.023}$	$0.229_{\pm 0.021}$
Fisher (replay)	$0.633_{\pm 0.027}$	$0.824_{\pm 0.013}$	$0.561_{\pm 0.011}$	$0.496_{\pm 0.085}$	$0.240_{\pm 0.167}$
+XGD	$0.644_{\pm 0.008}$	$0.827_{\pm 0.005}$	$0.691_{\pm 0.041}$	$0.543_{\pm 0.046}$	$0.241_{\pm 0.098}$
Localize-and-Stitch	$0.658_{\pm 0.005}$	$0.831_{\pm 0.005}$	$0.591_{\pm 0.002}$	$0.412_{\pm 0.010}$	$0.148_{\pm 0.016}$
+XGD	$0.659_{\pm 0.003}$	$0.826_{\pm 0.004}$	$0.649_{\pm 0.025}$	$0.457_{\pm 0.042}$	$0.174_{\pm 0.016}$
GradFix	0.636	0.838	0.604	0.430	0.157
DOGE/APGD	0.260	0.521	0.557	0.407	0.436
θ_0+GD	$0.661_{\pm 0.003}$	$0.833_{\pm 0.002}$	$0.562_{\pm 0.011}$	$0.386_{\pm 0.007}$	$0.114_{\pm 0.005}$
θ_0+ungated	$0.614_{\pm 0.021}$	$0.792_{\pm 0.016}$	$0.658_{\pm 0.049}$	$0.431_{\pm 0.016}$	$0.127_{\pm 0.007}$
θ_0+XGD	$0.656_{\pm 0.015}$	$0.825_{\pm 0.004}$	$0.698_{\pm 0.023}$	$0.465_{\pm 0.027}$	$0.189_{\pm 0.011}$

Model	CIFAR10	Food101	FashionMNIST	KMNIST	RenderedSST2
Pretrained θ_0	0.883	0.827	0.661	0.100	0.586
Single-task experts	0.975	0.862	0.948	0.982	0.713
Task arithmetic	$0.918_{\pm 0.000}$	$0.810_{\pm 0.000}$	$0.725_{\pm 0.000}$	$0.115_{\pm 0.000}$	$0.618_{\pm 0.000}$
+XGD	$0.897_{\pm 0.008}$	$0.805_{\pm 0.009}$	$0.718_{\pm 0.015}$	$0.229_{\pm 0.036}$	$0.633_{\pm 0.007}$
Breadcrumbs	$0.903_{\pm 0.009}$	$0.699_{\pm 0.048}$	$0.753_{\pm 0.001}$	$0.133_{\pm 0.024}$	$0.590_{\pm 0.010}$
+XGD	$0.903_{\pm 0.004}$	$0.761_{\pm 0.006}$	$0.752_{\pm 0.015}$	$0.255_{\pm 0.025}$	$0.634_{\pm 0.009}$
DARE-TIES	$0.835_{\pm 0.000}$	$0.404_{\pm 0.000}$	$0.772_{\pm 0.000}$	$0.229_{\pm 0.000}$	$0.586_{\pm 0.000}$
+XGD	$0.886_{\pm 0.007}$	$0.634_{\pm 0.032}$	$0.781_{\pm 0.005}$	$0.344_{\pm 0.030}$	$0.654_{\pm 0.004}$
TIES	$0.915_{\pm 0.000}$	$0.710_{\pm 0.000}$	$0.770_{\pm 0.000}$	$0.194_{\pm 0.000}$	$0.637_{\pm 0.000}$
+XGD	$0.918_{\pm 0.003}$	$0.755_{\pm 0.012}$	$0.771_{\pm 0.009}$	$0.295_{\pm 0.023}$	$0.666_{\pm 0.006}$
aTLAS	$0.879_{\pm 0.026}$	$0.797_{\pm 0.006}$	$0.687_{\pm 0.062}$	$0.215_{\pm 0.084}$	$0.608_{\pm 0.048}$
+XGD	$0.877_{\pm 0.029}$	$0.803_{\pm 0.006}$	$0.698_{\pm 0.053}$	$0.273_{\pm 0.077}$	$0.629_{\pm 0.044}$
TATR	$0.893_{\pm 0.008}$	$0.732_{\pm 0.042}$	$0.769_{\pm 0.002}$	$0.136_{\pm 0.006}$	$0.592_{\pm 0.006}$
+XGD	$0.881_{\pm 0.007}$	$0.763_{\pm 0.029}$	$0.753_{\pm 0.015}$	$0.283_{\pm 0.014}$	$0.649_{\pm 0.002}$
Fisher (replay)	$0.916_{\pm 0.026}$	$0.817_{\pm 0.067}$	$0.672_{\pm 0.029}$	$0.111_{\pm 0.008}$	$0.591_{\pm 0.003}$
+XGD	$0.906_{\pm 0.022}$	$0.821_{\pm 0.037}$	$0.712_{\pm 0.011}$	$0.212_{\pm 0.061}$	$0.622_{\pm 0.006}$
Localize-and-Stitch	$0.902_{\pm 0.014}$	$0.822_{\pm 0.007}$	$0.692_{\pm 0.009}$	$0.116_{\pm 0.007}$	$0.599_{\pm 0.010}$
+XGD	$0.894_{\pm 0.017}$	$0.814_{\pm 0.006}$	$0.703_{\pm 0.026}$	$0.209_{\pm 0.041}$	$0.608_{\pm 0.016}$
GradFix	0.885	0.766	0.739	0.091	0.575
DOGE/APGD	0.748	0.207	0.734	0.312	0.637
θ_0+GD	$0.890_{\pm 0.005}$	$0.828_{\pm 0.001}$	$0.667_{\pm 0.012}$	$0.110_{\pm 0.009}$	$0.589_{\pm 0.004}$
θ_0+ungated	$0.852_{\pm 0.023}$	$0.767_{\pm 0.046}$	$0.678_{\pm 0.029}$	$0.164_{\pm 0.018}$	$0.572_{\pm 0.026}$
θ_0+XGD	$0.868_{\pm 0.019}$	$0.794_{\pm 0.016}$	$0.707_{\pm 0.027}$	$0.273_{\pm 0.047}$	$0.629_{\pm 0.009}$

Table 14: CLIP-20 per-task results at $B = 16$. Tasks 1–10 of 20. Mean ± sample standard deviation over three buffer draws. Task metrics and reference rows are defined in Appendix E.

Model	SUN397	Cars	RESISC45	EuroSAT	SVHN
Pretrained θ_0	0.632	0.597	0.582	0.450	0.315
Single-task experts	0.749	0.783	0.949	0.991	0.972
Task arithmetic	$0.632_{\pm 0.011}$	$0.569_{\pm 0.029}$	$0.611_{\pm 0.019}$	$0.596_{\pm 0.014}$	$0.473_{\pm 0.008}$
+XGD	$0.636_{\pm 0.006}$	$0.589_{\pm 0.026}$	$0.657_{\pm 0.004}$	$0.752_{\pm 0.032}$	$0.622_{\pm 0.045}$
Breadcrumbs	$0.623_{\pm 0.000}$	$0.541_{\pm 0.000}$	$0.605_{\pm 0.000}$	$0.604_{\pm 0.000}$	$0.484_{\pm 0.000}$
+XGD	$0.632_{\pm 0.004}$	$0.580_{\pm 0.029}$	$0.655_{\pm 0.005}$	$0.765_{\pm 0.031}$	$0.638_{\pm 0.038}$
DARE-TIES	$0.465_{\pm 0.000}$	$0.318_{\pm 0.000}$	$0.426_{\pm 0.000}$	$0.367_{\pm 0.000}$	$0.508_{\pm 0.000}$
+XGD	$0.614_{\pm 0.020}$	$0.482_{\pm 0.020}$	$0.641_{\pm 0.029}$	$0.796_{\pm 0.026}$	$0.749_{\pm 0.021}$
TIES	$0.615_{\pm 0.022}$	$0.488_{\pm 0.040}$	$0.579_{\pm 0.030}$	$0.521_{\pm 0.033}$	$0.575_{\pm 0.004}$
+XGD	$0.646_{\pm 0.004}$	$0.575_{\pm 0.023}$	$0.674_{\pm 0.004}$	$0.769_{\pm 0.023}$	$0.708_{\pm 0.056}$
aTLAS	$0.644_{\pm 0.014}$	$0.634_{\pm 0.015}$	$0.719_{\pm 0.044}$	$0.796_{\pm 0.034}$	$0.513_{\pm 0.071}$
+XGD	$0.647_{\pm 0.011}$	$0.634_{\pm 0.011}$	$0.727_{\pm 0.041}$	$0.825_{\pm 0.039}$	$0.563_{\pm 0.052}$
TATR	$0.626_{\pm 0.020}$	$0.546_{\pm 0.057}$	$0.618_{\pm 0.025}$	$0.597_{\pm 0.010}$	$0.493_{\pm 0.030}$
+XGD	$0.631_{\pm 0.011}$	$0.582_{\pm 0.028}$	$0.658_{\pm 0.007}$	$0.763_{\pm 0.040}$	$0.628_{\pm 0.052}$
Fisher (replay)	$0.673_{\pm 0.033}$	$0.673_{\pm 0.079}$	$0.648_{\pm 0.068}$	$0.501_{\pm 0.067}$	$0.675_{\pm 0.212}$
+XGD	$0.669_{\pm 0.026}$	$0.663_{\pm 0.071}$	$0.683_{\pm 0.052}$	$0.717_{\pm 0.042}$	$0.712_{\pm 0.124}$
Localize-and-Stitch	$0.607_{\pm 0.022}$	$0.513_{\pm 0.073}$	$0.570_{\pm 0.040}$	$0.535_{\pm 0.069}$	$0.582_{\pm 0.062}$
+XGD	$0.626_{\pm 0.009}$	$0.567_{\pm 0.019}$	$0.629_{\pm 0.005}$	$0.720_{\pm 0.013}$	$0.615_{\pm 0.040}$
GradFix	$0.636_{\pm 0.001}$	$0.546_{\pm 0.015}$	$0.626_{\pm 0.002}$	$0.660_{\pm 0.030}$	$0.432_{\pm 0.059}$
DOGE/APGD	$0.274_{\pm 0.004}$	$0.173_{\pm 0.011}$	$0.392_{\pm 0.005}$	$0.615_{\pm 0.009}$	$0.755_{\pm 0.014}$
θ_0+GD	$0.630_{\pm 0.003}$	$0.586_{\pm 0.019}$	$0.597_{\pm 0.021}$	$0.554_{\pm 0.120}$	$0.352_{\pm 0.063}$
θ_0+ungated	$0.586_{\pm 0.010}$	$0.511_{\pm 0.023}$	$0.571_{\pm 0.014}$	$0.689_{\pm 0.020}$	$0.453_{\pm 0.008}$
θ_0+XGD	$0.618_{\pm 0.014}$	$0.517_{\pm 0.071}$	$0.627_{\pm 0.014}$	$0.756_{\pm 0.057}$	$0.532_{\pm 0.152}$

Model	GTSRB	MNIST	DTD	CIFAR100	STL10
Pretrained θ_0	0.325	0.483	0.439	0.617	0.971
Single-task experts	0.989	0.996	0.794	0.882	0.976
Task arithmetic	$0.465_{\pm 0.031}$	$0.761_{\pm 0.039}$	$0.471_{\pm 0.004}$	$0.670_{\pm 0.025}$	$0.959_{\pm 0.010}$
+XGD	$0.559_{\pm 0.042}$	$0.812_{\pm 0.022}$	$0.491_{\pm 0.004}$	$0.664_{\pm 0.026}$	$0.965_{\pm 0.003}$
Breadcrumbs	$0.495_{\pm 0.000}$	$0.807_{\pm 0.000}$	$0.477_{\pm 0.000}$	$0.646_{\pm 0.000}$	$0.952_{\pm 0.000}$
+XGD	$0.588_{\pm 0.030}$	$0.842_{\pm 0.017}$	$0.498_{\pm 0.006}$	$0.649_{\pm 0.023}$	$0.963_{\pm 0.003}$
DARE-TIES	$0.559_{\pm 0.000}$	$0.899_{\pm 0.000}$	$0.348_{\pm 0.000}$	$0.489_{\pm 0.000}$	$0.874_{\pm 0.000}$
+XGD	$0.699_{\pm 0.082}$	$0.907_{\pm 0.026}$	$0.465_{\pm 0.020}$	$0.631_{\pm 0.023}$	$0.943_{\pm 0.016}$
TIES	$0.601_{\pm 0.009}$	$0.892_{\pm 0.014}$	$0.460_{\pm 0.019}$	$0.636_{\pm 0.029}$	$0.939_{\pm 0.013}$
+XGD	$0.696_{\pm 0.054}$	$0.903_{\pm 0.004}$	$0.496_{\pm 0.010}$	$0.674_{\pm 0.002}$	$0.956_{\pm 0.004}$
aTLAS	$0.637_{\pm 0.079}$	$0.729_{\pm 0.030}$	$0.572_{\pm 0.012}$	$0.624_{\pm 0.010}$	$0.965_{\pm 0.003}$
+XGD	$0.674_{\pm 0.074}$	$0.811_{\pm 0.010}$	$0.572_{\pm 0.009}$	$0.629_{\pm 0.016}$	$0.964_{\pm 0.002}$
TATR	$0.488_{\pm 0.050}$	$0.796_{\pm 0.069}$	$0.466_{\pm 0.015}$	$0.649_{\pm 0.037}$	$0.955_{\pm 0.017}$
+XGD	$0.583_{\pm 0.073}$	$0.826_{\pm 0.043}$	$0.490_{\pm 0.005}$	$0.652_{\pm 0.034}$	$0.962_{\pm 0.009}$
Fisher (replay)	$0.384_{\pm 0.022}$	$0.705_{\pm 0.086}$	$0.448_{\pm 0.023}$	$0.729_{\pm 0.076}$	$0.961_{\pm 0.001}$
+XGD	$0.474_{\pm 0.031}$	$0.787_{\pm 0.044}$	$0.474_{\pm 0.011}$	$0.716_{\pm 0.042}$	$0.964_{\pm 0.003}$
Localize-and-Stitch	$0.441_{\pm 0.031}$	$0.782_{\pm 0.058}$	$0.438_{\pm 0.007}$	$0.588_{\pm 0.067}$	$0.947_{\pm 0.020}$
+XGD	$0.499_{\pm 0.017}$	$0.792_{\pm 0.038}$	$0.468_{\pm 0.005}$	$0.638_{\pm 0.019}$	$0.960_{\pm 0.010}$
GradFix	$0.428_{\pm 0.004}$	$0.688_{\pm 0.017}$	$0.474_{\pm 0.006}$	$0.640_{\pm 0.008}$	$0.957_{\pm 0.010}$
DOGE/APGD	$0.680_{\pm 0.045}$	$0.964_{\pm 0.002}$	$0.335_{\pm 0.004}$	$0.411_{\pm 0.002}$	$0.842_{\pm 0.017}$
θ_0+GD	$0.339_{\pm 0.024}$	$0.506_{\pm 0.024}$	$0.435_{\pm 0.008}$	$0.633_{\pm 0.014}$	$0.971_{\pm 0.000}$
θ_0+ungated	$0.413_{\pm 0.007}$	$0.620_{\pm 0.028}$	$0.420_{\pm 0.013}$	$0.578_{\pm 0.009}$	$0.958_{\pm 0.005}$
θ_0+XGD	$0.503_{\pm 0.066}$	$0.709_{\pm 0.205}$	$0.473_{\pm 0.015}$	$0.580_{\pm 0.070}$	$0.959_{\pm 0.013}$

Table 15: CLIP-20 per-task results at $B = 16$. Tasks 11–20 of 20. Mean ± sample standard deviation over three buffer draws. Task metrics and reference rows are defined in Appendix E.

Model	Flowers102	Pets	PCAM	FER2013	EMNIST
Pretrained θ_0	0.664	0.833	0.556	0.376	0.109
Single-task experts	0.886	0.925	0.887	0.711	0.942
Task arithmetic	$0.639_{\pm 0.027}$	$0.831_{\pm 0.009}$	$0.603_{\pm 0.033}$	$0.443_{\pm 0.013}$	$0.176_{\pm 0.011}$
+XGD	$0.659_{\pm 0.021}$	$0.834_{\pm 0.002}$	$0.747_{\pm 0.023}$	$0.513_{\pm 0.024}$	$0.233_{\pm 0.018}$
Breadcrumbs	$0.615_{\pm 0.000}$	$0.823_{\pm 0.000}$	$0.622_{\pm 0.000}$	$0.463_{\pm 0.000}$	$0.190_{\pm 0.000}$
+XGD	$0.641_{\pm 0.010}$	$0.833_{\pm 0.007}$	$0.768_{\pm 0.017}$	$0.523_{\pm 0.022}$	$0.244_{\pm 0.006}$
DARE-TIES	$0.357_{\pm 0.000}$	$0.652_{\pm 0.000}$	$0.647_{\pm 0.000}$	$0.464_{\pm 0.000}$	$0.202_{\pm 0.000}$
+XGD	$0.536_{\pm 0.052}$	$0.786_{\pm 0.040}$	$0.769_{\pm 0.031}$	$0.534_{\pm 0.006}$	$0.286_{\pm 0.022}$
TIES	$0.514_{\pm 0.037}$	$0.782_{\pm 0.020}$	$0.744_{\pm 0.002}$	$0.478_{\pm 0.005}$	$0.213_{\pm 0.002}$
+XGD	$0.594_{\pm 0.018}$	$0.811_{\pm 0.010}$	$0.783_{\pm 0.024}$	$0.533_{\pm 0.019}$	$0.275_{\pm 0.006}$
aTLAS	$0.716_{\pm 0.023}$	$0.851_{\pm 0.007}$	$0.556_{\pm 0.051}$	$0.539_{\pm 0.027}$	$0.396_{\pm 0.070}$
+XGD	$0.713_{\pm 0.015}$	$0.853_{\pm 0.004}$	$0.695_{\pm 0.035}$	$0.546_{\pm 0.016}$	$0.355_{\pm 0.039}$
TATR	$0.623_{\pm 0.051}$	$0.826_{\pm 0.017}$	$0.641_{\pm 0.072}$	$0.482_{\pm 0.035}$	$0.185_{\pm 0.022}$
+XGD	$0.646_{\pm 0.035}$	$0.831_{\pm 0.006}$	$0.761_{\pm 0.035}$	$0.511_{\pm 0.034}$	$0.242_{\pm 0.027}$
Fisher (replay)	$0.642_{\pm 0.015}$	$0.820_{\pm 0.002}$	$0.568_{\pm 0.008}$	$0.444_{\pm 0.070}$	$0.179_{\pm 0.027}$
+XGD	$0.653_{\pm 0.014}$	$0.827_{\pm 0.003}$	$0.730_{\pm 0.007}$	$0.517_{\pm 0.015}$	$0.230_{\pm 0.019}$
Localize-and-Stitch	$0.579_{\pm 0.067}$	$0.799_{\pm 0.026}$	$0.642_{\pm 0.054}$	$0.458_{\pm 0.040}$	$0.172_{\pm 0.018}$
+XGD	$0.645_{\pm 0.026}$	$0.822_{\pm 0.009}$	$0.711_{\pm 0.013}$	$0.484_{\pm 0.032}$	$0.205_{\pm 0.010}$
GradFix	$0.644_{\pm 0.009}$	$0.829_{\pm 0.008}$	$0.512_{\pm 0.015}$	$0.476_{\pm 0.054}$	$0.171_{\pm 0.005}$
DOGE/APGD	$0.257_{\pm 0.003}$	$0.485_{\pm 0.031}$	$0.609_{\pm 0.045}$	$0.423_{\pm 0.013}$	$0.378_{\pm 0.050}$
θ_0+GD	$0.654_{\pm 0.019}$	$0.830_{\pm 0.007}$	$0.553_{\pm 0.004}$	$0.413_{\pm 0.057}$	$0.125_{\pm 0.025}$
θ_0+ungated	$0.614_{\pm 0.015}$	$0.800_{\pm 0.018}$	$0.683_{\pm 0.074}$	$0.457_{\pm 0.029}$	$0.151_{\pm 0.007}$
θ_0+XGD	$0.645_{\pm 0.034}$	$0.823_{\pm 0.018}$	$0.649_{\pm 0.130}$	$0.478_{\pm 0.029}$	$0.221_{\pm 0.039}$

Model	CIFAR10	Food101	FashionMNIST	KMNIST	RenderedSST2
Pretrained θ_0	0.883	0.827	0.661	0.100	0.586
Single-task experts	0.975	0.862	0.948	0.982	0.713
Task arithmetic	$0.912_{\pm 0.007}$	$0.765_{\pm 0.045}$	$0.737_{\pm 0.012}$	$0.117_{\pm 0.004}$	$0.608_{\pm 0.010}$
+XGD	$0.900_{\pm 0.017}$	$0.799_{\pm 0.018}$	$0.753_{\pm 0.023}$	$0.231_{\pm 0.060}$	$0.638_{\pm 0.009}$
Breadcrumbs	$0.908_{\pm 0.000}$	$0.727_{\pm 0.000}$	$0.753_{\pm 0.000}$	$0.119_{\pm 0.000}$	$0.596_{\pm 0.000}$
+XGD	$0.898_{\pm 0.020}$	$0.780_{\pm 0.010}$	$0.763_{\pm 0.027}$	$0.253_{\pm 0.069}$	$0.645_{\pm 0.003}$
DARE-TIES	$0.835_{\pm 0.000}$	$0.404_{\pm 0.000}$	$0.772_{\pm 0.000}$	$0.229_{\pm 0.000}$	$0.586_{\pm 0.000}$
+XGD	$0.892_{\pm 0.012}$	$0.705_{\pm 0.069}$	$0.786_{\pm 0.019}$	$0.367_{\pm 0.029}$	$0.657_{\pm 0.007}$
TIES	$0.909_{\pm 0.011}$	$0.681_{\pm 0.051}$	$0.772_{\pm 0.002}$	$0.203_{\pm 0.017}$	$0.634_{\pm 0.005}$
+XGD	$0.918_{\pm 0.003}$	$0.757_{\pm 0.013}$	$0.787_{\pm 0.014}$	$0.345_{\pm 0.051}$	$0.669_{\pm 0.003}$
aTLAS	$0.869_{\pm 0.042}$	$0.796_{\pm 0.020}$	$0.737_{\pm 0.027}$	$0.164_{\pm 0.009}$	$0.596_{\pm 0.055}$
+XGD	$0.871_{\pm 0.037}$	$0.806_{\pm 0.017}$	$0.746_{\pm 0.020}$	$0.224_{\pm 0.011}$	$0.642_{\pm 0.029}$
TATR	$0.908_{\pm 0.009}$	$0.753_{\pm 0.072}$	$0.757_{\pm 0.020}$	$0.136_{\pm 0.013}$	$0.602_{\pm 0.018}$
+XGD	$0.899_{\pm 0.019}$	$0.782_{\pm 0.047}$	$0.762_{\pm 0.025}$	$0.237_{\pm 0.093}$	$0.639_{\pm 0.011}$
Fisher (replay)	$0.908_{\pm 0.028}$	$0.802_{\pm 0.018}$	$0.678_{\pm 0.016}$	$0.113_{\pm 0.005}$	$0.588_{\pm 0.016}$
+XGD	$0.912_{\pm 0.010}$	$0.815_{\pm 0.008}$	$0.724_{\pm 0.007}$	$0.159_{\pm 0.032}$	$0.617_{\pm 0.005}$
Localize-and-Stitch	$0.864_{\pm 0.040}$	$0.764_{\pm 0.051}$	$0.688_{\pm 0.035}$	$0.134_{\pm 0.011}$	$0.581_{\pm 0.025}$
+XGD	$0.895_{\pm 0.013}$	$0.805_{\pm 0.016}$	$0.725_{\pm 0.013}$	$0.195_{\pm 0.026}$	$0.623_{\pm 0.008}$
GradFix	$0.899_{\pm 0.013}$	$0.775_{\pm 0.007}$	$0.722_{\pm 0.022}$	$0.078_{\pm 0.015}$	$0.571_{\pm 0.025}$
DOGE/APGD	$0.721_{\pm 0.023}$	$0.199_{\pm 0.007}$	$0.726_{\pm 0.008}$	$0.258_{\pm 0.048}$	$0.643_{\pm 0.005}$
θ_0+GD	$0.887_{\pm 0.002}$	$0.821_{\pm 0.010}$	$0.678_{\pm 0.017}$	$0.107_{\pm 0.013}$	$0.595_{\pm 0.009}$
θ_0+ungated	$0.841_{\pm 0.014}$	$0.770_{\pm 0.002}$	$0.665_{\pm 0.015}$	$0.181_{\pm 0.008}$	$0.591_{\pm 0.036}$
θ_0+XGD	$0.840_{\pm 0.049}$	$0.789_{\pm 0.030}$	$0.697_{\pm 0.051}$	$0.216_{\pm 0.104}$	$0.629_{\pm 0.015}$

F AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and refactoring experiment code and with editing and polishing the text. AI tools also assisted with checking manuscript claims against saved experiment outputs and verifying bibliographic metadata. The authors are responsible for reviewing these revisions and for the final text, claims, and artifacts.