WHEN PRIVACY HURTS MERGEABILITY: GEOMETRY-AWARE MODEL MERGING UNDER DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging promises to construct a single multi-task model from independently fine-tuned task models without accessing the original task data. This makes it attractive when task data cannot be centralized, but released task models may still leak private fine-tuning data. Differential privacy (DP) provides a principled mechanism for limiting such leakage, yet its interaction with model merging remains poorly understood. In this paper, we study the geometry of differentially private model merging and identify two geometric factors associated with poor mergeability: 1) *local sharpness*, which increases the sensitivity of task losses to the parameter displacement induced by merging, and 2) *reference drift*, which captures how far private task models move from the shared pretrained initialization and can contribute to cross-task discrepancy. We propose DP-Merging, a geometry-aware framework that improves the mergeability of differentially private task models. DP-Merging uses a DP-compatible sharpness-aware objective to guide each private task model toward flatter loss regions, and a reference-based alignment regularizer to keep task models close to the shared pretrained initialization. Our analysis relates merge-induced loss increases to local loss sensitivity and parameter displacement. Experiments on vision and language tasks across multiple privacy budgets show that DP-Merging consistently improves private merged-model performance while preserving the privacy guarantees of the underlying DP fine-tuning procedures.

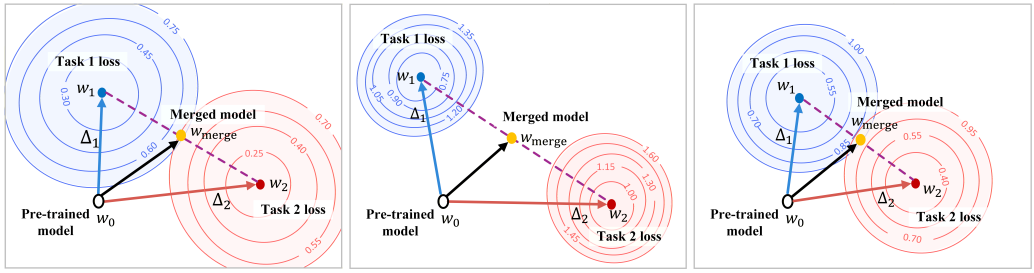
1 INTRODUCTION

Fine-tuning pretrained models has become the standard way to adapt foundation models to diverse downstream tasks (Liu et al., 2026; Wang et al., 2026). As task-specific models accumulate, storing and deploying them independently becomes increasingly costly. Multi-task learning can integrate these capabilities into a model, but requires joint access to data from all tasks and expensive retraining (Fifty et al., 2021; Agiza et al., 2024). Model merging has recently emerged as a practical alternative (Yang et al., 2026): it directly combines independently fine-tuned models in parameter space to obtain a unified model, without accessing the original task data. This makes model merging particularly attractive when task data are decentralized or cannot be shared.

Despite its potential, data-free merging is not inherently privacy-preserving. Although raw data are not shared during merging, the released task-specific weights and task vectors are derived from private data, posing a risk of sensitive information leakage (Zhang et al., 2024a; Wang et al., 2025; Yuan et al., 2025). Differential privacy (DP (Dwork et al., 2006)) provides a rigorous framework for limiting such leakage by bounding the influence of any single training example on the released models. A natural solution is therefore to fine-tune each task model under DP and then merge the resulting private models. However, as we show empirically, this straightforward solution can degrade beyond the utility loss of individual private models: even when private models retain reasonable task performance, their performance can drop substantially after merging. This raises a central question: *what makes differentially private models difficult to merge, and how can we improve their mergeability under privacy constraints?*

We argue that the difficulty is not merely that DP lowers the accuracy of the models being merged. Rather, parameter-space merging relies on geometric compatibility among task-specific solutions,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Non-private merging (b) Naive DP merging (c) DP-Merging (Ours)

Figure 1: Geometric illustration of model merging under DP. Task vectors are fine-tuned from w_0 and merged in parameter space. (a) Without DP, flat and compatible solutions merge effectively. (b) Naive DP can yield sharper solutions with weaker reference alignment, leading to high-loss merging. (c) DP-Merging improves flatness and reference alignment, yielding lower-loss merging under DP.

which can be disrupted by DP fine-tuning. Prior work suggests that merging often benefits when independently fine-tuned models remain geometrically compatible (Ainsworth et al., 2023). In particular, interpolation or averaging tends to incur smaller loss increase when the path between models stays in relatively flat, low-loss regions (Peña et al., 2023; Ito et al., 2025). In our experiments, privately fine-tuned solutions exhibit two geometric characteristics associated with poorer merging performance. Privately fine-tuned solutions can lie in sharper local regions, making the task loss more sensitive to merge-induced parameter displacement. Meanwhile, they may also deviate further from the shared pretrained initialization, increasing the mismatch among task-specific updates. We refer to these two factors as *local sharpness* and *reference drift*, respectively. These effects make private task models more sensitive to merge-induced displacement and less compatible in parameter space, leading to high-loss merged solutions, as illustrated in Figure 1.

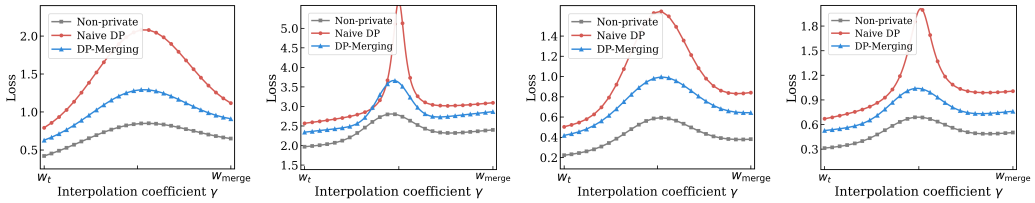
Motivated by this geometric view, we propose DP-Merging, a simple geometry-aware framework for differentially private model merging. Rather than designing a new post-hoc merging operator, DP-Merging applies two minimal interventions during private fine-tuning to restore the geometric conditions required by merging. First, a DP-compatible sharpness-aware objective encourages each private task model to lie in a flatter local region, reducing its sensitivity to merge-induced displacement. Second, a reference-anchored alignment regularizer keeps task vectors close to the shared pretrained initialization, limiting reference drift and reducing cross-task interference. Since the final merging step only processes DP-released task models, it remains a post-processing operation and incurs no additional privacy loss. We derive bounds on merge-induced loss increases that relate local sharpness and reference drift to merged-model utility, and empirically evaluate DP-Merging across vision and language tasks under multiple privacy budgets.

In summary, we make the following contributions:

- We identify a geometric failure mode of differentially private model merging, termed *DP-induced mergeability degradation*. We show that DP fine-tuning can make task-specific models harder to merge by increasing local sharpness and reference drift, which amplify parameter interference after merging.
- We propose DP-Merging, a geometry-aware private fine-tuning framework designed for subsequent model merging. It combines DP-compatible sharpness-aware updates with reference anchoring to jointly address local perturbation sensitivity and task-update magnitudes, without cross-task access during training. The resulting private models remain compatible with existing post-hoc merging rules.
- We derive merge-gap and merged-loss bounds connecting local sharpness, reference distances, and task-specific utility. Experiments on vision and language tasks across multiple privacy budgets demonstrate improvements over DP fine-tuning followed by merging.

2 Related Work

Non-private model merging. Existing work on model merging mainly studies the non-private setting. Representative approaches include direct weight-space averaging, such as Weight Averag-



(a) EuroSAT (b) SUN397 (c) SST-2 (d) QNLI

Figure 2: Interpolation loss landscapes from w_t to w_{merge}. We evaluate test loss along $w(\gamma) = (1-\gamma)w_t + \gamma w_{\text{merge}}$. DP-Merging consistently reduces the sharp loss barriers induced by standard DP, yielding smoother paths across vision and language tasks.

ing (Wortsman et al., 2022), task-vector composition such as Task Arithmetic (Ilharco et al., 2023), and more structured rules based on parameter importance or interference resolution, such as Fisher-weighted merging (Matena & Raffel, 2022), RegMean (Jin et al., 2023), TIES-Merging (Yadav et al., 2023), PCB Merging (DU et al., 2024), and WUDI-Merging (Cheng et al., 2025). These methods typically assume direct access to task-specific weights, parameter deltas, or importance statistics. In contrast, we study a privacy-constrained setting where task-specific models must be obtained under differential privacy before they can be released and merged.

Differentially private fine-tuning. DP fine-tuning of pretrained models has become increasingly practical, with most methods relying on per-example gradient clipping and Gaussian noise injection to improve the privacy–utility trade-off for a single released model (Yu et al., 2022; Park et al., 2023; Bu et al., 2024; Li et al., 2024). Other work also explores alternative private fine-tuning paradigms (Bu et al., 2024), such as forward-pass perturbation Du et al. (2023) or zeroth-order optimization (Zhang et al., 2024b). Closely related to our private fine-tuning procedure, Park et al. (2023) develop sharpness-aware training to improve individual model utility under DP. However, these methods are designed mainly to preserve the utility of individual private models. They do not address whether multiple privately fine-tuned task models remain geometrically compatible for post-hoc merging, which is the focus of our work.

Geometry of mergeability and sharpness. Prior work indicates that the success of model merging can be influenced not only by the merging method but also by geometric properties of the models, such as alignment in parameter space, low-loss connectivity, and landscape structure. Averaging or interpolation is more reliable when fine-tuned models lie in compatible low-loss regions or admit low-loss connectivity (Garipov et al., 2018; Draxler et al., 2018; Wortsman et al., 2022), while direct merging can fail when parameters are misaligned, for example due to permutation symmetries (Ainsworth et al., 2023; Ito et al., 2025; Zhang et al., 2025). Sharpness-aware optimization methods, such as Sharpness Aware Minimization (SAM (Foret et al., 2021)) seeks flatter solutions by optimizing losses under local parameter perturbations. Lee et al. (2025) apply this principle to reduce parameter interference in model merging. These studies clarify important geometric conditions for merging in non-private settings, but do not examine how privatization perturbs those conditions before composition.

3 Preliminaries

3.1 Problem Setup

Private Task-specific Fine-tuning. Let $w_0 \in \mathbb{R}^d$ denote a public pretrained model. We consider T downstream tasks with private datasets $\{\mathcal{D}_t\}_{t=1}^T$. Starting from w_0, each task independently applies a randomized fine-tuning algorithm $\mathcal{A}_t$ to obtain $w_t = \mathcal{A}_t(w_0, \mathcal{D}_t)$. For task t, we define the empirical loss as $\mathcal{L}_t(w) := \frac{1}{|\mathcal{D}_t|} \sum_{(x,y)\in\mathcal{D}_t} \ell(w; x, y)$, which serves as a proxy for the task risk.

Definition 1 ((ε, δ)-DP). *A randomized algorithm $\mathcal{A}$ is (ε, δ)-DP if, for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing by the addition or removal of one sample, and any measurable output set $\mathcal{S}$,*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta. \tag{1}$$

We require each task-specific algorithm $\mathcal{A}_t(w_0, \cdot)$ to satisfy this guarantee, with w_0 public and fixed. Under a fixed task assignment, disjoint task datasets, and independent task training, releasing $\{w_t\}_{t=1}^T$ preserves (ϵ, δ)-DP by parallel composition. Merging these outputs using only public information incurs no additional privacy loss by post-processing.

Post-hoc Model Merging. A merging rule constructs $w_{\text{merge}} = \mathcal{M}(w_1, \ldots, w_T)$ from the released task models. With task vectors $\Delta_t := w_t - w_0$, this can be written as $w_{\text{merge}} = w_0 + \mathcal{M}_\Delta(\Delta_1, \ldots, \Delta_T)$, where $\mathcal{M}_\Delta$ operates in task-vector space. Under a fixed privacy budget, our goal is to train task models that retain task-specific utility and remain compatible with subsequent merging, minimizing the average merged-model loss $\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(w_{\text{merge}})$.

3.2 Geometric Challenges of Private Mergeability

Private fine-tuning optimizes each task model independently, whereas merging evaluates these models after parameter composition. To distinguish task adaptation from the effect of merging, for task t, define the merge gap as $\text{Gap}_t(w_{\text{merge}}) := \mathcal{L}_t(w_{\text{merge}}) - \mathcal{L}_t(w_t)$. Then

$$\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(w_{\text{merge}}) = \underbrace{\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(w_t)}_{\text{task adaptation}} + \underbrace{\frac{1}{T}\sum_{t=1}^T \text{Gap}_t(w_{\text{merge}})}_{\text{merge gap}}. \tag{2}$$

Our goal is to limit merge-induced loss increases while retaining task-specific utility under a fixed ε.

Let $d_t := w_{\text{merge}} - w_t$. If $\mathcal{L}_t$ is twice continuously differentiable along the merging segment, then

$$\text{Gap}_t(w_{\text{merge}}) = \nabla\mathcal{L}_t(w_t)^\top d_t + \int_0^1 (1-s) d_t^\top \nabla^2 \mathcal{L}_t(w_t + s d_t) d_t \, \mathrm{d}s. \tag{3}$$

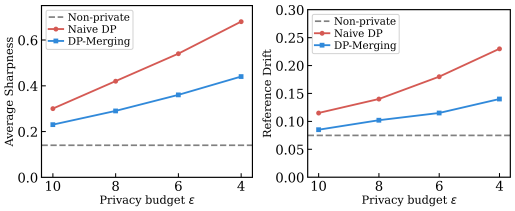
For any $\beta_t \geq 0$ that upper-bounds the largest Hessian eigenvalue along this segment,

$$\text{Gap}_t(w_{\text{merge}}) \leq \|\nabla\mathcal{L}_t(w_t)\|_2 \|d_t\|_2 + \frac{\beta_t}{2}\|d_t\|_2^2. \tag{4}$$

Here, $\|\nabla\mathcal{L}_t(w_t)\|_2$ measures first-order loss sensitivity, β_t bounds curvature along the merging path, and $\|d_t\|_2$ is the merging displacement. However, neither d_t nor the corresponding path is available during independent task training, since both depend on the other task models. This raises two challenges: encouraging robustness to an unknown merging displacement and constraining its magnitude without cross-task access.

Challenge 1: Local sharpness. Merging displaces each task model from its independently fine-tuned parameters. Sensitivity to parameter perturbations may therefore affect how well task utility is preserved after merging. Since the eventual merging direction is unknown during training, we examine local sensitivity using a gradient-direction sharpness proxy: $\text{Sharp}_t(w_t; \rho) = \mathcal{L}_t\left(w_t + \rho \frac{\nabla\mathcal{L}_t(w_t)}{\|\nabla\mathcal{L}_t(w_t)\|_2}\right) - \mathcal{L}_t(w_t)$. Figure 3a shows that Naive DP exhibits greater local sensitivity at tighter privacy budgets, while our method exhibits lower values. These observations motivate encouraging local perturbation robustness during private fine-tuning.

Challenge 2: Reference drift. The displacement term in Eq. 4 can be bounded using the distances of task models from the shared pretrained initialization. Consider weight averaging, where $w_{\text{merge}} = w_0 + \bar{\Delta}$ and $\bar{\Delta} = \frac{1}{T}\sum_{j=1}^T \Delta_j$. Then $w_{\text{merge}} - w_t = \bar{\Delta} - \Delta_t$, and $\|\bar{\Delta} - \Delta_t\|_2 \leq \frac{1}{T}\sum_{j=1}^T \|\Delta_j\|_2 + \|\Delta_t\|_2$. Thus, distances from w_0 provide an upper bound on merging displacement that can be regularized independently for each task. We quantify reference drift as $\text{Drift} = \frac{1}{T}\sum_{t=1}^T \frac{\|w_t - w_0\|_2}{\|w_0\|_2}$. As shown in Figure 3b, Naive DP exhibits larger reference drift at tighter privacy budgets in the evaluated settings. This motivates constraining reference drift while preserving useful task adaptation.



(a) Average Sharpness (b) Reference Drift

Figure 3: Empirical validation on the eight vision tasks using CLIP ViT-B/32.

4 DP-Merging

We propose DP-Merging, a geometry-aware private fine-tuning framework designed to produce task models for subsequent merging. The key idea is to encourage local perturbation robustness and constrain task-update magnitudes during independent private fine-tuning to support subsequent model merging.

Guided by Sec. 3.2, we combine two complementary components: *DP-compatible sharpness-aware fine-tuning*, which encourages robustness to local parameter perturbations, and *reference-anchored alignment*, which penalizes task-update magnitudes relative to the shared pretrained initialization w_0. The former targets local loss sensitivity; the latter regularizes an upper bound on merging displacement under weight averaging. Their combination couples local robustness with control of update magnitudes, while retaining task-specific adaptation.

4.1 DP-Compatible Sharpness-Aware Fine-Tuning

We first address the curvature term in Eq. 4. When a task model lies in a sharp local region, even a moderate parameter displacement can produce a large increase in task loss after merging. To mitigate this effect, DP-Merging follows a sharpness-aware principle Foret et al. (2021): instead of evaluating the training gradient only at the current model, it evaluates the gradient at a locally perturbed point. For a model w and minibatch B, define the clipped Gaussian gradient estimator

$$\tilde{g}(w;B) = 1/b(\sum_{z_i \in B} g_i/(\max\{1, \|g_i\|_2/C\}) + \xi), \quad g_i = \nabla_w \ell(w; z_i), \quad \xi \sim \mathcal{N}(0, \sigma^2 C^2 I), \tag{5}$$

C is the clipping threshold, σ is the noise multiplier, b is a fixed normalization constant. At iteration k for task t, we first compute the private gradient $\tilde{g}_{t,k} = \tilde{g}(w_{t,k}; B_{t,k})$ and use it to construct a local ascent perturbation

$$\epsilon_{t,k} = \rho_t \frac{\tilde{g}_{t,k}}{\|\tilde{g}_{t,k}\|_2}, \tag{6}$$

where ρ_t controls the neighborhood size. We then evaluate a private gradient at the perturbed point:

$$\tilde{h}_{t,k} = \tilde{g}(w_{t,k} + \epsilon_{t,k}; B_{t,k}). \tag{7}$$

The first gradient $\tilde{g}_{t,k}$ determines the perturbation direction, while the second gradient $\tilde{h}_{t,k}$ provides the update direction at the perturbed location. This procedure aims to encourage local robustness before merging. Both gradient evaluations are privatized using clipped Gaussian mechanisms and accounted for in the privacy analysis.

4.2 Reference-Anchored Alignment

Local perturbation robustness does not ensure that independently trained task models stay close to one another. We therefore anchor each task model to the shared initialization. At iteration k, DP-Merging updates task t by

$$w_{t,k+1} = w_{t,k} - \eta \left(\tilde{h}_{t,k} + 2\lambda(w_{t,k} - w_0) \right), \tag{8}$$

where $\lambda \geq 0$ controls the alignment strength. The second term is the gradient of $\lambda \|w_{t,k} - w_0\|_2^2$. Since it depends only on the current private training state and the public initialization, it incurs no additional privacy cost.

The anchor penalizes reference distances that bound the displacement term in Eq. 4. For uniform averaging, let $w_{\text{merge}} = \frac{1}{T} \sum_{s=1}^{T} w_s$. Then

$$\begin{aligned} \|w_{\text{merge}} - w_t\|_2 &= \left\| \frac{1}{T} \sum_{s=1}^{T} (w_s - w_t) \right\|_2 \leq \frac{1}{T} \sum_{s=1}^{T} \|w_s - w_t\|_2 \\ &\leq \frac{1}{T} \sum_{s=1}^{T} (\|w_s - w_0\|_2 + \|w_t - w_0\|_2) \end{aligned} \tag{9}$$

Thus, keeping each task model close to w_0 reduces an upper bound on the distance between the merged model and each task-specific solution. This alignment is induced only through the common reference point and does not require communication between tasks during fine-tuning.

After private fine-tuning, we obtain released task models $\{w_t^{\text{DP}}\}_{t=1}^{T}$. For clarity, we use weight averaging as the default merging rule: $w_{\text{merge}} = \frac{1}{T} \sum_{t=1}^{T} w_t^{\text{DP}}$. The same released models can also be used with other post-hoc merging rules. Since the final merging step only processes DP outputs, it is post-processing and preserves the privacy guarantees of task-specific fine-tuning.

5 Theoretical Analysis

5.1 Privacy guarantee of DP-Merging

Each iteration consists of two adaptive clipped Gaussian queries on one Poisson-sampled minibatch. We first account for their adaptive composition and then apply subsampling amplification to the joint mechanism. Perturbation construction, reference anchoring, optimizer updates, and model merging are post-processing of the private outputs. The proof is provided in Appendix C.

Algorithm 1 DP-Merging

Input: pretrained model w_0; private datasets $\{\mathcal{D}_t\}_{t=1}^T$; iterations K; learning rate η; perturbation radius ρ_t; alignment strength λ; clipping threshold C; noise multiplier σ; sampling rate q
Output: merged model w_{merge}
for each task $t = 1, \dots, T$ **do**
  Initialize task model $w_{t,0} \leftarrow w_0$
  for $k = 0, \dots, K-1$ **do**
    Sample minibatch $B_{t,k} \subset \mathcal{D}_t$ with sampling rate q
    Compute private gradient $\tilde{g}_{t,k} \leftarrow \tilde{g}(w_{t,k}; B_{t,k})$ using Eq. 5
    Construct perturbation $\epsilon_{t,k} \leftarrow \rho_t \tilde{g}_{t,k} / \|\tilde{g}_{t,k}\|_2$
    Compute perturbed private gradient $\tilde{h}_{t,k} \leftarrow \tilde{g}(w_{t,k} + \epsilon_{t,k}; B_{t,k})$
    Update with reference anchoring: $w_{t,k+1} \leftarrow w_{t,k} - \eta\big(\tilde{h}_{t,k} + 2\lambda(w_{t,k} - w_0)\big)$
  end for
  Obtain private task model $w_t^{\text{DP}} \leftarrow w_{t,K}$
end for
$w_{\text{merge}} \leftarrow \frac{1}{T}\sum_{t=1}^T w_t^{\text{DP}}$
return w_{merge}

Theorem 1 (**Privacy Guarantee**). *Under sample-level add/remove adjacency, assume disjoint task datasets with a fixed task assignment and independent task training. For task t, suppress task indices on the training parameters and run K iterations of the two-query mechanism in Section 4.1, with Poisson sampling rate q and per-query noise multiplier σ. Let $\varepsilon_{\text{SGM}}(\alpha; q, s)$ denote a valid RDP bound for the Poisson-subsampled Gaussian mechanism under add/remove adjacency. For any fixed $\delta \in (0,1)$ and nonempty finite set $\mathcal{O} \subset (1, \infty)$, w_t^{DP} is (ε_t, δ)-DP, where*

$$\varepsilon_t = \min_{\alpha \in \mathcal{O}} \left\{ K\varepsilon_{\text{SGM}}\left(\alpha; q, \frac{\sigma}{\sqrt{2}}\right) + \frac{\log(1/\delta)}{\alpha - 1} \right\}. \tag{10}$$

The joint release of all task models and their post-processed merged model is $(\max_t \varepsilon_t, \delta)$-DP.

The effective noise multiplier $\sigma/\sqrt{2}$ follows from adaptive Gaussian composition before Poisson amplification (Zhu et al., 2022). The detailed assumptions and proof are provided in Appendix C.1.

5.2 Mergeability Analysis

We relate local perturbation robustness and reference distances to merged-model utility under uniform averaging. Let $w_t := w_t^{\text{DP}}$, $w_{\text{merge}} = \frac{1}{T}\sum_t w_t$, $d_t := w_{\text{merge}} - w_t$, and $r_t := \|d_t\|_2$. For $\rho_t > 0$, define the worst-case local loss increment $S_t := \max_{\|e\|_2 \le \rho_t} [\mathcal{L}_t(w_t + e) - \mathcal{L}_t(w_t)]$. Unlike the directional diagnostic in Sec. 3.2, S_t considers all perturbations within the neighborhood.

Theorem 2 (**Sharpness-based merge-gap bound**). *Assume that each $\mathcal{L}_t$ is twice continuously differentiable with an M_t-Lipschitz Hessian on a convex neighborhood containing the radius-ρ_t ball around w_t and w_{merge}. Then*

$$\text{Gap}_t(w_{\text{merge}}) \le \|\nabla \mathcal{L}_t(w_t)\|_2 r_t + \left(\frac{S_t}{\rho_t^2} + \frac{M_t \rho_t}{6}\right) r_t^2 + \frac{M_t}{6} r_t^3. \tag{11}$$

The proof bounds the local Hessian eigenvalue by $\lambda_{\max}(\nabla^2 \mathcal{L}_t(w_t)) \le 2S_t/\rho_t^2 + M_t\rho_t/3$. Thus, local worst-case loss variation bounds the quadratic displacement cost, while the terms involving M_t account for curvature variation. The result does not require $r_t \le \rho_t$.

To connect displacement with reference anchoring, define $D := \frac{1}{T}\sum_t \|w_t - w_0\|_2^2$. Uniform averaging gives $\frac{1}{T}\sum_t r_t^2 = D - \|\frac{1}{T}\sum_t (w_t - w_0)\|_2^2 \le D$.

Corollary 1 (Reference-drift bound on merged loss). *Under Theorem 2, let $G := (\frac{1}{T}\sum_t \|\nabla \mathcal{L}_t(w_t)\|_2^2)^{1/2}$, $A := \max_t(S_t/\rho_t^2 + M_t\rho_t/6)$, and $M := \max_t M_t$. Then*

$$\frac{1}{T}\sum_t \mathcal{L}_t(w_{\text{merge}}) \le \frac{1}{T}\sum_t \mathcal{L}_t(w_t) + G\sqrt{D} + AD + \frac{M\sqrt{T}}{6} D^{3/2}. \tag{12}$$

The bound motivates combining local robustness with reference anchoring: S_t enters the sensitivity term, while D controls an upper bound on average squared merging displacement. Individual task losses and gradients remain explicit, so reducing drift alone does not ensure better utility. Complete proofS see the Appendix C.2, C.3.

Table 1: Multi-task accuracy (%) on the 8-task vision benchmark with ViT-B/32, ε=4. "+ Ours" trains task models with DP-Merging before applying the corresponding merging method.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg Acc
DP + Individual	$59.2_{\pm 0.32}$	$67.7_{\pm 0.41}$	$86.1_{\pm 0.52}$	$89.7_{\pm 0.35}$	$87.5_{\pm 0.49}$	$88.6_{\pm 0.69}$	$89.3_{\pm 0.55}$	$69.4_{\pm 0.43}$	79.7
DP + Weight Averaging	$45.3_{\pm 0.55}$	$43.4_{\pm 0.61}$	$51.9_{\pm 0.48}$	$52.7_{\pm 0.50}$	$43.2_{\pm 0.47}$	$33.8_{\pm 0.68}$	$67.1_{\pm 0.87}$	$30.9_{\pm 0.74}$	46.0
+ Ours	$48.2_{\pm 0.49}$	$49.5_{\pm 0.56}$	$56.3_{\pm 0.44}$	$57.5_{\pm 0.43}$	$43.2_{\pm 0.58}$	$37.4_{\pm 0.45}$	$69.0_{\pm 0.62}$	$36.3_{\pm 0.44}$	$49.7_{+3.7}$
DP + RegMean	$45.8_{\pm 0.62}$	$44.9_{\pm 0.51}$	$58.3_{\pm 0.63}$	$58.6_{\pm 0.57}$	$59.1_{\pm 0.47}$	$48.3_{\pm 0.55}$	$73.7_{\pm 0.51}$	$32.0_{\pm 0.60}$	52.6
+ Ours	$49.4_{\pm 0.49}$	$48.6_{\pm 0.46}$	$63.1_{\pm 0.55}$	$62.4_{\pm 0.40}$	$63.6_{\pm 0.35}$	$53.1_{\pm 0.48}$	$76.8_{\pm 0.42}$	$38.2_{\pm 0.43}$	$56.9_{+4.3}$
DP + Task Arithmetic	$45.2_{\pm 0.66}$	$44.9_{\pm 0.56}$	$56.7_{\pm 0.64}$	$58.9_{\pm 0.58}$	$60.2_{\pm 0.55}$	$59.7_{\pm 0.63}$	$77.3_{\pm 0.81}$	$30.4_{\pm 0.75}$	54.2
+ Ours	$48.9_{\pm 0.48}$	$49.2_{\pm 0.41}$	$59.6_{\pm 0.37}$	$62.4_{\pm 0.40}$	$67.1_{\pm 0.36}$	$62.5_{\pm 0.49}$	$79.6_{\pm 0.56}$	$35.1_{\pm 0.59}$	$58.1_{+3.9}$
DP + TIES-Merging	$49.8_{\pm 0.54}$	$48.6_{\pm 0.50}$	$50.7_{\pm 0.61}$	$59.7_{\pm 0.46}$	$66.5_{\pm 0.57}$	$72.1_{\pm 0.48}$	$78.3_{\pm 0.52}$	$34.2_{\pm 0.63}$	57.5
+ Ours	$52.9_{\pm 0.41}$	$53.7_{\pm 0.39}$	$56.0_{\pm 0.45}$	$64.1_{\pm 0.33}$	$72.1_{\pm 0.28}$	$78.0_{\pm 0.42}$	$81.4_{\pm 0.50}$	$38.6_{\pm 0.47}$	$62.1_{+4.6}$
DP + PCB Merging	$45.1_{\pm 0.51}$	$43.4_{\pm 0.59}$	$50.5_{\pm 0.61}$	$57.2_{\pm 0.54}$	$61.7_{\pm 0.35}$	$70.9_{\pm 0.70}$	$77.0_{\pm 0.48}$	$30.3_{\pm 0.55}$	54.5
+ Ours	$49.5_{\pm 0.33}$	$48.2_{\pm 0.47}$	$56.4_{\pm 0.52}$	$63.4_{\pm 0.35}$	$65.0_{\pm 0.40}$	$76.3_{\pm 0.49}$	$79.1_{\pm 0.38}$	$35.4_{\pm 0.38}$	$59.2_{+4.7}$
DP + WUDI-Merging	$47.6_{\pm 0.47}$	$45.5_{\pm 0.51}$	$58.5_{\pm 0.63}$	$59.3_{\pm 0.52}$	$66.4_{\pm 0.48}$	$57.1_{\pm 0.57}$	$78.2_{\pm 0.57}$	$39.1_{\pm 0.55}$	56.5
+ Ours	$51.5_{\pm 0.28}$	$48.6_{\pm 0.46}$	$63.1_{\pm 0.50}$	$63.9_{\pm 0.44}$	$70.4_{\pm 0.49}$	$61.2_{\pm 0.45}$	$82.7_{\pm 0.48}$	$44.8_{\pm 0.36}$	$60.8_{+4.3}$

Table 2: Multi-task accuracy (%) on the 8-task vision benchmark with ViT-L/14, $\varepsilon = 4$.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg Acc
DP + Individual	$64.8_{\pm 0.37}$	$74.2_{\pm 0.50}$	$93.1_{\pm 0.33}$	$92.9_{\pm 0.23}$	$92.2_{\pm 0.19}$	$95.4_{\pm 0.34}$	$96.7_{\pm 0.48}$	$75.6_{\pm 0.51}$	85.5
DP + Weight Averaging	$50.3_{\pm 0.46}$	$49.4_{\pm 0.53}$	$57.4_{\pm 0.38}$	$57.7_{\pm 0.44}$	$50.2_{\pm 0.41}$	$39.8_{\pm 0.56}$	$72.5_{\pm 0.35}$	$35.1_{\pm 0.44}$	51.6
+ Ours	$54.4_{\pm 0.42}$	$54.4_{\pm 0.45}$	$62.1_{\pm 0.35}$	$62.2_{\pm 0.36}$	$55.4_{\pm 0.38}$	$45.4_{\pm 0.45}$	$80.1_{\pm 0.29}$	$44.2_{\pm 0.47}$	$57.3_{+5.7}$
DP + RegMean	$50.3_{\pm 0.58}$	$49.8_{\pm 0.53}$	$61.0_{\pm 0.43}$	$64.6_{\pm 0.52}$	$63.2_{\pm 0.47}$	$53.4_{\pm 0.38}$	$79.7_{\pm 0.50}$	$41.6_{\pm 0.55}$	57.9
+ Ours	$53.3_{\pm 0.46}$	$54.8_{\pm 0.51}$	$65.9_{\pm 0.34}$	$70.6_{\pm 0.40}$	$69.2_{\pm 0.29}$	$55.0_{\pm 0.56}$	$82.7_{\pm 0.49}$	$49.6_{\pm 0.57}$	$62.7_{+4.8}$
DP + Task Arithmetic	$50.2_{\pm 0.51}$	$50.9_{\pm 0.61}$	$61.7_{\pm 0.55}$	$73.5_{\pm 0.43}$	$75.4_{\pm 0.26}$	$54.7_{\pm 0.50}$	$82.3_{\pm 0.33}$	$45.4_{\pm 0.68}$	61.8
+ Ours	$54.1_{\pm 0.47}$	$55.5_{\pm 0.42}$	$64.4_{\pm 0.52}$	$78.9_{\pm 0.37}$	$84.1_{\pm 0.25}$	$62.5_{\pm 0.43}$	$84.1_{\pm 0.28}$	$51.3_{\pm 0.56}$	$66.7_{+4.9}$
DP + TIES-Merging	$44.8_{\pm 0.40}$	$45.3_{\pm 0.65}$	$46.8_{\pm 0.59}$	$64.4_{\pm 0.36}$	$71.2_{\pm 0.37}$	$58.7_{\pm 0.64}$	$83.9_{\pm 0.48}$	$50.2_{\pm 0.69}$	58.1
+ Ours	$50.8_{\pm 0.43}$	$52.6_{\pm 0.55}$	$51.7_{\pm 0.47}$	$69.7_{\pm 0.42}$	$75.3_{\pm 0.30}$	$61.1_{\pm 0.51}$	$87.8_{\pm 0.26}$	$52.8_{\pm 0.46}$	$62.7_{+4.6}$
DP + PCB Merging	$50.7_{\pm 0.57}$	$49.6_{\pm 0.48}$	$51.5_{\pm 0.56}$	$62.2_{\pm 0.50}$	$67.7_{\pm 0.34}$	$55.9_{\pm 0.45}$	$82.6_{\pm 0.25}$	$42.1_{\pm 0.61}$	57.8
+ Ours	$54.5_{\pm 0.54}$	$54.2_{\pm 0.56}$	$56.2_{\pm 0.46}$	$67.9_{\pm 0.40}$	$71.2_{\pm 0.32}$	$61.3_{\pm 0.55}$	$87.2_{\pm 0.38}$	$47.4_{\pm 0.52}$	$62.5_{+4.7}$
DP + WUDI-Merging	$51.7_{\pm 0.59}$	$51.1_{\pm 0.72}$	$63.4_{\pm 0.78}$	$64.3_{\pm 0.56}$	$71.8_{\pm 0.45}$	$63.1_{\pm 0.48}$	$83.0_{\pm 0.40}$	$44.2_{\pm 0.43}$	61.6
+ Ours	$54.9_{\pm 0.50}$	$54.7_{\pm 0.66}$	$66.8_{\pm 0.62}$	$68.5_{\pm 0.48}$	$76.6_{\pm 0.43}$	$68.3_{\pm 0.47}$	$85.1_{\pm 0.32}$	$50.3_{\pm 0.42}$	$65.6_{+4.0}$

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTINGS

Datasets and Models. We evaluate our method on *vision and language tasks.* For vision tasks, following Yang et al. (2024); Cheng et al. (2025), we study multi-task model merging across eight image classification datasets, namely SUN397 (Xiao et al., 2016), Cars (Krause et al., 2013), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), SVHN (Yuval, 2011), GTSRB (Stallkamp et al., 2011), MNIST (LeCun, 1998), and DTD (Cimpoi et al., 2014), and adopt CLIP-based ViT backbones (Radford et al., 2021), including ViT-B/32, ViT-B/16, and ViT-L/14. For language tasks, we use the GLUE benchmark (Wang et al., 2018) and evaluate with RoBERTa-Base and RoBERTa-Large (Liu et al., 2019).

Baselines. We compare with representative merging methods under the same DP setting, including Weight Averaging Wortsman et al. (2022), RegMean Jin et al. (2023), Task Arithmetic Ilharco et al. (2023), TIES-Merging Yadav et al. (2023), PCB Merging DU et al. (2024), and WUDI-Merging Cheng et al. (2025). All baselines independently train task-specific models with the same DP optimizer and privacy budget, and then merge the released DP models without accessing private data. The merging step is post-processing and therefore incurs no additional privacy loss. We apply the same operators to the task models produced by DP-Merging to evaluate whether our merge-aware fine-tuning improves mergeability across merging rules.

Implementation Details. We implement sample-level DP fine-tuning with Opacus Yousefpour et al. (2021) and use AdamW Loshchilov et al. (2017) as the default optimizer. Unless otherwise specified, each task model is trained with clipping norm $C = 0.2$, batch size 16, learning rate $1e-5$, weight decay $1e-2$, and 10 training epochs/steps. Privacy loss is computed using the Opacus privacy accountant and reported as (ε, δ)-DP, where $\delta = 1/N$ and $\varepsilon \in \{1, 2, 4, 8\}$. For DP-Merging, the perturbation radius $\rho = 0.05$ and the alignment coefficient $\lambda = 0.05$. Additional implementation details are provided in Appendix B.3.

6.2 Main Results

Evaluation on Visual Tasks. We evaluate DP-Merging on the 8-task vision benchmark under $\varepsilon = 4$. Table 1 and Table 2 report the main results with ViT-B/32 and ViT-L/14, and additional results with ViT-B/16 are provided in Appendix **??**. Across both backbones, merging standard DP task models yields substantially lower average accuracy than evaluating the individual task models, suggesting that DP fine-tuning hurts task-vector mergeability. DP-Merging consistently improves the merged accuracy under all merging operators. For example, WUDI-Merging improves from 56.5% to 60.8% on ViT-B/32, while Task Arithmetic improves from 61.8% to 66.7% on ViT-L/14 when applied to DP-Merging task models.

Evaluation on Language Tasks. Table 3 and Table 4 report results on eight GLUE tasks with RoBERTa-Large and RoBERTa-Base at $\varepsilon = 4$. Compared with standard DP baselines, DP-Merging consistently achieves higher average accuracy across different merging operators. For RoBERTa-Large, average scores increase from 71.8–74.8 to 74.0–76.9. For RoBERTa-Base, they increase from 71.4–74.5 to 73.8–76.4. WUDI-Merging achieves the highest average score with DP-Merging on both backbones, reaching 76.9 and 76.4. The improvements across model sizes and merging methods extend the vision results to language understanding tasks.

Table 3: Multi-task performance of RoBERTa-Large models on GLUE benchmark, $\varepsilon = 4$.

Method	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
DP + Individual	$60.5_{\pm 0.22}$	$79.1_{\pm 0.19}$	$81.0_{\pm 0.29}$	$89.2_{\pm 0.12}$	$82.3_{\pm 0.20}$	$77.1_{\pm 0.21}$	$93.8_{\pm 0.14}$	$82.1_{\pm 0.15}$	80.6
DP + Weight Averaging	$56.1_{\pm 0.40}$	$73.8_{\pm 0.46}$	$67.9_{\pm 0.41}$	$79.5_{\pm 0.35}$	$78.6_{\pm 0.32}$	$69.3_{\pm 0.46}$	$91.2_{\pm 0.28}$	$67.6_{\pm 0.44}$	73.0
+ Ours	$58.2_{\pm 0.32}$	$76.6_{\pm 0.33}$	$69.2_{\pm 0.30}$	$79.3_{\pm 0.25}$	$81.6_{\pm 0.24}$	$69.9_{\pm 0.37}$	$91.8_{\pm 0.25}$	$68.2_{\pm 0.36}$	$74.4_{+1.4}$
DP + RegMean	$55.6_{\pm 0.40}$	$73.8_{\pm 0.34}$	$65.7_{\pm 0.43}$	$79.2_{\pm 0.38}$	$78.6_{\pm 0.35}$	$67.9_{\pm 0.45}$	$91.0_{\pm 0.25}$	$62.5_{\pm 0.47}$	71.8
+ Ours	$58.8_{\pm 0.34}$	$78.7_{\pm 0.29}$	$68.6_{\pm 0.35}$	$79.5_{\pm 0.21}$	$80.9_{\pm 0.26}$	$69.3_{\pm 0.36}$	$91.9_{\pm 0.28}$	$64.6_{\pm 0.32}$	$74.0_{+2.3}$
DP + Task Arithmetic	$56.1_{\pm 0.46}$	$75.5_{\pm 0.38}$	$67.2_{\pm 0.43}$	$78.9_{\pm 0.32}$	$81.5_{\pm 0.31}$	$69.9_{\pm 0.48}$	$91.2_{\pm 0.50}$	$61.7_{\pm 0.45}$	72.8
+ Ours	$59.6_{\pm 0.26}$	$77.8_{\pm 0.22}$	$68.9_{\pm 0.31}$	$81.7_{\pm 0.25}$	$82.8_{\pm 0.36}$	$70.9_{\pm 0.30}$	$92.1_{\pm 0.36}$	$63.8_{\pm 0.34}$	$74.7_{+1.9}$
DP + TIES-Merging	$60.5_{\pm 0.35}$	$78.3_{\pm 0.29}$	$70.4_{\pm 0.38}$	$80.6_{\pm 0.22}$	$80.8_{\pm 0.46}$	$70.6_{\pm 0.34}$	$90.5_{\pm 0.24}$	$63.0_{\pm 0.33}$	74.3
+ Ours	$61.7_{\pm 0.32}$	$81.6_{\pm 0.18}$	$73.3_{\pm 0.21}$	$82.5_{\pm 0.27}$	$82.6_{\pm 0.29}$	$71.8_{\pm 0.22}$	$92.6_{\pm 0.16}$	$66.4_{\pm 0.25}$	$76.6_{+2.2}$
DP + PCB Merging	$56.2_{\pm 0.47}$	$74.2_{\pm 0.39}$	$76.6_{\pm 0.36}$	$77.5_{\pm 0.44}$	$81.8_{\pm 0.30}$	$70.6_{\pm 0.34}$	$91.1_{\pm 0.45}$	$60.2_{\pm 0.32}$	73.5
+ Ours	$58.9_{\pm 0.34}$	$75.7_{\pm 0.23}$	$79.4_{\pm 0.25}$	$79.8_{\pm 0.21}$	$83.1_{\pm 0.35}$	$73.4_{\pm 0.28}$	$92.6_{\pm 0.14}$	$62.6_{\pm 0.31}$	$75.5_{+2.2}$
DP + WUDI-Merging	$57.2_{\pm 0.15}$	$75.8_{\pm 0.29}$	$78.9_{\pm 0.32}$	$79.6_{\pm 0.30}$	$82.6_{\pm 0.28}$	$73.5_{\pm 0.43}$	$91.3_{\pm 0.32}$	$59.5_{\pm 0.41}$	74.8
+ Ours	$59.8_{\pm 0.22}$	$77.9_{\pm 0.13}$	$82.1_{\pm 0.24}$	$83.3_{\pm 0.23}$	$82.7_{\pm 0.25}$	$75.1_{\pm 0.27}$	$92.8_{\pm 0.19}$	$61.3_{\pm 0.24}$	$76.9_{+2.1}$

Table 4: Multi-task performance of RoBERTa-Base models on GLUE benchmark, $\varepsilon = 4$.

Method	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
DP + Individual	$59.2_{\pm 0.46}$	$77.7_{\pm 0.21}$	$80.1_{\pm 0.38}$	$88.7_{\pm 0.19}$	$81.5_{\pm 0.24}$	$76.7_{\pm 0.43}$	$92.7_{\pm 0.17}$	$79.4_{\pm 0.35}$	79.5
DP + Weight Averaging	$55.3_{\pm 0.49}$	$73.4_{\pm 0.28}$	$67.4_{\pm 0.31}$	$78.7_{\pm 0.26}$	$78.2_{\pm 0.29}$	$68.8_{\pm 0.37}$	$90.5_{\pm 0.27}$	$66.1_{\pm 0.55}$	72.3
+ Ours	$58.2_{\pm 0.44}$	$76.1_{\pm 0.26}$	$69.0_{\pm 0.29}$	$79.1_{\pm 0.27}$	$81.3_{\pm 0.25}$	$69.7_{\pm 0.33}$	$91.6_{\pm 0.28}$	$68.2_{\pm 0.42}$	$74.2_{+1.9}$
DP + RegMean	$55.3_{\pm 0.38}$	$73.5_{\pm 0.30}$	$65.6_{\pm 0.57}$	$78.6_{\pm 0.29}$	$78.1_{\pm 0.26}$	$67.4_{\pm 0.38}$	$90.7_{\pm 0.28}$	$62.0_{\pm 0.39}$	71.4
+ Ours	$58.6_{\pm 0.32}$	$78.2_{\pm 0.27}$	$68.8_{\pm 0.48}$	$79.5_{\pm 0.25}$	$80.6_{\pm 0.29}$	$69.2_{\pm 0.34}$	$91.9_{\pm 0.26}$	$63.8_{\pm 0.38}$	$73.8_{+2.4}$
DP + Task Arithmetic	$55.2_{\pm 0.57}$	$74.9_{\pm 0.29}$	$66.7_{\pm 0.40}$	$78.0_{\pm 0.37}$	$80.2_{\pm 0.28}$	$69.7_{\pm 0.36}$	$90.3_{\pm 0.31}$	$61.4_{\pm 0.53}$	72.1
+ Ours	$59.6_{\pm 0.52}$	$77.8_{\pm 0.27}$	$68.9_{\pm 0.28}$	$80.9_{\pm 0.26}$	$82.0_{\pm 0.39}$	$71.5_{\pm 0.32}$	$91.6_{\pm 0.30}$	$63.2_{\pm 0.44}$	$74.4_{+2.4}$
DP + TIES-Merging	$59.8_{\pm 0.38}$	$78.6_{\pm 0.30}$	$70.7_{\pm 0.28}$	$79.7_{\pm 0.31}$	$81.2_{\pm 0.27}$	$70.1_{\pm 0.35}$	$91.3_{\pm 0.29}$	$64.2_{\pm 0.56}$	74.5
+ Ours	$61.0_{\pm 0.56}$	$80.6_{\pm 0.28}$	$72.4_{\pm 0.29}$	$80.9_{\pm 0.27}$	$82.1_{\pm 0.30}$	$71.2_{\pm 0.33}$	$92.6_{\pm 0.28}$	$65.8_{\pm 0.45}$	$75.8_{+1.4}$
DP + PCB Merging	$55.7_{\pm 0.36}$	$73.6_{\pm 0.39}$	$76.5_{\pm 0.31}$	$77.2_{\pm 0.28}$	$81.7_{\pm 0.30}$	$70.3_{\pm 0.47}$	$91.0_{\pm 0.27}$	$60.1_{\pm 0.63}$	73.3
+ Ours	$58.6_{\pm 0.33}$	$74.2_{\pm 0.20}$	$78.3_{\pm 0.35}$	$79.5_{\pm 0.30}$	$82.9_{\pm 0.28}$	$72.4_{\pm 0.34}$	$92.3_{\pm 0.29}$	$61.6_{\pm 0.41}$	$75.0_{+1.7}$
DP + WUDI-Merging	$56.7_{\pm 0.37}$	$75.5_{\pm 0.28}$	$78.5_{\pm 0.30}$	$79.3_{\pm 0.29}$	$82.4_{\pm 0.27}$	$73.1_{\pm 0.36}$	$91.2_{\pm 0.31}$	$59.0_{\pm 0.37}$	74.5
+ Ours	$59.3_{\pm 0.34}$	$76.4_{\pm 0.29}$	$81.4_{\pm 0.28}$	$82.1_{\pm 0.34}$	$83.2_{\pm 0.22}$	$74.6_{\pm 0.26}$	$92.5_{\pm 0.35}$	$61.3_{\pm 0.32}$	$76.4_{+1.9}$

Table 5: Performance under different ε on visual and language tasks using TIES-Merging.

Privacy Budget ε	Visual Tasks (ViT-B/32)			Language Tasks (RoBERTa-Large)		
	DP (AdamW)	DP-Merging	Gain	DP (AdamW)	DP-Merging	Gain
1.0	46.1	**52.2**	+6.1	64.6	**68.4**	+3.8
2.0	52.8	**58.6**	+5.8	68.3	**70.6**	+2.3
4.0	57.5	**62.1**	+4.6	74.3	**76.6**	+2.3
8.0	62.0	**66.7**	+4.7	76.3	**77.8**	+1.5

6.3 Analysis under Different Privacy Budgets

We vary the privacy budget over $\varepsilon \in \{1, 2, 4, 8\}$ while keeping TIES-Merging as the merging rule. As shown in Table 5, DP-Merging outperforms standard DP fine-tuning with AdamW at every evaluated budget on both vision and language tasks. The largest gains occur at $\varepsilon = 1$: average accuracy increases from 46.1% to 52.2% on ViT-B/32, and the average score increases from 64.6 to 68.4 on RoBERTa-Large.

6.4 Ablation Study

We conduct ablation studies to understand the contribution of each component in DP-Merging and analyze its sensitivity to key hyperparameters. All experiments use $\varepsilon = 4$ and default settings.

Effect of each component. DP-Merging contains two key components: sharpness-aware fine-tuning and reference-anchored alignment. To evaluate their individual contributions, we compare the full method with two variants: (1) removing the alignment component, and (2) removing the sharpness-aware fine-tuning. As shown in Table 6, both components improve performance over naive DP. The sharpness-aware fine-tuning mainly improves local robustness by encouraging flatter task-specific solutions, while the alignment component enhances cross-task compatibility by reducing discrepancies among task vectors. Combining both components yields the best performance, demonstrating that sharpness and alignment are complementary.

Table 6: Ablation study of different merging algorithms on the 8-task vision benchmark with ViT-B/32. "w/o Alignment" represents without the reference-anchored alignment component, and "w/o Flatness" represents without sharpness-aware fine-tuning.

Variant	Weight Averaging	Task Arithmetic	TIES-Merging	PCB-Merging
DP-AdamW	46.0	54.2	57.5	54.5
DP-Merging	49.7	58.1	62.1	59.2
DP-Merging w/o Alignment	48.1	55.6	60.7	57.4
DP-Merging w/o Flatness	47.5	56.2	59.8	57.9

Sensitivity to the flatness radius ρ. We study the influence of the flatness radius ρ on merged accuracy. When varying ρ, the alignment strength λ is fixed to its default value. As shown in Table 7, moderate values of ρ consistently improve merged accuracy, while overly large values may hurt task-specific adaptation or over-constrain different tasks.

Sensitivity to the alignment strength λ. We study the effect of alignment strength λ on merged accuracy (flatness radius ρ fixed). As shown in Table 8, moderate values of λ consistently improve merged accuracy, while excessively large values may limit flexibility across tasks.

Table 7: Sensitivity to the flatness radius ρ on ViT-B/32 using TIES-Merging.

ρ	0	0.01	0.03	0.05	0.10
Avg Acc	59.8	60.4	61.6	**62.1**	61.8

Table 8: Sensitivity to the alignment strength λ on ViT-B/32 using TIES-Merging.

λ	0	0.01	0.05	0.10	0.50
Avg Acc	60.8	61.2	**62.2**	62.0	61.6

Robustness to merging operators. Since our method improves the private fine-tuning stage rather than designing a new merging rule, we evaluate whether it works with different data-free merging operators. As shown in Table 9, DP-Merging consistently improves the final merged model across different representative merging operators, demonstrating that our method enhances the intrinsic mergeability of task-specific models.

Table 9: Robustness to different merging operators on the 8-task vision benchmark with ViT-B/32.

Training	Weight Averaging	Task Arithmetic	TIES-Merging	PCB-Merging	WUDI-Merging
DP(AdamW)	46.0	54.2	57.5	54.5	56.5
DP-Merging (Ours)	49.7	58.1	62.1	59.2	60.8

7 Conclusion

We studied model merging under differential privacy, focusing on the geometric properties of independently fine-tuned private task models. Our empirical observations identify local sharpness and reference drift as two factors associated with poor mergeability. Motivated by these findings, we proposed DP-Merging, which combines DP-compatible sharpness-aware fine-tuning with reference-anchored regularization. Our analysis bounds the merge-induced loss increase using local loss variation and parameter displacement, and connects reference distances to average merged loss. Experiments across vision and language benchmarks demonstrate improvements over standard private fine-tuning across the evaluated privacy budgets and merging methods, with both private gradient evaluations included in the privacy accounting. These results suggest that private fine-tuning and model merging should be designed together. Future work could develop merging rules tailored to the geometry and uncertainty of DP-trained task updates. Jointly designing private fine-tuning procedures and merging rules may further improve the privacy–utility trade-off.

REFERENCES

Ahmed Agiza, Marina Neseem, and Sherief Reda. Mtlora: Low-rank adaptation approach for efficient multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16196–16205, 2024.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term fine-tuning of foundation models. In *International Conference on Machine Learning*, pp. 4730–4751. PMLR, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Guodong DU, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, and Min Zhang. Parameter competition balancing for model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2665–2679, 2023.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Linear mode connectivity between multiple models modulo permutation symmetries. In *Forty-second International Conference on Machine Learning*, 2025.

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

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, pp. 554–561, 2013.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yeoreum Lee, Jinwook Jung, and Sungyong Baik. Mitigating parameter interference in model merging via sharpness-aware fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xianzhi Li, Ran Zmigrod, Zhiqiang Ma, Xiaomo Liu, and Xiaodan Zhu. Fine-tuning language models with differential privacy through adaptive noise allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8368–8375, 2024.

Jin Liu, Yinbin Miao, Ning Xi, and Junkang Liu. Rethinking loRA for privacy-preserving federated learning in large models. In *The Fourteenth International Conference on Learning Representations*, 2026.

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, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Ilya Mironov. Rényi differential privacy. In *Proc. IEEE computer security foundations symposium (CSF)*, pp. 263–275, 2017.

Jinseong Park, Hoki Kim, Yujin Choi, and Jaewook Lee. Differentially private sharpness-aware training. In *International Conference on Machine Learning*, pp. 27204–27224. PMLR, 2023.

Fidel A Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20237–20246, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, pp. 1453–1460. IEEE, 2011.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Lijin Wang, Jingjing Wang, Tianshuo Cong, Xinlei He, Zhan Qin, and Xinyi Huang. From purity to peril: Backdooring merged models from" harmless" benign components. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6339–6358, 2025.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Taming momentum: Rethinking optimizer states through low-rank approximation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 119:3–22, 2016.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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):1–41, 2026.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

Zenghui Yuan, Yangming Xu, Jiawen Shi, Pan Zhou, and Lichao Sun. Merge hijacking: Backdoor attacks to model merging of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32688–32703, 2025.

Netzer Yuval. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *International Conference on Machine Learning*, pp. 77090–77106. PMLR, 2025.

Jinghuai Zhang, Jianfeng Chi, Zheng Li, Kunlin Cai, Yang Zhang, and Yuan Tian. Badmerging: Backdoor attacks against model merging. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4450–4464, 2024a.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Dpzero: Private fine-tuning of language models without backpropagation. In *International Conference on Machine Learning*, pp. 59210–59246. PMLR, 2024b.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pp. 4782–4817. PMLR, 2022.

8 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including **SUN397, Cars, RESISC45, EuroSAT, SVHN, GTSRB, MNIST, DTD, and the GLUE tasks CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST2, and STSB**, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

9 REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the results presented in this paper. Detailed experimental settings, including training procedures, model configurations, hyperparameters, and hardware specifications, are provided in Section 6.1 and Appendix B. The source code and scripts required to reproduce our experiments will be made publicly available upon acceptance.

10 AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript writing and polishing, including sentence rephrasing, grammar checking, improving readability and clarity, and checking the consistency of theoretical arguments and mathematical notation.

The LLMs were not used to originate the research ideas or design the methods or experiments. The research and analyses were conducted by the authors. The authors reviewed and revised all AI-assisted content to ensure its accuracy and consistency with the underlying research. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

When Privacy Hurts Mergeability: Geometry-Aware Model Merging under Differential Privacy

Supplementary Material

LIST OF APPENDICES

A: Notation
B: Implementation of Experiments
- B.1 Datasets
- B.2 Models
- B.3 Implementation Details
- B.4 More Results

C: Implementation of Theoretical Analysis
- C.1 Privacy guarantee of DP-Merging
- C.2 Mergeability Analysis
- C.3 Reference Drift and Merged-Loss Bound
- C.4 Private Perturbation Approximation

D: Limitations and Future Work

A NOTATION

Table 10 summarizes the main notation used throughout the paper.

Table 10: Summary of notation.

Notation	Description
w_0	Public pretrained initialization shared by all downstream tasks.
T	Number of downstream tasks.
$t \in [T]$	Task index.
$\mathcal{D}_t$	Private training dataset of task t.
$\ell(w; x, y)$	Sample-wise loss evaluated at model parameter w on example (x, y).
$\mathcal{L}_t(w)$	Empirical loss of task t on $\mathcal{D}_t$.
$\mathcal{A}_t$	Randomized DP fine-tuning algorithm for task t.
w_t	Task-specific model obtained by privately fine-tuning w_0 on $\mathcal{D}_t$.
$\Delta_t = w_t - w_0$	Task vector of task t relative to the public initialization.
w_{merge}	Merged model obtained from the released task-specific models.
$\mathcal{M}$	Model-merging operator, such as Weight Averaging, RegMean, Task Arithmetic, TIES-Merging, PCB, or WUDI.
H_t	Hessian or local curvature matrix of $\mathcal{L}_t$ around w_t.
$\lambda_{\max}(H_t)$	Largest eigenvalue of H_t, used as a local sharpness measure.
$\|w_t - w_0\|_2$	Reference drift of task model w_t from the shared initialization w_0.
C	Per-sample gradient clipping threshold in DP fine-tuning.
σ	Noise multiplier of the Gaussian mechanism.
q	Sampling rate used in DP fine-tuning.
(ε, δ)	Differential privacy parameters.
ρ	SAM radius used in the DP-aware flat-minima search component.
λ	Coefficient of the reference-based alignment regularizer.

B IMPLEMENTATION OF EXPERIMENTS

B.1 DATASETS

We evaluate DP-Merging on both vision and language benchmarks. For all tasks, the training data are treated as private and are only used during DP fine-tuning. The merging stage only accesses the released DP task models and does not use the original training data.

Vision tasks. For vision experiments, we follow the standard multi-task model-merging setting used in prior work (Yang et al., 2024; Cheng et al., 2025). We consider eight image classification tasks: SUN397 (Xiao et al., 2016) for scene recognition, Cars (Krause et al., 2013) for fine-grained car classification, RESISC45 (Cheng et al., 2017) and EuroSAT (Helber et al., 2019) for remote-sensing and land-cover recognition, SVHN (Yuval, 2011) and MNIST (LeCun, 1998) for digit recognition, GTSRB (Stallkamp et al., 2011) for traffic-sign recognition, and DTD (Cimpoi et al., 2014) for texture classification. These tasks cover diverse visual domains and therefore provide a broad testbed for evaluating whether DP task models remain mergeable across heterogeneous classification problems. For each dataset, we independently fine-tune a private task model from the same pretrained vision backbone under sample-level differential privacy. The released DP task models are then merged into a single model without accessing the original training data. We report classification accuracy on each task and use the average accuracy across all eight tasks as the main vision metric.

Language tasks. For language experiments, we use eight tasks from the GLUE benchmark (Wang et al., 2018): CoLA for linguistic acceptability, SST-2 for sentiment classification, MRPC and QQP for paraphrase detection, STS-B for semantic textual similarity, MNLI and RTE for natural language inference, and QNLI for question-answering natural language inference. For each task, we fine-tune a separate private language model under sample-level differential privacy and merge the released DP task models without accessing the original GLUE training data. Following the standard GLUE protocol, we report Matthew's correlation for CoLA, Spearman correlation for STS-B, F1/accuracy for MRPC and QQP, and accuracy for SST-2, MNLI, QNLI, and RTE. All task scores are converted to a 0–100 scale, and the average score over the eight tasks is used as the main language metric.

B.2 MODELS

Vision models. For vision experiments, we use CLIP-pretrained Vision Transformer backbones (Radford et al., 2021), including ViT-B/32, ViT-B/16, and ViT-L/14. ViT-B/32 and ViT-B/16 share the same base-size Transformer architecture with 12 layers and hidden dimension 768, but use different patch sizes of 32×32 and 16×16, respectively. ViT-L/14 is a larger backbone with 24 layers, hidden dimension 1024, and 14×14 image patches. For each backbone, all task-specific models are initialized from the same public CLIP checkpoint and independently fine-tuned on each private vision dataset under sample-level DP. During merging, we merge the shared visual encoder parameters of the released DP task models. Since the eight vision datasets have different label spaces, task-specific classification heads are kept separate for evaluation. Thus, the merged vision model consists of one shared visual encoder and the corresponding task head for each dataset.

Language models. For language experiments, we use RoBERTa-Base and RoBERTa-Large (Liu et al., 2019). RoBERTa-Base has 12 Transformer layers with hidden dimension 768, while RoBERTa-Large has 24 Transformer layers with hidden dimension 1024. For each GLUE task, all task-specific models are initialized from the same public RoBERTa checkpoint and independently fine-tuned under sample-level DP. During merging, we merge the shared RoBERTa encoder parameters of the released DP task models. Because GLUE tasks have different label spaces and output formats, task-specific prediction heads are kept separate for evaluation. The merged language model, therefore, uses one shared encoder together with the corresponding task head for each evaluation task.

B.3 IMPLEMENTATION DETAILS

This section provides additional implementation details that are omitted from the main text due to space limits. All experiments are conducted on NVIDIA RTX 5090 GPUs.

Training schedules. All task-specific models are fine-tuned using sample-level differential privacy with gradient clipping $C = 0.2$, batch size 16, learning rate 1×10^{-5}, weight decay 1×10^{-2}, and 10 training epochs/steps. We use AdamW as the optimizer with parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. For clarity, different task types follow the same DP fine-tuning defaults; any deviations from this default schedule are explicitly noted in the corresponding experiment description in the main text.

Merging hyperparameters. For all experiments, we use fixed merging hyperparameters for each merging method to ensure fair comparison: For Task Arithmetic, the scaling coefficient is 0.3. For TIES-Merging, pruning density is 0.2. For PCB Merging and WUDI-Merging, we use default configurations from the original papers unless otherwise specified.

The same merging hyperparameters are used for both the standard DP fine-tuning baselines and DP-Merging. For DP-Merging specifically, the perturbation radius $\rho = 0.05$ and the alignment coefficient $\lambda = 0.05$ are used consistently across all tasks.

Table 11: Multi-task accuracy (%) on the 8-task vision benchmark with ViT-B/16, $\varepsilon = 4$.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg Acc
DP + Individual	$59.8_{\pm 0.40}$	$68.2_{\pm 0.52}$	$88.1_{\pm 0.45}$	$86.9_{\pm 0.30}$	$86.2_{\pm 0.26}$	$89.4_{\pm 0.33}$	$90.7_{\pm 0.19}$	$70.6_{\pm 0.57}$	80.0
DP + Weight Averaging	$45.3_{\pm 0.61}$	$43.4_{\pm 0.58}$	$51.4_{\pm 0.50}$	$51.7_{\pm 0.47}$	$44.2_{\pm 0.29}$	$32.8_{\pm 0.39}$	$67.5_{\pm 0.32}$	$30.1_{\pm 0.96}$	45.8
+ Ours	$49.4_{\pm 0.55}$	$48.4_{\pm 0.54}$	$57.1_{\pm 0.46}$	$56.2_{\pm 0.49}$	$49.4_{\pm 0.32}$	$39.4_{\pm 0.33}$	$75.1_{\pm 0.26}$	$38.6_{\pm 0.89}$	$51.7_{+5.9}$
DP + RegMean	$45.3_{\pm 0.61}$	$43.8_{\pm 0.76}$	$55.9_{\pm 0.48}$	$58.6_{\pm 0.32}$	$58.2_{\pm 0.41}$	$47.4_{\pm 0.55}$	$73.7_{\pm 0.27}$	$36.0_{\pm 0.72}$	52.4
+ Ours	$48.3_{\pm 0.66}$	$47.8_{\pm 0.51}$	$59.9_{\pm 0.29}$	$64.6_{\pm 0.45}$	$63.6_{\pm 0.37}$	$49.4_{\pm 0.30}$	$76.7_{\pm 0.30}$	$44.0_{\pm 0.66}$	$56.8_{+4.4}$
DP + Task Arithmetic	$45.2_{\pm 0.78}$	$44.9_{\pm 0.63}$	$56.7_{\pm 0.45}$	$68.4_{\pm 0.51}$	$70.5_{\pm 0.44}$	$49.7_{\pm 0.25}$	$77.3_{\pm 0.33}$	$40.4_{\pm 0.84}$	56.6
+ Ours	$49.1_{\pm 0.54}$	$48.5_{\pm 0.69}$	$59.4_{\pm 0.40}$	$72.6_{\pm 0.34}$	$78.1_{\pm 0.22}$	$56.5_{\pm 0.59}$	$79.1_{\pm 0.38}$	$46.3_{\pm 0.75}$	$61.2_{+4.6}$
DP + TIES-Merging	$39.8_{\pm 0.57}$	$38.7_{\pm 0.71}$	$40.7_{\pm 0.53}$	$59.5_{\pm 0.27}$	$66.2_{\pm 0.35}$	$52.1_{\pm 0.46}$	$78.3_{\pm 0.31}$	$44.2_{\pm 0.76}$	52.4
+ Ours	$45.8_{\pm 0.62}$	$46.6_{\pm 0.47}$	$45.7_{\pm 0.35}$	$63.8_{\pm 0.41}$	$69.2_{\pm 0.39}$	$55.1_{\pm 0.60}$	$81.3_{\pm 0.24}$	$46.2_{\pm 0.80}$	$56.7_{+4.3}$
DP + PCB Merging	$45.7_{\pm 0.63}$	$43.9_{\pm 0.70}$	$46.5_{\pm 0.38}$	$57.2_{\pm 0.29}$	$61.7_{\pm 0.46}$	$50.3_{\pm 0.38}$	$77.0_{\pm 0.27}$	$37.1_{\pm 0.88}$	52.4
+ Ours	$49.4_{\pm 0.52}$	$49.0_{\pm 0.58}$	$51.2_{\pm 0.28}$	$61.3_{\pm 0.33}$	$65.1_{\pm 0.37}$	$55.8_{\pm 0.31}$	$81.1_{\pm 0.30}$	$38.6_{\pm 0.54}$	$56.4_{+4.0}$
DP + WUDI-Merging	$46.7_{\pm 0.67}$	$45.5_{\pm 0.62}$	$58.6_{\pm 0.39}$	$59.3_{\pm 0.32}$	$66.4_{\pm 0.44}$	$57.1_{\pm 0.36}$	$78.2_{\pm 0.41}$	$39.1_{\pm 0.69}$	56.4
+ Ours	$49.6_{\pm 0.54}$	$48.1_{\pm 0.57}$	$61.3_{\pm 0.31}$	$63.8_{\pm 0.35}$	$71.5_{\pm 0.38}$	$61.3_{\pm 0.29}$	$81.6_{\pm 0.34}$	$45.7_{\pm 0.61}$	$60.4_{+4.0}$

B.4 MORE RESULTS

Results on ViT-B/16. Table 11 reports the multi-task accuracy (%) of ViT-B/16 on the eight-task vision benchmark with $\varepsilon = 4$. DP + Individual achieves an average accuracy of 80.0%, whereas directly merging the privately fine-tuned models with Weight Averaging substantially reduces the average accuracy to 45.8%. Incorporating DP-Merging consistently improves all evaluated merging methods, increasing the average accuracy from 45.8% to 51.7% for Weight Averaging, from 52.4% to 56.8% for RegMean, from 56.6% to 61.2% for

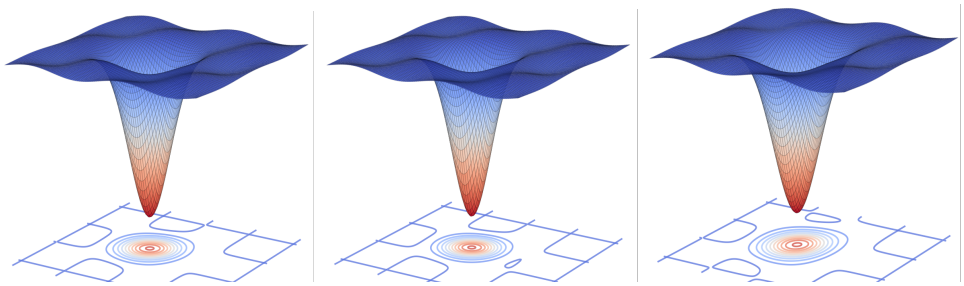
Task Arithmetic, from 52.4% to 56.7% for TIES-Merging, from 52.4% to 56.4% for PCB Merging, and from 56.4% to 60.4% for WUDI-Merging. Among the merged models, DP-Merging + Task Arithmetic achieves the highest average accuracy of 61.2%. These consistent gains across diverse merging strategies demonstrate that DP-Merging improves the mergeability of differentially private task models.

Comparison under private and non-private fine-tuning. We compare the effects of sharpness-aware fine-tuning and reference anchoring under non-private and private training. As shown in Table 12, SAM improves non-private merged accuracy by 0.5–1.0 percentage points over AdamW, while adding reference anchoring yields a further 0.1–0.3 percentage-point gain. Under $\varepsilon = 4$, DP-Merging improves over DP-AdamW by 3.7–4.7 percentage points across the four merging operators. These results show that the benefits of the combined training strategy extend to both settings, with larger absolute gains under DP in the evaluated experiments.

Table 12: Merged accuracy (%) under non-private and private fine-tuning on the 8-task vision benchmark with ViT-B/32. Anchor denotes reference anchoring.

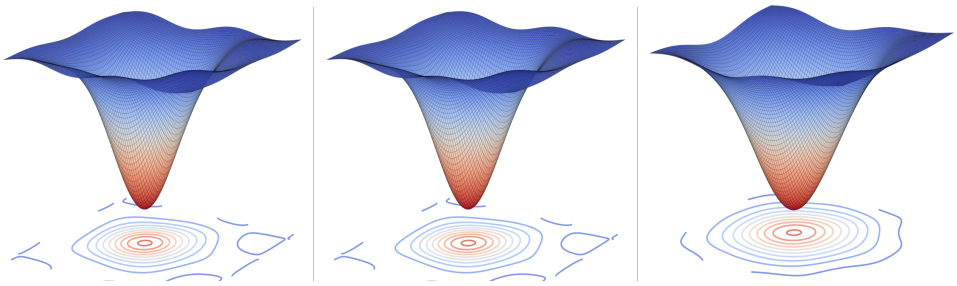
Training	Privacy	Weight Averaging	Task Arithmetic	TIES-Merging	PCB-Merging
AdamW	Non-private	64.8	67.9	72.2	68.1
SAM	Non-private	65.5	68.4	73.2	68.9
SAM + Anchor	Non-private	65.6	68.6	73.3	69.2
DP(AdamW)	$\varepsilon = 4$	46.0	54.2	57.5	54.5
DP-Merging (Ours)	$\varepsilon = 4$	49.7	58.1	62.1	59.2

Loss Landscape Visualization. To illustrate the effect of different DP fine-tuning and merging strategies on the geometry of merged models, we visualize the loss landscapes of the final merged ViT models under privacy budget $\varepsilon = 4$. Figure 4 shows the landscapes of models merged using Task Arithmetic after Naive DP fine-tuning, while Figure 5 shows the landscapes of models merged using Task Arithmetic after DP-Merging. All visualized models are final merged models for three ViT variants (ViT-B/32, ViT-B/16, ViT-L/14). Comparing the two sets of landscapes, we observe that DP-Merging produces wider and smoother low-loss regions, indicating improved stability and geometric compatibility of the merged models.



(a) ViT-B/32 (b) ViT-B/16 (c) ViT-L/14

Figure 4: Loss landscapes of ViT models that were first Naive DP fine-tuned under $\varepsilon = 4$ and then merged using Task Arithmetic (ViT-B/32, ViT-B/16, ViT-L/14).



(a) ViT-B/32 (b) ViT-B/16 (c) ViT-L/14

Figure 5: Loss landscapes of the merged ViT models (ViT-B/32, ViT-B/16, and ViT-L/14) under $\varepsilon = 4$, obtained using DP-Merging with Task Arithmetic.

C Implementation of Theoretical Analysis

C.1 Privacy Guarantee of DP-Merging

We prove **Theorem 1** under sample-level add/remove adjacency. Neighboring datasets differ by the addition or removal of one training sample. The task assignment is fixed, so a neighboring change affects only one task dataset. Each task is trained independently using its own data and the public initialization w_0. Fix a task t and suppress its training-parameter indices. The iteration count K, sampling rate q, clipping threshold C, noise multiplier σ, normalization constant $b > 0$, and other training hyperparameters are fixed independently of the private data. Each per-sample loss depends only on that sample and the model parameters. Apart from the two private gradient queries, all training operations use only their outputs and public information.

Private gradient mechanism. Define $\bar{g}(w;z) = \frac{\nabla_w \ell(w;z)}{\max\{1, \|\nabla_w \ell(w;z)\|_2 / C\}}$, $\|\bar{g}(w;z)\|_2 \le C$. At iteration k, each sample is independently included in B_k with probability q. The two queries are

$$\tilde{g}_k = \frac{1}{b}\left(\sum_{z \in B_k} \bar{g}(w_k; z) + \xi_k^{(1)}\right), \tag{13}$$

$$\tilde{h}_k = \frac{1}{b}\left(\sum_{z \in B_k} \bar{g}(w_k + \epsilon_k; z) + \xi_k^{(2)}\right), \tag{14}$$

where $\xi_k^{(1)}, \xi_k^{(2)} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2 C^2 I)$, $\epsilon_k = \rho_t \frac{\tilde{g}_k}{\|\tilde{g}_k\|_2}$. We set $\epsilon_k = 0$ when $\|\tilde{g}_k\|_2 = 0$. Both queries reuse B_k. Sampling and Gaussian noise are fresh across iterations, and the gradient sum is zero for an empty minibatch. The normalization constant b is public and fixed across neighboring inputs.

Dominating pairs. For probability distributions P and Q, define the hockey-stick divergence

$$H_\gamma(P\|Q) = \int (dP - \gamma\, dQ)_+, \qquad \gamma > 0.$$

We write $(P', Q') \preceq (P, Q)$ if

$$H_\gamma(P'\|Q') \le H_\gamma(P\|Q) \quad \text{for all } \gamma > 0.$$

For $\alpha > 1$ and $P \ll Q$, the identity

$$\exp((\alpha-1)D_\alpha(P\|Q)) = \alpha(\alpha-1)\int_0^\infty \gamma^{\alpha-2} H_\gamma(P\|Q)\, d\gamma$$

shows that domination also bounds Rényi divergence. Consequently, $(P', Q') \preceq (P, Q)$ implies $D_\alpha(P'\|Q') \le D_\alpha(P\|Q)$ for the distributions considered below.

We use two properties of dominating pairs: adaptive composition yields a product dominating pair, and Poisson subsampling transforms a dominating pair (P, Q) for the larger-to-smaller neighboring direction into $((1-q)Q + qP, Q)$ (Zhu et al., 2022). The adaptive-composition property requires the conditional bound for each query to hold uniformly over previous outputs.

Step 1: conditional sensitivity. Condition on all private outputs preceding iteration k. The current parameters w_k and optimizer state are then fixed. First regard the sampled minibatch as the input of an unsampled two-query mechanism.

For any fixed parameter vector v, let $f_v(B) = \sum_{z \in B} \bar{g}(v; z)$. For add/remove neighbors $B = B' \cup \{z\}$, $\|f_v(B) - f_v(B')\|_2 = \|\bar{g}(v; z)\|_2 \le C$. Thus, the first query is a Gaussian mechanism with noise multiplier σ. Its division by b scales both sensitivity and noise equally and does not change this bound.

For the second query, fix any value y of the first private output. The perturbation $\epsilon(y)$ and query location $w_k + \epsilon(y)$ are fixed under both neighboring inputs. Consequently, the second query also has sensitivity at most C before normalization and uses independent Gaussian noise with covariance $\sigma^2 C^2 I$. This bound holds for every y.

The conditioning uses the same output value y when comparing neighboring inputs. It does not assume that two executions on neighboring datasets produce identical first outputs.

Step 2: adaptive Gaussian composition. Each conditional query is dominated by the Gaussian pair $P_0 = \mathcal{N}(1, \sigma^2), Q_0 = \mathcal{N}(0, \sigma^2)$. Adaptive composition therefore bounds the joint two-query transcript by $(P_0 \otimes P_0,\ Q_0 \otimes Q_0)$.

To simplify this pair, apply the invertible transformation

$$U = \frac{X_1 + X_2}{2}, V = \frac{X_1 - X_2}{2}.$$

Under $P_0 \otimes P_0$, the independent coordinates U and V have distributions $U \sim \mathcal{N}(1, \sigma^2/2), V \sim \mathcal{N}(0, \sigma^2/2)$. Under $Q_0 \otimes Q_0$, they instead satisfy $U \sim \mathcal{N}(0, \sigma^2/2), V \sim \mathcal{N}(0, \sigma^2/2)$. The common independent V component does not affect the likelihood ratio. The product pair is therefore equivalent for privacy analysis to $P_s = \mathcal{N}(1, s^2), Q_s = \mathcal{N}(0, s^2), s = \frac{\sigma}{\sqrt{2}}$.

Hence, the unsampled adaptive transcript is dominated by a Gaussian pair with effective noise multiplier $\sigma/\sqrt{2}$. This does not assert that the actual adaptive transcript is jointly Gaussian. As a consistency check, its RDP bound is

$$D_\alpha(P_s \| Q_s) = \frac{\alpha}{2s^2} = \frac{\alpha}{\sigma^2},$$

matching the sum of the two individual Gaussian RDP bounds.

Step 3: Poisson amplification. The two queries share one Poisson minibatch. We therefore apply subsampling amplification once to their joint mechanism, after the adaptive composition above.

For $q \in (0, 1]$ and $s > 0$, define $P_{q,s} = (1 - q)Q_s + qP_s$. Let $\mathcal{M}_k$ denote the complete sampled two-query mechanism, conditional on a fixed preceding private transcript. For every neighboring pair $\mathcal{D} = \mathcal{D}' \cup \{z\}$, Poisson amplification gives

$$\left(\mathcal{M}_k(\mathcal{D}), \mathcal{M}_k(\mathcal{D}')\right) \preceq (P_{q,s}, Q_s).$$

Here, $\mathcal{M}_k(\mathcal{D})$ denotes the distribution of the two-query transcript. Using the identity

$$H_\gamma(Q \| P) = 1 - \gamma + \gamma H_{1/\gamma}(P \| Q),$$

we also obtain

$$\left(\mathcal{M}_k(\mathcal{D}'), \mathcal{M}_k(\mathcal{D})\right) \preceq (Q_s, P_{q,s}).$$

Thus, both directions of add/remove adjacency are covered by

$$\varepsilon_{\mathrm{SGM}}(\alpha; q, s) = \max\left\{D_\alpha(P_{q,s} \| Q_s),\ D_\alpha(Q_s \| P_{q,s})\right\}, \qquad \alpha > 1. \tag{15}$$

One complete DP-Merging iteration consequently has RDP cost at most

$$\varepsilon_{\mathrm{SGM}}\left(\alpha; q, \frac{\sigma}{\sqrt{2}}\right).$$

For an explicit expression, let $X \sim Q_s$ and define

$$r_{q,s}(X) = 1 - q + q \exp\left(\frac{2X - 1}{2s^2}\right).$$

Since $r_{q,s} = dP_{q,s}/dQ_s$,

$$D_\alpha(P_{q,s} \| Q_s) = \frac{1}{\alpha - 1} \log \mathbb{E}_{X \sim Q_s}\left[r_{q,s}(X)^\alpha\right], \tag{16}$$

$$D_\alpha(Q_s \| P_{q,s}) = \frac{1}{\alpha - 1} \log \mathbb{E}_{X \sim Q_s}\left[r_{q,s}(X)^{1-\alpha}\right]. \tag{17}$$

These expressions are valid for every $\alpha > 1$.

For integer $\alpha \geq 2$, expanding the forward moment yields

$$D_\alpha(P_{q,s} \| Q_s) = \frac{1}{\alpha - 1} \log\left[\sum_{j=0}^{\alpha} \binom{\alpha}{j} (1 - q)^{\alpha - j} q^j \exp\left(\frac{j(j-1)}{2s^2}\right)\right].$$

With $s = \sigma/\sqrt{2}$, the exponent becomes $j(j-1)/\sigma^2$. This factor results from adaptive Gaussian composition, rather than a direct sensitivity calculation for a nonadaptive concatenated query.

Step 4: composition over iterations. The one-iteration bound holds uniformly over all previous private transcripts. Adaptive RDP composition (Mironov, 2017) therefore gives

$$\varepsilon_t^{\mathrm{RDP}}(\alpha) \le K\varepsilon_{\mathrm{SGM}}\left(\alpha; q, \frac{\sigma}{\sqrt{2}}\right).$$

Although K, q, σ suppress the task index, this bound is evaluated using the configuration of task t.

The final task model is a function of the private transcript and public information. It satisfies the same RDP bound by post-processing. For any fixed $\delta \in (0,1)$, the standard RDP-to-DP conversion gives

$$\varepsilon_t = \min_{\alpha\in\mathcal{O}} \left\{ K\varepsilon_{\mathrm{SGM}}\left(\alpha; q, \frac{\sigma}{\sqrt{2}}\right) + \frac{\log(1/\delta)}{\alpha - 1} \right\},$$

where $\mathcal{O} \subset (1,\infty)$ is a nonempty finite set. Thus, w_t^{DP} is (ε_t, δ)-DP.

Post-processing within training. Constructing ϵ_k from the first private gradient incurs no additional privacy loss. The reference-anchoring term $2\lambda(w_k - w_0)$ depends only on the current private state and the public initialization. Restoring the unperturbed parameters, applying the optimizer update, and updating optimizer state likewise incur no additional privacy loss when they use only private query outputs and public information.

In contrast, evaluating the second gradient accesses the private minibatch again. It is not post-processing and is explicitly included in the joint two-query bound.

Parallel composition and merging. A neighboring change affects only one task dataset. Since tasks are trained independently, parallel composition implies that the collection $\{w_t^{\mathrm{DP}}\}_{t=1}^T$ is

$$\left(\max_{t\in[T]} \varepsilon_t, \delta\right)\text{-DP}.$$

Any merged model computed solely from this collection and public information is post-processing. Consequently, the joint release of all task models and their merged model satisfies the same guarantee. This proves Theorem 1.

Accounting in implementation. The actual noise multiplier in each gradient query is σ. For privacy accounting, each complete SAM iteration is represented by one Poisson-subsampled Gaussian dominating mechanism with sampling rate q and effective noise multiplier $\sigma/\sqrt{2}$. The composition count is the number K of complete SAM iterations. This accounting representation neither changes the training noise nor introduces an additional query.

A numerical accountant may use any valid upper bound on Eq. equation 15, provided that both neighboring directions are covered. The two queries must not be accounted for as independently subsampled mechanisms, since they reuse the same minibatch. Noise multipliers are calibrated separately for each training configuration to meet the common target privacy budget.

C.2 Local Sharpness and Merge-Gap Bound

We prove Theorem 2. The analysis is conditional on the returned task models and does not require a particular optimizer. Throughout, vector norms are Euclidean norms and matrix norms are operator norms.

Notation. Let $w_t := w_t^{\mathrm{DP}}$ and $w_{\mathrm{merge}} := \frac{1}{T}\sum_{t=1}^T w_t$. Define $d_t := w_{\mathrm{merge}} - w_t$, $r_t := \|d_t\|_2$, $g_t := \nabla\mathcal{L}_t(w_t)$, and $H_t := \nabla^2\mathcal{L}_t(w_t)$. For $\rho_t > 0$, let

$$S_t := \max_{\|e\|_2\le\rho_t} \left[\mathcal{L}_t(w_t + e) - \mathcal{L}_t(w_t)\right]. \tag{18}$$

The maximum exists by continuity on the compact perturbation ball, and $S_t \ge 0$ because $e = 0$ is feasible.

Taylor remainder. Under the assumptions of Theorem 2, $\mathcal{L}_t$ has an M_t-Lipschitz Hessian on a convex neighborhood containing the perturbation ball and w_{merge}. For a displacement v whose segment lies in this neighborhood,

$$\mathcal{L}_t(w_t + v) = \mathcal{L}_t(w_t) + g_t^\top v + \frac{1}{2}v^\top H_t v + R_t(v), \tag{19}$$

where

$$R_t(v) = \int_0^1 (1-s)v^\top \left[\nabla^2\mathcal{L}_t(w_t + sv) - H_t\right] v\,\mathrm{d}s. \tag{20}$$

Hessian Lipschitzness implies

$$|R_t(v)| \le M_t\|v\|_2^3 \int_0^1 s(1-s)\,\mathrm{d}s = \frac{M_t}{6}\|v\|_2^3. \tag{21}$$

Lemma 1 (Curvature bound from local sharpness)**.** *Under the assumptions of Theorem 2,*

$$\lambda_{\max}(H_t) \le \frac{2S_t}{\rho_t^2} + \frac{M_t\rho_t}{3}. \tag{22}$$

Proof. Let v_t be a unit eigenvector of H_t corresponding to $\lambda_{\max}(H_t)$. Since both $\rho_t v_t$ and $-\rho_t v_t$ are feasible perturbations,

$$\frac{\mathcal{L}_t(w_t+\rho_t v_t)+\mathcal{L}_t(w_t-\rho_t v_t)}{2} - \mathcal{L}_t(w_t) \le S_t. \tag{23}$$

Applying Eq. equation 19 to these two perturbations cancels the linear terms:

$$\begin{aligned}
&\frac{\mathcal{L}_t(w_t+\rho_t v_t)+\mathcal{L}_t(w_t-\rho_t v_t)}{2} - \mathcal{L}_t(w_t) \\
&= \frac{\rho_t^2}{2}\lambda_{\max}(H_t) + \frac{R_t(\rho_t v_t)+R_t(-\rho_t v_t)}{2} \\
&\ge \frac{\rho_t^2}{2}\lambda_{\max}(H_t) - \frac{M_t\rho_t^3}{6}.
\end{aligned} \tag{24}$$

Combining Eqs. equation 23 and equation 24 and rearranging proves the result. □

The symmetric perturbations cancel the first-order contribution, so the lemma does not require w_t to be stationary. It concerns the worst-case neighborhood increment S_t, rather than a single directional diagnostic.

Proof of Theorem 2. Applying Eq. equation 19 with $v = d_t$ gives

$$\mathrm{Gap}_t(w_{\mathrm{merge}}) = g_t^\top d_t + \frac{1}{2} d_t^\top H_t d_t + R_t(d_t).$$

By Cauchy–Schwarz, symmetry of H_t, and Eq. equation 21,

$$\mathrm{Gap}_t(w_{\mathrm{merge}}) \le \|g_t\|_2 r_t + \frac{\lambda_{\max}(H_t)}{2} r_t^2 + \frac{M_t}{6} r_t^3.$$

Substituting Lemma 1 yields

$$\mathrm{Gap}_t(w_{\mathrm{merge}}) \le \|g_t\|_2 r_t + \left(\frac{S_t}{\rho_t^2} + \frac{M_t\rho_t}{6}\right) r_t^2 + \frac{M_t}{6} r_t^3. \tag{25}$$

This proves Theorem 2. The result does not require $r_t \le \rho_t$; the Hessian regularity assumption covers the entire merging segment. □

C.3 Reference Drift and Merged-Loss Bound

Reference distances and task dispersion. Let $u_t := w_t - w_0$, $\bar{u} := \frac{1}{T}\sum_{t=1}^T u_t$, $D := \frac{1}{T}\sum_{t=1}^T \|u_t\|_2^2$, and $V := \frac{1}{T}\sum_{t=1}^T r_t^2$. Under uniform averaging, $d_t = \bar{u} - u_t$. Therefore,

$$\begin{aligned}
V &= \frac{1}{T}\sum_{t=1}^T \|u_t - \bar{u}\|_2^2 \\
&= \frac{1}{T}\sum_{t=1}^T \|u_t\|_2^2 - \|\bar{u}\|_2^2 = D - \|\bar{u}\|_2^2 \le D.
\end{aligned} \tag{26}$$

Thus, squared reference distances bound average squared merging displacement. They need not equal it: identical task updates have zero merging displacement even when their reference distances are large.

Proof of Corollary 1. Let $G := (\frac{1}{T}\sum_t \|g_t\|_2^2)^{1/2}$, $A := \max_t(S_t/\rho_t^2 + M_t\rho_t/6)$, and $M := \max_t M_t$. Cauchy–Schwarz gives

$$\frac{1}{T}\sum_{t=1}^T \|g_t\|_2 r_t \le G\sqrt{V} \le G\sqrt{D}. \tag{27}$$

Since $A \ge 0$,

$$\frac{1}{T}\sum_{t=1}^T \left(\frac{S_t}{\rho_t^2} + \frac{M_t\rho_t}{6}\right) r_t^2 \le AV \le AD. \tag{28}$$

Moreover, $\max_t r_t \le (\sum_t r_t^2)^{1/2} = \sqrt{TV}$, so

$$\frac{1}{T}\sum_{t=1}^{T} r_t^3 \le \Big(\max_t r_t\Big) V \le \sqrt{T}\, V^{3/2} \le \sqrt{T}\, D^{3/2}. \tag{29}$$

Averaging Eq. equation 25 and using these inequalities yields

$$G_{\mathrm{merge}} \le G\sqrt{D} + AD + \frac{M\sqrt{T}}{6} D^{3/2}.$$

Finally,

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(w_{\mathrm{merge}}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(w_t) + G_{\mathrm{merge}},$$

which proves Corollary 1. □

Illustration with quadratic task losses. Consider $\mathcal{L}_t(w) = c_t + \frac{1}{2}(w - a_t)^\top H_t (w - a_t)$ with $H_t \succeq 0$. If each task model is at its optimum, $w_t = a_t$, then

$$\mathrm{Gap}_t(w_{\mathrm{merge}}) = \frac{1}{2} d_t^\top H_t d_t, \qquad S_t = \frac{\rho_t^2}{2}\lambda_{\max}(H_t). \tag{30}$$

Hence $\mathrm{Gap}_t(w_{\mathrm{merge}}) \le (S_t/\rho_t^2) r_t^2$, with equality when d_t lies in a leading eigenspace of H_t. This is the zero-remainder case of Theorem 2.

For $H_t = hI$ with $h > 0$, the gap equals $hr_t^2/2$. Thus, small curvature alone does not bound the gap without a constraint on displacement. Likewise, displacement alone does not determine the gap when curvatures differ.

Anchoring and task adaptation. For the same quadratic loss, consider the anchored task objective $\mathcal{L}_t(w) + \lambda\|w - w_0\|_2^2$ with $\lambda > 0$. Its unique minimizer satisfies

$$H_t(w_t^\lambda - a_t) + 2\lambda(w_t^\lambda - w_0) = 0,$$

and therefore

$$w_t^\lambda - w_0 = (H_t + 2\lambda I)^{-1} H_t (a_t - w_0). \tag{31}$$

Write $H_t = Q_t \,\mathrm{diag}(h_{t,i}) Q_t^\top$ and $z_t := Q_t^\top (a_t - w_0)$. Then

$$\|w_t^\lambda - w_0\|_2^2 = \sum_i \left(\frac{h_{t,i}}{h_{t,i} + 2\lambda}\right)^2 z_{t,i}^2, \tag{32}$$

$$\mathcal{L}_t(w_t^\lambda) - c_t = \frac{1}{2}\sum_i h_{t,i} \left(\frac{2\lambda}{h_{t,i} + 2\lambda}\right)^2 z_{t,i}^2. \tag{33}$$

For this fixed quadratic problem, reference distance is nonincreasing in λ, whereas excess task loss is nondecreasing. As $\lambda \to \infty$, all solutions approach w_0 and their merging displacements vanish, but their task losses approach $\mathcal{L}_t(w_0)$.

This example isolates the trade-off introduced by anchoring; it is not an analysis of the full private robust objective or its optimization. It explains why task losses must remain explicit in the merged-utility bound. Since the Hessians are fixed, the example makes no claim that sharpness-aware training changes their curvature.

C.4 Private Perturbation Approximation

We relate the exact neighborhood quantity S_t to the loss increment at a perturbation constructed from a clipped noisy gradient.

Lemma 2 (Directional approximation error)**.** *Fix w and $\rho > 0$. Assume that $\mathcal{L}$ has an L_{sm}-Lipschitz gradient on $\{w + e : \|e\|_2 \le \rho\}$. Let $g := \nabla\mathcal{L}(w)$ and let $\tilde{g}$ be any gradient estimate. Define*

$$\hat{e} := \begin{cases} \rho\tilde{g}/\|\tilde{g}\|_2, & \tilde{g} \neq 0, \\ 0, & \tilde{g} = 0, \end{cases}$$

and

$$S := \max_{\|e\|_2 \le \rho} [\mathcal{L}(w + e) - \mathcal{L}(w)], \qquad \widehat{S} := \mathcal{L}(w + \hat{e}) - \mathcal{L}(w).$$

Then

$$0 \le S - \widehat{S} \le L_{\mathrm{sm}}\rho^2 + 2\rho\|\tilde{g} - g\|_2. \tag{34}$$

Proof. Let $e^\star$ attain the maximum defining S. The smoothness inequalities imply

$$\mathcal{L}(w+e^\star) \le \mathcal{L}(w) + g^\top e^\star + \frac{L_{\mathrm{sm}}}{2}\|e^\star\|_2^2$$

and

$$\mathcal{L}(w+\hat{e}) \ge \mathcal{L}(w) + g^\top \hat{e} - \frac{L_{\mathrm{sm}}}{2}\|\hat{e}\|_2^2.$$

Subtracting gives

$$S - \widehat{S} \le g^\top(e^\star - \hat{e}) + L_{\mathrm{sm}}\rho^2. \tag{35}$$

By construction, $\hat{e}$ maximizes $\tilde{g}^\top e$ over the perturbation ball, including when $\tilde{g} = 0$. Thus,

$$\begin{aligned} g^\top(e^\star - \hat{e}) &= (g-\tilde{g})^\top(e^\star - \hat{e}) + \tilde{g}^\top(e^\star - \hat{e}) \\ &\le \|g-\tilde{g}\|_2\|e^\star - \hat{e}\|_2 \\ &\le 2\rho\|g-\tilde{g}\|_2. \end{aligned}$$

Substitution into Eq. equation 35 proves the upper bound. The lower bound follows because $\hat{e}$ is feasible in the definition of S. □

Sources of gradient error. For $g_i := \nabla_w \ell(w; z_i)$, define

$$g_B := \frac{1}{b}\sum_{z_i \in B} g_i, \qquad g_B^C := \frac{1}{b}\sum_{z_i \in B} \frac{g_i}{\max\{1, \|g_i\|_2/C\}}.$$

The private estimator is $\tilde{g} = g_B^C + \xi/b$. The triangle inequality yields

$$\begin{aligned} \|\tilde{g} - \nabla\mathcal{L}(w)\|_2 &\le \|g_B - \nabla\mathcal{L}(w)\|_2 + \|g_B^C - g_B\|_2 + \frac{\|\xi\|_2}{b} \\ &\le \|g_B - \nabla\mathcal{L}(w)\|_2 + \frac{1}{b}\sum_{z_i \in B}(\|g_i\|_2 - C)_+ + \frac{\|\xi\|_2}{b}. \end{aligned} \tag{36}$$

These terms account for sampling and normalization error, clipping, and injected noise. The inequality holds for each realization and does not assume an unbiased estimator. For $\xi \sim \mathcal{N}(0, \sigma^2 C^2 I_d)$,

$$\mathbb{E}\|\xi\|_2 \le \sqrt{\mathbb{E}\|\xi\|_2^2} = \sigma C\sqrt{d}.$$

The approximation can therefore be loose when gradient-estimation error is large.

Corollary 2 (Sufficient condition for loss ascent)**.** *Under Lemma 2, let $E := \|\tilde{g} - g\|_2$. Then*

$$\mathcal{L}(w+\hat{e}) - \mathcal{L}(w) \ge \rho\|g\|_2 - 2\rho E - \frac{L_{\mathrm{sm}}\rho^2}{2}. \tag{37}$$

In particular, the loss strictly increases if $2E + L_{\mathrm{sm}}\rho/2 < \|g\|_2$.

Proof. Let e_g maximize $g^\top e$ over $\|e\|_2 \le \rho$. Then $g^\top e_g = \rho\|g\|_2$. Since $\hat{e}$ maximizes $\tilde{g}^\top e$ over the same ball,

$$\begin{aligned} \rho\|g\|_2 - g^\top\hat{e} &= (g-\tilde{g})^\top(e_g - \hat{e}) + \tilde{g}^\top(e_g - \hat{e}) \\ &\le \|g-\tilde{g}\|_2\|e_g - \hat{e}\|_2 \\ &\le 2\rho E. \end{aligned}$$

The smoothness lower bound gives

$$\mathcal{L}(w+\hat{e}) - \mathcal{L}(w) \ge g^\top\hat{e} - \frac{L_{\mathrm{sm}}}{2}\|\hat{e}\|_2^2.$$

Combining these inequalities proves Eq. equation 37. The stated condition makes its right-hand side strictly positive. □

This condition compares the gradient signal with estimation error and the finite-radius smoothness term. It is sufficient rather than necessary: its failure does not imply loss descent. The result concerns the perturbation step, not convergence of the full training procedure.

Connection to the merge-gap bound. At a returned model w_t, suppose a gradient estimate $\tilde{g}_t$ defines $\hat{e}_t$ as above. Let $\widehat{S}_t := \mathcal{L}_t(w_t + \hat{e}_t) - \mathcal{L}_t(w_t)$ and $E_t := \|\tilde{g}_t - \nabla\mathcal{L}_t(w_t)\|_2$. If the gradient is $L_{\mathrm{sm},t}$-Lipschitz on the perturbation ball, then

$$S_t \le \widehat{S}_t + L_{\mathrm{sm},t}\rho_t^2 + 2\rho_t E_t.$$

Substituting into Eq. equation 25 gives

$$\begin{aligned}\mathrm{Gap}_t(w_{\mathrm{merge}}) \le \|g_t\|_2 r_t \\ + \left(\frac{\widehat{S}_t}{\rho_t^2} + L_{\mathrm{sm},t} + \frac{2E_t}{\rho_t} + \frac{M_t\rho_t}{6}\right) r_t^2 + \frac{M_t}{6} r_t^3.\end{aligned} \tag{38}$$

This relation makes explicit the approximation terms required when replacing worst-case local sharpness with a single directional evaluation. It does not introduce an additional query into the training algorithm.

Scope of the analysis. The mergeability bounds concern returned models under uniform averaging and apply to AdamW outputs without assuming a gradient-descent recursion. The perturbation results do not establish that the second private gradient is an unbiased gradient of the exact robust loss; the algorithm reuses the minibatch that determines the perturbation.

The results relate individual utility, local loss sensitivity, and reference distances. They do not imply that reducing any one quantity alone improves merged performance. Large displacements, rapidly varying curvature, or substantial gradient-estimation error can make the bounds loose. The privacy guarantee is established separately in Appendix C.1.

D Limitations and Future Work

While our DP-Merging and multi-task strategies demonstrate consistent improvements across vision and NLP benchmarks, several limitations remain. First, the effectiveness of parameter merging depends on task-specific compatibility, and its performance on highly heterogeneous tasks or models with substantially different architectures is not fully explored. Second, the experiments are limited to standard benchmarks and mid- to large-scale models (ViT-L/14, RoBERTa-Large), leaving the applicability to larger models or more complex, real-world datasets untested. Third, while average scores demonstrate clear overall gains, individual task performance can vary, with some tasks showing only marginal improvements or slight declines, highlighting the need for more task-specific analysis.

Future work includes developing adaptive merging strategies that account for task similarity and parameter sensitivity, extending the approach to larger and more diverse model architectures, and exploring dynamic task balancing and uncertainty-aware weighting mechanisms during training to improve stability and consistency across tasks.