SCouT: Softly Coupled Fine-Tuning to Preserve Specialization while Improving Mergeability

Anonymous authors
Paper under double-blind review

Abstract

The pre-training–fine-tuning paradigm is widely used to adapt large-scale models to downstream tasks, but maintaining separate task-specific models incurs substantial storage and deployment overhead. Model merging addresses this by combining multiple specialists into a single multitask model. However, independently optimized specialists can be poorly compatible, leading to a substantial drop in merged-model performance. We propose **S**oftly **Cou**pled Fine-**T**uning (**SCouT**), which introduces merge-aware coupling during fine-tuning to improve mergeability while preserving task-specific performance. The coupling strength interpolates between independent fine-tuning and conventional multitask learning, recovering independent fine-tuning at zero coupling and approaching a shared multitask solution under strong coupling. Our analysis identifies a favorable intermediate regime in which weak coupling improves post-merge risk linearly in the coupling parameter while affecting specialist risk only quadratically, explaining why mergeability can improve much faster than specialization degrades. Experiments across model architectures and merging rules show that SCouT substantially improves post-merge performance while retaining strong specialists. These results suggest that weak cross-task coupling can steer fine-tuning toward more compatible solutions while still largely preserving their standalone specialization.[1]

1 Introduction

The pre-training–fine-tuning paradigm has become standard for adapting large-scale models to diverse downstream tasks (Bommasani et al., 2021; Radford et al., 2021). However, this paradigm typically produces a separate set of task-specific parameters for every downstream task, leading to increasing storage and deployment costs as the number of tasks grows (Houlsby et al., 2019). Multitask learning with hard parameter sharing (hard MTL) offers an alternative by jointly training a shared model across tasks. However, it typically requires access to data from all tasks, which may be impractical in distributed or privacy-sensitive settings, and shared parameterization can cause negative transfer when task objectives interfere, particularly under limited model capacity (Standley et al., 2020; Yu et al., 2020).

These limitations motivate *model merging*, which combines independently fine-tuned task-specific models into a single multitask model without additional joint training or access to the original task data. We refer to each task-specific model as a *specialist*, and to the model obtained by combining these specialists as the *merged model*. Simple methods such as weight averaging and Task Arithmetic can suffer from interference between task updates (Ilharco et al., 2023; Yadav et al., 2023). More advanced post-hoc methods mitigate this by reweighting parameters, resolving conflicts, or optimizing the merge itself (Matena & Raffel, 2022; Yadav et al., 2023; Cheng et al., 2025). However, they act only after fine-tuning, when the specialists are already fixed. If these specialists lie in poorly compatible regions of the loss landscape, post-hoc merging may still fail to find a low-loss multitask solution (Lu et al., 2026). We refer to the resulting degradation relative to the individual specialists as the *merging gap*.

[1] Code is available at: `https://github.com/SCouT-release/SCouT`.

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.

Recent work has therefore explored *merger-friendly fine-tuning*, modifying the training process to produce specialists that are easier to merge. Existing approaches improve mergeability through model linearization or training-time perturbations (Ortiz-Jimenez et al., 2023; Jin et al., 2025; Lee et al., 2025; Yang et al., 2026b). However, each specialist is still optimized largely in isolation, so independent training cannot explicitly react to incompatibilities that only become visible when multiple specialists are considered together (Wang et al., 2024; Marczak et al., 2025). In this work, we step back and ask a complementary question: *Can specialists improve their mergeability by exchanging only model parameters during fine-tuning without sacrificing specialization?*

We answer this question with Softly Coupled Fine-Tuning (SCouT), which introduces merge-aware coupling between task-specific models during training. Each specialist continues to optimize its own task loss, while the coupling softly coordinates their updates to improve compatibility for later merging. We motivate this coupling strategy through a local second-order analysis of the merging gap. The coupling strength then places SCouT in an intermediate regime between independent fine-tuning and hard MTL, with zero coupling recovering independent specialists and strong coupling driving them toward a shared solution.

SCouT bridges independent fine-tuning and hard MTL, but its value lies beyond simple interpolation. Our analysis shows that weak coupling can improve mergeability much faster than it degrades specialization, revealing a favorable intermediate regime in which task-specific strength and merge compatibility can coexist. In particular, our main contributions are:

- We propose SCouT, which coordinates task-specific models during fine-tuning through a merge-aware coupling term derived from a Taylor approximation of the merging gap. By exchanging model weights without sharing raw task data, SCouT is parallelizable and scalable. The resulting specialists remain individually strong while becoming more compatible for post-hoc merging.
- We theoretically characterize the specialization–mergeability trade-off and show that weak coupling can reduce post-merge risk linearly while increasing specialist risk only quadratically in the coupling strength. This asymmetric response establishes a favorable regime in which mergeability improves while the specialist advantage over hard MTL is largely preserved.
- We empirically verify the theoretical characterization of the specialization–mergeability trade-off. Moreover, we evaluate SCouT across model architectures and merging rules, and show that it consistently improves post-merge performance over independent fine-tuning and merger-friendly baselines, while retaining stronger specialist performance than hard MTL methods.

2 Related Work

2.1 Model Merging

We consider model merging of specialists trained on different downstream tasks and distinguish post-hoc merging from merger-friendly fine-tuning (Yang et al., 2026a).

Post-hoc merging. Post-hoc methods combine already fine-tuned specialists without modifying their training. Weight averaging and Task Arithmetic directly combine model parameters or task vectors (Wortsman et al., 2022; Ilharco et al., 2023), while more advanced methods use parameter importance (Matena & Raffel, 2022; Jin et al., 2023), resolve conflicting or redundant updates (Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025), or optimize the merge itself (Yang et al., 2024; Cheng et al., 2025; Yang et al., 2025). These post-hoc approaches are complementary to SCouT; we therefore evaluate SCouT together with several post-hoc merging rules to test whether its benefits persist.

Merger-friendly fine-tuning. A growing line of work modifies fine-tuning itself to produce more merge-compatible specialists. Fine-Tuning in the Tangent Space (FTTS) and subsequent methods improve weight disentanglement through model linearization or restricted adaptation (Ortiz-Jimenez et al., 2023; Jin et al., 2025; Tang et al., 2024; Liu et al., 2024). Sharpness-Aware Fine-Tuning (SAFT) encourages flatter task-specific solutions through sharpness-aware optimization (Lee et al., 2025), while MergOPT introduces merge-induced perturbations during training (Yang et al., 2026b). These methods primarily modify the optimization of each specialist to improve its eventual mergeability. In contrast, SCouT explicitly coordinates multiple specialists during fine-tuning through

their prescribed merged model, making cross-specialist interaction itself the mechanism for improving mergeability.

2.2 Multitask Learning

Multitask learning is commonly divided into hard and soft parameter sharing (Zhang & Yang, 2022). Hard MTL optimizes a shared parameterization across tasks, with methods such as loss balancing and gradient manipulation designed to mitigate negative transfer (Kendall et al., 2018; Yu et al., 2020; Liu et al., 2021; Navon et al., 2022). Soft parameter-sharing methods (Soft MTL) instead retain task-specific parameters while enabling controlled information transfer through regularization (Duong et al., 2015; Nassif et al., 2020; Argyriou et al., 2008) or learned sharing mechanisms (Misra et al., 2016; Sun et al., 2020; Wallingford et al., 2022; Wan & Vlaski, 2026). SCouT is closest to regularization-based soft parameter sharing, but differs in its goal: the coupling is defined with respect to a prescribed downstream merged model, rather than task-similarity metrics or consensus. Consequently, SCouT is designed to preserve task-specific specialists while explicitly improving their subsequent merge.

3 Problem Statement and Method

3.1 Preliminaries

Notation. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a neural network parameterized by $\theta \in \Theta \subseteq \mathbb{R}^P$, where P is the number of model parameters, and let θ_0 denote a pretrained initialization. We consider K downstream tasks indexed by $k \in [K] \triangleq \{1, \dots, K\}$. Task k is associated with a dataset $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{N_k}$, and empirical risk $J_k(\theta) \triangleq \frac{1}{N_k} \sum_{i=1}^{N_k} \ell_k(f_\theta(x_{k,i}), y_{k,i})$, where $\ell_k(\cdot, \cdot)$ denotes the task-specific loss.

We use θ_k to denote generic parameters of specialist k. The solutions obtained by independent fine-tuning (independent FT) and hard parameter-sharing multi-task learning (hard MTL) are

$$\textbf{Independent FT:} \quad \theta_k^o \in \arg\min_{\theta \in \Theta} J_k(\theta), \qquad \textbf{Hard MTL:} \quad \theta^{\mathrm{MTL}} \in \arg\min_{\theta \in \Theta} \sum_{k=1}^{K} J_k(\theta). \tag{1}$$

Here, θ_k^o denotes an independently fine-tuned specialist for task k, while θ^{MTL} denotes a single shared solution optimized jointly across all tasks. We further define the stacked specialist parameters $\vartheta \triangleq \mathrm{col}\{\theta_1, \dots, \theta_K\}$.

Task Arithmetic. Although many post-hoc merging methods exist, we use Task Arithmetic (TA) as the primary merging rule to motivate SCouT because its simple parameter-space construction makes the compatibility of specialists explicit (Ilharco et al., 2023). For specialist k, TA defines the task vector $\tau_k \triangleq \theta_k - \theta_0$. Given fixed merging coefficients α_k, the corresponding merged model is

$$\theta_{\mathrm{merge}} = \theta_0 + \sum_{k=1}^{K} \alpha_k \tau_k. \tag{2}$$

We demonstrate in the numerical results that the benefits of SCouT extend beyond the linear Task Arithmetic in (2), despite the fact that it is used to motivate its formulation.

3.2 Mitigating the Merging Gap

Independent fine-tuning optimizes each specialist in isolation and therefore provides no direct control over the performance of the merged model. Therefore, we first characterize the empirical-risk degradation induced by merging, and then use this characterization to derive a regularizing coupling term as a training-time coordinator.

Suppose independent fine-tuning yields the task-specific minimizers $\theta_1^o, \dots, \theta_K^o$. To quantify the incompatibility among the independently fine-tuned specialists, we define the *merging gap* as

$$\mathcal{G}_{\mathrm{merge}}(\vartheta^o) \triangleq \sum_{k=1}^{K} \left[J_k(\theta_{\mathrm{merge}}^o) - J_k(\theta_k^o) \right], \tag{3}$$

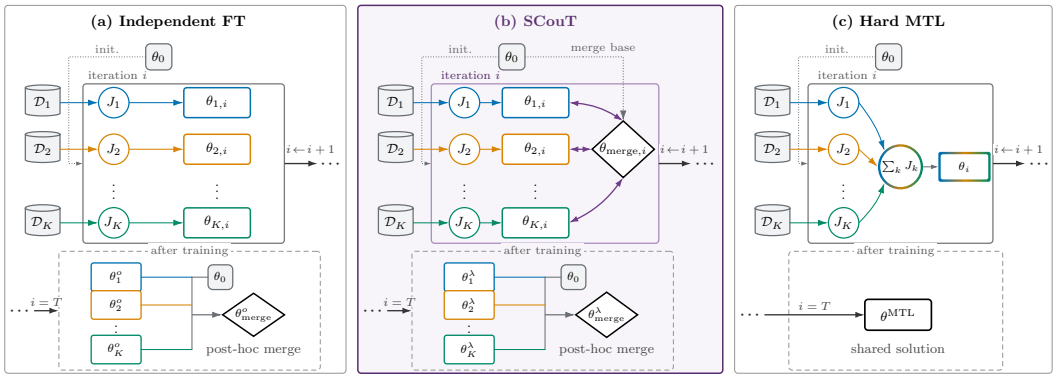

Figure 1: Optimization recursions of independent FT, SCouT, and hard MTL. Each upper block shows a generic training iteration i, repeated until the final step T. After training, the independent-FT and SCouT specialists are combined by post-hoc merging.

where $\vartheta^o = \mathrm{col}\{\theta_1^o, \ldots, \theta_K^o\}$. This quantity measures the aggregate empirical-risk degradation incurred when the individually optimized specialists are replaced by their merged model.

To relate this degradation to the local geometry of the task losses, define $\Delta_k^o \triangleq \theta_{\mathrm{merge}}^o - \theta_k^o$. A second-order Taylor expansion of J_k around θ_k^o gives

$$J_k(\theta_{\mathrm{merge}}^o) = J_k(\theta_k^o) + \nabla J_k(\theta_k^o)^\top \Delta_k^o + \frac{1}{2}(\Delta_k^o)^\top H_k \Delta_k^o + \mathcal{O}(\|\Delta_k^o\|^3), \tag{4}$$

where $H_k \triangleq \nabla^2 J_k(\theta_k^o)$. Since θ_k^o is a local minimizer, we have $\nabla J_k(\theta_k^o) = 0$ and $H_k \succeq 0$. Hence,

$$\mathcal{G}_{\mathrm{merge}}(\vartheta^o) = \frac{1}{2}\sum_{k=1}^K \|\theta_{\mathrm{merge}}^o - \theta_k^o\|_{H_k}^2 + \mathcal{O}\left(\sum_{k=1}^K \|\theta_{\mathrm{merge}}^o - \theta_k^o\|^3\right). \tag{5}$$

Thus, merging is particularly costly when the merged model departs from a specialist along directions of high task-specific curvature. Moreover, when $\theta_{\mathrm{merge}}^o$ lies sufficiently close to θ_k^o, the higher-order Taylor remainder is negligible relative to the quadratic term. For a practical curvature-free surrogate, suppose locally that $H_k \preceq \beta_k I \preceq \beta I$. Neglecting the higher-order remainder then yields

$$\mathcal{G}_{\mathrm{merge}}(\vartheta^o) \lesssim \frac{\beta}{2}\sum_{k=1}^K \|\theta_{\mathrm{merge}}^o - \theta_k^o\|^2. \tag{6}$$

The above analysis characterizes only the fixed independently fine-tuned solutions. To promote the same compatibility during training, we extend the specialist-to-merge discrepancy to the trainable parameters $\vartheta = \mathrm{col}\{\theta_1, \ldots, \theta_K\}$ and use it as a regularizer alongside the task empirical risks. This converts the post-hoc merging gap characterization into the softly coupled objective introduced next.

3.3 Softly Coupled Fine-tuning

Motivated by the merging gap characterization above, we define SCouT through the regularized objective

$$\textbf{SCouT:} \qquad \vartheta^\lambda \in \arg\min_{\vartheta} \sum_{k=1}^K J_k(\theta_k) + \frac{\lambda}{2}\sum_{k=1}^K \|\theta_{\mathrm{merge}} - \theta_k\|_{A_k}^2, \tag{7}$$

where $A_k \succeq 0$ is a fixed symmetric matrix characterizing the task-dependent coupling geometry, and $\lambda \geq 0$ controls the specialization–mergeability trade-off. The curvature-weighted form suggested by the merging gap analysis is recovered by choosing $A_k = H_k$, where Hessians are available. In practice, A_k may instead be approximated by the empirical Fisher information, while $A_k = I$ yields the computationally efficient Euclidean surrogate in (6).

The explicit optimization recursion and the complete training and post-hoc merging procedure of SCouT are provided in Appendix B.2.

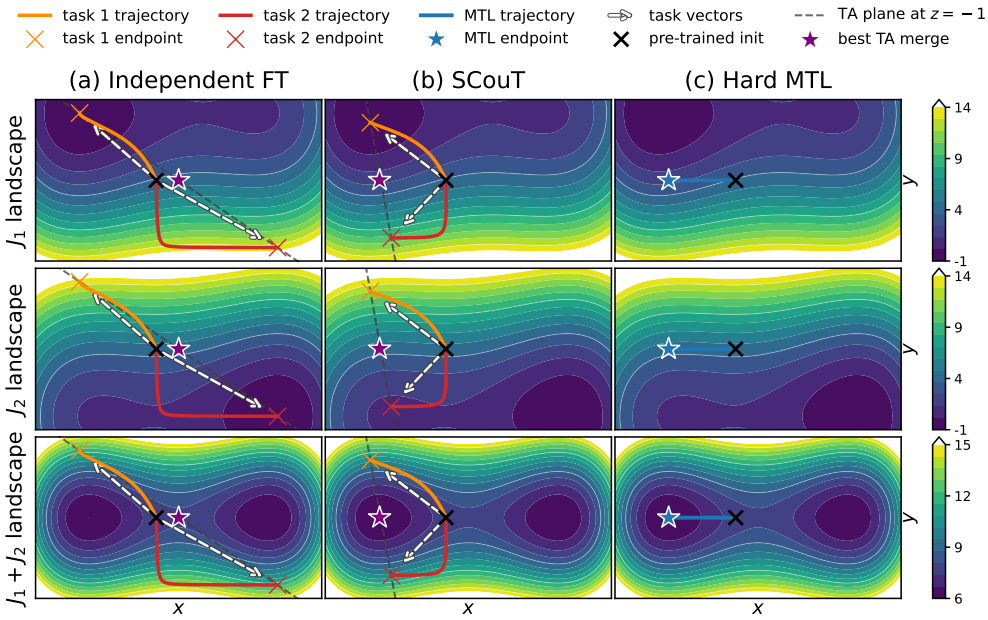

Figure 2: Toy 2D loss landscapes on the slice $z = -1$. Rows show J_1, J_2, and $J_1 + J_2$; columns compare the gradient-descent trajectories and endpoints of independent FT, SCouT, and hard MTL. Here $J_k(x, y, z) = (x^2 - 1)^2 + 3(y - s_k)^2 + 2(z + 1)^2 + s_k x$, with $s_k = (-1)^{k+1}$. In (a) and (b), the purple star marks the best TA merge from (2) over $\alpha_k \in [-2, 2]$, selected to minimize $J_1 + J_2$.

The softly coupled objective (7) places SCouT in an intermediate regime between independent FT and hard MTL. Starting from the same pre-trained initialization, each specialist of SCouT maintains a separate parameterization, as in independent FT, while their updates are coordinated through the merged model. In contrast, hard MTL enforces a single shared parameterization across tasks. Figure 1 summarizes these three optimization structures.

To illustrate the geometric consequence of this intermediate regime, Figure 2 considers a simple two-task problem whose aggregate risk $J_1 + J_2$ contains two low-loss basins separated by a high-loss ridge. Although stylized, this example highlights a broader limitation of post-hoc merging: once independently fine-tuned specialists are fixed, the merging is restricted to a parameter-defined search family, within which task heterogeneity and unfavorable loss geometry can prevent even optimally chosen merging coefficients from reaching a low-loss multitask solution (Ainsworth et al., 2023; Lee et al., 2025; Lu et al., 2026). SCouT instead coordinates the task-specific optimization trajectories during fine-tuning, *steering the specialists toward a more merge-compatible region while allowing them to remain distinct*. Hard MTL reaches a low-loss multitask solution through a different mechanism by enforcing a single shared parameterization, thereby sacrificing task-specific specialization at the parameter level.

4 Theoretical Analysis

Since the SCouT formulation interpolates between independent fine-tuning and multitask learning by means of the coupling parameter λ, the critical question is how the specialist and merged model performance vary as a function of λ. To this end, we characterize the performance for small coupling strengths λ, and establish that for suitably chosen $\lambda > 0$, merged performance can be substantially improved with negligible impact on specialist performance. Proofs of these statements and a standard stationarity guarantee are provided in Appendix A.

4.1 Limiting Regimes

Let $\mathcal{F}_\lambda(\vartheta)$ denote the softly coupled objective. We define the *specialist risk* $\mathcal{J}(\vartheta)$ and *coupling* $\mathcal{C}(\vartheta)$ as follows:

$$\mathcal{F}_\lambda(\vartheta) \triangleq \mathcal{J}(\vartheta) + \lambda \mathcal{C}(\vartheta), \qquad \mathcal{J}(\vartheta) \triangleq \sum_{k=1}^{K} J_k(\theta_k), \qquad \mathcal{C}(\vartheta) \triangleq \frac{1}{2} \sum_{k=1}^{K} \|\theta_{\text{merge}} - \theta_k\|_{A_k}^2 . \tag{8}$$

We first characterize the two limiting points of SCouT as the coupling strength approaches zero or infinity.

Proposition 1 (Limiting regimes). *At $\lambda = 0$, the softly coupled objective reduces to the independent-FT objective:*

$$\arg\min_{\vartheta} \mathcal{F}_0(\vartheta) = \arg\min_{\{\theta_k\}_{k=1}^K} \sum_{k=1}^K J_k(\theta_k). \tag{9}$$

In the limit $\lambda \to \infty$, the coupling constraint is enforced, yielding the limiting problem

$$\arg\min_{\vartheta} \mathcal{F}_\infty(\vartheta) = \arg\min_{\vartheta:\, \mathcal{C}(\vartheta)=0} \mathcal{J}(\vartheta). \tag{10}$$

If, additionally, $A_k \succ 0$ for every $k \in [K]$ and $\sum_{k=1}^K \alpha_k = 1$, then the limiting constrained problem reduces to hard MTL.

4.2 Specialist and Post-merge Risk Response

After characterizing the limiting regimes, we now study the intermediate regime and ask whether a finite coupling strength can improve mergeability before substantially degrading specialization. To quantify post-merge performance, define the *post-merge risk* as

$$\mathcal{M}(\vartheta) \triangleq \sum_{k=1}^K J_k(\theta_{\mathrm{merge}}). \tag{11}$$

The following theorem provides the central theoretical result of this work, which characterizes the local variation of the specialist and post-merge risks around the independent FT solutions.

Theorem 1 (**Asymmetric specialization–mergeability response**). *Suppose that each $J_k(\cdot)$ is three times continuously differentiable in neighborhoods of θ_k^o and $\theta_{\mathrm{merge}}^o$, with $\nabla J_k(\theta_k^o) = 0$ and $H_k \triangleq \nabla^2 J_k(\theta_k^o) \succeq 0$. Let $H \triangleq \mathrm{blkdiag}\{H_1, \ldots, H_K\}$, and let $\mathrm{Range}(H)$ denote the range of H.*

Then, there exists $\epsilon_\lambda > 0$ such that, for every $0 \le \lambda < \epsilon_\lambda$, $\mathcal{F}_\lambda$ has a unique stationary point ϑ^λ near ϑ^o within the affine subspace

$$\vartheta^\lambda \in \vartheta^o + \mathrm{Range}(H). \tag{12}$$

As $\lambda \to 0$, this local solution ϑ^λ satisfies

$$\Delta\mathcal{M}(\lambda) \triangleq \mathcal{M}(\vartheta^\lambda) - \mathcal{M}(\vartheta^o) = -\lambda\chi + \mathcal{O}(\lambda^2), \qquad \chi \triangleq \left\langle \nabla_\vartheta \mathcal{M}(\vartheta^o), H^\dagger \nabla_\vartheta \mathcal{C}(\vartheta^o) \right\rangle, \tag{13}$$

$$\Delta\mathcal{J}(\lambda) \triangleq \mathcal{J}(\vartheta^\lambda) - \mathcal{J}(\vartheta^o) = \frac{\lambda^2}{2}\Upsilon + \mathcal{O}(\lambda^3), \qquad \Upsilon \triangleq \nabla_\vartheta \mathcal{C}(\vartheta^o)^\top H^\dagger \nabla_\vartheta \mathcal{C}(\vartheta^o). \tag{14}$$

Corollary 1 (Weak-coupling advantage). *Suppose that the conditions of Theorem 1 hold. If independent FT has a strict aggregate specialist advantage over hard MTL, $\sum_{k=1}^K J_k(\theta^{\mathrm{MTL}}) > \mathcal{J}(\vartheta^o)$, and the initial coupling direction improves the post-merge risk, i.e., $\chi > 0$, then there exists $\lambda_{\mathrm{sim}} > 0$ such that, for every $\lambda \in (0, \lambda_{\mathrm{sim}})$,*

$$\mathcal{M}(\vartheta^\lambda) < \mathcal{M}(\vartheta^o), \qquad \mathcal{J}(\vartheta^\lambda) < \sum_{k=1}^K J_k(\theta^{\mathrm{MTL}}). \tag{15}$$

Theorem 1 reveals an asymmetric local risk response to coupling. Starting from independent FT ($\lambda = 0$), the post-merge risk changes linearly in λ as $-\lambda\chi$, whereas the specialist risk changes quadratically as $\frac{\lambda^2}{2}\Upsilon$. Hence, when the initial coupling direction decreases the post-merge risk ($\chi > 0$), sufficiently weak coupling improves mergeability meaningfully while incurring negligible specialization cost. Combined with a strict specialist advantage of independent FT over hard MTL, Corollary 1 guarantees a nonempty intermediate regime that improves the independent-FT merge while retaining better specialist performance than hard MTL.

5 Experiments

We now empirically evaluate SCouT along four dimensions: (i) whether its empirical risks exhibit the weak-coupling response as predicted in Theorem 1; (ii) how the coupling strength controls the specialization–mergeability trade-off across model scales; (iii) how SCouT compares with merger-friendly fine-tuning and hard MTL; and (iv) whether its gains persist across post-hoc merging rules.

5.1 Experimental Setup

SCouT is tested on ten image-classification tasks using pre-trained CLIP ViT backbones (Radford et al., 2021), including ViT-B/32 and ViT-L/14. All methods are initialized from the same pre-trained checkpoint. For SCouT, we adopt a simple *isotropic coupling* and *uniform training-time merging operator* throughout the main experiments:

$$A_k = I, \qquad \theta_{\mathrm{merge},i} = \theta_0 + \frac{1}{K}\sum_{k=1}^{K}(\theta_{k,i} - \theta_0)\,. \tag{16}$$

Due to space limitations, full details on the datasets, training protocols, hyperparameter selection, and baseline implementations are provided in Appendix B.

5.2 Weak-coupling Risk Response

Theorem 1 predicts a quadratic specialist-risk response and a linear post-merge response as $\lambda \to 0$. We test whether this asymmetry remains visible in practical SCouT, where practical training algorithms may not have converged to local minima as assumed in the theorem. Using ViT-B/32, we conduct experiments over five random seeds under fixed training configurations, measure both risk changes relative to $\lambda = 0$, and fit their mean values over $\lambda \in [0, 0.1]$ using $\Delta\mathcal{J}(\lambda) = a_2\lambda^2$ and $\Delta\mathcal{M}(\lambda) = b_1\lambda$.

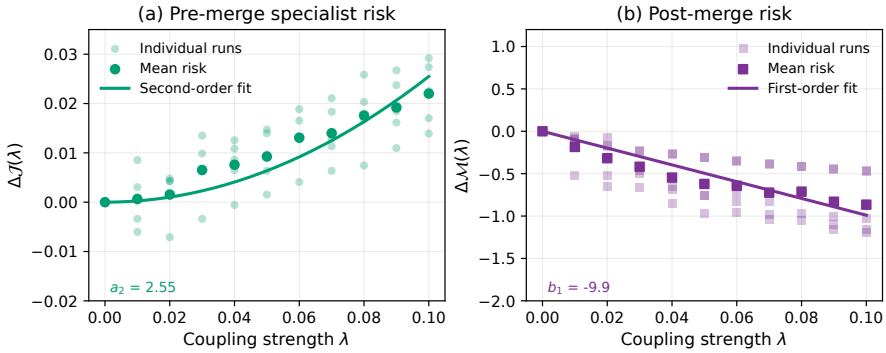

Figure 3: Training risk responses in the weak-coupling regime, fitted using their theoretically predicted leading-order scaling.

Figure 3 shows that the predicted leading-order forms closely match the empirical trends: specialist risk increases quadratically ($a_2 = 2.55$), whereas post-merge risk decreases linearly ($b_1 = -9.9$) despite variability across runs. These distinct response orders support the predicted asymmetry. Complete risk sweeps and further details are provided in Appendix C.1.

5.3 Performance Comparison and Analysis

Having examined the local weak-coupling response, we next study how λ controls the broader specialization–mergeability trade-off and compare SCouT with competing methods. For specialist-based methods, we evaluate each specialist on its corresponding task and average the accuracies across tasks (*Specialist Acc.*). We also report the merged model's average accuracy (*Merged-model Acc.*), and normalized accuracy (*Norm. Acc.*), defined as the mean task-wise ratio of merged-model to specialist accuracy. For hard-MTL baselines, we report shared-model accuracy (*Shared-model Acc.*). TA is the merging rule used for the λ-sweep and training-strategy comparison.

Effect of coupling strength λ. Figure 4 examines how the coupling strength λ controls the specialization–mergeability trade-off through a λ-sweep. Consistent with the limiting-regime analysis, $\lambda = 0$ recovers independent FT, while large λ drives the specialists toward the hard MTL regime. For ViT-B/32, all tested values from $\lambda = 5 \times 10^{-3}$ to $\lambda = 20$ improve post-merge accuracy over the independent-FT merge while retaining higher pre-merge accuracy than hard MTL. This result is consistent with the weak-coupling advantage characterized in Corollary 1. As λ increases, pre-merge accuracy decreases gradually, whereas post-merge accuracy improves substantially, qualitatively reflecting the asymmetric response to coupling described in Theorem 1. For the larger

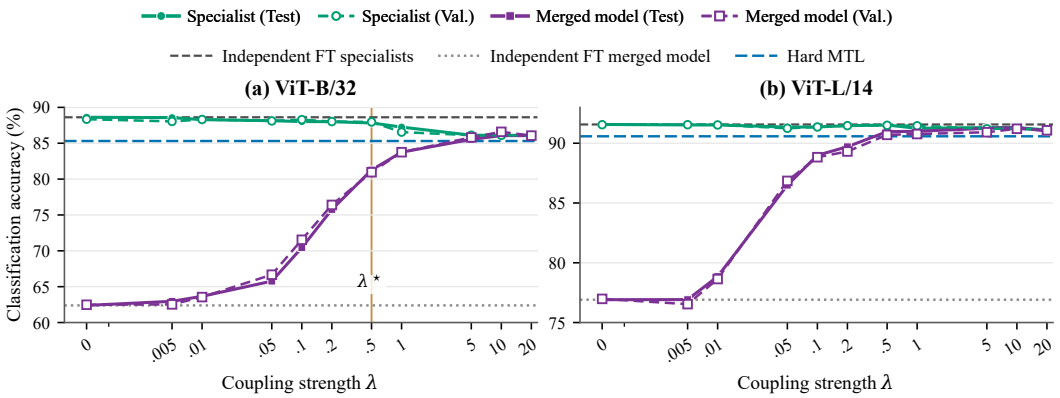

Figure 4: Effect of the coupling strength λ on SCouT for (a) ViT-B/32 and (b) ViT-L/14. Curves report the average pre-merge specialist accuracy and post-merge accuracy obtained with TA, while horizontal lines indicate the corresponding independent FT and hard MTL baselines. All results are averaged over five random seeds.

ViT-L/14, the gap between independent FT and hard MTL is markedly narrower, suggesting that the increased model capacity alleviates negative transfer under hard sharing. Nevertheless, a favorable intermediate regime remains. We therefore focus the remaining analysis on ViT-B/32, where limited capacity makes the specialization–mergeability trade-off more pronounced and provides a more revealing testbed for improving compact multitask models without relying on model scaling.

The validation and test curves exhibit closely aligned trends across the λ-sweep. We can select λ using validation data only, among configurations whose pre-merge accuracy decreases by at most 2.5% relative to independent FT, we choose the one with the highest post-merge validation accuracy. This criterion selects $\lambda^\star = 0.5$ for ViT-B/32.

Table 1: Comparison of training strategies on ViT-B/32 and ViT-L/14.

Training strategy	ViT-B/32			ViT-L/14		
	Specialist Acc.	Merged/shared-model Acc.	Norm. Acc.	Specialist Acc.	Merged/shared-model Acc.	Norm. Acc.
Pre-trained	43.75	-	-	62.00	-	-
Independent FT	88.62 (0%)	61.66 (0%)	69.80	91.57	76.91	83.22
FTTS (Ortiz-Jimenez et al., 2023)	83.29	67.39	80.98	88.56	77.09	87.03
FT-Attention (Jin et al., 2025)	86.89	69.14	79.57	90.92	77.86	85.62
SAFT (Lee et al., 2025)	87.53	62.86	72.18	91.06	79.26	86.11
MergOPT (Yang et al., 2026b)	86.67	64.21	74.08	91.03	75.92	82.69
Hard MTL	–	85.31	–	–	90.58	–
Hard MTL + UW (Kendall et al., 2018)	–	85.62	–	–	89.94	–
Hard MTL + PCGrad (Yu et al., 2020)	–	85.53	–	–	90.09	–
SCouT ($\lambda = 0.01$)	88.29 (-0.37%)	63.66 (+3.24%)	72.14	91.53	78.85	86.73
SCouT ($\lambda^\star = 0.5$)	87.83 (-0.89%)	81.27 (+31.80%)	92.28	91.53	90.97	99.43
SCouT ($\lambda = 5$)	86.12 (-2.82%)	85.54 (+38.73%)	99.33	91.10	91.01	99.91

Comparison of training strategies. Table 1 shows that SCouT provides a stronger specialization–mergeability trade-off than the other fine-tuning methods. On ViT-B/32, relative to independent FT, SCouT reduces pre-merge specialist accuracy by only 0.89% while increasing TA merged-model accuracy from 61.66% to 81.27%, a 31.80% relative improvement. It also outperforms FTTS, FT-Attention, SAFT, and MergOPT in both specialist and merged accuracy, indicating that the gain in mergeability is achieved without a large sacrifice in specialist performance. Although the selected merged SCouT model remains 4.04–4.35 points below the hard-MTL baselines, SCouT retains separate specialists with higher specialist accuracy, while substantially narrowing the merging gap.

Table 2: Post-merge accuracies on ViT-B/32 across fine-tuning strategies and merging methods.

Training strategy	Weight Avg. (Wortsman et al., 2022)		Fisher (Matena & Raffel, 2022)		TA (Ilharco et al., 2023)		TIES (Yadav et al., 2023)		WUDI (Cheng et al., 2025)	
	Merged Acc.	Norm. Acc.	Merged Acc.	Norm. Acc.	Merged Acc.	Norm. Acc.	Merged Acc.	Norm. Acc.	Merged Acc.	Norm. Acc.
Independent FT	58.81 (0%)	67.89	64.56 (0%)	73.07	61.66 (0%)	69.80	64.25 (0%)	73.71	76.79 (0%)	87.84
FTTS	67.29	80.92	69.56	83.67	67.39	80.98	65.91	79.23	71.81	86.11
FT-Attention	59.41	69.01	67.25	77.82	69.14	79.57	68.26	78.68	74.43	85.33
SAFT	58.78	67.73	66.92	76.91	62.86	72.18	64.69	74.39	76.95	88.52
MergOPT	60.33	69.82	66.12	76.41	64.21	74.08	65.86	75.94	78.21	90.07
SCouT ($\lambda^\star = 0.5$)	**81.27** (+38.19%)	**92.28**	**81.77** (+26.66%)	**92.78**	**81.27** (+31.80%)	**92.28**	**79.63** (+23.94%)	**90.36**	**81.71** (+6.41%)	**92.90**

Compatibility with post-hoc merging rules. Table 2 evaluates SCouT across different post-hoc merging rules. SCouT achieves the highest post-merge and normalized accuracy under all five methods, consistently outperforming the other merger-friendly strategies. Even with WUDI, the strongest merger for independent FT, SCouT improves post-merge accuracy from 76.79% to 81.71%. This shows that stronger post-hoc merging alone cannot fully compensate for poorly aligned specialists; by improving compatibility during fine-tuning, SCouT provides a better starting point for a wide range of merging rules. Further discussion is provided in Appendix C.2.

Loss-landscape analysis. Figure 5 illustrates how coupling reshapes the geometry of the specialist solutions. With independent FT ($\lambda = 0$), the TA search family is restricted by the resulting specialist directions, and its best merge attains an average loss of 0.546. As λ increases, the specialists are positioned such that the same post-hoc search reaches progressively lower-loss regions, reducing the best-merge loss to 0.510 at $\lambda = 0.5$ and 0.453 at $\lambda = 5$. This supports the central motivation of SCouT: improving mergeability during fine-tuning provides a more favorable geometry for subsequent post-hoc merging.

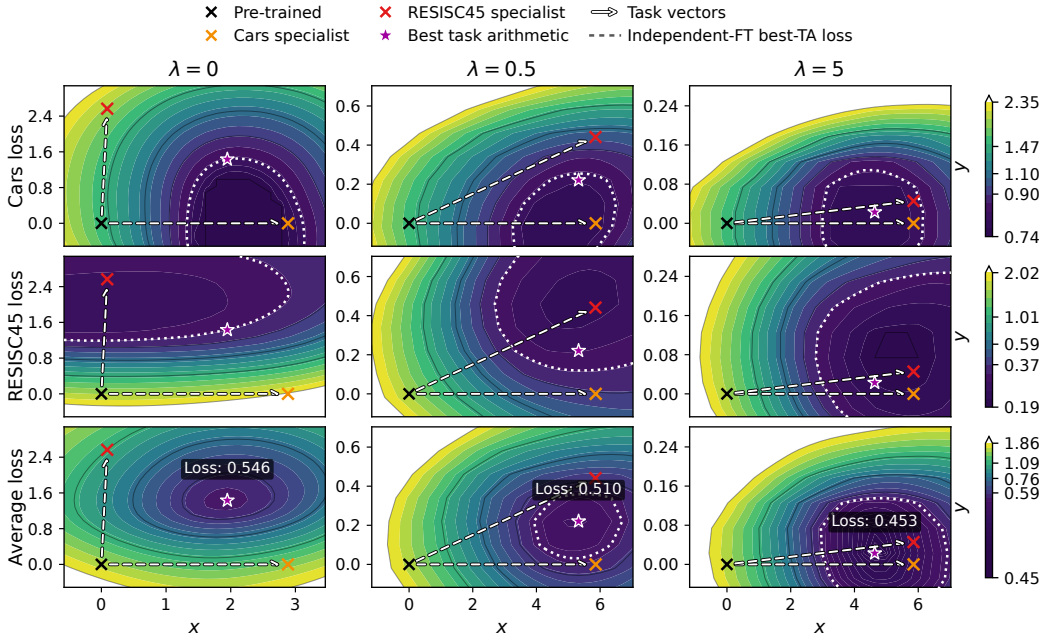

Figure 5: Loss-landscape visualization for Cars and RESISC45 at different coupling strengths $\lambda \in \{0, 0.5, 5\}$. The landscape basis is computed using Gram-Schmidt orthonormalization of the two specialist task vectors (Yang et al., 2026c). Markers denote the pre-trained model, specialists, and the best TA merge. The white dotted contour marks the loss level achieved by the best validation-selected TA merge of independently fine-tuned specialists ($\lambda = 0$).

6 Conclusion

We introduced Softly Coupled Fine-Tuning (SCouT), a merge-aware training framework that coordinates task-specific specialists during training while preserving their individual objectives. By varying the coupling strength, SCouT provides a tunable specialization–mergeability trade-off. Our theoretical and empirical analyses show that weak coupling can improve post-merge performance much faster than it degrades specialist performance. This asymmetry yields a favorable regime that substantially improves model compatibility while preserving specialization. For example, on ViT-B/32, a validation-selected coupling strength improves post-merge accuracy by 31.80% while reducing specialist accuracy by only 0.89%, relative to independent fine-tuning. This makes SCouT particularly suitable for deployment scenarios that require both high-performing task-specific specialists and a single merged model for efficient multitask inference.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. Technical report, Stanford Center for Research on Foundation Models, 2021.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, 2000.

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 *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10121–10143, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *International Workshop on Paraphrasing (IWP)*, 2005.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pp. 845–850, 2015.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

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

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs. Quora, 2017. URL `https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs`.

Ruochen Jin, Bojian Hou, Jiancong Xiao, Weijie J. Su, and Li Shen. Fine-tuning attention modules only: Enhancing weight disentanglement in task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018.

Steven G. Krantz and Harold R. Parks. *The Implicit Function Theorem: History, Theory, and Applications*. Birkhäuser, 2013.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. PyTorch distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12):3005–3018, 2020.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18878–18890, 2021.

Tian Yu Liu, Aditya Golatkar, and Stefano Soatto. Tangent transformers for composition, privacy and removal. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Juanwu Lu, Anand Bhaskar, Brian Axelrod, Ekaterina Tolstaya, and Tristan Emrich. Model merging on loss landscapes: A geometric perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7644–7653, 2026.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43177–43199, 2025.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17703–17716, 2022.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3994–4003, 2016.

Roula Nassif, Stefan Vlaski, Cédric Richard, and Ali H Sayed. Learning over multitask graphs—Part I: Stability analysis. *IEEE Open Journal of Signal Processing*, 1:28–45, 2020.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Graphics, Visualization, Computer Vision and Image Processing*, pp. 722–729, 2008.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 66727–66754, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132, 2020.

Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8728–8740, 2020.

Anke Tang, Li Shen, Yong Luo, Yibing Zhan, Han Hu, Bo Du, Yixin Chen, and Dacheng Tao. Parameter-efficient multi-task model fusion with partial linearization. In *International Conference on Learning Representations (ICLR)*, 2024.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention*, pp. 210–218. Springer, 2018.

Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7561–7570, 2022.

Zirui Wan and Stefan Vlaski. Multitask learning with learned task relationships. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP Workshop at Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 353–355, 2018.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1112–1122, 2018.

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 (ICML)*, volume 162, pp. 23965–23998, 2022.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in neural information processing systems (NeurIPS)*, volume 36, pp. 7093–7115, 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 *International Conference on Learning Representations (ICLR)*, 2024.

Enneng Yang, Anke Tang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, and Jie M. Zhang. Continual model merging without data: Dual projections for balancing stability and plasticity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

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, 2026a.

Enneng Yang, Qun Yang, Peng Wang, Anke Tang, Guibing Guo, Xiaochun Cao, and Li Shen. MergOPT: A merge-aware optimizer for robust model merging. In *International Conference on Learning Representations (ICLR)*, 2026b.

Sihan Yang, Kexuan Shi, and Weiyang Liu. Orthogonal model merging. In *International Conference on Machine Learning (ICML)*, 2026c.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 5824–5836, 2020.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

APPENDIX CONTENTS

A Theoretical Analysis and Proof **14**

A.1 Proof of Proposition 1 14

A.2 Optimization Guarantee 15

A.3 Proof of Theorem 1 17

A.4 Proof of Corollary 1 18

B Experimental Details **19**

B.1 Implementation Details 19

B.2 Optimization and Algorithm 20

B.3 Baselines 21

C Additional Experimental Analysis **22**

C.1 Empirical Analysis of the Local Risk Response 22

C.2 Compatibility with Post-hoc Merging Rules 22

C.3 Loss Landscape Visualization 23

D Evaluation on Natural Language Processing Tasks **25**

D.1 Experimental Setup 25

D.2 Results and Analysis 25

E Training Cost Comparison **26**

F AI Usage Statement **27**

A THEORETICAL ANALYSIS AND PROOF

A.1 PROOF OF PROPOSITION 1

Proof. For each $\lambda \geq 0$, let

$$\vartheta_\star^\lambda \in \arg\min_\vartheta \mathcal{F}_\lambda(\vartheta). \tag{17}$$

Because $\mathcal{F}_0(\vartheta) = \mathcal{J}(\vartheta)$, we immediately have

$$\arg\min_\vartheta \mathcal{F}_0(\vartheta) = \arg\min_\vartheta \mathcal{J}(\vartheta), \tag{18}$$

which corresponds to independent fine-tuning.

We next consider $\lambda \to \infty$. Define the zero-penalty set

$$\bar{\Theta} \triangleq \{\vartheta : \mathcal{C}(\vartheta) = 0\}. \tag{19}$$

This set is nonempty because $\theta_1 = \cdots = \theta_K = \theta_0$ belongs to $\bar{\Theta}$.

For any $\bar{\vartheta} \in \bar{\Theta}$, the optimality of $\vartheta_\star^\lambda$ gives

$$\mathcal{J}(\vartheta_\star^\lambda) + \lambda \mathcal{C}(\vartheta_\star^\lambda) \leq \mathcal{J}(\bar{\vartheta}) + \lambda \mathcal{C}(\bar{\vartheta}) = \mathcal{J}(\bar{\vartheta}). \tag{20}$$

Let $\mathcal{J}^{\inf}$ be a lower bound of $\mathcal{J}$. Because $\mathcal{C} \geq 0$,

$$0 \leq \mathcal{C}(\vartheta_\star^\lambda) \leq \frac{\mathcal{J}(\bar{\vartheta}) - \mathcal{J}^{\inf}}{\lambda}. \tag{21}$$

Consequently, $\mathcal{C}(\vartheta_\star^\lambda) \to 0$ as $\lambda \to \infty$.

Let $\vartheta_\star^\infty$ be any accumulation point, and consider a subsequence along which $\vartheta_\star^\lambda \to \vartheta_\star^\infty$. By continuity of $\mathcal{C}$,

$$\mathcal{C}(\vartheta_\star^\infty) = 0, \tag{22}$$

so $\vartheta_\star^\infty \in \bar{\Theta}$. Moreover, dropping the nonnegative penalty term from the optimality inequality gives

$$\mathcal{J}(\vartheta_\star^\lambda) \le \mathcal{J}(\bar{\vartheta}), \qquad \forall\, \bar{\vartheta} \in \bar{\Theta}. \tag{23}$$

Taking the limit along the convergent subsequence and using the continuity of $\mathcal{J}$ yields

$$\mathcal{J}(\vartheta_\star^\infty) \le \mathcal{J}(\bar{\vartheta}), \qquad \forall\, \bar{\vartheta} \in \bar{\Theta}. \tag{24}$$

Therefore,

$$\vartheta_\star^\infty \in \arg\min_{\vartheta \in \bar{\Theta}} \mathcal{J}(\vartheta). \tag{25}$$

If $A_k \succ 0$ for every k, then $\mathcal{C}(\vartheta) = 0$ if and only if

$$\theta_1 = \cdots = \theta_K = \theta_{\text{merge}}. \tag{26}$$

Let their common value be θ. The merging rule gives

$$\theta = \theta_0 + \sum_{k=1}^K \alpha_k(\theta - \theta_0) \tag{27}$$

$$= \theta_0 + \left(\sum_{k=1}^K \alpha_k\right)(\theta - \theta_0). \tag{28}$$

When $\sum_{k=1}^K \alpha_k = 1$, every consensus parameter θ is feasible. Hence, the limiting constrained problem is

$$\arg\min_\theta \sum_{k=1}^K J_k(\theta), \tag{29}$$

which is precisely hard MTL. □

A.2 Optimization Guarantee

While the main text focuses on how the performance of SCouT varies with the coupling strength λ, we complement that analysis with an optimization guarantee for each fixed λ. Under standard smoothness assumptions, the softly coupled objective has a Lipschitz-continuous gradient, allowing gradient descent to attain the standard stationarity rate stated below.

Proposition 2 (Stationarity)**.** *Suppose that each task loss $J_k(\cdot)$ is differentiable with a Lipschitz-continuous gradient, and $\mathcal{F}_\lambda(\cdot)$ is bounded below by $\mathcal{F}_\lambda^{\inf}$. Let $\delta_\mathcal{F}$ denote a Lipschitz constant of $\nabla\mathcal{F}_\lambda$. If the iterates are generated by gradient descent with $0 < \mu_i \le 1/\delta_\mathcal{F}$, then, for every $T \ge 1$,*

$$\min_{0 \le i \le T-1} \|\nabla \mathcal{F}_\lambda(\vartheta_i)\|^2 \le \frac{2\left[\mathcal{F}_\lambda(\vartheta_0) - \mathcal{F}_\lambda^{\inf}\right]}{\sum_{i=1}^T \mu_i}. \tag{30}$$

In particular, a constant step size $0 < \mu \le 1/\delta_\mathcal{F}$ yields the standard $\mathcal{O}(1/T)$ stationarity rate.

Proof. For fixed merging coefficients $\{\alpha_k\}_{k=1}^K$, the Task Arithmetic merge can be written as

$$\theta_{\text{merge}} = \sum_{\ell=1}^K \alpha_\ell \theta_\ell + \left(1 - \sum_{\ell=1}^K \alpha_\ell\right)\theta_0. \tag{31}$$

Let $\mathbb{1} \in \mathbb{R}^K$ denote the all-ones column vector. The stacked discrepancies are

$$r(\vartheta) \triangleq \text{col}\{\theta_{\text{merge}} - \theta_1, \ldots, \theta_{\text{merge}} - \theta_K\} = (B \otimes I_P)\vartheta + c, \tag{32}$$

where

$$B \triangleq \mathbb{1}\,[\alpha_1 \quad \cdots \quad \alpha_K] - I_K, \qquad c \triangleq \left(1 - \sum_{\ell=1}^{K} \alpha_\ell\right)(\mathbb{1} \otimes \theta_0). \tag{33}$$

Define the block-diagonal matrix

$$\mathcal{A} \triangleq \mathrm{blkdiag}\{A_1, \ldots, A_K\}. \tag{34}$$

The coupling regularizer can then be expressed as

$$\mathcal{C}(\vartheta) = \frac{1}{2} r(\vartheta)^\top \mathcal{A} r(\vartheta). \tag{35}$$

Because each A_k is fixed and symmetric, differentiation gives

$$\nabla \mathcal{C}(\vartheta) = (B^\top \otimes I_P)\mathcal{A}\left[(B \otimes I_P)\vartheta + c\right]. \tag{36}$$

Hence, for any ϑ and ϑ',

$$\|\nabla \mathcal{C}(\vartheta) - \nabla \mathcal{C}(\vartheta')\| = \left\|(B^\top \otimes I)\mathcal{A}(B \otimes I)(\vartheta - \vartheta')\right\| \tag{37}$$

$$\leq \delta_A \|\vartheta - \vartheta'\|, \tag{38}$$

where $\|\cdot\|$ denotes the induced spectral norm for matrices

$$\delta_A \triangleq \left\|(B^\top \otimes I)\mathcal{A}(B \otimes I)\right\|. \tag{39}$$

Thus, $\mathcal{C}$ has a δ_A-Lipschitz continuous gradient. Moreover,

$$\delta_A \leq \|B^\top B\| \max_{k \in [K]} \|A_k\|. \tag{40}$$

Since the task parameters occupy disjoint blocks and each ∇J_k is δ_k-Lipschitz,

$$\|\nabla \mathcal{J}(\vartheta) - \nabla \mathcal{J}(\vartheta')\|^2 = \sum_{k=1}^{K} \|\nabla J_k(\theta_k) - \nabla J_k(\theta_k')\|^2 \tag{41}$$

$$\leq \left(\max_{k \in [K]} \delta_k\right)^2 \sum_{k=1}^{K} \|\theta_k - \theta_k'\|^2. \tag{42}$$

Therefore, $\mathcal{J}$ has a $\max_{k \in [K]} \delta_k$-Lipschitz continuous gradient. Since

$$\mathcal{F}_\lambda(\vartheta) = \mathcal{J}(\vartheta) + \lambda \mathcal{C}(\vartheta), \tag{43}$$

a valid Lipschitz constant for its gradient is

$$\delta_{\mathcal{F}} = \max_{k \in [K]} \delta_k + \lambda \delta_A. \tag{44}$$

The gradient-descent update can be written compactly as

$$\vartheta_i = \vartheta_{i-1} - \mu_i \nabla \mathcal{F}_\lambda(\vartheta_{i-1}). \tag{45}$$

By the descent lemma,

$$\mathcal{F}_\lambda(\vartheta_i) \leq \mathcal{F}_\lambda(\vartheta_{i-1}) + \nabla \mathcal{F}_\lambda(\vartheta_{i-1})^\top(\vartheta_i - \vartheta_{i-1}) + \frac{\delta_{\mathcal{F}}}{2}\|\vartheta_i - \vartheta_{i-1}\|^2 \tag{46}$$

$$= \mathcal{F}_\lambda(\vartheta_{i-1}) - \mu_i \left(1 - \frac{\delta_{\mathcal{F}} \mu_i}{2}\right) \|\nabla \mathcal{F}_\lambda(\vartheta_{i-1})\|^2. \tag{47}$$

Since $\mu_i \leq 1/\delta_{\mathcal{F}}$,

$$1 - \frac{\delta_{\mathcal{F}} \mu_i}{2} \geq \frac{1}{2}, \tag{48}$$

and therefore

$$\mathcal{F}_\lambda(\vartheta_i) \leq \mathcal{F}_\lambda(\vartheta_{i-1}) - \frac{\mu_i}{2} \|\nabla \mathcal{F}_\lambda(\vartheta_{i-1})\|^2. \tag{49}$$

Summing from $i = 1$ to T yields

$$\sum_{i=1}^{T} \mu_i \left\|\nabla \mathcal{F}_\lambda(\vartheta_{i-1})\right\|^2 \leq 2\left[\mathcal{F}_\lambda(\vartheta_0) - \mathcal{F}_\lambda(\vartheta_T)\right] \tag{50}$$

$$\leq 2\left[\mathcal{F}_\lambda(\vartheta_0) - \mathcal{F}_\lambda^{\inf}\right]. \tag{51}$$

Since the minimum is no larger than the corresponding weighted average,

$$\min_{0\leq i\leq T-1} \left\|\nabla \mathcal{F}_\lambda(\vartheta_i)\right\|^2 \leq \frac{2\left[\mathcal{F}_\lambda(\vartheta_0) - \mathcal{F}_\lambda^{\inf}\right]}{\sum_{i=1}^{T}\mu_i}. \tag{52}$$

□

A.3 Proof of Theorem 1

Proof. Let the columns of $U_k \in \mathbb{R}^{P\times r_k}$ form an orthonormal basis of $\mathrm{Range}(H_k)$,

$$\mathrm{span}(U_k) = \mathrm{Range}(H_k) = \mathcal{N}(H_k)^\perp. \tag{53}$$

Since each $H_k \succeq 0$ is symmetric, we can write its reduced eigendecomposition as

$$H_k = U_k \Lambda_k U_k^\top, \tag{54}$$

where $\Lambda_k \in \mathbb{R}^{r_k\times r_k}$ is positive definite. Define $U \triangleq \mathrm{blkdiag}\{U_1, \ldots, U_K\}$. We now analyze the objective (7) on the locally active affine subspace $\vartheta^o + \mathrm{Range}(U)$, in which $\mathrm{Range}(U) = \mathrm{Range}(H)$.

Define the stationarity mapping of the reduced objective

$$\widetilde{\mathcal{S}}(u,\lambda) \triangleq \nabla_u \mathcal{J}(\vartheta^o + Uu) + \lambda \nabla_u \mathcal{C}(\vartheta^o + Uu) \tag{55}$$

$$= U^\top\left[\nabla_\vartheta \mathcal{J}(\vartheta^o + Uu) + \lambda \nabla_\vartheta \mathcal{C}(\vartheta^o + Uu)\right], \tag{56}$$

where $u \in \mathbb{R}^r$ is a coordinate vector and $r \triangleq \sum_{k=1}^K r_k$. Because the independently fine-tuned specialists are stationary points of their respective local objectives,

$$\widetilde{\mathcal{S}}(0,0) = U^\top \nabla_\vartheta \mathcal{J}(\vartheta^o) = 0. \tag{57}$$

Moreover,

$$\nabla_u \widetilde{\mathcal{S}}(u,\lambda) = U^\top \nabla^2_\vartheta \mathcal{J}(\vartheta^o + Uu)U + \lambda U^\top \nabla^2_\vartheta \mathcal{C}(\vartheta^o + Uu)U. \tag{58}$$

At $(0,0)$, we have

$$\nabla_u \widetilde{\mathcal{S}}(0,0) = U^\top H U, \tag{59}$$

we can then define

$$H_{\mathcal{U}} \triangleq U^\top H U = \begin{bmatrix} U_1^\top H_1 U_1 & 0 & \cdots & 0 \\ 0 & U_2^\top H_2 U_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & U_K^\top H_K U_K \end{bmatrix} = \mathrm{blkdiag}\left\{U_1^\top H_1 U_1, \ldots, U_K^\top H_K U_K\right\}. \tag{60}$$

All the projected Hessians are positive definite; therefore, $H_{\mathcal{U}} \succ 0$ and $H_{\mathcal{U}}$ is invertible. Since each J_k is three times continuously differentiable and $\mathcal{C}$ is a quadratic function for fixed $A_k \succeq 0$, the mapping $\widetilde{\mathcal{S}}$ is twice continuously differentiable. The implicit function theorem (Krantz & Parks, 2013; Bonnans & Shapiro, 2000) therefore guarantees that there exists $\epsilon_\lambda, \epsilon_u > 0$ and a unique mapping

$$\lambda \longmapsto u^\lambda, \qquad |\lambda| < \epsilon_\lambda, \tag{61}$$

with $\|u^\lambda\| < \epsilon_u$, such that

$$u^0 = 0, \qquad \widetilde{\mathcal{S}}(u^\lambda, \lambda) = 0. \tag{62}$$

Thus, u^λ locally parameterizes the stationary points of the reduced objective as the coupling strength varies around $\lambda = 0$. Now, we can convert back to the original parameter space. Define

$$\vartheta^\lambda \triangleq \vartheta^o + Uu^\lambda, \qquad g_R \triangleq \nabla_\vartheta \mathcal{C}(\vartheta^o). \tag{63}$$

Differentiating (62) with respect to λ gives

$$0 = \left[U^\top \nabla^2_\vartheta \mathcal{J}(\vartheta^\lambda) U + \lambda U^\top \nabla^2_\vartheta \mathcal{C}(\vartheta^\lambda) U\right] \frac{\mathrm{d}u^\lambda}{\mathrm{d}\lambda} + U^\top \nabla_\vartheta \mathcal{C}(\vartheta^\lambda). \tag{64}$$

Evaluating at $\lambda = 0$ yields

$$H_{\mathcal{U}} \left.\frac{\mathrm{d}u^\lambda}{\mathrm{d}\lambda}\right|_{\lambda=0} + U^\top g_R = 0.$$

Therefore,

$$\left.\frac{\mathrm{d}u^\lambda}{\mathrm{d}\lambda}\right|_{\lambda=0} = -H_{\mathcal{U}}^{-1} U^\top g_R. \tag{65}$$

Since $\vartheta^\lambda = \vartheta^o + U u^\lambda$, the corresponding direction in the original parameter space is

$$\begin{aligned} \left.\frac{\mathrm{d}\vartheta^\lambda}{\mathrm{d}\lambda}\right|_{\lambda=0} &= -U H_{\mathcal{U}}^{-1} U^\top g_R \\ &= -H^\dagger g_R, \end{aligned} \tag{66}$$

where the last equality follows from $H^\dagger = U H_{\mathcal{U}}^{-1} U^\top$. Consequently,

$$\vartheta^\lambda = \vartheta^o - \lambda H^\dagger g_R + \mathcal{O}(\lambda^2). \tag{67}$$

Thus, an infinitesimal increase in the coupling strength moves the independent specialists in the curvature-adjusted coupling direction $-H^\dagger g_R$. We first consider its effect on the post-merge multi-task risk. A first-order Taylor expansion around ϑ^o gives

$$\begin{aligned} \mathcal{M}(\vartheta^\lambda) &= \mathcal{M}(\vartheta^o) + \nabla_\vartheta \mathcal{M}(\vartheta^o)^\top (\vartheta^\lambda - \vartheta^o) + \mathcal{O}\left(\|\vartheta^\lambda - \vartheta^o\|^2\right) \\ &= \mathcal{M}(\vartheta^o) - \lambda \nabla_\vartheta \mathcal{M}(\vartheta^o)^\top H^\dagger g_R + \mathcal{O}(\lambda^2). \end{aligned} \tag{68}$$

Since $\chi \triangleq \left\langle \nabla_\vartheta \mathcal{M}(\vartheta^o), H^\dagger g_R \right\rangle$, we obtain (13)

$$\mathcal{M}(\vartheta^\lambda) = \mathcal{M}(\vartheta^o) - \lambda\chi + \mathcal{O}(\lambda^2).$$

We next consider the aggregate specialist risk. Since $\nabla_\vartheta \mathcal{J}(\vartheta^o) = 0$, a second-order Taylor expansion yields

$$\begin{aligned} \mathcal{J}(\vartheta^\lambda) &= \mathcal{J}(\vartheta^o) + \frac{1}{2}(\vartheta^\lambda - \vartheta^o)^\top H (\vartheta^\lambda - \vartheta^o) + \mathcal{O}\left(\|\vartheta^\lambda - \vartheta^o\|^3\right) \\ &= \mathcal{J}(\vartheta^o) + \frac{\lambda^2}{2} g_R^\top H^\dagger g_R + \mathcal{O}(\lambda^3), \end{aligned} \tag{69}$$

where we used $H^\dagger H H^\dagger = H^\dagger$. Since $H^\dagger \succeq 0$,

$$g_R^\top H^\dagger g_R \geq 0.$$

Thus, the post-merge improvement is approximately linear in λ, whereas the aggregate specialist cost is quadratic. □

A.4 Proof of Corollary 1

Proof. By the assumed hard-MTL gap,

$$\gamma = \sum_{k=1}^K J_k(\theta^{\mathrm{MTL}}) - \mathcal{J}(\vartheta^o) > 0. \tag{70}$$

Since

$$\mathcal{J}(\vartheta^\lambda) - \mathcal{J}(\vartheta^o) = \mathcal{O}(\lambda^2), \tag{71}$$

there exists $\lambda_J > 0$ such that

$$\mathcal{J}(\vartheta^\lambda) < \mathcal{J}(\vartheta^o) + \gamma = \sum_{k=1}^K J_k(\theta^{\mathrm{MTL}}) \tag{72}$$

for every $0 < \lambda < \lambda_J$.

Likewise, since $\chi > 0$, there exists $\lambda_M > 0$ such that

$$\mathcal{M}(\vartheta^{\lambda}) < \mathcal{M}(\vartheta^{o}) \tag{73}$$

for every $0 < \lambda < \lambda_M$.

Taking

$$\lambda_{\text{sim}} \triangleq \min\{\epsilon_\lambda, \lambda_M, \lambda_J\} \tag{74}$$

establishes both inequalities simultaneously. □

B Experimental Details

B.1 Implementation Details

All experiments are conducted on the same hardware platform using two NVIDIA L40S GPUs, each with 45 GB of memory. The experimental settings are detailed below.

Datasets and models. We conduct image-classification experiments on ten downstream datasets using pre-trained CLIP ViT backbones. For each dataset, we use the standard training, validation, and test splits whenever available. Table 3 summarizes the datasets, number of classes, and split sizes.

Table 3: Summary of the image-classification datasets used in our experiments.

Dataset	Classes	Train	Validation	Test
CIFAR-100 (Krizhevsky, 2009)	100	45,000	5,000	10,000
Flowers102 (Nilsback & Zisserman, 2008)	102	1,020	1,020	6,149
PCAM (Veeling et al., 2018)	2	30,000	10,000	10,000
FER2013 (Goodfellow et al., 2015)	7	28,709	3,589	3,589
Cars (Krause et al., 2013)	196	7,330	814	8,041
DTD (Cimpoi et al., 2014)	47	3,384	376	1,880
GTSRB (Stallkamp et al., 2011)	43	23,976	2664	12,630
RESISC45 (Cheng et al., 2017)	45	17,010	1,890	6,300
SUN397 (Xiao et al., 2010)	397	17,865	1,985	19,850
SVHN (Netzer et al., 2011)	10	68,257	5,000	26,032

For datasets without a publicly available test split, we use the original validation split for testing and construct a new validation split from the training data. Classification accuracy is used as the evaluation metric for all datasets.

Training protocol. For the image-classification experiments, we use OpenAI-pretrained CLIP models (Radford et al., 2021), loaded through the OpenCLIP implementation (Ilharco et al., 2021), with two backbone sizes: ViT-B/32 and ViT-L/14. For each dataset, we construct a classification head from the corresponding class names and hand-crafted prompt templates. The text encoder remains frozen throughout training, such that the resulting text embeddings define a fixed linear classification head, while only the image encoder is fine-tuned. All methods use identical dataset-specific heads, data splits, preprocessing, and evaluation protocols.

Unless otherwise specified, all fine-tuning experiments use AdamW with a learning rate of 10^{-5}, batch size 128, weight decay 0.1, and a cosine learning-rate schedule with 200 warm-up steps, and are trained for 2000 optimization steps. For SCouT, the coupling update is applied at every optimization step unless otherwise stated, and the coupling strength λ is selected through a validation sweep. For all post-merge evaluations, the merging coefficients are selected using the validation set and subsequently fixed for evaluation on the held-out test set.

B.2 Optimization and Algorithm

Under the TA merging rule in (2), θ_{merge} is affine in the specialist parameters. Differentiating the softly coupled objective in (7) therefore gives the following gradient-descent recursion:

$$\begin{cases} \theta_{k,i} = \theta_{k,i-1} - \mu_i \left[\nabla J_k(\theta_{k,i-1}) + \lambda \left(\alpha_k \sum_{\ell=1}^{K} A_\ell \left(\theta_{\text{merge},i-1} - \theta_{\ell,i-1} \right) - A_k \left(\theta_{\text{merge},i-1} - \theta_{k,i-1} \right) \right) \right] \\ \theta_{\text{merge},i} \triangleq \theta_0 + \sum_{\ell=1}^{K} \alpha_\ell \left(\theta_{\ell,i} - \theta_0 \right), \end{cases}$$

where μ_i denotes the step size at iteration i. The full pseudocode of SCouT is presented in Algorithm 1. SCouT starts from the same pre-trained model θ_0 for all tasks. At each iteration, the current specialists are first combined to construct the training-time merged model (line 4). For each task, SCouT then computes the task-specific gradient and the merge-aware coupling gradient (lines 7–8), combines them, and updates the specialist parameters using the base optimizer. After training, the resulting specialists $\{\theta_k^\lambda\}_{k=1}^K$ can be combined using the prescribed merging rule to obtain a final multitask model, for example, TA gives:

$$\theta_{\text{merge}}^\lambda = \theta_0 + \sum_{k=1}^{K} \alpha_k \left(\theta_k^\lambda - \theta_0 \right). \tag{75}$$

Remark. Although the explicit recursion above is derived for TA, SCouT is not restricted to this merging rule. Any differentiable merging operator can be used to define θ_{merge} and backpropagate the corresponding coupling gradient. Moreover, we regularize the parameter discrepancy between each specialist and the merged model, rather than directly minimizing the post-merge risk $\sum_k J_k(\theta_{\text{merge}})$. The latter would require additional forward and backward evaluations of the merged model on every task at each iteration. In contrast, our coupling depends only on the current model parameters, providing a simple merge-aware coordination mechanism without additional task-loss evaluations.

Algorithm 1 SCouT: Softly Coupled Fine-Tuning

Require: Pretrained parameters θ_0, task datasets $\{\mathcal{D}_k\}_{k=1}^K$, coupling strength λ, merging coefficients $\{\alpha_k\}_{k=1}^K$, coupling matrices $\{A_k\}_{k=1}^K$, number of steps T, base optimizer (e.g., SGD or AdamW).

Ensure: Fine-tuned specialists $\{\theta_k^\lambda\}_{k=1}^K$

1: Initialize $\theta_{k,0} \leftarrow \theta_0, \forall k \in [K]$
2: **for** $i = 1, \ldots, T$ **do**
3: Sample mini-batches $\mathcal{B}_{k,i} \sim \mathcal{D}_k, \forall k \in [K]$
4: Construct merged parameters:

$$\boldsymbol{\theta}_{\text{merge},i-1} = \theta_0 + \sum_{\ell=1}^{K} \alpha_\ell (\boldsymbol{\theta}_{\ell,i-1} - \theta_0)$$

5: Compute $\boldsymbol{d}_{i-1} \triangleq \sum_{\ell=1}^{K} A_\ell (\boldsymbol{\theta}_{\text{merge},i-1} - \boldsymbol{\theta}_{\ell,i-1})$
6: **for** $k = 1, \ldots, K$ **do**
7: Compute task gradient: $\boldsymbol{g}_{k,i}^{\text{task}} \triangleq \nabla J_k(\boldsymbol{\theta}_{k,i-1}; \mathcal{B}_{k,i})$
8: Compute coupling gradient:

$$\boldsymbol{g}_{k,i}^{\text{cpl}} \triangleq \alpha_k \boldsymbol{d}_{i-1} - A_k (\boldsymbol{\theta}_{\text{merge},i-1} - \boldsymbol{\theta}_{k,i-1})$$

9: Set $\boldsymbol{g}_{k,i} \triangleq \boldsymbol{g}_{k,i}^{\text{task}} + \lambda \boldsymbol{g}_{k,i}^{\text{cpl}}$
10: **end for**
11: Update all specialists simultaneously: $\boldsymbol{\theta}_{k,i} = \text{Optimizer}(\boldsymbol{\theta}_{k,i-1}, \boldsymbol{g}_{k,i}), \ \forall k \in [K]$
12: **end for**
13: **return** $\theta_k^\lambda \leftarrow \boldsymbol{\theta}_{k,T}, \forall k \in [K]$

B.3 BASELINES

Independent fine-tuning baselines. In addition to standard independent fine-tuning, we consider some representative merger-friendly methods: Fine-Tuning in the Tangent Space (FTTS) (Ortiz-Jimenez et al., 2023), attention-only fine-tuning (FT-Attention) (Jin et al., 2025), Sharpness-Aware Fine-Tuning (SAFT) (Lee et al., 2025) and Merge-Aware Optimizer (MergOPT) (Yang et al., 2026b).

FTTS (Ortiz-Jimenez et al., 2023) linearizes the pretrained model around θ_0:

$$f_\theta^{\mathrm{lin}}(x) \triangleq f_{\theta_0}(x) + \nabla_\theta f_{\theta_0}(x)^\top (\theta - \theta_0), \tag{76}$$

and fine-tunes each task by minimizing the corresponding loss $J_k^{\mathrm{lin}}(\theta)$ induced by the linearized model. FT-Attention (Jin et al., 2025) instead performs nonlinear fine-tuning only on the attention modules, exploiting their observed kernel-like behavior to improve weight disentanglement while reducing training cost and preserving specialist performance.

SAFT (Lee et al., 2025) promotes flatter task-specific solutions through sharpness-aware fine-tuning:

$$\theta_k^{\mathrm{SAFT}} \in \arg\min_\theta J_k(\theta + \widehat{\epsilon}_k(\theta)), \qquad \widehat{\epsilon}_k(\theta) \triangleq \rho \frac{\theta^2 \nabla J_k(\theta)}{\|\nabla J_k(\theta)\|}, \tag{77}$$

where $\rho > 0$ controls the perturbation magnitude, and is selected on the validation set. MergOPT (Yang et al., 2026b) instead models the perturbation as a merge-induced parameter shift,

$$\theta_k^{\mathrm{MergOPT}} \in \arg\min_\theta \sup_{(\alpha,K,\tau)} J_k(\theta + \zeta_k(\alpha, K, \tau)), \qquad \zeta_k(\alpha, K, \tau) \triangleq (\alpha - 1)\tau_k + \alpha \sum_{j \neq k} \tau_j, \tag{78}$$

and trains each specialist to be robust to such plausible merging offsets. Following the practical implementation, the merging configurations used to construct these perturbations are sampled from the prescribed feasible set during training rather than obtained by explicitly solving an inner maximization.

Hard MTL baselines. We compare against conventional joint training, uncertainty weighting (UW) (Kendall et al., 2018), and PCGrad (Yu et al., 2020), all of which optimize a single shared parameter vector θ. Conventional joint training solves

$$\theta^{\mathrm{MTL}} \in \arg\min_\theta \sum_{k=1}^K J_k(\theta). \tag{79}$$

UW (Kendall et al., 2018) learns task-dependent uncertainty parameters:

$$\theta^{\mathrm{UW}} \in \arg\min_{\theta, \{\sigma_k\}} \sum_{k=1}^K \left(\frac{1}{2\sigma_k^2} J_k(\theta) + \log \sigma_k \right), \tag{80}$$

where $\{\sigma_k\}$ are optimized jointly with θ.

For PCGrad (Yu et al., 2020), with $g_k \triangleq \nabla J_k(\theta)$, conflicting gradients are projected as

$$\widetilde{g}_k \triangleq g_k - \frac{g_k^\top g_\ell}{\|g_\ell\|^2} g_\ell, \qquad \text{if } g_k^\top g_\ell < 0, \tag{81}$$

before forming the shared update.

Post-hoc merging rules. We evaluate weight averaging (Wortsman et al., 2022), Task Arithmetic (TA) (Ilharco et al., 2023), Fisher merging (Matena & Raffel, 2022), TIES (Yadav et al., 2023), and WUDI (Cheng et al., 2025). Unless required otherwise by the original method, merging hyperparameters are selected using the same validation protocol over $\alpha \in [0, 2]$.

Weight averaging (Wortsman et al., 2022) is given by

$$\theta^{\mathrm{Avg}} \triangleq \frac{1}{K} \sum_{k=1}^K \theta_k. \tag{82}$$

Fisher merging (Matena & Raffel, 2022) instead performs parameter-wise averaging weighted by the estimated diagonal empirical Fisher information, assigning greater weight to task-specific parameter values with higher estimated importance:

$$\theta^{\mathrm{Fisher}} = \theta_0 + \alpha_{\mathrm{Fisher}} \left(\frac{\sum_{k=1}^{K} F_k \theta_k}{\sum_{k=1}^{K} F_k} - \theta_0 \right), \tag{83}$$

where F_k is the diagonal empirical Fisher estimate for task k. We further apply a global scaling coefficient α_{Fisher} to the resulting update, and α_{Fisher} is selected on the validation set.

For TA (Ilharco et al., 2023), with $\tau_k \triangleq \theta_k - \theta_0$, we use a common scaling coefficient for computational efficiency:

$$\theta^{\mathrm{TA}} = \theta_0 + \alpha \sum_{k=1}^{K} \tau_k, \tag{84}$$

where $\alpha \in [0, 2]$ is selected on the validation set and then fixed for test evaluation.

TIES (Yadav et al., 2023) reduces interference by trimming small task-vector entries, electing a parameter-wise aggregate sign, and merging only sign-consistent updates. With the resulting task vectors $\{\widetilde{\tau}_k\}$,

$$\theta^{\mathrm{TIES}} = \theta_0 + \alpha_{\mathrm{TIES}} \sum_{k=1}^{K} \widetilde{\tau}_k, \tag{85}$$

where the trimming ratio is set to 0.2 as in the original paper and α_{TIES} is selected on the validation set.

WUDI (Cheng et al., 2025) reduces task-vector interference through layer-wise projection. It constructs task-specific parameter subspaces, removes interfering components, and aggregates the corrected updates into τ_m.

C Additional Experimental Analysis

C.1 Empirical Analysis of the Local Risk Response

Theorem 1 characterizes the local stationary solution of the softly coupled objective around the independently fine-tuned solution ϑ^o, restricted to the active Hessian subspace. In practical SCouT, however, all task models are initialized from the common pretrained model θ_0, with coupling applied throughout fine-tuning. Moreover, the theoretical reference ϑ^o assumes exact task-wise minimizers, whereas practical fine-tuning is stopped after a finite budget once performance stabilizes, rather than being driven to exact minimization and potential overfitting. Therefore, the practical checkpoints are not guaranteed to follow the local stationary conditions in Theorem 1. We thus examine whether the predicted local behavior is still reflected empirically by the practical SCouT. Specifically, we ask: *Does practical SCouT exhibit the asymmetric local risk response predicted by Theorem 1?*

Because Theorem 1 concerns empirical training risks rather than test accuracy, we evaluate the specialist and post-merge training risks across coupling strengths. The TA coefficients are fixed at $\alpha_k = 1/K$, rather than selected using validation data. Figure 6 reports results over five random seeds and shows a pronounced asymmetric response to coupling, consistent with the accuracy trends in the main text.

To test the predicted leading-order scaling, we fit the mean risk changes over the weak-coupling interval $\lambda \in [0, 0.1]$ using

$$\Delta\mathcal{J}(\lambda) = a_2 \lambda^2, \qquad \Delta\mathcal{M}(\lambda) = b_1 \lambda. \tag{86}$$

The resulting fits are shown in Figure 3 in the main text.

C.2 Compatibility with Post-hoc Merging Rules

Table 4 reports the validation-selected merging coefficients for each training method and merging rule. The optimal merge scale varies substantially across fine-tuning strategies, indicating that training alters the geometry of the task updates supplied to the merger. Because each coefficient is

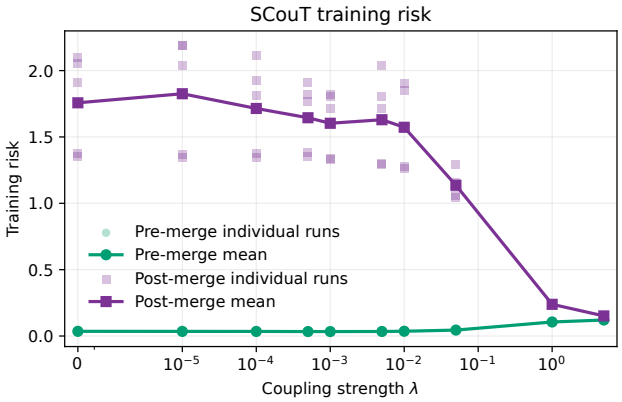

Figure 6: Specialist and post-merge training risks across coupling strengths. Markers denote individual seeds, and solid lines denote their means.

independently selected on the validation set over $\alpha \in [0, 2]$, SCouT's gains cannot be explained by a favorable shared scaling choice.

For Fisher merging, SCouT selects $\alpha = 1.15$, close to the unscaled solution $\alpha = 1$, whereas the other methods require substantially larger scaling (1.60–2.00). For TA, SCouT selects $\alpha = 0.10 = 1/K$, meaning that the uniform task-vector average used during training is already validation-optimal. TIES instead prefers a larger coefficient ($\alpha = 1.50$), reflecting its different effective update direction after pruning and sign-conflict resolution. Hence, coefficient magnitude itself should not be interpreted as a measure of mergeability.

Table 5 further shows that the preferred TA coefficient converges to the prescribed training-time value as coupling increases: it is 0.20 at $\lambda = 0$ but becomes $0.10 = 1/K$ from $\lambda = 0.2$ onward. This suggests that coupling shapes the specialists toward the prescribed merge operator, potentially eliminating the need for an additional post-hoc search over the merge scale.

Table 4: Merge coefficients on ViT-B/32 for different fine-tuning strategies and post-hoc merging rules. The coefficients are selected on the validation set and fixed for test evaluation.

Training strategy	**Fisher** α_{Fisher}	**TA** α	**TIES** α_{TIES}
Independent FT	1.90	0.20	0.80
FTTS	1.60	0.15	0.40
FT-Attention	2.00	0.30	1.30
SAFT	1.95	0.20	0.85
MergOPT	2.00	0.20	0.95
SCouT ($\lambda^\star = 0.5$)	1.15	0.10	1.50

Table 5: Validation-selected TA merge coefficients for SCouT trained with different coupling strengths λ on ViT-B/32.

Coupling strength λ	0	0.005	0.01	0.05	0.1	0.2	0.5	1	5	10	20
TA coefficient α	0.20	0.20	0.20	0.20	0.20	0.10	0.10	0.10	0.10	0.10	0.10

C.3 Loss Landscape Visualization

Figures 7 and 8 extend the loss-landscape analysis in the main text to CIFAR100–Flowers102 and GTSRB–SVHN. The same qualitative behavior is observed across both task pairs: as the coupling strength increases, the two specialist task vectors become more aligned and the best TA merge moves into a lower average-loss region. Moreover, for $\lambda > 0$, the resulting TA merge reaches loss levels

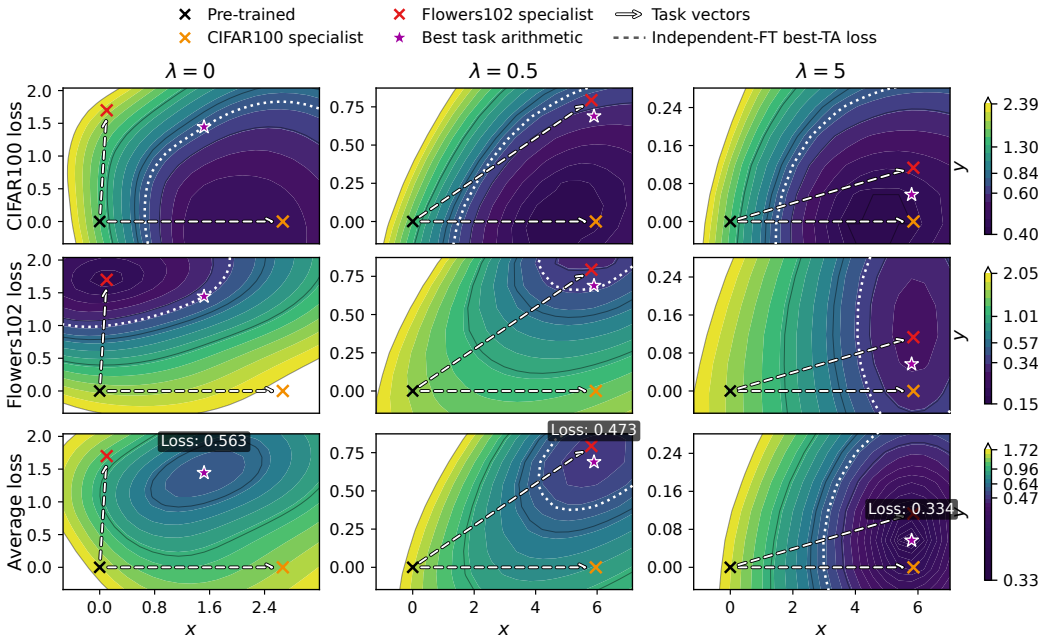

Figure 7: Loss-landscape visualization for CIFAR100 and Flowers102 at different coupling strengths.

below the best-TA loss obtained from independently fine-tuned specialists, as indicated by the white reference contours. These results reinforce the main-text observation that SCouT does not merely bring the specialists closer, but reshapes their task vectors toward a more compatible geometry in which the prescribed post-hoc merge can access lower-loss solutions.

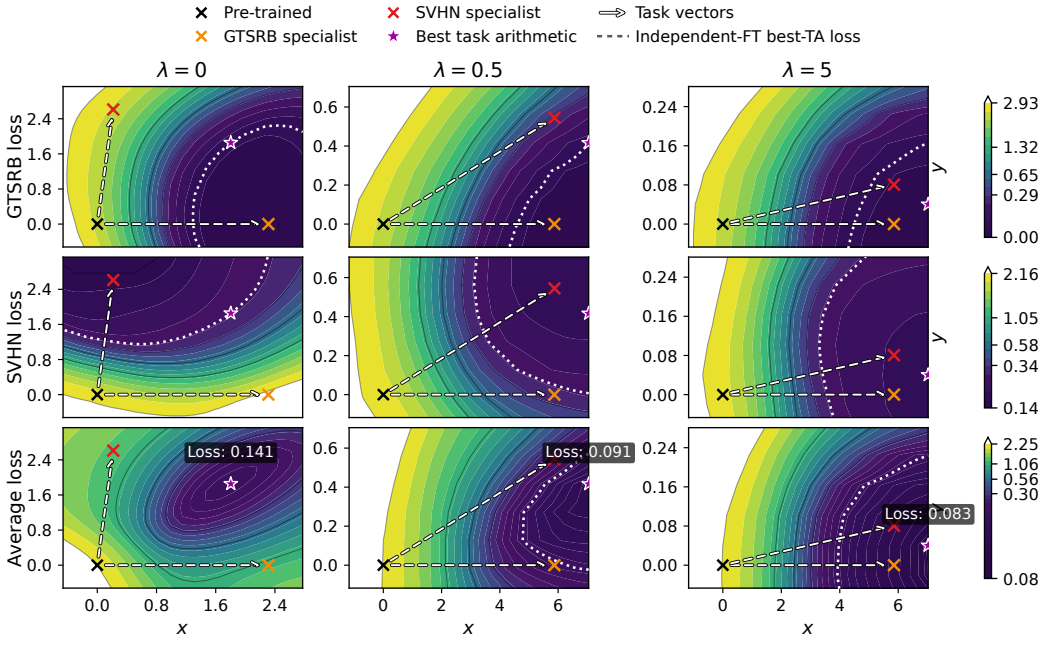

Figure 8: Loss-landscape visualization for GTSRB and SVHN at different coupling strengths.

D Evaluation on Natural Language Processing Tasks

To evaluate the generality of SCouT beyond vision, we further conduct experiments on natural language processing (NLP) tasks from the GLUE benchmark.

D.1 Experimental Setup

Unless otherwise stated, we follow the same general evaluation protocol as in the vision experiments. All experiments are conducted on the same hardware platform with two NVIDIA L40S GPUs.

Datasets and models. We evaluate on six classification tasks from the GLUE benchmark (Wang et al., 2018): SST-2 (Socher et al., 2013), CoLA (Warstadt et al., 2019), MRPC (Dolan & Brockett, 2005), QQP (Iyer et al., 2017), QNLI (Rajpurkar et al., 2016), and MNLI (Williams et al., 2018). These tasks span single-sentence classification, linguistic acceptability, paraphrase detection, and natural-language inference. We use the pretrained RoBERTa-base model (Liu et al., 2019) as the common encoder initialization for all tasks. Inputs are tokenized using the corresponding RoBERTa tokenizer and truncated to a maximum sequence length of 128.

As in the vision experiments, the official GLUE validation sets are reserved for final evaluation, while a validation subset is held out from each task's original training set for model and hyperparameter selection. Table 6 summarizes the resulting data splits.

Table 6: Summary of the GLUE tasks used in our experiments. For MNLI, the two evaluation sizes correspond to the matched and mismatched validation sets, respectively.

Task	Classes	Train	Validation	Test
SST-2 (Socher et al., 2013)	2	62,349	5,000	872
CoLA (Warstadt et al., 2019)	2	7,696	855	1,043
MRPC (Dolan & Brockett, 2005)	2	3,301	367	408
QQP (Iyer et al., 2017)	2	358,846	5,000	40,430
QNLI (Rajpurkar et al., 2016)	2	99,743	5,000	5,463
MNLI (Williams et al., 2018)	3	387,702	5,000	9,815 / 9,832

Evaluation metrics. Following the standard GLUE evaluation protocol, we report Matthews correlation for CoLA; the mean of accuracy and F1 score for MRPC and QQP; and accuracy for SST-2, QNLI, and MNLI. For MNLI, matched and mismatched accuracies are reported separately, and their arithmetic mean is used as the task-level score. All task scores are expressed on a 0–100 scale, and the aggregated GLUE score is computed as the unweighted average across the six tasks.

Task-specific classification heads. Unlike CLIP, pretrained RoBERTa does not provide task-specific classification heads for the GLUE tasks. We therefore associate each task with a separate classification head consisting of dropout, a fully connected layer with $\tanh$ activation, a second dropout layer, and an output projection. For each task, the same head initialization is used across all compared methods. The encoder and task-specific head are jointly optimized during fine-tuning, whereas the heads are neither coupled nor merged. After model merging, the merged encoder is evaluated together with each task's corresponding learned head, without any additional adaptation.

Training protocol. Unless otherwise specified, all methods are trained using AdamW with a learning rate of 2×10^{-5}, batch size 32, weight decay 0.01, and a linear learning-rate schedule with a warm-up ratio of 0.06. Gradients are clipped independently for each task to a maximum ℓ_2-norm of 1.0, and training is performed for 8000 optimization steps. As in the CLIP experiments, SCouT uses the same isotropic coupling and uniform TA coefficients shown in (16).

D.2 Results and Analysis

As in the CLIP experiments, we report the specialist score, the merged- or shared-model score, and the normalized score, which measures merged-model performance relative to that of the corresponding specialists.

Effect of coupling strength. Figure 9 shows the same overall trend as in the vision experiments. As λ increases, the specialist score decreases only gradually, whereas the merged-model score improves substantially and approaches the hard-MTL baseline. The validation and test curves exhibit closely matched trends, indicating that the improvement is not specific to the evaluation split.

We select $\lambda^\star = 0.1$ using only the validation results, restricting the relative decrease in specialist score compared with independent FT to at most 1.5% while achieving a substantial improvement in the merged-model score.

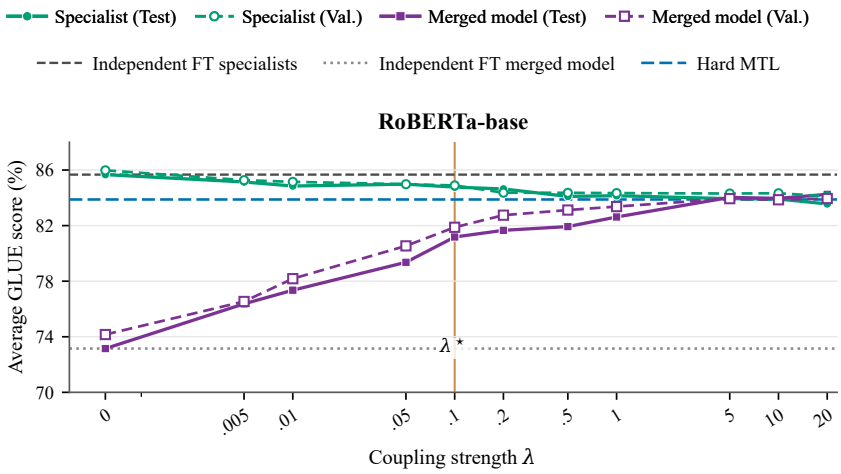

Figure 9: Effect of the coupling strength λ on SCouT for RoBERTa-base on GLUE.

Comparison of training strategies. As shown in Table 7, SCouT at $\lambda^\star = 0.1$ improves the merged score from 73.15 to 81.18, an absolute gain of 8.03 points, while reducing the specialist score by only 0.90 points (1.05% relative). SCouT also exceeds all other specialist-based baselines in merged performance. With stronger coupling ($\lambda = 5$), SCouT attains a merged score of 84.03 and a normalized score close to 100, matching the hard-MTL baselines.

Table 7: Comparison of training strategies on RoBERTa-base for GLUE.

Training strategy	Specialist Score	Merged/shared-model Score	Norm. Score
Independent FT	85.67 (0%)	73.15 (0%)	83.99
FTTS	76.94	67.97	87.87
FT-Attention	84.13	72.74	85.05
SAFT	85.78	70.40	79.65
MergOPT	85.53	74.16	85.60
Hard MTL	–	83.88	–
Hard MTL + UW	–	83.62	–
Hard MTL + PCGrad	–	84.19	–
SCouT ($\lambda = 0.01$)	84.86 (-0.95%)	77.35 (+5.74%)	90.45
SCouT ($\lambda^\star = 0.1$)	84.77 (-1.05%)	81.18 (+10.98%)	95.51
SCouT ($\lambda = 5$)	83.94 (-2.02%)	84.03 (+14.87%)	100.35

E Training Cost Comparison

We compare training efficiency using the same task assignment and data exposure. Each global step processes one equal-sized minibatch from each of the K tasks. For specialist-based methods—Independent FT, FTTS, FT-Attention, SAFT, MergOPT, and SCouT—the K models are assigned round-robin to G GPUs. With $K = 10$ and $G = 2$, each GPU holds five specialists and

updates them sequentially. Global-step time is the maximum elapsed time for one global training step across GPUs. For SCouT, after computing task-specific gradients, merged parameters are obtained through bucketed all-reduce Li et al. (2020). Hard MTL instead uses one shared encoder replica per GPU, accumulates gradients across local tasks, and synchronizes them using Distributed Data Parallel Li et al. (2020).

For ViT-B/32, we report the mean wall-clock time per global step (time/step), peak GPU memory usage (VRAM), and training time relative to Independent FT. Timing includes data transfer, forward and backward passes, method-specific operations, optimizer updates, and inter-GPU communication, but excludes data loading and model construction. Peak VRAM is the maximum allocated memory across GPUs after warm-up.

Table 8: Training cost comparison of fine-tuning methods for ViT-B/32. All methods are evaluated under the same configuration using two GPUs.

Method	Time/step (s) ↓	Rel. time ↓	VRAM (GB) ↓
Independent FT	0.696	$1.00\times$	9.4
FTTS	1.407	$2.02\times$	14.5
FT-Attention	0.560	$0.80\times$	6.1
SAFT	1.374	$1.98\times$	9.8
MergOPT	0.787	$1.13\times$	9.8
SCouT	**0.775**	$\mathbf{1.11}\times$	**11.7**
Hard MTL	0.684	$0.98\times$	5.8

As shown in Table 8, SCouT introduces only an 11% runtime overhead relative to Independent FT despite performing cross-task coupling. Under the same task assignment, its additional VRAM usage ($+2.3$ GB) primarily arises from retaining multiple task gradients until the coupling update, together with temporary coupling and optimizer buffers. This overhead decreases as K/G, the number of tasks assigned to each GPU, becomes smaller—either by reducing the number of tasks or by using more GPUs. To examine this effect under our two-GPU setup, we compare the VRAM usage of Independent FT and SCouT across different numbers of tasks ($K \in \{2, 4, 6, 8, 10\}$).

Table 9: Peak VRAM usage (GB) of Independent FT and SCouT as the number of tasks K varies. All experiments use ViT-B/32 and two GPUs.

Method	$K = 10$ $(K/G = 5)$	$K = 8$ $(K/G = 4)$	$K = 6$ $(K/G = 3)$	$K = 4$ $(K/G = 2)$	$K = 2$ $(K/G = 1)$
Independent FT	9.4	8.3	7.3	6.2	5.1
SCouT	**11.7**	**10.1**	**8.5**	**6.9**	**5.3**

Table 9 shows that SCouT's peak VRAM decreases from 11.7 GB at $K/G = 5$ to 5.3 GB at $K/G = 1$, approaching Independent FT's 5.1 GB. This confirms that its overhead mainly comes from retaining gradients for multiple local specialists and can be reduced by distributing tasks across more GPUs.

F AI USAGE STATEMENT

We used generative AI tools to polish the writing, identify relevant literature, and assist with debugging and refining author-designed experimental code. The authors reviewed all AI-assisted text and code, verified the cited sources and experimental results, and take full responsibility for the final content.