

DAMM: Decoupled Adaptive Model Merging with Coordinate-Wise Trust Regions

Anonymous Authors¹

Abstract

Merging task-specific experts from a shared backbone enables efficient multi-task deployment, yet existing methods often underperform due to cross-task parameter conflicts. We propose DAMM (Decoupled Adaptive Model Merging), a high-fidelity model merging framework that partitions the parameter space into a shared mergeable subspace and task-specific residuals, thereby reducing interference while preserving task functionality. At test time, DAMM employs a distillation-based procedure to jointly learn an adaptive mask and a shared subspace. The resulting shared update is constrained by coordinate-wise trust regions defined by the task-specific experts, which suppresses harmful extrapolation and stabilizes merging across diverse tasks. This decoupling also makes the remaining task-specific residuals easy to store: although DAMM keeps residuals for each task, they are highly quantization-resilient. On a 20-task ViT-B/32 merge, 2-bit residual quantization retains 96.4% of full-precision performance (86.3% vs. 89.5%) at a cost of only 26.0 MB in extra storage. Across 20 vision and 7 NLP tasks spanning ViT, RoBERTa, and GPT-2, DAMM matches over 98% of the corresponding task-specific accuracy and outperforms strong merging baselines by up to 11% in absolute accuracy. Code will be released upon acceptance.

1. Introduction

In recent years, large-scale pretrained models have provided general-purpose representations and strong initialization, enabling the community to obtain high-performing task-specific expert models at relatively low cost via fine-tuning (Ilharco et al., 2022b; Wortsman et al., 2022b; Paul & Chen,

2022; Zhou et al., 2025). As the number of tasks continues to grow, the paradigm of “one fine-tuned expert per task, stored and deployed separately” imposes an increasingly pronounced systemic burden: many task experts share the same pretrained backbone capability, yet exist as independent parameter copies, causing the costs of model storage, distribution, and version maintenance to grow with task scale, and fragmenting capabilities across multiple isolated experts, which hinders efficient reuse.

Against this backdrop, model merging provides a direct path: it aims to integrate multiple task experts into a single model, thereby reusing existing expert capabilities without retraining and reducing the storage and deployment overhead of maintaining many models in parallel (Wortsman et al., 2022a; Matena & Raffel, 2022; Ortiz-Jimenez et al., 2023; Dimitriadis et al., 2023). In common settings, task-specific capability can be represented as the parameter delta of a fine-tuned model relative to the pretrained model (i.e., a task vector). However, when multiple task vectors are aggregated directly, *parameter conflicts* often arise: different tasks produce inconsistent or even opposite updates along certain parameter dimensions, pushing the merged weights away from the low-loss basin around the pretrained solution and leading to substantial performance degradation (Ainsworth et al., 2022; Xu et al., 2024; Stoica et al., 2023).

To mitigate parameter conflicts, prior work follows two main directions. Conflict-aware pruning improves robustness by removing conflict parameters (Yadav et al., 2023; Yu et al., 2024; Huang et al., 2024; Du et al., 2024; Wang et al., 2024). Coefficient-based fusion learns task- or layer-wise weights, enabling more flexible composition in a *data-free* scenario (Matena & Raffel, 2022; Jin et al., 2022; Yang et al., 2023; Akiba et al., 2025).

Building on these paradigms, several recent studies in the supplementary material have proposed more specialized strategies to further refine the merging process. For instance, SASA enhances representation capacity by generating high-rank updates through sparse fine-tuning in the spectral domain, while CASS adopts a pruning-centric perspective, utilizing contribution-based metrics to mask specific functional units like attention heads. However, when tasks are highly heterogeneous, stronger suppression typically stabilizes ag-

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

gregation but can also attenuate task-specific knowledge (Marczak et al., 2025; Stoica et al., 2023; Qu & Horvath, 2024). Moreover, when labeled data are unavailable, adaptation merging without explicit geometric constraints can drift away from the low-loss region near experts, leading to performance drops (Frankle et al., 2020; Ortiz-Jimenez et al., 2023; Sun et al., 2025b). This leaves an open gap for approaches that preserve expert specialization while maintaining stable merging under data-free constraints.

To address these challenges, we propose DAMM, which formulates model merging as a dual process of structured parameter selection and constrained updates. DAMM learns a structural mask to partition parameters into two distinct components: a shared subspace for cross-task fusion and a task-specific residual for isolating non-mergeable features. This decomposition aggregates cross-task consensus while shielding conflict-prone dimensions, effectively mitigating the interference caused by indiscriminate mixing (Marczak et al., 2025; Stoica et al., 2023; Qu & Horvath, 2024). Furthermore, we introduce an expert-induced *coordinate-wise trust region* (CWTR) to stabilize the merging process. By limiting update magnitudes within the mergeable space, this constraint prevents the model from drifting away from the low-loss regions of individual experts (Frankle et al., 2020; Ortiz-Jimenez et al., 2023; Sun et al., 2025b). Finally, we observe that the task-private residuals exhibit a highly compact distribution. This property allows for high-fidelity performance even under low-bit representations, significantly reducing the incremental storage cost for multi-task scaling.

In summary, our main contributions are three-fold:

- We propose DAMM, a high-fidelity merging framework that decomposes expert updates into a *shared mergeable component* and *task-specific residuals*. This structured decoupling explicitly isolates conflict-prone dimensions during aggregation, mitigating cross-task parameter conflicts while better preserving task-specialized functionality.
- We introduce an expert-induced CWTR constraint to guide the optimization of the shared mergeable update. By restricting updates to remain within the range supported by the input experts, the constraint reduces uncontrolled drift from the expert low-loss region, improving stability under substantial task heterogeneity.
- On 27 tasks (20 vision, 7 NLP) across ViT, RoBERTa, and GPT-2, DAMM attains $> 98\%$ expert-relative accuracy on average and improves over strong merging baselines by up to 11% in absolute accuracy.

2. Related Work

Weight Aggregation and Conflict Mitigation. Early model merging mainly uses weighted averaging to balance tasks (Wortsman et al., 2022a; Ilharco et al., 2022a). Later work improves robustness via importance-aware weighting (Matena & Raffel, 2022) and fine-grained adaptation at the layer/sample/parameter level (Yang et al., 2023; Lee et al., 2025; Ye et al., 2025; Jin et al., 2022; Camacho et al., 2024). To further mitigate interference, pruning and sign-consensus methods have been proposed to filter conflicting parameters (Yadav et al., 2023; Yu et al., 2024; Huang et al., 2024; Du et al., 2024; Sun et al., 2025a). However, these approaches still yield a single merged model. Under systematic conflicts, they enforce stronger compromises and may discard task-specific parameters, degrading fidelity. Recent work therefore separates task updates in weight space: AWD (Xiong et al., 2024) extracts a redundant component to encourage task-vector orthogonality. In contrast, DAMM decomposes experts into a mergeable shared subspace and task-specific residuals, isolating interference without irreversible parameter removal.

Representation Shift. Model merging can perturb fine-tuned representation distributions, leading to representation bias and performance degradation (Du et al., 2024; Yan et al., 2025; Nobari et al., 2025). Existing solutions typically fall into two categories: (i) *post-hoc repair* methods that recalibrate statistics or add alignment steps (e.g., REPAIR (Jordan et al., 2022) and Surgery (Yang et al., 2024a;b)), and (ii) *procedural interventions* that suppress conflicts during merging (e.g., MaTS (Tam et al., 2023) and APGD (Wei et al., 2025)). Unlike pipelines that depend on separate repair/alignment stages, DAMM internalizes representation stability within a test-time adaptation objective (Wang et al., 2020; 2022; Du et al., 2025; Tan et al., 2025). Concretely, it adapts the merged model on unlabeled streams by aligning task-general knowledge into the shared space while limiting unnecessary changes to task-specific components, thereby mitigating representation drift with low overhead.

Geometric Constraints. From a loss-landscape perspective, merging succeeds when the merged solution lies in a low-loss region supported by the expert models. Re-basin (Ainsworth et al., 2022) and its variants (Imfeld et al., 2023; Xu et al., 2024; Rinaldi et al., 2025) remove geometric barriers via permutation alignment, while FW Merging (Chen et al., 2025) uses the Frank-Wolfe algorithm to iteratively search for feasible solutions in restricted model spaces. Building on this line of work, DAMM introduces expert-defined feasible regions and enforces per-parameter convex-hull constraints, so that test-time adaptation updates remain within expert-induced parameter regions by construction. Compared to coarse-grained global constraints, this fine-grained geometric control improves merging stability

and robustness under strong distribution shifts.

3. Decoupled Adaptive Model Merging

3.1. Problem Setup and Task Vectors

Consider a pretrained model with parameters $\theta_0 \in \mathbb{R}^d$. Fine-tuning θ_0 on T downstream tasks yields a set of task-specific experts $\{\theta_t\}_{t=1}^T$. We define the task vector for task t as:

$$\tau_t = \theta_t - \theta_0. \quad (1)$$

Task-conditional setting. We focus on a task-conditional scenario where each unlabeled test stream is associated with a known task identity t . Consequently, our objective is not to derive a single unified parameter vector θ_{merged} , but rather to construct a family of task-conditional models $\{\theta_t^{\text{DAMM}}\}_{t=1}^T$ with a shared subspace. While task identity acquisition is a critical component of end-to-end systems, it remains orthogonal to our proposed parameterization and adaptation mechanisms. To isolate the intrinsic effects of model merging and test-time adaptation, we follow the standard convention of assuming task identities are provided at inference time.

Task-agnostic linear merging baseline. A common task-agnostic baseline merges multiple experts into a single parameter vector θ_{merged} , which is defined as follows:

$$\theta_{\text{merged}} = \theta_0 + \sum_{t=1}^T \lambda_t \tau_t, \quad \sum_{t=1}^T \lambda_t = 1, \quad \lambda_t \geq 0. \quad (2)$$

This formulation compresses disparate task updates into a unified parameter space, often leading to suboptimal performance when tasks conflict.

3.2. Decoupled Parameterization

A key challenge in cross-task merging is that task vectors $\{\tau_t\}_{t=1}^T$ may conflict in the parameter space. As a result, naive averaging can cancel useful information and degrade expert performance. To mitigate this issue, we decompose each task vector, coordinate-wise, into a mergeable shared component and a non-mergeable task-specific residual.

We introduce a learnable binary mask $M \in \{0, 1\}^d$ to select which coordinates are shared. When $M_i = 1$, the i -th coordinate is assigned to the shared component; when $M_i = 0$, it remains task-specific. For task t , DAMM defines the task-conditional parameters as follows:

$$\theta_t^{\text{DAMM}} = \theta_0 + M \odot V_{\text{shared}} + (1 - M) \odot \tau_t, \quad (3)$$

where $V_{\text{shared}} \in \mathbb{R}^d$ is a shared space and \odot denotes element-wise multiplication. To provide a warm start, V_{shared} is initialized as the mean of the task vectors, $\frac{1}{T} \sum_{t=1}^T \tau_t$, and is subsequently refined on the unlabeled target stream.

Parameter efficiency To achieve coordinate-wise control, a straightforward alternative involves maintaining per-task gating vectors $u_t \in \mathbb{R}^d$, which is expressed as follows:

$$\theta = \theta_0 + \sum_{t=1}^T u_t \odot \tau_t. \quad (4)$$

This strategy incurs an additional parameter cost of $\mathcal{O}(Td)$ along with substantial adaptation overhead. In contrast, DAMM utilizes a single shared vector V_{shared} and mask logits as detailed in Section 3.3. This design yields a total overhead of $\mathcal{O}(d)$ that remains independent of the number of tasks T .

3.3. Adaptive Mask Learning with STE

Since binary masks are non-differentiable, we parameterize the selection process using learnable logits $\phi \in \mathbb{R}^d$. We first compute a continuous approximation of the mask by applying the sigmoid function σ to these logits. The soft mask is defined as:

$$M_s = \sigma(\phi). \quad (5)$$

Based on this continuous representation, we derive a discrete hard mask through a fixed thresholding operation. This transformation is expressed as:

$$M_h = \mathbb{I}(M_s > 0.5), \quad (6)$$

where \mathbb{I} denotes the indicator function. To facilitate discrete selection in the forward pass while maintaining gradient flow for backpropagation, we employ the straight-through estimator (STE), and the resulting mask is formulated as:

$$M = M_s + \text{sg}[M_h - M_s], \quad (7)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation which prevents gradient updates from flowing through the M_h .

3.4. Coordinate-Wise Trust Regions

Unlabeled test-time adaptation can cause drift in V_{shared} , which may degrade the fidelity of $\{\theta_t^{\text{DAMM}}\}_{t=1}^T$. To mitigate this effect, we impose a CWTR induced by $\{\tau_t\}_{t=1}^T$ and project each update back to the region. For each coordinate $i \in \{1, \dots, d\}$, we first aggregate the extrema across $\{\tau_t\}_{t=1}^T$ as follows:

$$\tau_i^{\min} = \min_t \tau_{t,i}, \quad \tau_i^{\max} = \max_t \tau_{t,i}. \quad (8)$$

We then define the center and radius as follows:

$$c_i = \frac{\tau_i^{\min} + \tau_i^{\max}}{2}, \quad r_i = \frac{\tau_i^{\max} - \tau_i^{\min}}{2}, \quad (9)$$

Based on these, we construct an α -scaled interval:

$$\mathcal{C}_i(\alpha) = [c_i - \alpha r_i, c_i + \alpha r_i], \quad \mathcal{C}(\alpha) = \prod_{i=1}^d \mathcal{C}_i(\alpha). \quad (10)$$

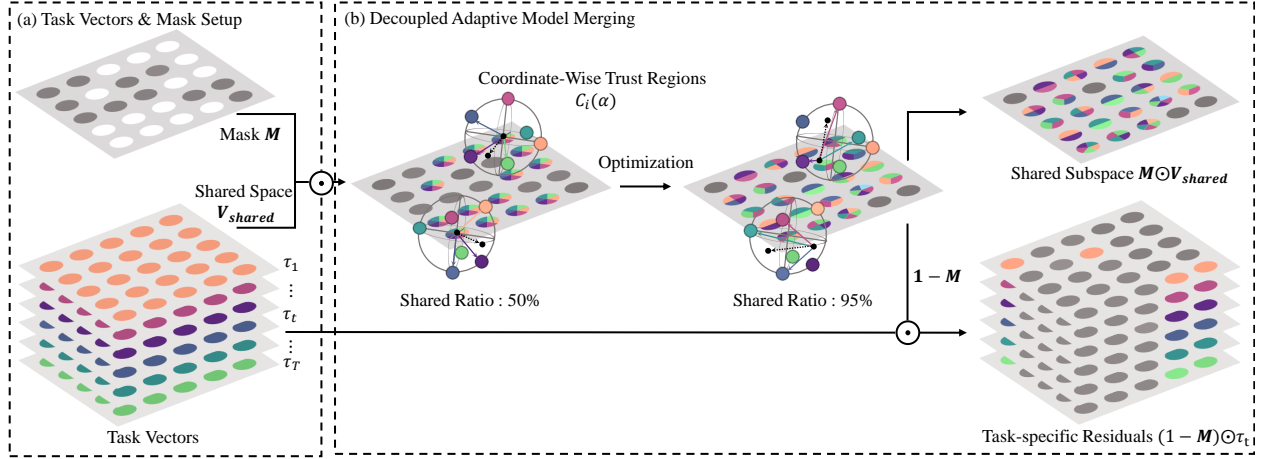


Figure 1. Overview of the DAMM framework. (a) A learned binary mask M partitions parameters into shared and task-specific regions to mitigate inter-task interference. (b) The shared space V_{shared} is optimized subject to expert constraints, facilitating cross-task transfer.

Here, $\alpha > 0$ scales the trust region, with $\alpha < 1$ shrinking the region and $\alpha > 1$ expanding it.

Motivation and advantages. This construction is motivated by two considerations. First, any coordinate-wise convex combination of $\{\tau_t\}_{t=1}^T$ with non-negative weights summing to one must lie between the per-coordinate minimum and maximum across tasks. Therefore, \mathcal{C} contains the coordinate-wise convex hull and provides a conservative safe set for unlabeled adaptation. Second, unlike task- or layer-level merging with a single coefficient per task or layer, CWTR allow parameter-level flexibility. They enable sharing and adapting transferable coordinates while constraining conflicting ones within expert-induced bounds, thereby reducing negative transfer.

Projected update. Given the current V_{shared} , the optimization process begins with an unconstrained gradient step as follows:

$$\tilde{V} = V_{\text{shared}} - \eta \nabla \mathcal{L}(V_{\text{shared}}), \quad (10)$$

and the resulting intermediate vector is subsequently projected onto $\mathcal{C}(\alpha)$ in the following manner:

$$V_{\text{shared}} \leftarrow \Pi_{\mathcal{C}(\alpha)}(\tilde{V}), \quad (11)$$

$$[\Pi_{\mathcal{C}(\alpha)}(v)]_i = \min\{\max\{v_i, c_i - \alpha r_i\}, c_i + \alpha r_i\}. \quad (12)$$

This projection is a closed-form element-wise clipping operation with a computational complexity of $\mathcal{O}(d)$.

3.5. Objective: Distillation and Regularization

On the unlabeled target stream, the shared parameters V_{shared} and the mask logits ϕ are jointly optimized.

Efficient teacher selection. Under the task-conditional setting, the corresponding task expert θ_t serves as the teacher for distillation. Consequently, each update step requires only a single teacher forward pass. This approach ensures that the teacher computation remains constant and independent of the total number of tasks T .

Feature distillation. To enable V_{shared} to effectively acquire and integrate shared knowledge across multiple tasks, the model is trained to mimic the feature representations of task-specific experts. Let $g(x; \theta) \in \mathbb{R}^m$ denote the final output feature of the visual encode. The distillation process is performed by matching these features through an ℓ_1 loss as follows:

$$\mathcal{L}_{\text{distill}} = \|g(x; \theta_t^{\text{DAMM}}) - g(x; \theta_t)\|_1, \quad (13)$$

where the teacher feature $g(x; \theta_t)$ is treated as a constant during optimization to prevent gradient propagation through the teacher network.

Density regularization. The mask density determines the capacity of the shared subspace. Specifically, an insufficient number of shared coordinates leads to the incomplete utilization of shared knowledge across tasks, whereas an excessive number may result in parameter interference and compromise the fidelity of specific tasks. To balance these competing factors, the expected shared ratio is encouraged to remain near a target density ρ as follows:

$$\mathcal{L}_{\text{dens}} = (\text{mean}(M_s) - \rho)^2. \quad (14)$$

Entropy regularization. In single-stage joint optimization, M_{soft} may saturate too early toward 0/1, degrading

STE gradients and leading to suboptimal collapsed masks. To keep sufficient uncertainty and plasticity during optimization, a binary entropy regularizer is incorporated in the following manner:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{d} \sum_{i=1}^d \left(M_{s,i} \log M_{s,i} + (1 - M_{s,i}) \log(1 - M_{s,i}) \right). \quad (15)$$

Overall objective. The final optimization objective is defined as the weighted sum of the aforementioned components as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{dens}} \mathcal{L}_{\text{dens}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}}. \quad (16)$$

where λ_{dens} and λ_{ent} are hyperparameters controlling the regularization strength.

4. Experiments

We evaluate DAMM across diverse NLP and CV benchmarks to verify its efficacy across various architectures and task scales. We compare our approach against two categories of baselines: (i) Static Merging (e.g., Weight Averaging (Wortsman et al., 2022a), Task Arithmetic (TA) (Ilharco et al., 2022a), and Ties-Merging (Yadav et al., 2023), CART (Choi et al., 2024), Consensus (Wang et al., 2024), TSV-M (Gargiulo et al., 2025)); and (ii) Adaptive Merging (e.g., AdaMerging (Yang et al., 2023) and AdaRank (Lee et al., 2025)). Due to space constraints, we report additional comparisons to Fisher Merging (Matena & Raffel, 2022) and RegMean (Jin et al., 2022) in Appendix A. Comprehensive details regarding task setups, the full suite of baselines, and hyperparameter configurations are also deferred to Appendix A.

4.1. Main Results on Natural Language Processing

Setup. We evaluate DAMM on two model families: RoBERTa-base (encoder-only) (Liu et al., 2019) and GPT-2 (decoder-only) (Radford et al., 2019). Following FusionBench (Tang et al., 2024), we select 7 GLUE tasks and use the publicly released single-task fine-tuned checkpoints as experts under a unified *data-free* protocol. More experimental details are provided in the appendix.

Results and Analysis. Table 1 shows that DAMM achieves the highest average among the compared methods on both backbones (84.0% on RoBERTa; 75.9% on GPT-2), close to the per-task expert oracle (84.8% and 76.8%, respectively). The largest gains appear on CoLA, where baselines can collapse (e.g., AdaMerging drops to -3.6% on RoBERTa), while DAMM reaches 60.7%, comparable to

the individual expert (60.2%). Relative to the strongest adaptive baseline, CART+AdaRank (Choi et al., 2024), DAMM improves the average by +9.8 *percentage points* (pp) on RoBERTa (74.2%→84.0%) and +11.0 pp on GPT-2 (64.9%→75.9%), indicating consistent benefits across two different architectures. DAMM remains slightly below per-task experts on a few tasks (e.g., GPT-2 QQP/RTE), but produces a more balanced merged model.

4.2. Main Results on Computer Vision

Setup. We evaluate DAMM on two ViT backbones, ViT-B/32 and ViT-L/14 (Dosovitskiy, 2020). Following the standard multi-task model merging protocol (Wang et al., 2020), we merge expert checkpoints independently fine-tuned on diverse downstream classification tasks. To vary task diversity and merging difficulty, we consider task suites of 8, 14, and 20 tasks.

Results and Analysis. Table 2 shows that DAMM achieves the best average performance among merged models across all settings and remains close to the individual-expert. In the 20-task suite, DAMM preserves 98.7% and 99.6% of the oracle average accuracy (ViT-B/32: 89.5 vs. 90.7; ViT-L/14: 93.3 vs. 93.7), indicating strong scalability as the number of tasks grows. In contrast, static merging degrades substantially as the number of tasks increases. For example, on ViT-B/32, CART drops by 7.9 pp from 8 to 20 tasks (84.7%→76.8%). In the same setting, DAMM remains stable and widens its margin over CART from 5.9 pp to 12.7 pp. Among adaptive baselines, DAMM also improves over CART+AdaRank, with a 3.1 pp gain on ViT-B/32 in the 20-task setting (89.5% vs. 86.4%). Per-task results and additional comparisons are provided in Appendix B (Tables 7–12).

4.3. Hold-out Compositional Generalization

A central goal of DAMM is to learn a task-agnostic shared space V_{shared} from in-distribution (ID) experts. This space is then transferred to out-of-distribution (OOD) tasks unseen during the construction of V_{shared} . We evaluate this capability via task hold-out generalization.

Setup. For any task t , the merged task vector is given by $\Delta W_t = M \odot V_{\text{shared}} + (1 - M) \odot \tau_t$, where M is the mask learned in Eq. 7, and $(1 - M) \odot \tau_t$ is the ID or OOD task-specific residual associated with the current inference task. For ViT-B/32, we run 6 random 8-of-20 splits (8 ID / 12 OOD). For GPT-2, we run 7 random 4-of-7 splits (4 ID / 3 OOD). We further evaluate the effect of CWTR in the hold-out compositional generalization (Tables 13 and 14)).

ViT-B/32: CWTR is critical for held-out compositional generalization. On ViT-B/32, CWTR is essential for

Table 1. Multi-task performance on 7 NLP tasks with RoBERTa (R) and GPT-2 (G) backbones. Individual denotes single-task experts. The best results are marked in **bold**.

Method	CoLA		SST2		MRPC		QQP		MNLI		QNLI		RTE		Average	
	R	G	R	G	R	G	R	G	R	G	R	G	R	G	R	G
Individual	60.2	40.8	94.0	91.2	89.2	80.4	91.4	89.6	87.2	82.0	92.7	88.3	79.1	65.3	84.8	76.8
Weight Averaging	18.1	12.1	81.9	52.5	77.9	51.0	79.6	76.7	43.8	59.3	71.1	57.6	61.7	44.8	62.0	50.6
TA	23.3	-0.2	86.6	83.6	78.7	69.6	84.0	81.8	63.7	71.9	73.0	70.5	61.0	47.3	67.2	60.6
Ties-Merging	25.0	3.3	83.5	81.8	78.7	68.4	85.2	82.8	60.7	74.3	75.8	69.6	42.2	47.7	64.4	61.1
CART	30.9	11.4	92.0	86.2	80.9	54.7	79.5	81.8	57.7	70.1	77.7	76.2	71.1	52.4	70.0	61.8
AdaMerging	-3.6	5.9	92.7	79.8	77.2	70.8	82.2	81.0	78.8	68.5	79.6	67.6	66.4	46.2	67.6	60.0
TA+AdaRank	14.0	6.2	91.5	88.2	77.7	62.8	79.6	79.4	78.1	75.4	84.1	81.3	67.2	49.8	70.3	63.3
CART+AdaRank	36.4	11.5	92.8	88.5	77.2	64.7	79.6	79.8	77.5	74.5	87.5	84.7	68.2	50.2	74.2	64.9
DAMM (Ours)	60.7	38.6	94.0	91.2	89.0	80.2	88.4	88.2	86.4	81.0	92.2	88.2	77.3	64.3	84.0	75.9

Table 2. Average multi-task performance on 8, 14, 20 vision tasks with merged ViT-B/32 and ViT-L/14. The best results are marked in **bold**.

Method	ViT-B/32			ViT-L/14		
	8	14	20	8	14	20
Pretrained	48.0	59.6	56.0	65.0	68.4	65.4
Individual	91.1	89.9	90.7	94.4	93.5	94.2
<i>Static Merging Methods</i>						
Weight Averaging	65.9	64.3	60.9	79.6	76.8	71.7
TA	69.2	65.4	61.0	84.5	79.6	74.2
Ties-Merging	72.4	65.2	62.9	86.1	79.5	75.8
Consensus-Ties	74.8	68.2	62.9	87.2	81.5	78.8
Consensus-TA	75.2	70.0	65.0	86.6	81.9	77.6
TSV-M	83.8	79.5	76.7	91.2	88.3	87.3
CART	84.7	79.5	76.8	92.6	88.0	87.9
<i>Adaptive Merging Methods</i>						
AdaMerging	80.1	76.7	69.2	90.8	88.0	86.8
TA+AdaRank	87.9	82.1	81.4	92.9	89.4	89.1
CART+AdaMerging	85.9	82.3	82.7	93.1	90.4	91.3
CART+AdaRank	89.2	86.2	86.4	93.4	91.4	91.8
DAMM (Ours)	90.6	89.0	89.5	94.0	93.0	93.2

OOD composition ((Table 3)). Across 6 splits, the mean OOD accuracy increases from 24.0 ± 6.5 without CWTR ($17.1\text{--}31.3\%$) to 80.3 ± 3.8 with CWTR ($76.3\text{--}85.7\%$). For example, in Exp1, OOD accuracy improves from 30.9% to 82.9%. The gains are also consistent for the mean accuracy over all 20 tasks, which rises from 42.5 ± 6.2 to 84.2 ± 2.0 , and for ID tasks, which improve from 70.3 ± 10.0 to 90.0 ± 1.4 . A likely cause of these sharp OOD failures without CWTR is that an unconstrained V_{shared} can drift away from expert-supported regions, leading to represen-

tation shift when it is composed with τ_t under the fixed mask M . By constraining V_{shared} coordinate-wise to stay at the expert-induced envelope, CWTR makes the resulting composition more stable on held-out tasks.

Table 3. Hold-out generalization performance on ViT-B/32. Mean accuracy (%) for ID, OOD, and all tasks over 6 random splits, with and without CWTR. The best results are marked in **bold**.

DAMM	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5	Sp.6	Mean \pm Std
8 ID, w/o	76.6	80.2	70.3	63.7	76.5	54.2	70.3 ± 10.0
8 ID, w/	90.1	90.5	88.6	92.2	88.9	89.9	90.0 ± 1.4
12 OOD, w/o	30.9	20.0	19.9	31.3	24.9	17.1	24.0 ± 6.5
12 OOD, w/	82.9	76.3	85.7	76.6	82.2	78.1	80.3 ± 3.8
All 20, w/o	49.2	44.0	40.0	44.2	45.5	32.0	42.5 ± 6.2
All 20, w/	85.8	82.0	86.9	82.8	84.9	82.8	84.2 ± 2.0

GPT-2: CWTR yields minor gains and mainly improves robustness. On GPT-2, CWTR has a small effect (Table 4). The mean OOD accuracy is essentially unchanged (75.5 ± 1.9 without CWTR vs. 75.6 ± 1.9 with CWTR), with ranges of $72.3\text{--}77.6\%$ and $72.1\text{--}77.7\%$, respectively. The mean accuracy over all 7 tasks also differs only slightly, increasing from 73.5 ± 1.7 to 73.8 ± 1.3 . Consistent with our experimental observations, GPT-2 fusion relies more on a high-density shared component across most layers, while residuals contribute little in many layers; consequently, V_{shared} tends to stay near a good solution region in the full parameter space even without CWTR, so CWTR mainly provides a mild robustness constraint.

4.4. Effect of Decoupling on Merging

We ablate decoupling to assess how removing the explicit residuals affects merging, and we also examine the role of CWTR in stabilizing this non-decoupled variant.

Table 4. Hold-out generalization performance on GPT-2. Mean accuracy (%) for ID, OOD, and all tasks over 7 random splits, with and without CWTR. The best results are marked in **bold**.

DAMM	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5	Sp.6	Sp.7	Mean±Std
4 ID w/o	74.8	67.2	74.1	72.0	68.8	73.7	73.7	72.0±2.9
4 ID w/	74.2	68.0	74.0	72.6	71.5	73.5	73.7	72.5±2.2
3 OOD w/o	76.5	76.8	74.1	72.3	76.6	74.6	77.6	75.5±1.9
3 OOD w/	76.3	77.1	74.2	72.1	76.5	75.0	77.7	75.6±1.9
All 7 w/o	75.5	71.3	74.1	72.2	72.1	74.1	75.4	73.5±1.7
All 7 w/	75.1	71.9	74.1	72.4	73.6	74.2	75.4	73.8±1.3

Non-decoupled merging. We consider a baseline without decoupling that applies the shared space V_{shared} alone and does not use the task-specific residuals $(1 - M) \odot \tau_t$. Table 5 reports mean accuracy (%) over 6 random 8-of-20 splits, and per-task results are provided in Appendix B (Table 15). Merging without decoupling is unstable on held-out tasks. Without CWTR, the mean OOD accuracy is 14.4 ± 5.9 , and *All 20* drops to 22.0% in Exp6. Adding CWTR substantially improves robustness, increasing the mean OOD accuracy to 49.6 ± 6.5 and the *All 20* mean to 59.8 ± 3.7 , with Exp6 improving from 22.0% to 58.6%. The per-task results in Appendix B further show that CWTR mitigates severe failures on multiple OOD tasks. Despite these gains, merging without decoupling remains far below the complete DAMM model under the same split protocol (Table 3: 82.0–86.9% on *All 20*). This gap suggests that CWTR mainly prevents degenerate merged solutions, while decoupling and task-specific residuals are critical for reliable composition across diverse experts.

Table 5. Hold-out generalization performance on ViT-B/32. Mean accuracy (%) for ID, OOD, and all tasks over 6 random splits, evaluated under non-decoupled configurations with and without CWTR. The best results are marked in **bold**.

Non-decoupled	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5	Sp.6	Mean±Std
8 ID, w/o	62.1	64.5	58.6	42.4	65.8	37.1	55.1±12.5
8 ID, w/	73.1	78.9	68.2	78.1	78.3	74.2	75.1±4.4
12 OOD, w/o	23.0	9.9	12.6	19.4	9.4	12.0	14.4±5.9
12 OOD, w/	55.4	38.3	50.7	49.4	55.6	48.1	49.6±6.5
All 20, w/o	38.6	31.7	31.0	28.6	32.0	22.0	30.6±5.4
All 20, w/	62.5	54.5	57.7	60.9	64.7	58.6	59.8±3.7

4.5. Sensitivity Analysis of Fusion Hyperparameters

To understand the stability of our fusion procedure and guide hyperparameter selection, we conduct a sensitivity analysis on the two key controls: the density target p and the trust-region scale α . Figure 3 analyzes how p and α

affect both the average accuracy (Avg Acc) and the mean fusion ratio on ViT-B/32 across eight tasks (SUN397, Cars, RESISC45, EuroSAT, SVHN, GTSRB, MNIST, and DTD). Decreasing p steadily improves Avg Acc, reaching its best performance around $p \in [0.4, 0.7]$, but it simultaneously suppresses fusion: the Mean Fusion Ratio remains high for $p \geq 0.6$ and drops sharply once $p \leq 0.5$. Varying α reveals a more pronounced trade-off: while larger α monotonically increases the Mean Fusion Ratio toward saturation ($\approx 99\%$), Avg Acc is stable for $\alpha \leq 1$ and then degrades rapidly for $\alpha > 10$. Overall, these trends indicate a clear performance–fusion trade-off, with a robust operating region around $p \approx 0.6$ – 0.7 and $\alpha \leq 1$ that maintains near-peak accuracy while preserving substantial fusion. Unless otherwise specified, we use $p = 0.8$ and $\alpha = 1$ in all experiments.

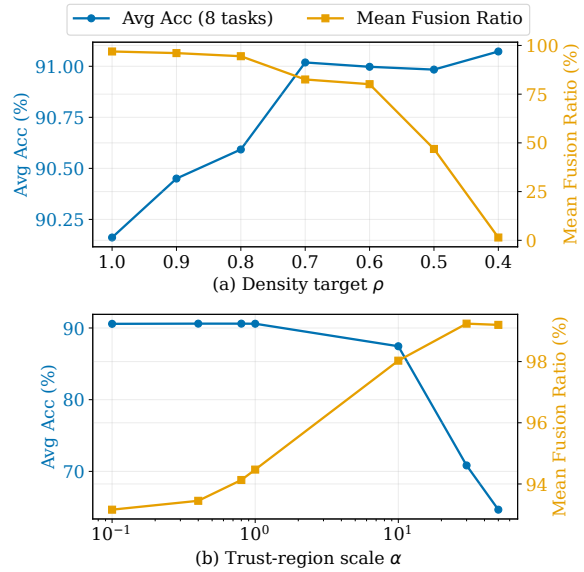


Figure 3. Hyperparameter sensitivity of the density target p and trust-region scale α

4.6. Analyzing Fusion Ratios and Sensitive Layers

To identify where task-specific capacity is most critical, we examine layer-wise fusion amenability as task density increases. We define the *fusion ratio* as the fraction of fusible coordinates per layer, and the *mean fusion ratio* as its average across all layers.

Figure 2 shows that ViT-B/32 exhibits extensive sharing, with the mean fusion ratio consistently exceeding 93%. However, increasing task density induces a steady decline in fusibility ($94.47\% \rightarrow 93.97\%$), primarily driven by MLP and projection modules. For instance, `resblocks.7.mlp_proj` exhibits a significant drop of $\Delta = -7.7\%$, identifying it as a primary bottleneck for reliable merging. Notably, this sensitivity is capacity-dependent: comparative analysis with ViT-L/14 (see Appendix B) reveals that larger

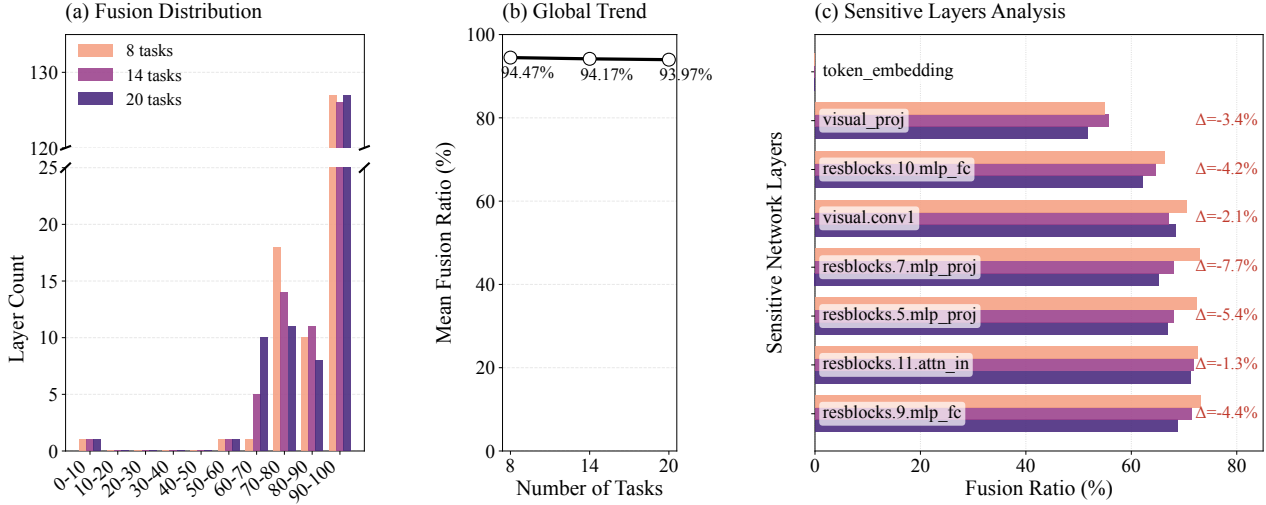


Figure 2. Fusion dynamics of ViT-B/32 across task densities: (a) layer-wise fusion-ratio distribution; (b) decreasing mean fusion ratio as task density increases; (c) the most sensitive layers with significant shifts, Δ denotes the ratio change from 8 to 20 tasks (red for decrease).

models maintain a more robust and expansive shared subspace ($M \odot V_{\text{shared}}$) under high task density. This suggests that increased model capacity inherently mitigates parameter conflicts through more flexible overlap, whereas smaller models require more structured task-specific isolation to maintain stability.

4.7. Quantization of Task Vectors and Residuals

DAMM stores a task-specific residual for each task. This makes per-task storage a scalability bottleneck. To mitigate this issue, we quantize the task representation using uniform per-tensor asymmetric min-max quantization. We evaluate ViT-B/32 on 8, 14, and 20 tasks. We compare Full Quantization (FQ) of task vectors with Residual Quantization (RQ) of task-specific residuals. As shown in Table 6, RQ is more robust at 2-bit. It improves over FQ by 0.22, 0.29, and 0.32 pp in average accuracy. This is expected because residuals typically have a narrower effective range, which reduces distortion under aggressive quantization. The gap vanishes at 3-bit and above. Quantizing residuals to 2-bit gives a theoretical $16\times$ reduction in per-task residual storage relative to FP32, excluding quantization metadata. It also remains close to CART+AdaRank. The gaps are 0.40, 0.03, and 0.15 pp for 8, 14, and 20 tasks. For ViT-B/32 (≈ 344 MB FP32), using mean fusion ratios of 94.47%/94.17%/93.97%, the 2-bit residual overhead totals only 9.5/17.6/26.0 MB for 8/14/20 tasks, which is negligible relative to the backbone.

5. Conclusion

In this paper, we propose DAMM, a high-fidelity model merging framework. To mitigate cross-task interference, DAMM decomposes model parameters into a mergeable

Table 6. Quantization performance (Avg. Acc %) on ViT-B/32. FQ and RQ denote Full and Residual Quantization, respectively.

Tasks	Method	2-bit	3-bit	4-bit	8-bit
8	FQ	88.58	90.66	90.61	90.59
	RQ	88.80	90.66	90.60	90.59
14	FQ	85.88	89.04	89.02	89.00
	RQ	86.17	89.04	89.02	89.01
20	FQ	85.93	89.49	89.48	89.47
	RQ	86.25	89.48	89.50	89.47

shared subspace and lightweight task-specific residuals. Evaluated on 20 vision and 7 NLP tasks across architectures, DAMM achieves performance close to task-specific experts and outperforms strong baselines. Our analysis shows that most layers exhibit high fusion ratios, while a small set of MLP modules becomes increasingly sensitive as the number of tasks grows, indicating that task-specific corrections are localized. Ultimately, DAMM establishes a practical roadmap for continual model merging, demonstrating that a shared subspace augmented with task-specific residuals can effectively sustain high performance on new, OOD tasks.

Impact Statement

The DAMM framework enhances multi-task integration in both performance and resource efficiency by reframing model merging as a structured decoupling process. By isolating a universal shared subspace and retaining only minimal task-specific residuals, our approach achieves strong performance recovery while substantially reducing the storage overhead associated with large-scale model ensembles. This

supports broader goals in environmental sustainability and Green AI. Moreover, this “shared-plus-residual” paradigm provides a robust foundation for continual model merging, enabling foundation-model systems to evolve and integrate new functionalities at minimal resource cost. As a foundational study focused on architectural optimization, we do not anticipate any immediate negative societal impacts.

References

- Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries, 2022. URL <https://arxiv.org/abs/2209.04836>, 2022.
- Akiba, T., Shing, M., Tang, Y., Sun, Q., and Ha, D. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204, 2025.
- Bossard, L., Guillaumin, M., and Van Gool, L. Food-101—mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.
- Camacho, A. M. O., Horoi, S., Wolf, G., and Belilovsky, E. Non-uniform parameter-wise model merging. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 5946–5954. IEEE, 2024.
- Chen, H. M., Hu, S. X., Luk, W., Hospedales, T., and Fan, H. Fw-merging: Scaling model merging with frank-wolfe optimization. *arXiv preprint arXiv:2503.12649*, 2025.
- Cheng, G., Han, J., and Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- Choi, J., Kim, D., Lee, C., and Hong, S. Revisiting weight averaging for model merging. *arXiv preprint arXiv:2412.12153*, 2024.
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.
- Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.
- Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.
- Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.
- Dagan, I., Glickman, O., and Magnini, B. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.
- Dimitriadis, N., Frossard, P., and Fleuret, F. Pareto manifold learning: Tackling multiple tasks via ensembles of single-task models. In *International Conference on Machine Learning*, pp. 8015–8052. PMLR, 2023.
- Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*, 2005.
- Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Du, G., Lee, J., Li, J., Jiang, R., Guo, Y., Yu, S., Liu, H., Goh, S. K., Tang, H.-K., He, D., et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.
- Du, Y., Wang, X., Chen, C., Ye, J., Wang, Y., Li, P., Yan, M., Zhang, J., Huang, F., Sui, Z., et al. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9413–9422, 2025.
- Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.
- Gargiulo, A. A., Crisostomi, D., Bucarelli, M. S., Scardapane, S., Silvestri, F., and Rodola, E. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18695–18705, 2025.
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013.
- Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

- Huang, C., Ye, P., Chen, T., He, T., Yue, X., and Ouyang, W. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024.
- Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022a.
- Ilharco, G., Wortsman, M., Gadre, S. Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., and Schmidt, L. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35: 29262–29277, 2022b.
- Imfeld, M., Galdi, J., Giordano, M., Hofmann, T., Anagnostidis, S., and Singh, S. P. Transformer fusion with optimal transport. *arXiv preprint arXiv:2310.05719*, 2023.
- Jin, X., Ren, X., Preotiuc-Pietro, D., and Cheng, P. Data-less knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.
- Jordan, K., Sedghi, H., Saukh, O., Entezari, R., and Neyshabur, B. Repair: Renormalizing permuted activations for interpolation repair. *arXiv preprint arXiv:2211.08403*, 2022.
- Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Lee, C., Choi, J., Lee, C., Kim, D., and Hong, S. Adarank: Adaptive rank pruning for enhanced model merging. *arXiv preprint arXiv:2503.22178*, 2025.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Marczak, D., Magistri, S., Cygert, S., Twardowski, B., Bagdanov, A. D., and van de Weijer, J. No task left behind: Isotropic model merging with common and task-specific subspaces. *arXiv preprint arXiv:2502.04959*, 2025.
- Matena, M. S. and Raffel, C. A. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.
- Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.
- Nobari, A. H., Alim, K., ArjomandBigdeli, A., Srivastava, A., Ahmed, F., and Azizan, N. Activation-informed merging of large language models. *arXiv preprint arXiv:2502.02421*, 2025.
- Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.
- Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.
- Paul, S. and Chen, P.-Y. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.
- Qu, X. and Horvath, S. Vanishing feature: Diagnosing model merging and beyond. *arXiv preprint arXiv:2402.05966*, 2024.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- Rinaldi, F., Capitani, G., Bonicelli, L., Crisostomi, D., Bolelli, F., Ficarra, E., Rodola, E., Calderara, S., and Porrello, A. Update your transformer to the latest release: Re-basin of task vectors. *arXiv preprint arXiv:2505.22697*, 2025.
- Sharma, L., Graesser, L., Nangia, N., and Evci, U. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In

- 550 *Proceedings of the 2013 conference on empirical methods*
551 *in natural language processing*, pp. 1631–1642, 2013.
- 552 Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. The
553 german traffic sign recognition benchmark: a multi-class
554 classification competition. In *The 2011 international joint*
555 *conference on neural networks*, pp. 1453–1460. IEEE,
556 2011.
- 557 Stoica, G., Bolya, D., Bjorner, J., Ramesh, P., Hearn, T., and
558 Hoffman, J. Zipit! merging models from different tasks
559 without training. *arXiv preprint arXiv:2305.03053*, 2023.
- 560 Sun, W., Li, Q., Geng, Y.-a., and Li, B. Cat merging: A
561 training-free approach for resolving conflicts in model
562 merging. *arXiv preprint arXiv:2505.06977*, 2025a.
- 563 Sun, W., Li, Q., Wang, W., Geng, Y., and Li, B. Task
564 arithmetic in trust region: A training-free model merging
565 approach to navigate knowledge conflicts. In *Proceedings*
566 *of the 33rd ACM International Conference on Multimedia*,
567 pp. 5178–5187, 2025b.
- 568 Tam, D., Bansal, M., and Raffel, C. Merging by match-
569 ing models in task parameter subspaces. *arXiv preprint*
570 *arXiv:2312.04339*, 2023.
- 571 Tan, M., Chen, G., Wu, J., Zhang, Y., Chen, Y., Zhao,
572 P., and Niu, S. Uncertainty-calibrated test-time model
573 adaptation without forgetting. *IEEE Transactions on*
574 *Pattern Analysis and Machine Intelligence*, 2025.
- 575 Tang, A., Shen, L., Luo, Y., Hu, H., Du, B., and Tao, D. Fu-
576 sionbench: A comprehensive benchmark of deep model
577 fusion. *arXiv preprint arXiv:2406.03280*, 2024.
- 578 Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and
579 Welling, M. Rotation equivariant cnns for digital pathol-
580 ogy. In *International Conference on Medical image com-*
581 *puting and computer-assisted intervention*, pp. 210–218.
582 Springer, 2018.
- 583 Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell,
584 T. Tent: Fully test-time adaptation by entropy minimiza-
585 tion. *arXiv preprint arXiv:2006.10726*, 2020.
- 586 Wang, K., Dimitriadis, N., Ortiz-Jimenez, G., Fleuret, F.,
587 and Frossard, P. Localizing task information for im-
588 proved model merging and compression. *arXiv preprint*
589 *arXiv:2405.07813*, 2024.
- 590 Wang, Q., Fink, O., Van Gool, L., and Dai, D. Continual test-
591 time domain adaptation. In *Proceedings of the IEEE/CVF*
592 *Conference on Computer Vision and Pattern Recognition*,
593 pp. 7201–7211, 2022.
- 594 Warstadt, A., Singh, A., and Bowman, S. R. Neural network
595 acceptability judgments. *Transactions of the Association*
596 *for Computational Linguistics*, 7:625–641, 2019.
- 597 Wei, Y., Tang, A., Shen, L., Hu, Z., Yuan, C., and Cao, X.
598 Modeling multi-task model merging as adaptive projec-
599 tive gradient descent. *arXiv preprint arXiv:2501.01230*,
600 2025.
- 601 Williams, A., Nangia, N., and Bowman, S. A broad-
602 coverage challenge corpus for sentence understanding
603 through inference. In *Proceedings of the 2018 confer-*
604 *ence of the North American chapter of the association for*
computational linguistics: human language technologies,
volume 1 (long papers), pp. 1112–1122, 2018.
- Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R.,
Gontijo-Lopes, R., Morcos, A. S., Namkoong, H.,
Farhadi, A., Carmon, Y., Kornblith, S., 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, 2022a.
- Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith,
S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A.,
Namkoong, H., et al. Robust fine-tuning of zero-shot
models. In *Proceedings of the IEEE/CVF conference on*
computer vision and pattern recognition, pp. 7959–7971,
2022b.
- Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a
novel image dataset for benchmarking machine learning
algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., and Oliva,
A. Sun database: Exploring a large collection of scene
categories. *International Journal of Computer Vision*,
119(1):3–22, 2016.
- Xiong, F., Cheng, R., Chen, W., Zhang, Z., Guo, Y.,
Yuan, C., and Xu, R. Multi-task model merging
via adaptive weight disentanglement. *arXiv preprint*
arXiv:2411.18729, 2024.
- Xu, Z., Yuan, K., Wang, H., Wang, Y., Song, M., and Song, J.
Training-free pretrained model merging. In *Proceedings*
of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition, pp. 5915–5925, 2024.
- Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal,
M. Ties-merging: Resolving interference when merging
models. *Advances in Neural Information Processing*
Systems, 36:7093–7115, 2023.
- Yan, K., Zhang, M., Cui, S., Qu, Z., Jiang, B., Liu, F., and
Zhang, C. Calm: Consensus-aware localized merging
for multi-task learning. *arXiv preprint arXiv:2506.13406*,
2025.

- Yang, E., Wang, Z., Shen, L., Liu, S., Guo, G., Wang, X., and Tao, D. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023.
- Yang, E., Shen, L., Wang, Z., Guo, G., Chen, X., Wang, X., and Tao, D. Representation surgery for multi-task model merging. *arXiv preprint arXiv:2402.02705*, 2024a.
- Yang, E., Shen, L., Wang, Z., Guo, G., Wang, X., Cao, X., Zhang, J., and Tao, D. Surgeryv2: Bridging the gap between model merging and multi-task learning with deep representation surgery. *arXiv preprint arXiv:2410.14389*, 2024b.
- Yann, L. The mnist database of handwritten digits. *R*, 1998.
- Ye, P., Huang, C., Shen, M., Chen, T., Huang, Y., and Ouyang, W. Dynamic model merging with mixture of weights. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.
- Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 16(12):9851–9915, 2025.

A. Experimental Details

A.1. Implementation Details

All experiments are conducted on a single NVIDIA RTX A6000 GPU (48GB) using a unified codebase. All task experts directly utilize publicly available fine-tuned checkpoints. Each task is represented by a task vector, defined as the parameter difference between the expert and a shared pre-trained backbone. During the optimization process, all task-specific classification heads are frozen; we only optimize the adaptive masks and shared subspaces following the objective in Eq. 16. We set the regularization strengths to $\lambda_{\text{dens}} = 10^{-2}$ and $\lambda_{\text{ent}} = 10^{-4}$, with a target density of $\rho = 0.8$. Optimization is performed using the Adam optimizer ($\beta = (0.9, 0.999)$) for 6000 iterations, with distinct learning rates of 1×10^{-4} for shared parameters and 1×10^{-3} for mask parameters. We employ a batch size of 8 per task, a validation batch size of 64, and automatic mixed precision (bfloat16) to improve throughput.

A.2. Benchmarks and Evaluation Protocol

Computer Vision. For vision benchmarks, we follow standard model merging evaluation protocols using CLIP Vision Transformer backbones (ViT-B/32 and ViT-L/14) (Radford et al., 2021). We evaluate across three scaling scenarios, ensuring fair comparison by adopting benchmark configurations from prior work (Lee et al., 2025):

- **8-task benchmark:** Consists of Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), SVHN (Netzer et al., 2011), GTSRB (Stallkamp et al., 2011), MNIST (Yann, 1998), SUN397 (Xiao et al., 2016), and RESISC45 (Cheng et al., 2017).
- **14-task benchmark:** Extends the 8-task benchmark with CIFAR100 (Krizhevsky et al., 2009), STL10 (Coates et al., 2011), Flowers102 (Nilsback & Zisserman, 2008), OxfordIIITPet (Parkhi et al., 2012), PCAM (Veeling et al., 2018), and FER2013 (Goodfellow et al., 2013).
- **20-task benchmark:** Further incorporates EMNIST (Cohen et al., 2017), CIFAR10 (Krizhevsky et al., 2009), Food101 (Bossard et al., 2014), FashionMNIST (Xiao et al., 2017), RenderedSST2 (Socher et al., 2013), and KMNIST (Clanuwa et al., 2018).

Natural Language Processing. Following the setting from FusionBench (Tang et al., 2024), we merge seven fine-tuned weights for text classification tasks: CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013), MRPC (Dolan & Brockett, 2005), QQP (Sharma et al., 2019), MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), and RTE (Dagan et al., 2005). We report the Matthews correlation coefficient for CoLA and accuracy for all other tasks.

B. Detailed Experimental Results

A comparative analysis of Figure 2 and Figure 4 shows a clear capacity-dependent difference in merging dynamics between ViT-B/32 (86M) and ViT-L/14 (304M). For ViT-B/32, the mean merging ratio decreases monotonically as task density increases (94.47% \rightarrow 94.17% \rightarrow 93.97%), and the most sensitive layers exhibit consistently negative changes from 8 to 20 tasks (e.g., resblocks.7.mlp_proj: $\Delta = -7.7\%$; resblocks.5.mlp_proj: $\Delta = -5.4\%$). In contrast, ViT-L/14 maintains a higher and more robust merging ratio with a non-monotonic trend (93.32% \rightarrow 95.7% \rightarrow 94.62%). This spike at 14 tasks is likely influenced by the particular task set composition, but the overall behavior remains more robust as task density increases. ViT-L/14 also shows both negative and positive shifts in sensitive layers, most notably token_embedding with $\Delta = +49.8\%$. Overall, these results suggest that increased model capacity supports a larger V_{shared} , making merging less sensitive to higher task density.

Table 7. Multi-Task performance comparison on 8 Vision Tasks with Merged ViT-B/32.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
Pretrained	62.3	59.7	60.7	45.5	31.4	32.6	48.5	43.8	48.1
Individual	79.5	78.2	95.9	99.9	97.6	99.1	99.7	79.0	91.1
Traditional MTL	73.9	74.4	93.9	98.2	95.8	98.9	99.5	77.9	89.1
Weight Averaging	65.2	63.4	71.5	71.9	64.2	52.8	87.5	50.7	65.9
Fisher Merging	68.6	69.2	70.7	66.4	72.9	51.1	87.9	59.9	68.3
RegMean	65.3	63.5	75.6	78.6	78.1	67.4	93.7	52.0	71.8
Task Arithmetic	55.2	54.9	66.7	78.9	80.2	69.7	97.3	50.4	69.2
Ties-Merging	59.8	58.6	70.7	79.7	86.2	72.1	98.3	54.2	72.5
Consensus-Ties	62.5	61.8	76.3	81.6	82.0	80.5	97.3	56.0	74.8
Consensus-TA	63.9	62.2	76.1	84.2	84.2	76.6	97.4	57.5	75.2
TSV-M	67.2	70.8	86.3	94.6	91.0	92.3	99.3	68.9	83.8
CART	68.5	73.0	88.3	95.8	87.8	93.4	99.1	72.1	84.7
AdaMerging	64.5	68.1	79.2	93.8	87.0	91.9	97.5	59.1	80.1
AdaMerging++	66.6	68.3	82.2	94.2	89.6	89.0	98.3	60.6	81.1
TA+AdaRank	71.1	79.1	91.3	97.2	94.2	98.3	99.2	72.7	87.9
CART+AdaMerging	69.5	75.1	89.3	95.7	93.0	96.8	98.9	68.4	85.8
CART+AdaRank	72.1	78.9	93.3	98.4	95.6	98.8	99.4	76.9	89.2
DAMM (Ours)	78.1	76.0	95.3	99.8	97.5	99.2	99.7	79.2	90.6

Table 8. Multi-Task performance comparison on 14 Vision Tasks with Merged ViT-B/32. The best results are highlighted in **bold**.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST
Pretrained	62.3	59.7	60.7	45.5	31.4	32.6	48.5
Individual	79.5	78.2	95.9	99.9	97.6	99.1	99.7
Weight Averaging	64.2	60.7	67.2	64.6	49.4	43.5	76.2
Task Arithmetic	63.9	59.5	67.5	67.7	52.9	47.0	80.8
Ties-Merging	65.1	61.8	68.3	63.7	51.3	45.9	80.0
Consensus-Ties	63.6	58.8	69.7	71.9	56.2	61.2	88.3
Consensus-TA	62.8	54.8	68.5	76.0	69.3	63.0	93.5
TSV-M	66.3	62.1	81.2	91.7	82.4	83.6	98.8
CART	68.3	60.6	86.1	91.3	72.7	82.6	98.1
AdaMerging	64.3	68.5	81.7	92.6	86.6	90.8	97.5
TA+AdaRank	69.2	77.3	91.3	95.9	94.1	97.1	99.1
CART+AdaMerging	67.4	72.5	87.8	96.0	90.9	95.6	98.6
CART+AdaRank	70.7	77.0	91.1	98.7	94.4	97.8	99.3
DAMM (Ours)	77.1	75.2	94.6	99.6	97.4	99.2	99.7

Method	DTD	CIFAR100	FER2013	Flowers102	OxfordIIITPet	PCAM	STL10
Pretrained	43.8	64.2	39.0	66.3	87.4	60.6	97.1
Individual	79.0	89.3	72.9	90.4	91.3	87.9	98.0
Weight Averaging	47.2	69.8	41.6	68.2	88.1	61.9	97.2
Task Arithmetic	48.2	69.6	42.9	67.6	87.5	63.2	96.7
Ties-Merging	48.7	69.7	42.4	68.1	88.0	62.1	97.2
Consensus-Ties	51.8	67.9	45.4	65.7	86.2	72.3	45.3
Consensus-TA	52.4	66.6	45.3	68.3	86.9	77.0	95.6
TSV-M	64.6	72.0	62.3	75.3	90.4	84.5	97.2
CART	67.7	75.6	60.9	81.6	90.2	79.7	97.6
AdaMerging	60.2	67.3	53.1	73.8	87.9	53.8	96.3
TA+AdaRank	72.5	75.6	45.4	82.1	92.2	59.5	97.5
CART+AdaMerging	71.2	71.5	60.0	80.5	87.8	75.6	96.3
CART+AdaRank	75.6	79.4	61.4	89.1	91.7	83.0	97.7
DAMM (Ours)	78.6	88.0	70.4	87.9	91.4	88.8	98.2

Table 9. Multi-Task performance comparison on 20 Vision Tasks with Merged ViT-B/32. The best results are highlighted in **bold**.

Method	SUN397	Cars	RESISC 45	EuroSAT	SVHN	GTSRB	MNIST	DTD	CIFAR 100	FER 2013
Pretrained	62.3	59.7	60.7	45.5	31.4	32.6	48.5	43.8	64.2	39.0
Individual	79.5	78.2	95.9	99.9	97.6	99.1	99.7	79.0	89.3	72.9
Weight Averaging	59.6	46.0	56.3	41.3	70.0	64.6	64.0	69.3	66.9	66.5
Task Arithmetic	64.1	59.4	64.6	56.6	47.3	41.4	70.5	46.2	69.2	41.0
Ties-Merging	64.5	57.0	68.8	59.4	48.7	48.0	78.3	49.5	70.6	43.3
Consensus-Ties	64.4	58.9	67.2	54.4	51.1	47.9	77.5	48.4	67.6	96.3
Consensus-TA	63.6	52.5	65.3	64.1	63.1	52.6	88.0	49.1	65.6	42.0
TSV-M	64.3	52.0	75.9	87.1	75.2	76.8	94.6	61.1	68.1	58.2
CART	65.3	38.1	81.3	88.7	70.0	77.4	96.2	64.6	73.7	59.9
AdaMerging	62.1	66.3	78.7	92.1	72.7	90.6	93.6	57.6	66.3	48.4
TA+AdaRank	68.1	74.4	90.7	95.6	92.0	96.0	96.9	68.5	75.6	43.8
CART+AdaMerging	67.3	71.2	86.3	96.6	88.3	95.0	96.4	71.2	72.4	54.4
CART+AdaRank	69.5	75.7	91.7	97.6	93.6	96.8	97.4	73.4	77.6	61.2
DAMM (Ours)	76.3	74.5	93.9	99.0	97.4	99.1	99.7	78.4	87.7	69.4

Method	Flowers 102	Oxford IIITPet	PCAM	STL10	EMNIST	CIFAR10	Food101	Fashion MNIST	Rendered SST2	KMNIST
Pretrained	66.3	87.4	60.6	97.1	17.2	89.8	82.6	63.0	58.6	9.8
Individual	90.4	91.3	87.9	98.0	99.8	97.9	89.1	95.3	74.4	98.6
Weight Averaging	87.6	62.2	40.8	31.6	92.8	81.1	70.8	60.5	8.5	47.5
Task Arithmetic	66.7	87.7	62.4	96.9	32.9	92.7	81.1	70.7	60.4	8.7
Ties-Merging	71.6	85.3	64.4	96.0	39.9	93.5	75.9	72.7	64.7	12.4
Consensus-Ties	67.1	86.9	67.0	42.8	41.0	92.4	79.4	74.9	60.8	11.4
Consensus-TA	66.4	85.9	72.6	95.4	53.4	92.3	75.1	74.7	62.7	15.6
TSV-M	71.2	88.6	84.5	96.4	95.3	93.8	77.3	85.4	70.2	57.2
CART	77.6	87.7	74.8	97.0	93.1	94.8	77.1	86.5	68.6	63.4
AdaMerging	65.7	87.0	54.6	96.6	21.5	90.5	80.7	82.9	65.0	10.8
TA+AdaRank	75.5	91.9	60.7	97.3	95.9	94.6	83.8	91.0	69.3	66.3
CART+AdaMerging	79.6	86.4	80.0	96.9	95.1	91.5	79.8	82.7	70.0	92.5
CART+AdaRank	85.6	91.7	83.8	97.5	96.6	94.8	83.6	91.0	73.5	96.0
DAMM (Ours)	86.5	91.6	88.1	98.3	99.7	97.9	88.1	94.4	71.2	98.3

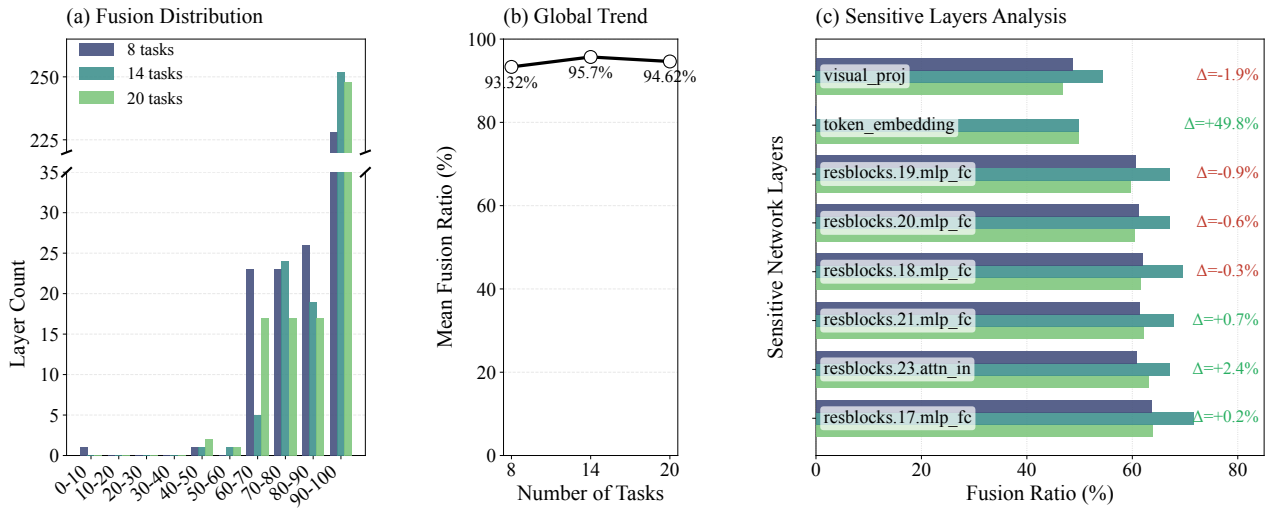


Figure 4. Fusion dynamics of ViT-L/14 across task densities: (a) layer-wise fusion-ratio distribution; (b) decreasing mean fusion ratio as task density increases; (c) the most sensitive layers with significant shifts, Δ denotes the ratio change from 8 to 20 tasks (red for decrease, green for increase).

DAMM: Decoupled Adaptive Model Merging with Coordinate-Wise Trust Regions

Table 10. Multi-Task performance comparison on 8 Vision Tasks with Merged ViT-L/14. The best results are highlighted in **bold**.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
Pretrained	68.3	77.8	71.0	62.4	58.4	50.6	76.4	55.3	65.0
Individual	84.4	92.2	97.3	99.9	98.1	99.2	99.8	84.4	94.4
Traditional MTL	80.8	90.6	96.3	96.3	97.6	99.1	99.6	84.4	93.1
Weight Averaging	72.1	81.6	82.6	91.9	78.2	70.7	97.1	62.8	79.6
Fisher Merging	69.2	88.6	87.5	93.5	80.6	74.8	93.3	70.0	82.2
RegMean	73.3	81.8	86.1	97.0	88.0	84.2	98.5	60.8	83.7
Task Arithmetic	73.9	82.1	86.6	94.1	87.9	86.7	98.9	65.6	84.5
Ties-Merging	76.5	85.0	89.3	96.3	90.3	83.3	99.0	68.9	86.1
Consensus-Ties	74.9	83.6	88.7	96.6	90.5	93.2	99.1	71.1	87.2
Consensus-TA	74.5	82.2	88.8	94.2	92.6	93.3	99.2	67.8	86.6
TSV-M	78.0	90.0	93.4	99.0	94.8	96.3	99.5	78.8	91.2
CART	79.3	90.4	95.4	99.3	96.1	98.3	99.6	82.5	92.6
AdaMerging	79.0	90.3	90.8	96.2	93.4	98.0	99.0	79.9	90.8
AdaMerging++	79.4	90.3	91.6	97.4	93.4	97.5	99.0	79.2	91.0
TA+AdaRank	80.4	92.4	94.5	98.8	96.6	99.1	99.4	82.3	92.9
CART+AdaMerging	80.1	91.5	94.7	99.3	96.8	98.9	99.5	83.6	93.1
CART+AdaRank	80.6	92.1	96.0	99.7	97.0	98.8	99.4	83.8	93.4
DAMM (Ours)	83.5	91.3	96.7	99.8	98.1	99.3	99.8	83.5	94.0

Table 11. Multi-Task performance comparison on 14 Vision Tasks with Merged ViT-L/14. The best results are highlighted in **bold**.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST
Pretrained	68.3	77.8	71.0	62.4	58.4	50.6	76.4
Individual	84.4	92.2	97.3	99.9	98.1	99.2	99.8
Weight Averaging	70.9	79.7	78.0	84.1	72.8	61.7	93.9
Fisher Merging	69.2	88.6	87.5	93.5	80.6	74.8	93.3
RegMean	73.3	81.8	86.1	97.0	88.0	84.2	98.5
Task Arithmetic	72.1	76.4	81.1	89.3	81.7	75.4	97.9
Ties-Merging	74.0	78.8	83.7	90.7	83.0	70.7	98.1
Consensus-Ties	72.1	75.6	84.6	95.4	87.8	83.4	97.9
Consensus-TA	73.5	76.8	85.4	91.4	87.3	84.7	98.7
TSV-M	75.8	86.1	92.3	98.0	93.6	94.3	99.5
CART	77.9	86.0	94.1	98.8	92.8	95.9	99.5
AdaMerging	76.4	91.2	91.0	97.7	94.5	97.2	98.9
AdaMerging++	79.4	90.3	91.6	97.4	93.4	97.5	99.0
TA+AdaRank	78.7	92.6	94.8	98.4	95.7	98.5	99.0
CART+AdaMerging	79.8	91.8	94.5	98.2	95.2	98.1	99.1
CART+AdaRank	79.7	92.0	95.0	98.8	96.5	98.6	99.3
DAMM (Ours)	82.5	91.1	96.9	99.7	98.1	99.1	99.8

Method	DTD	CIFAR100	FER2013	Flowers102	OxfordIIITPet	PCAM	STL10
Pretrained	55.3	75.8	38.2	79.1	93.6	51.2	99.4
Individual	84.4	93.3	77.0	98.0	95.6	90.3	99.5
Weight Averaging	59.7	82.7	42.5	80.5	94.7	74.2	99.4
Fisher Merging	70.0	72.5	45.0	74.2	89.5	65.4	97.5
RegMean	60.8	75.6	52.4	81.6	90.2	79.7	97.6
Task Arithmetic	60.1	81.1	46.7	77.5	95.1	81.1	98.8
Ties-Merging	62.1	82.7	49.6	66.6	94.7	80.1	98.9
Consensus-Ties	65.5	80.4	47.5	76.7	94.4	80.7	98.5
Consensus-TA	62.9	80.7	51.4	76.9	95.3	82.6	98.6
TSV-M	74.5	85.6	69.0	87.9	96.1	83.9	99.5
CART	78.7	87.2	66.1	90.3	96.0	79.0	99.6
AdaMerging	79.5	84.3	49.5	95.1	95.5	82.4	99.1
AdaMerging++	79.2	87.1	66.0	89.2	92.5	87.0	98.9
TA+AdaRank	79.9	86.9	52.1	93.3	96.1	86.8	99.4
CART+AdaMerging	82.0	86.8	75.2	94.5	96.5	74.7	99.4
CART+AdaRank	81.9	86.0	71.8	94.4	96.4	89.9	99.5
DAMM (Ours)	83.5	92.7	75.4	96.6	95.7	91.1	99.6

Table 12. Multi-Task performance comparison on 20 Vision Tasks with Merged ViT-L/14. The best results are highlighted in **bold**.

Method	SUN397	Cars	RESISC 45	EuroSAT	SVHN	GTSRB	MNIST	DTD	CIFAR 100	FER 2013
Pretrained	68.3	77.8	71.0	62.4	58.4	50.6	76.4	55.3	75.8	38.2
Individual	84.4	92.2	97.3	99.9	98.1	99.2	99.8	84.4	93.3	77.0
Weight Averaging	70.3	78.6	76.2	79.0	70.8	59.0	92.6	58.1	82.6	40.6
Task Arithmetic	71.3	76.6	77.8	82.9	75.6	65.4	95.8	59.3	81.8	41.9
Ties-Merging	72.5	75.4	79.3	82.6	78.8	64.7	96.6	60.2	80.7	44.8
Consensus-TA	72.6	76.2	82.4	86.9	82.1	76.7	97.4	61.6	80.5	45.4
Consensus-Ties	72.1	71.5	80.6	85.1	82.5	78.5	96.1	63.1	77.4	44.5
TSV-M	74.4	81.1	90.6	96.3	90.0	90.8	97.3	71.4	82.4	63.9
CART	76.3	75.3	92.4	97.9	89.9	94.1	98.5	76.1	84.8	62.5
AdaMerging	75.2	90.7	91.4	97.6	88.6	97.0	97.7	74.0	83.2	47.9
TA+AdaRank	77.3	91.7	94.7	97.4	93.2	98.1	98.0	75.9	85.0	54.4
CART+AdaMerging	79.3	91.1	93.8	98.4	93.9	97.5	97.7	81.4	85.9	74.1
CART+AdaRank	79.9	91.5	94.7	98.6	94.8	98.2	97.4	80.4	86.5	70.7
DAMM (Ours)	82.2	90.7	96.4	99.7	98.0	99.2	99.7	82.8	92.6	73.9

Method	Flowers 102	Oxford IIITPet	PCAM	STL10	EMNIST	CIFAR10	Food101	Fashion MNIST	Rendered SST2	KMNIST
Pretrained	79.1	93.6	51.2	99.4	15.6	95.6	92.3	66.9	68.9	10.4
Individual	98.0	95.6	90.3	99.5	99.8	99.2	95.5	95.8	85.4	98.8
Weight Averaging	80.0	94.5	71.0	99.4	36.3	97.3	92.5	76.3	67.4	11.5
Task Arithmetic	78.2	94.9	76.1	99.0	55.2	97.4	90.9	80.5	66.8	17.5
Ties-Merging	69.1	94.7	75.4	98.7	75.5	97.3	90.3	82.6	69.1	28.8
Consensus-TA	77.8	95.4	81.5	98.9	82.7	97.1	90.9	84.5	70.6	34.4
Consensus-Ties	74.8	94.6	78.9	98.3	79.7	96.3	87.5	81.6	65.9	43.9
TSV-M	85.6	95.9	85.0	99.3	99.3	97.9	92.3	91.0	82.9	77.7
CART	87.9	95.8	80.7	99.3	98.5	98.3	92.6	91.8	80.0	85.8
AdaMerging	95.1	95.4	50.3	99.1	96.3	97.2	92.7	89.7	82.5	94.3
TA+AdaRank	92.0	95.9	66.3	99.3	97.5	97.8	93.8	91.6	85.7	97.0
CART+AdaMerging	95.7	96.5	79.2	99.4	98.1	97.8	93.4	91.5	85.4	96.7
CART+AdaRank	93.1	96.4	89.8	99.4	98.3	98.2	94.0	92.1	85.0	97.5
DAMM (Ours)	95.9	95.8	87.8	99.6	99.8	99.3	95.1	95.2	81.7	98.5

Table 13. Performance comparison of DAMM on ViT-B/32 over 6 random 8-of-20 task selections. We report accuracy of $\mathbf{V}_{\text{shared}}$, without (w/o) and with (w/) projecting it onto the expert-defined per-parameter convex hull induced by the 8 fused experts. Gray-shaded cells denote the 8 ID expert tasks used to learn the structured mask and shared subspace, while the remaining 12 tasks are held out to evaluate OOD generalization.

Task Name	Exp. 1		Exp. 2		Exp. 3		Exp. 4		Exp. 5		Exp. 6	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Cars	2.6	61.2	4.0	56.0	60.3	76.4	2.3	59.4	1.7	58.6	49.7	75.9
DTD	76.1	79.2	14.2	62.5	73.9	79.2	11.8	62.6	13.8	61.7	8.8	62.1
EuroSAT	26.7	96.0	15.6	92.8	63.0	99.8	31.5	96.1	6.8	90.9	51.1	99.8
GTSRB	97.5	99.1	9.1	82.4	95.4	99.1	8.7	85.5	11.1	85.0	9.3	80.1
MNIST	70.4	98.6	21.6	95.0	30.3	97.3	98.6	99.7	25.2	97.0	93.5	99.7
RESISC45	58.3	95.4	11.6	82.6	12.0	86.5	46.3	95.3	7.0	82.8	12.8	84.1
SUN397	9.8	71.2	11.9	69.8	4.9	71.3	15.3	78.0	3.8	69.6	8.9	77.8
SVHN	94.9	97.5	20.4	80.8	20.0	88.8	91.5	97.5	10.2	82.5	20.1	85.0
CIFAR100	78.6	88.5	79.9	88.5	9.7	80.1	70.6	88.5	17.8	82.2	7.0	79.1
STL10	95.9	98.1	96.4	98.1	29.4	98.2	95.3	98.1	69.5	97.7	31.3	98.2
Flowers102	13.4	79.3	67.3	89.0	7.3	80.6	51.4	89.1	7.2	78.0	44.5	89.2
OxfordPet	55.0	91.3	59.8	91.1	8.5	91.8	40.6	91.3	25.0	90.8	3.7	91.1
PCAM	54.6	84.4	86.7	88.3	60.3	84.7	50.2	81.9	84.8	88.3	84.6	88.3
FER2013	56.3	71.2	57.2	71.1	17.6	54.1	13.0	52.2	59.8	71.1	15.5	51.4
EMNIST	22.6	95.9	98.3	99.8	10.5	98.6	19.6	89.9	99.4	99.8	94.3	99.7
CIFAR10	67.6	97.5	95.6	98.0	27.8	96.4	60.6	97.4	95.5	97.9	23.7	96.2
Food101	9.2	85.9	8.6	85.2	13.5	88.5	6.4	85.6	16.8	88.4	7.3	88.5
FashMNIST	33.4	86.8	21.8	86.4	93.5	94.9	29.6	85.7	93.2	94.8	14.3	82.5
SST2	49.9	70.2	50.0	70.1	64.8	72.8	49.9	71.1	65.0	72.7	49.9	70.1
KMNIST	10.5	67.8	11.0	52.0	98.0	98.5	11.5	52.0	97.6	98.5	9.2	46.6
Avg. (8 ID)	76.6	90.1	80.2	90.5	70.3	88.6	63.7	92.2	76.5	88.9	54.2	89.9
Avg. (12 OOD)	30.9	82.9	20.0	76.3	19.9	85.7	31.3	76.6	24.9	82.2	17.1	78.1
Avg. (All 20)	49.2	85.8	44.0	82.0	40.0	86.9	44.2	82.8	45.5	84.9	32.0	82.8

Table 14. Performance comparison of DAMM on GPT-2 over 7 random 4-of-7 task selections. We report accuracy of $\mathbf{V}_{\text{shared}}$, without (w/o) and with (w/) projecting it onto the expert-defined per-parameter convex hull induced by the 4 fused experts. Gray-shaded cells denote the 4 ID expert tasks used to learn the structured mask and shared subspace, while the remaining 3 tasks are held out to evaluate OOD generalization.

Task Name	Exp. 1		Exp. 2		Exp. 3		Exp. 4		Exp. 5		Exp. 6		Exp. 7	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
CoLA	41.3	39.1	27.3	27.6	41.9	42.0	28.4	30.4	40.8	41.9	33.1	32.9	40.0	41.6
SST-2	90.0	89.5	91.2	91.1	91.1	91.1	90.0	90.1	91.2	91.3	91.4	90.9	89.7	89.6
MRPC	80.2	80.6	61.0	63.7	74.0	73.3	80.4	80.6	55.9	65.4	80.6	80.6	75.7	74.3
QQP	87.5	87.7	89.4	89.5	89.5	89.5	89.3	89.4	87.3	87.3	89.5	89.6	89.4	89.4
MNLI	81.0	81.6	77.5	77.6	81.1	81.3	81.2	81.3	76.1	76.3	77.1	77.3	81.1	81.2
QNLI	82.8	83.2	88.2	88.2	80.7	80.6	88.1	88.3	87.6	88.1	82.1	82.3	88.4	88.1
RTE	65.7	64.3	64.6	65.7	60.3	60.7	47.6	46.6	66.1	65.0	64.6	65.3	63.5	63.9
Avg. (4ID)	74.8	74.2	67.2	68.0	74.1	74.0	72.0	72.6	68.8	71.5	73.7	73.5	73.7	73.7
Avg. (3OOD)	76.5	76.3	76.8	77.1	74.1	74.2	72.3	72.0	76.6	76.5	74.6	75.0	77.6	77.7
Avg. (All7)	75.5	75.1	71.3	71.9	74.1	74.1	72.2	72.4	72.1	73.6	74.1	74.1	75.4	75.4

Table 15. Performance comparison of full-parameter merging on ViT-B/32 over 6 random 8-of-20 task selections. We report accuracy of $\mathbf{V}_{\text{merged}}$, without (w/o) and with (w/) projecting it onto the expert-defined per-parameter convex hull induced by the 8 fused experts. Gray-shaded cells denote the 8 ID expert tasks used for merging, while the remaining 12 tasks are held out to evaluate OOD generalization.

Task Name	Exp. 1		Exp. 2		Exp. 3		Exp. 4		Exp. 5		Exp. 6	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Cars	1.2	22.6	1.3	21.8	33.7	58.7	0.7	24.4	1.1	19.2	3.1	58.1
DTD	58.5	55.1	8.0	38.7	55.3	55.5	6.3	38.9	5.1	35.0	3.2	36.9
EuroSAT	22.3	37.3	14.6	24.2	39.3	62.2	25.4	35.3	13.9	35.0	25.7	58.3
GTSRB	98.3	85.3	5.9	29.8	98.4	84.0	5.2	29.2	3.8	29.2	3.5	24.3
MNIST	34.9	80.9	18.4	30.5	12.1	29.6	89.8	98.5	19.1	36.9	84.8	97.2
RESISC45	34.7	61.9	8.5	40.4	4.5	38.4	23.3	62.2	6.9	47.9	3.8	35.7
SUN397	2.5	54.5	3.3	55.3	1.1	52.8	3.5	61.3	0.9	55.2	1.1	62.1
SVHN	94.8	92.7	14.9	28.5	16.9	35.0	85.2	93.2	10.0	27.0	14.4	38.8
CIFAR100	56.2	75.7	57.2	76.4	4.0	42.2	26.4	73.3	5.8	59.1	3.1	44.2
STL10	89.2	97.0	92.5	97.2	17.5	91.5	78.7	96.8	41.9	91.0	14.8	90.9
Flowers102	4.2	50.3	29.1	59.2	2.4	57.6	17.3	54.6	1.0	56.3	8.9	56.6
OxfordPet	22.8	85.3	22.5	85.6	3.0	83.9	14.8	84.8	3.8	75.2	2.4	83.7
PCAM	63.1	58.0	84.1	83.1	49.8	59.9	51.5	59.6	81.6	84.6	82.7	85.0
FER2013	42.0	31.7	42.0	34.4	14.4	34.4	11.6	35.9	32.0	38.2	15.6	34.0
EMNIST	6.9	23.7	98.0	99.2	4.8	14.2	18.8	22.8	96.9	99.1	89.1	97.5
CIFAR10	54.1	91.4	90.9	96.3	20.4	72.6	42.4	90.5	83.4	96.2	13.6	71.9
Food101	3.6	69.9	3.8	68.2	2.5	78.0	2.5	69.4	1.4	79.4	1.1	78.5
FashMNIST	24.7	63.5	19.1	61.7	92.2	89.6	10.0	60.2	89.8	90.8	10.6	51.5
SST2	49.9	51.9	49.9	54.0	50.0	58.2	49.9	53.9	50.1	67.4	49.9	54.4
KMNIST	8.4	10.3	10.7	6.4	97.5	59.1	8.6	12.5	91.1	71.1	9.7	10.2
Avg. (8 ID)	62.1	73.1	64.5	78.9	58.6	68.2	42.4	78.1	65.8	78.3	37.1	74.2
Avg. (12 OOD)	23.0	55.4	9.9	38.3	12.6	50.7	19.4	49.4	9.4	55.6	12.0	48.1
Avg. (All 20)	38.6	62.5	31.7	54.5	31.0	57.7	28.6	60.9	32.0	64.7	22.0	58.6