From Parameter to Representation: A Closed-Form Approach for Controllable Model Merging

Jialin Wu[1], Jian Yang[2], Handing Wang[3], Jiajun Wen[1], Zhiyong Yu[1]*

[1]Department of Computer, Rocket Force University of Engineering
[2]Department of Engineering, Rocket Force University of Engineering
[3]School of Artificial Intelligence, Xidian University
wujialin11@nudt.edu.cn, yangjian@nudt.edu.cn, hdwang@xidian.edu.cn, wenjiajun11@nudt.edu.cn, yutouzy@163.com

Abstract

Model merging combines expert models for multitask performance but faces challenges from parameter interference. This has sparked recent interest in controllable model merging, giving users the ability to explicitly balance performance trade-offs. Existing approaches employ a compile-then-query paradigm, performing a costly offline multi-objective optimization to enable fast, preference-aware model generation. This offline stage typically involves iterative search or dedicated training, with complexity that grows exponentially with the number of tasks. To overcome these limitations, we shift the perspective from parameter-space optimization to a direct correction of the model's final representation. Our approach models this correction as an optimal linear transformation, yielding a closed-form solution that replaces the entire offline optimization process with a single-step, architecture-agnostic computation. This solution directly incorporates user preferences, allowing a Pareto-optimal model to be generated on-the-fly with complexity that scales linearly with the number of tasks. Experimental results show our method generates a superior Pareto front with more precise preference alignment and drastically reduced computational cost.

Code — https://github.com/CREAHDD/ReACT
Extended version — https://arxiv.org/abs/2511.10943

Introduction

The shift towards fine-tuning large pre-trained models has led to a surge in specialized models customized for specific tasks. Merging these models has become an effective way to combine their strengths into a single, versatile network without needing expensive retraining or access to the original training data. Early approaches, like averaging weights (Matena and Raffel 2022; Jin et al. 2023) or performing task vector arithmetic (Ilharco et al. 2023; Yadav et al. 2023; Yang et al. 2024b), aimed to produce a static, merged model. However, these *one-size-fits-all* methods often suffer from interference between task-specific parameters, leaving its performance far behind that of the individual experts. This revealed a need for a more flexible and controllable approach to model merging, letting users explicitly adjust the balance based on their preferences.

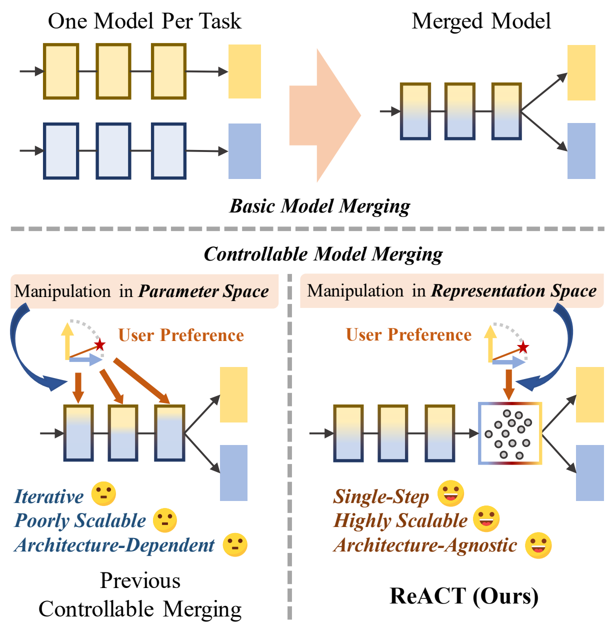


Figure 1: Conceptual distinction between basic model merging (top) and controllable merging (bottom), exemplified with a two-task scenario. While prior approaches to controllable merging (bottom left) rely on slow, iterative optimization in parameter space, our method (bottom right) achieves direct control through an efficient, single-step correction in representation space.

Recent methods like Pareto Merging (PM) (Chen and Kwok 2025) and MAP (Li et al. 2025) enable user control by framing model merging as a multi-objective optimization (MOO) problem. Their shared *compile-then-query* paradigm, however, imposes severe overhead in computation, memory, and data. The offline *compiling* stage is computationally prohibitive: PM requires complex iterative training that can become unstable, while MAP employs an evolutionary search whose complexity grows exponentially with more tasks. This cost is further magnified when combining PM with advanced backbones like AdaMerging (Yang et al. 2024b), which necessitates a more complex

*Corresponding author.

joint optimization. In terms of resources, MAP must store all original task vectors, a memory requirement that scales poorly, and often relies on labeled data which may be unavailable. Moreover, this entire costly preparation of both MAP and PM must be repeated from scratch whenever the set of expert models changes, hindering practical scalability and adaptation.

To address these concerns, we propose **ReACT** (**Re**presentation **A**nalytical **C**ontrol **T**ransformation), a fundamentally different, *on-the-fly* analytical approach. As conceptually illustrated in Figure 1, ReACT shift the paradigm from costly parameter-space optimization to a direct correction in the model's final representation space, based on our key insight that performance degradation stems from a global linear distortion. We therefore reframe the problem as finding an optimal linear correction map, regularized by an orthogonal prior. Crucially, this formulation admits a closed-form solution that naturally extends to the multi-objective case via linear scalarization, bypassing iterative optimization. The resulting analytical framework is inherently architecture-agnostic and highly modular, allowing expert models to be added or removed without costly re-optimization. ReACT enables on-the-fly generation of preference-aware models for real-time exploration and is highly data-efficient, achieving near-best performance with only a fraction of unlabeled test data. Our main contributions are threefold:

- We reframe controllable merging as a representation correction problem, identifying the primary bottleneck as a simple linear distortion rather than a complex parameter conflict.
- We introduce an *on-the-fly* analytical method, deriving what is, to our knowledge, the first *closed-form solution* for controllable model merging that bypasses iterative optimization entirely.
- Extensive experiments show our method achieves a state-of-the-art Pareto front, superior preference alignment, and drastically reduced computational cost, maintaining this strong performance even when using only a fraction of the unlabeled data required by competing methods.

Related Work

Model Merging. Model merging aims to consolidate multiple task-specific models, fine-tuned from a common pre-trained base, into a single network. We categorize prior work based on the nature of the merged model they produce. A primary line of work focuses on creating a single, static merged model. This includes simple weight averaging (Singh and Jaggi 2020; Matena and Raffel 2022; Jin et al. 2023) and more sophisticated methods operating on task vectors—the parameter shifts from pre-training (Ilharco et al. 2023). Subsequent refinements have focused on mitigating parameter conflicts through techniques like pruning and rescaling (Yadav et al. 2023; Yang et al. 2024b; Yu et al. 2024; Du et al. 2024). However, these methods produce a *one-size-fits-all* model, offering no mechanism to control performance trade-offs among tasks. Another category introduces dynamic, task-specific modules to improve performance. For instance, Representation Surgery (Yang et al. 2024a) learns a lightweight MLP to correct representation bias for each task, while other methods generate task-specific masks or adapters (Wang et al. 2024; Huang et al. 2024; Lu et al. 2024; Qi et al. 2025). While effective, these approaches typically require loading a single task's module during inference, preventing a smooth, controllable trade-off across multiple tasks simultaneously.

The works most relevant to ours are those that enable controllable merging, often framing it as a Multi-Objective Optimization (MOO) problem. Pareto Merging (PM) (Chen and Kwok 2025) learns a low-rank representation of the Pareto set via a complex optimization, while MAP (Li et al. 2025) employs an evolutionary algorithm to search for solutions. Both pioneering methods follow a *compile-then-query* paradigm that operates in the parameter space. Despite their innovation, this paradigm imposes a significant upfront computational cost and suffers from poor scalability as the number of tasks increases. In contrast, our work proposes an *on-the-fly* analytical framework that operates in the representation space.

Multi-Objective Optimization in Deep Learning. Multi-Objective Optimization (MOO) is a foundational framework for balancing competing objectives, not only in model merging but also in the broader field of multi-task learning (MTL). In the context of MTL, a dominant line of work focuses on training-based methods that operate during training or train a preference-aware module directly, such as GradNorm (Chen et al. 2018), PCGrad (Yu et al. 2020) and PaLoRA (Dimitriadis, Frossard, and Fleuret 2025). These approaches iteratively manipulate task gradients or sample preference vectors at each training step to steer the model towards the Pareto front.

However, a significant drawback of these training-time methods is their computational expense and reliance on the full training pipeline. Our work, in contrast, operates in a distinct, post-hoc paradigm. We apply MOO principles to already-trained models, sidestepping the costly iterative training process entirely. Furthermore, by formulating the problem in the representation space, we derive a closed-form solution, eliminating even the post-hoc optimization or search required by other controllable merging methods.

Preliminaries

Model Merging and Task Vectors

Model merging techniques typically operate on a set of T task-specific models, $\{\theta_1, \theta_2, ..., \theta_T\}$, that are fine-tuned from a common pre-trained model, θ_0. A powerful concept in this area is the *task vector* (Ilharco et al. 2023), which captures the parameter shift for a specific task:

$$\Delta_t = \theta_t - \theta_0 \tag{1}$$

A simple merged model, θ_{merge}, can then be constructed by applying a linear combination of these task vectors to the pre-trained base:

$$\theta_{\text{merge}} = \theta_0 + \sum_{t=1}^{T} \alpha_t \Delta_t \tag{2}$$

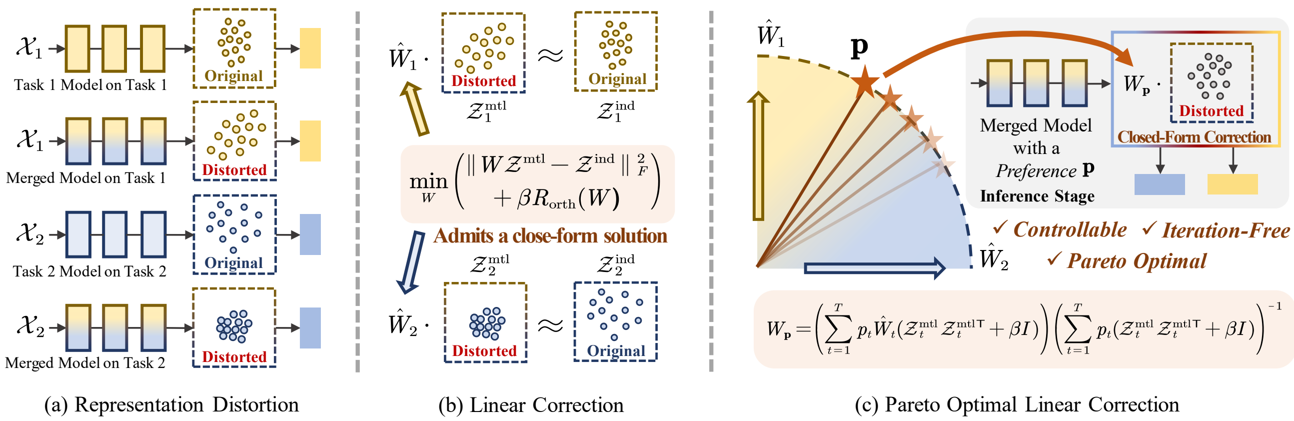


Figure 2: An overview of our proposed method, illustrated with a two-task ($T = 2$) example. (a) We first identify that model merging causes representation distortion: the feature distribution of a merged model deviates from that of an individual task model. (b) We propose to correct this by finding a linear correction matrix $\hat{W}_t$ for each task, which has a closed-form solution. (c) Finally, we derive a Pareto-optimal transformation $W_{\mathbf{p}}$ by analytically aggregating the individual corrections based on a user preference vector $\mathbf{p}$.

where α_t are merging coefficients. More advanced methods determine these coefficients α_t sophistically, sometimes even optimizing them layer-wise or per-parameter (Yang et al. 2024b; Du et al. 2024; Wang et al. 2024). The pre-merged model θ_{merge} can be obtained via various methods, from the simple linear combination in Eq. 2 to more advanced techniques like TIES-Merging or AdaMerging. Our framework operates as a post-hoc correction on any such pre-merged model.

Controllable Merging as MOO

The goal of controllable merging is to find solutions that optimally trade off performance across multiple tasks. This can be naturally formulated as a Multi-Objective Optimization (MOO) problem. Given T tasks, we aim to simultaneously minimize a vector of loss functions, $\{L_1(\theta), L_2(\theta), ..., L_T(\theta)\}$, where $L_t(\theta)$ is the loss for task t using model parameters θ.

A solution θ^* is considered *Pareto-optimal* if it is not possible to improve performance on one task without degrading performance on at least one other task. The set of all such Pareto-optimal solutions forms the *Pareto front* (Fleischer 2003).

A common technique to find points on the Pareto front is *Linear Scalarization*. This method transforms the multi-objective problem into a single-objective one by taking a weighted sum of the individual objectives, guided by a user-defined preference vector $\mathbf{p} = [p_1, ..., p_T]^\intercal$, where $p_t \geq 0$ and $\sum p_t = 1$:

$$\min_{\theta} \sum_{t=1}^{T} p_t L_t(\theta) \tag{3}$$

Crucially, existing methods attempt to solve this MOO problem directly in the high-dimensional parameter space of θ. In contrast, our work will reformulate the objective in the model's final representation space, leading to a more direct and efficient solution.

Method

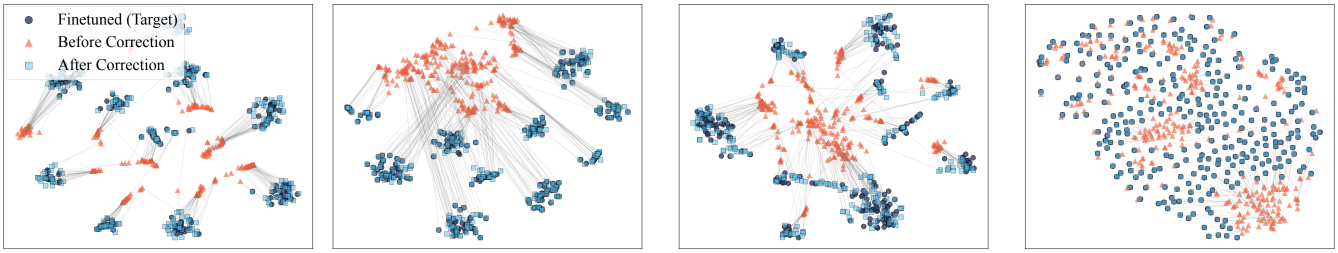
Instead of performing expensive optimization in the parameter space, our method, ReACT, directly corrects the model's final representation. For a given task t, our approach is founded on the hypothesis that the discrepancy between the representations from the merged model, $\mathcal{Z}_t^{\text{mtl}}$, and those from the ideal single-task expert, $\mathcal{Z}_t^{\text{ind}}$, is primarily a linear distortion. This allows us to model the correction as a simple linear transformation. We show that finding the optimal transformation is a multi-objective problem for which we can derive a closed-form, Pareto-optimal solution for any user preference. Figure 2 provides an overview of our framework. The following sections first formalize this correction for a single task and then extend it to the multi-objective setting.

Linear Representation Correction

Prior work like Representation Surgery (Yang et al. 2024a) employs sample-specific non-linear functions for this correction. However, as hypothesized and visualized in Figure 3, this misalignment between $\mathcal{Z}_t^{\text{mtl}}$ and $\mathcal{Z}_t^{\text{ind}}$ is structurally simple. We therefore hypothesize it is dominated by a global linear distortion (e.g., rotation and scaling) rather than a sample-specific non-linear warp, as shown in Figure 3. Based on this key insight, we propose a simpler, more elegant correction. We model the correction as a linear transformation matrix W_t for each task t:

$$\hat{\mathcal{Z}}_t^{\text{mtl}} = W_t \mathcal{Z}_t^{\text{mtl}} \tag{4}$$

To find the optimal transformation matrix $W_t \in \mathbb{R}^{D_{\text{rep}} \times D_{\text{rep}}}$, we use a set of N calibration samples. We collect their representations into matrices $\mathcal{Z}_t^{\text{mtl}}, \mathcal{Z}_t^{\text{ind}} \in \mathbb{R}^{D_{\text{rep}} \times N}$ and minimize



(a) MNIST (b) EuroSAT (c) SVHN (d) Cars

Figure 3: Visualization of representation correction. t-SNE plots for four tasks from an 8-model merge show the merged model's representations (orange) are severely misaligned with ideal single-task targets (dark blue). Our method effectively pulls the corrected features (light blue) back to the target clusters, visually confirming that a simple linear transformation is sufficient to correct the discrepancy. More t-SNE plots see Appendix D.4.

the squared Frobenius norm between the corrected and ideal representations:

$$\min_{W_t} \|W_t \mathcal{Z}_t^{\text{mtl}} - \mathcal{Z}_t^{\text{ind}}\|_F^2 \tag{5}$$

Optimal Orthogonal Regularization

Directly solving Eq. 5 via ordinary least squares can overfit when the correction data, available at test time, is scarce. To improve robustness, we regularize the solution towards the optimal orthogonal transformation W_t^{orth}, which is solvable via the Orthogonal Procrustes problem (by finding the SVD of the cross-covariance matrix $\mathcal{Z}_t^{\text{ind}} {\mathcal{Z}_t^{\text{mtl}}}^{\mathsf{T}} = U_t S_t V_t^{\mathsf{T}}$ and setting $W_t^{\text{orth}} = U_t V_t^{\mathsf{T}}$). This structural prior prevents W_t from distorting the geometric structure of the representation space. The regularized objective is:

$$\min_{W_t} \|W_t \mathcal{Z}_t^{\text{mtl}} - \mathcal{Z}_t^{\text{ind}}\|_F^2 + \beta \|W_t - W_t^{\text{orth}}\|_F^2 \tag{6}$$

where β is a hyperparameter. This objective has a closed-form solution (see Appendix B.1 for derivation):

$$\hat{W}_t = (\mathcal{Z}_t^{\text{ind}} {\mathcal{Z}_t^{\text{mtl}}}^{\mathsf{T}} + \beta W_t^{\text{orth}})(\mathcal{Z}_t^{\text{mtl}} {\mathcal{Z}_t^{\text{mtl}}}^{\mathsf{T}} + \beta I)^{-1} \tag{7}$$

This provides the optimal correction for a single task.

Pareto-Optimal Representation Correction

To handle multiple tasks and user preferences $\mathbf{p} = [p_1, \cdots, p_T]^{\mathsf{T}}$, we extend the single-task objective (Eq. 6) to a multi-objective optimization (MOO) problem:

$$\min_{W} \{L_1(W), \cdots, L_T(W)\}, \tag{8}$$

where each $L_t(W)$ is the loss from Eq. 6. We apply Linear Scalarization to find Pareto-optimal solutions, resulting in a single preference-weighted objective:

$$\min_{W} \sum_{t=1}^{T} p_t L_t(W) \tag{9}$$

Crucially, since each loss function $L_t(W)$ is a convex quadratic function of W, the weighted objective in Eq. 9 also possesses this property. This allows us to derive a unique, analytical solution, which we formalize in the following proposition (proof in Appendix B.2):

Proposition 1. *For any preference* $\mathbf{p}$*, the unique Pareto-optimal solution* $W_{\mathbf{p}}$ *to the multi-objective problem in Eq. 8 has a closed-form expression given by:*

$$W_{\mathbf{p}} = \left(\sum_{t=1}^{T} p_t (\mathcal{Z}_t^{ind} {\mathcal{Z}_t^{mtl}}^{\mathsf{T}} + \beta W_t^{orth}) \right) \left(\sum_{t=1}^{T} p_t (\mathcal{Z}_t^{mtl} {\mathcal{Z}_t^{mtl}}^{\mathsf{T}} + \beta I) \right)^{-1} \tag{10}$$

The solution $W_{\mathbf{p}}$ is the core of our framework, enabling instant generation of a preference-aware model. Moreover, its structure reveals a principled aggregation mechanism. By substituting the single-task optimal solution $\hat{W}_t$ from Eq. 7 back into Eq. 10, we can algebraically rearrange $W_{\mathbf{p}}$ into the following intuitive form:

$$W_{\mathbf{p}} = \left(\sum_{t=1}^{T} p_t \hat{W}_t C_t \right) \left(\sum_{t=1}^{T} p_t C_t \right)^{-1} \tag{11}$$

where $C_t = \mathcal{Z}_t^{\text{mtl}} {\mathcal{Z}_t^{\text{mtl}}}^{\mathsf{T}} + \beta I$. This form clearly shows that our solution is a weighted average of the individual correction maps $\hat{W}_t$, where each map is weighted not just by the user preference p_t, but by a data-dependent matrix C_t. This is fundamentally different from a naive average, $W_{\text{naive}} = \sum p_t \hat{W}_t$, to which our solution only simplifies if all C_t are identical. Since C_t is the regularized autocorrelation matrix of the features, our method naturally gives greater influence to corrections for tasks with more pronounced, high-variance feature structures. We will validate the empirical superiority of this data-aware weighting in our ablation study. To summarize our method and its efficient implementation, we present the detailed pseudocode. Algorithm 1 details the one-time, offline computation of reusable components. Algorithm 2 then illustrates the two-stage online workflow: first, a near-instant assembly of a preference-specific corrector $W_{\mathbf{p}}$, and then the use of this corrector for inference.

Complexity and Scalability

The analytical, non-iterative nature of our framework leads to its exceptional computational efficiency. The one-time setup for T tasks is dominated by a highly parallelizable representation extraction, followed by matrix computations of $O(TD_{\text{rep}}^3)$ on features of dimension D_{rep}. This approach bypasses the prohibitive evaluation costs and large $O(Td)$ storage of MAP (Li et al. 2025), where d denotes the full model parameter count, as well as the extensive, less parallelizable iterative training of PM (Chen and Kwok 2025). Consequently, our method requires only $O(d + TD_{\text{rep}}^2)$ storage, while adapting to a new preference $\mathbf{p}$ is nearly instantaneous at $O(TD_{\text{rep}}^2 + D_{\text{rep}}^3)$. Since D_{rep} is a small constant defined by the model architecture (e.g., 512 for ViT-B/32), this cubic complexity arises from a single matrix inversion that executes in *milliseconds*. This offers a uniquely practical solution for real-time generation of preference-aligned models. A more detailed complexity analysis is provided in the Appendix C.

Algorithm 1: One-Time Component Computation (Offline)

Input: Merged backbone f_{merge}; Task-specific models $\{(f_t, h_t)\}_{t=1}^T$; Calibration data $\{\mathcal{D}_t^{\text{calib}}\}_{t=1}^T$; Hyperparameter β.

Output: Pre-computed components $\{\hat{W}_t, C_t\}_{t=1}^T$.

for $t = 1, \ldots, T$ **do**
  Extract representations: $\mathcal{Z}_t^{\text{mtl}} \leftarrow f_{\text{merge}}(\mathcal{D}_t^{\text{calib}})$, $\mathcal{Z}_t^{\text{ind}} \leftarrow f_t(\mathcal{D}_t^{\text{calib}})$;
  Compute cross-covariance: $S_t \leftarrow \mathcal{Z}_t^{\text{ind}} \mathcal{Z}_t^{\text{mtl}\top}$;
  Solve Orthogonal Procrustes problem via SVD:
    $U_t, \_, V_t^\top \leftarrow \text{SVD}(S_t)$;
    $W_t^{\text{orth}} \leftarrow U_t V_t^\top$;
  Pre-compute components based on Eq. 11:
    $C_t \leftarrow \mathcal{Z}_t^{\text{mtl}} \mathcal{Z}_t^{\text{mtl}\top} + \beta I$;
    $\hat{W}_t \leftarrow (S_t + \beta W_t^{\text{orth}}) C_t^{-1}$;
end for
return $\{\hat{W}_t, C_t\}_{t=1}^T$

Algorithm 2: Online Preference Adaptation and Inference

Requires: Pre-computed components $\{\hat{W}_t, C_t\}_{t=1}^T$ from Alg. 1; Merged backbone f_{merge}; A data point x for a target task t with head h_t.

 # Step 1: Assemble corrector for preference $\mathbf{p}$ (run once per preference)
$W_{\mathbf{p}} \leftarrow \left(\sum_{i=1}^T p_i \hat{W}_i C_i\right) \left(\sum_{i=1}^T p_i C_i\right)^{-1}$;
return $W_{\mathbf{p}}$
 # Step 2: Use assembled corrector $W_{\mathbf{p}}$ for inference (run per data point)
Extract representation: $z^{\text{mtl}} \leftarrow f_{\text{merge}}(x)$;
Apply correction: $\hat{z}^{\text{mtl}} \leftarrow W_{\mathbf{p}} z^{\text{mtl}}$;
Predict using task head: $y_t \leftarrow h_t(\hat{z}^{\text{mtl}})$;
return y_t

Experiments

Our experiments are designed to answer four key questions:

Q1 Does our method achieve state-of-the-art performance and controllability?

Q2 Does it offer a more efficient and scalable paradigm?

Q3 Is its empirical success rooted in its core design principles?

Q4 How sufficient is our linear correction compared to a non-linear alternative?

Experimental Setting

Experiment setup. Following the protocols of Task Arithmetic (Ilharco et al. 2023) and Pareto Merging (Chen and Kwok 2025), we utilize the publicly available models provided by Task Arithmetic. This includes their pre-trained CLIP (Radford et al. 2021) ViT-B/32 visual encoder and the eight individual models already fine-tuned on diverse image classification tasks: SUN397 (Xiao et al. 2010), Cars (Krause et al. 2013), RESISC45 (Cheng, Han, and Lu 2017), EuroSAT (Helber et al. 2019), SVHN (Netzer et al. 2011), GTSRB (Stallkamp et al. 2012), MNIST (LeCun et al. 1998), and DTD (Cimpoi et al. 2014). We position our method against a hierarchy of baselines: performance upper bounds (individually fine-tuned models (Individual), Multi-Task Learning (MTL)), non-controllable merging methods (Task Arithmetic (TA), TIES-Merging (TIES) (Yadav et al. 2023), AdaMerging (AM), AdaMerging++ (AMPP) (Yang et al. 2024b)), and the direct state-of-the-art in controllable merging, Pareto Merging (PM). For fair comparison, all results for our method are generated using the AM backbone unless specified otherwise. Please see Appendix A.1 and A.2 for the introduction of the tasks and implementation details.

Evaluation metrics. Our evaluation relies on three metrics. The primary metric for task-specific performance is classification accuracy. To assess the overall quality of the solution set approximating the Pareto front, we employ Hypervolume (HV) (Zitzler and Thiele 1999). To measure how precisely a model's performance aligns with a given user preference, we use Uniformity (Mahapatra and Rajan 2020). To ensure a fair comparison across tasks with disparate difficulty, both metrics are computed on accuracies normalized relative to their individual model's performance. Detailed formulations for HV and Uniformity are provided in the Appendix A.3.

Main Results

To answer **Q1**, we first show in Table 1 that our method provides a substantial uplift over the non-controllable backbones it enhances, such as AMPP (85.4% vs. 81.1%). We then compare against leading controllable methods. As evaluating all possible user preferences is infeasible, we test controllability under three representative scenarios reflecting common use cases: equal preference (uniform weights), priority preference (a single task is given 50% weight, with the remaining weight distributed evenly among the other tasks), and one-hot preference (a single task receives all weight). Following the protocol in (Chen and Kwok 2025),

Pref.	Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Average
–	Individual	75.3	77.7	96.1	99.7	97.5	98.7	99.7	79.4	90.5
	MTL	73.9	74.4	93.9	98.2	95.8	98.9	99.5	77.9	88.9
–	TA (Ilharco et al. 2023)	55.2	54.9	66.7	78.9	80.2	69.7	97.3	50.4	69.1
	TIES (Yadav et al. 2023)	59.5	60.0	71.7	78.2	86.3	72.9	98.2	52.8	72.4
	AM (Yang et al. 2024b)	64.5	68.1	79.2	93.8	87.0	91.9	97.5	59.1	80.1
	AMPP (Yang et al. 2024b)	66.6	68.3	82.2	94.2	89.6	89.0	98.3	60.6	81.1
equal	MAP (Li et al. 2025)	60.0	58.8	85.8	69.5	83.5	73.4	87.8	53.2	71.5
	AM+PM (Chen and Kwok 2025)	70.1	**74.2**	87.3	**96.5**	90.2	95.6	98.5	66.7	84.9
	AM+Ours	71.0	70.0	88.2	95.4	90.9	97.1	98.6	68.0	84.9
	AM+Ours (10% unlabeled test data)	70.0	69.4	87.6	94.6	90.7	96.6	98.6	66.2	84.2
	AMPP+PM (Chen and Kwok 2025)	70.6	73.9	87.5	96.7	90.8	96.7	98.6	67.2	85.2
	AMPP+Ours	**72.0**	70.4	**88.5**	95.8	**91.7**	**97.3**	**98.6**	**68.5**	**85.4**
	AMPP+Ours (10% unlabeled test data)	71.0	69.3	88.0	95.2	91.4	96.8	98.6	67.1	84.7
priority	AM+PM (Chen and Kwok 2025)	71.1	**74.2**	89.0	**97.6**	92.1	97.4	**99.0**	64.0	85.5
	AM+Ours	72.5	72.8	91.2	97.1	92.3	98.1	98.8	73.7	87.1
	AM+Ours (10% unlabeled test data)	71.5	71.5	90.4	96.3	92.0	97.4	98.8	71.1	86.1
	AMPP+PM (Chen and Kwok 2025)	72.1	73.7	88.8	97.5	92.2	97.5	**99.0**	66.1	85.9
	AMPP+Ours	**73.1**	73.0	**91.4**	97.5	**92.7**	**98.2**	98.8	**74.0**	**87.3**
	AMPP+Ours (10% unlabeled test data)	72.3	71.8	90.4	96.8	92.5	97.6	98.8	71.0	86.4
one-hot	AM+PM (Chen and Kwok 2025)	70.9	59.9	89.5	96.9	85.4	97.2	99.1	62.5	82.7
	AM+Ours	72.8	74.7	93.1	98.8	93.4	98.6	**99.3**	78.2	88.6
	AM+Ours (10% unlabeled test data)	72.0	72.1	91.1	96.3	92.8	97.8	99.0	70.6	86.5
	AMPP+PM (Chen and Kwok 2025)	71.7	62.7	89.9	97.0	88.3	97.2	99.1	63.2	83.6
	AMPP+Ours	**73.4**	**75.3**	**93.3**	**99.0**	**93.6**	**98.7**	**99.3**	**78.3**	**88.9**
	AMPP+Ours (10% unlabeled test data)	72.7	72.2	91.3	97.0	93.1	97.9	99.1	71.5	86.7

Table 1: Test accuracies (%) on eight datasets when merging eight ViT-B/32 models. We compare our method against non-controllable baselines and state-of-the-art controllable approaches under three preference scenarios. Our method consistently achieves the highest average accuracy in each group, with the best results highlighted in **bold**.

we report MAP's result only under equal preference. Our primary comparison is against Pareto Merging (PM), focusing on its strongest data-based variants (AM+PM and AMPP+PM). This provides the most direct comparison, as our method is also a data-driven correction. However, our approach acts as a lightweight post-hoc step, in contrast to PM's complex joint optimization with these backbones. The results reveal a clear trend. Our method is competitive with PM in the balanced setting, but its advantage grows as preferences become more focused, culminating in the one-hot scenario. Here our method achieves a striking +5.3% average accuracy gain over PM (88.9% vs. 83.6%), showing a superior ability to satisfy specific targets without the severe performance degradation of prior work. Notably, this performance is so data-efficient that our method using just 10% of the calibration data still outperforms the fully-resourced PM baseline in the crucial priority and one-hot scenarios.

To specifically test controllability against PM, we designed a challenging experiment within the 8-task merge. We focused on a 3-task subspace (Cars, RESISC45, DTD) and allocated 60% of the total preference weight to them while sampling preferences across their simplex uniformly, and distributing the rest equally among the other five tasks. Table 2 reports the HV and Uniformity (U) from this experiment, measured on both the 3-task subspace (@3) and the full 8-task space (@8). Our method achieves substantially higher scores across all metrics in this challenging scenario. The strong @8 scores show this precision is achieved without disproportionately sacrificing the non-prioritized tasks, demonstrating a more robust and holistic trade-off capability. The high @3 scores confirm its precise control, and Figure 5 provides the visual evidence: for our method (b), the accuracy of each task peaks sharply at its corresponding preference corner (e.g., Cars accuracy is highest at corner C). In contrast, PM's response (a) is more diffuse, with accuracy peaks that are less pronounced and often misaligned with the intended preference.

To further probe this fine-grained control, we examine the direct trade-offs between pairs of tasks. We generated performance curves by smoothly varying the preference weight between two tasks (from 0 to 1), while setting the weights for all other six tasks to zero. Figure 4 visualizes these pairwise Pareto frontiers. The results highlight two critical advantages of our method. First, our approach consistently produces superior frontiers, offering a better accuracy trade-off than PM. Second, and more importantly, our method demonstrates robust controllability where PM fails. While our method produces smooth, predictable trade-off curves

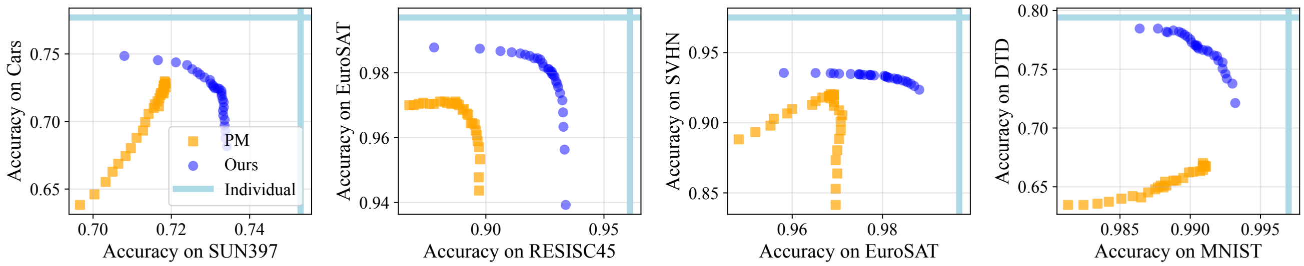
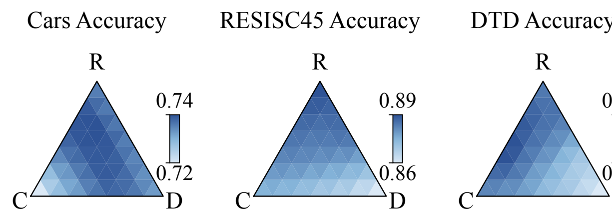


Figure 4: Pairwise performance trade-offs within an 8-task merge (AMPP backbone). Each subplot shows the accuracy on a task pair as preferences are varied between them. Our method (blue) consistently achieves a superior and more stable Pareto front compared to Pareto Merging (orange), which fails to produce controllable responses on several critical task pairs (e.g., SUN397-Cars, MNIST-DTD).

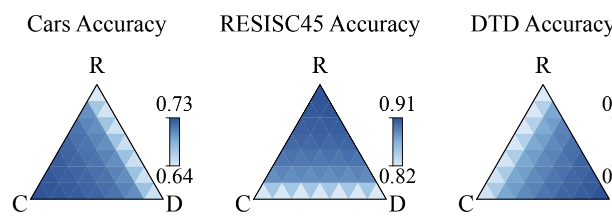
Method	HV@3 ↑	HV@8 ↑	U@3 ↑	U@8 ↑
AM+PM	76.77	60.90	53.47	31.98
AM+Ours	**83.95**	**68.49**	**63.30**	**44.77**
AMPP+PM	74.92	61.18	53.41	26.59
AMPP+Ours	**83.94**	**70.33**	**63.13**	**41.58**

Table 2: Performance comparison against Pareto Merging (PM). We report the Hypervolume (HV) and Uniformity (U) for both 3-task (@3) and 8-task (@8) settings.

across all pairs, PM's performance collapses on several critical pairings (e.g., SUN397-Cars, MNIST-DTD), failing to generate a controllable response. This shows our method's significantly enhanced reliability in satisfying specific user preferences.



(a) PM



(b) Ours

Figure 5: Visual evidence for the superior U@3 metrics in Table 2. On a 3-task (C: Cars, R: RESISC45, D: DTD) subspace of an 8-task merge, our method (b) shows sharp, predictable accuracy peaks. This ideal control landscape directly explains its higher Uniformity. In contrast, PM (a) yields a misaligned response, resulting in lower scores.

Efficiency and Scalability Analysis

To answer **Q2**, we analyze our method's efficiency and scalability, focusing on the critical cost of enabling preference-awareness. Our method's efficiency stems from its decoupled design, which operates on a pre-merged backbone in a distinct, subsequent step. This contrasts with approaches like PM that require a costly joint optimization of their preference-aware mechanism alongside the backbone merge itself. Figure 6 plots the computation time for the generation of the preference-aware model (for MAP) or the auxiliary module (for PM and Ours), confirming our method's linear time complexity. For an 8-task merge, our method completes in just 0.056 GPU hours, achieving an approximately 36x speedup over PM and a 208x speedup over MAP. For PM, we plot its reported costs for 2 and 8 tasks—the only configurations detailed in their work—and connect them with a dashed line to illustrate the scaling trend of its iterative optimization framework. This efficiency advantage is critical for practical applications and enables on-the-fly generation of tailored models.

Ablation and Mechanism Validation

To answer **Q3**, we analyze our framework's key components and hyperparameters.

Efficacy of optimal orthogonal regularization. We first study the effect of the regularization strength, β, which balances data-fitting with structural preservation. As Figure 7 shows, ablating the regularization term ($\beta = 0$) leaves only the least-squares objective. This leads to overfitting on the calibration data, causing poor performance, especially when data is scarce. Introducing the regularization ($\beta > 0$) provides a substantial performance boost by pulling the solution towards an optimal orthogonal transformation, which prevents the representation space from distorting. The method is also robust to the specific choice of β, as performance remains stable across a wide range of values. The regularization enables high data efficiency. Even with only 10% of the test data, our method achieves near-optimal accuracy. Finally, the regularized solution slightly outperforms a pure or-

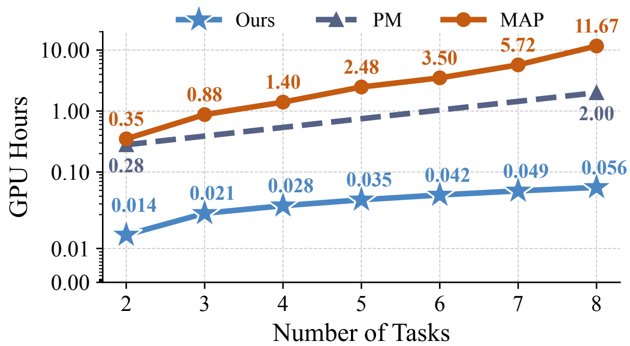


Figure 6: Scalability comparison. We plot the time for the preference-aware stage against the number of tasks. Note the logarithmic scale on the y-axis, which highlights the orders-of-magnitude difference in efficiency.

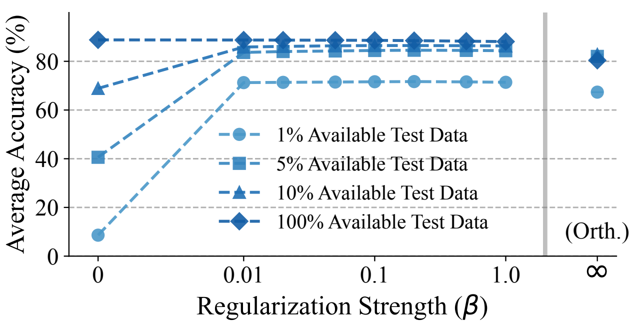


Figure 7: Accuracy vs. regularization strength β. Our method is robust to the choice of β and highly data-efficient, achieving near-optimal performance with only 10% of test data.

thogonal transformation ($\beta \to \infty$, noted as **Orth.**), confirming that our method effectively combines a strong structural prior with valuable data-driven correction.

Aggregation strategies. We compare our Pareto-optimal aggregation against two intuitive yet sub-optimal baselines for preference-aware weighting. The first (Naive) simply applies a weighted average to the optimal linear transformation of each task. The second (Polar) uses polar decomposition to separately weight the rotation and scaling components of each transformation before recombination. As shown in Table 3, while simpler heuristics can narrowly focus on the priority task, our principled aggregation achieves a much better overall trade-off. This result empirically validates our analysis: the data-aware weighting matrices C_t, which account for the feature structure of each task, are critical for finding a globally optimal solution, unlike Naive and Polar strategies.

Sufficiency of Linear Correction

To answer **Q4**, we compare our linear correction to the non-linear MLP-based from Representation Surgery (RS). As shown in Figure 8, with minimal data (1%), RS appears more robust, as our linear solution is sensitive to noise in

Method	**Equal Pref.**	**Priority Pref.**		
	NAcc.↑	**Pri. NAcc.↑**	**Non-Pri. NAcc.↑**	**HV↑**
Naive	92.2	97.1	87.7	67.6
Polar	91.5	**97.2**	84.9	64.0
Ours	**93.5**	96.0	**92.6**	**70.2**

Table 3: Comparison of aggregation strategies under equal and priority preference settings. We report Normalized Accuracy (NAcc.), Priority-task NAcc., Non-Priority NAcc., and Hypervolume (HV).

Figure 8: Linear vs. non-linear correction. Comparison of our global linear map against the non-linear MLP corrector from Representation Surgery (RS) across varying amounts of available test data. The dashed line indicates the accuracy of the merged model (AM) before any correction is applied.

this low-data regime. However, our method's performance rapidly surpasses the RS with just 5% of the data and widens its lead thereafter. This demonstrates that our linear approach is more data-efficient, better capturing the underlying global distortion once a minimum viable sample set is available.

Conclusion

We introduced ReACT, a new *on-the-fly* framework for controllable model merging that bypasses costly parameter-space optimization by directly correcting the model's representations. Our key insight is that this correction can be modeled as a simple, regularized linear map, for which we derived a closed-form, Pareto-optimal solution in the representation alignment space. This analytical approach leads to a highly efficient and scalable method. Experiments confirmed ReACT achieves a state-of-the-art Pareto front and superior preference alignment in terms of task performance.

However, our method has limitations: the linear scalarization may not cover the entirety of a concave Pareto front in the performance space, and it assumes predominantly linear distortions requiring a small unlabeled calibration dataset. We hope our work encourages further exploration of simple, analytical solutions in representation space for complex model combination problems, including extending to models with more complex representation structures and exploring calibration-free alternatives.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62071481.

References

Chen, W.; and Kwok, J. T. 2025. Pareto Merging: Multi-Objective Optimization for Preference-Aware Model Merging. In *Forty-Second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, July 13-19, 2025*. OpenReview.net.

Chen, Z.; Badrinarayanan, V.; Lee, C.; and Rabinovich, A. 2018. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 793–802. PMLR.

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

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Dimitriadis, N.; Frossard, P.; and Fleuret, F. 2025. Pareto Low-Rank Adapters: Efficient Multi-Task Learning with Preferences. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Du, G.; Lee, J.; Li, J.; Jiang, R.; Guo, Y.; Yu, S.; Liu, H.; Goh, S. K.; Tang, H.; He, D.; and Zhang, M. 2024. Parameter Competition Balancing for Model Merging. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Fleischer, M. 2003. The measure of Pareto optima applications to multi-objective metaheuristics. In *International Conference on Evolutionary Multi-Criterion Optimization*, 519–533. Springer.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Huang, C.; Ye, P.; Chen, T.; He, T.; Yue, X.; and Ouyang, W. 2024. EMR-Merging: Tuning-Free High-Performance Model Merging. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jin, X.; Ren, X.; Preotiuc-Pietro, D.; and Cheng, P. 2023. Dataless Knowledge Fusion by Merging Weights of Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, L.; Zhang, T.; Bu, Z.; Wang, S.; He, H.; Fu, J.; Wu, Y.; Bian, J.; Chen, Y.; and Bengio, Y. 2025. MAP: Low-compute Model Merging with Amortized Pareto Fronts via Quadratic Approximation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Lu, Z.; Fan, C.; Wei, W.; Qu, X.; Chen, D.; and Cheng, Y. 2024. Twin-Merging: Dynamic Integration of Modular Expertise in Model Merging. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Mahapatra, D.; and Rajan, V. 2020. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, 6597–6607. PMLR.

Matena, M.; and Raffel, C. 2022. Merging Models with Fisher-Weighted Averaging. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4. Granada.

Qi, B.; Li, F.; Wang, Z.; Gao, J.; Li, D.; Ye, P.; and Zhou, B. 2025. Less Is More: Efficient Model Merging with Binary Task Switch. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15265–15274.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021,* *18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Singh, S. P.; and Jaggi, M. 2020. Model Fusion via Optimal Transport. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332.

Wang, K.; Dimitriadis, N.; Ortiz-Jiménez, G.; Fleuret, F.; and Frossard, P. 2024. Localizing Task Information for Improved Model Merging and Compression. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C. A.; and Bansal, M. 2023. TIES-Merging: Resolving Interference When Merging Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yang, E.; Shen, L.; Wang, Z.; Guo, G.; Chen, X.; Wang, X.; and Tao, D. 2024a. Representation Surgery for Multi-Task Model Merging. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yang, E.; Wang, Z.; Shen, L.; Liu, S.; Guo, G.; Wang, X.; and Tao, D. 2024b. AdaMerging: Adaptive Model Merging for Multi-Task Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yu, L.; Yu, B.; Yu, H.; Huang, F.; and Li, Y. 2024. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient Surgery for Multi-Task Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zitzler, E.; and Thiele, L. 1999. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4): 257–271.