Geometry-Aware Directional Alignment for Coherent Model Merging

Zhikang Chen[1,†] **Sen Cui**[2] **Deheng Ye**[3] **Min Zhang**[4] **Gang Niu**[5]
Yu Zhang[7,†] **Masashi Sugiyama**[5,6] **Tingting Zhu**[1]

[1]University of Oxford [2]Tsinghua University [3]Nanyang Technological University
[4]East China Normal University [5]RIKEN [6]University of Tokyo
[7]Southern University of Science and Technology

[†]Corresponding authors: reub0770@ox.ac.uk, zhangy7@sustech.edu.cn

Abstract

Model merging integrates independently fine-tuned models without joint retraining. Existing methods combine parameters through decomposition, coefficient optimization, or subspace learning, often assuming compatible directional structures. Independently trained models can instead exhibit misaligned directions in parameter and feature spaces, causing destructive interference and degraded performance. Insights from Neural Collapse further suggest that class features follow structured but model-specific directional patterns. We identify directional alignment as a principle for effective model merging and propose Merging with Directional Alignment (MDA), a unified geometric framework that enforces directional consistency in both spaces. Parameter-space alignment reconstructs a shared low-rank representation and aligns it with an ETF-inspired geometric reference. Feature-space alignment jointly optimizes fusion coefficients and task-specific rotations. Our analysis characterizes structural coherence and generalization under the stated assumptions. Experiments across vision, language, and healthcare QA benchmarks demonstrate consistent improvements across model scales and task settings.

1 Introduction

Fine-tuning has become the standard approach for adapting large pretrained models to downstream tasks [Hu et al., 2022, Muqeeth et al., 2024]. However, as the number of tasks grows, maintaining a separate set of tuned parameters or adapters for each task leads to storage costs that scale linearly with the task count. In practical scenarios such as edge deployment and multi-task services, this overhead quickly becomes prohibitive. To address this issue, model merging [Ilharco et al., 2022] has emerged. By combining task-specific fine-tuned weights into a single shared model, it significantly reduces storage requirements while retaining task performance.

Existing model merging methods can be broadly categorized into two families: the data-free approach and the data-based approach. Data-free approaches mostly exploit the algebraic properties of the parameter space, employing strategies such as weighted averaging [Ilharco et al., 2022, Matena and Raffel, 2022], subspace projection [Wei et al., 2025], or matrix decomposition [Gargiulo et al., 2025] to alleviate task interference. These methods do not require access to task data, which makes them attractive in terms of privacy and scalability. In contrast, data-based approaches leverage real or synthetic data to learn or optimize fusion coefficients [Yang et al., 2023a, Tang et al., 2023], thus preserving the performance of every task after merging. Although both paradigms have clear advantages, the former is broadly applicable as it only leverages the intrinsic properties of parameters, and the latter offers stronger performance guarantees, most of the existing work has primarily focused

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

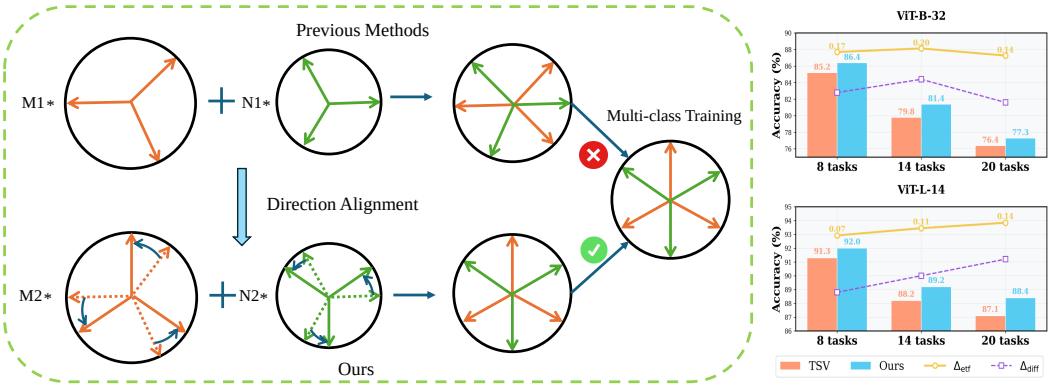


Figure 1: **Left:** Comparison with prior methods. Our approach builds a simplex Equiangular Tight Frame (ETF) to align task-specific parameter directions, achieving better directional consistency and more closely approximating multi-class joint training. M1, N1, M2, N2 denote fusion coefficients rather than model or classifier weights. **Right:** Correlation between directional deviation Δ_{ETF} (from the ideal ETF) and the performance gap Δ_{diff} across tasks. Compared with *task singular vectors* (TSV) [Gargiulo et al., 2025], our method shows a clear relationship: larger deviations from the ETF structure consistently correspond to larger performance gaps.

on fusion coefficient optimization. What has been largely overlooked is the role of directional information, i.e., how the direction of high-dimensional parameter or feature space fundamentally shapes the behavior of the merged model. This raises an important question:

> *Should we jointly optimize both aspects for more effective model fusion, since vectors in the parameter or feature space are fundamentally characterized by both magnitude and direction?*

We address directional incompatibility in both parameter and feature spaces. A task vector is the difference between fine-tuned and pretrained parameters [Ilharco et al., 2022]. Independently trained models often encode task-specific information along different dominant parameter directions; combining these directions can introduce destructive interference. We use the simplex *Equiangular Tight Frame* (ETF) as an idealized geometric reference for directional separation. In feature space, Neural Collapse [Papyan et al., 2020] motivates structured directional templates for classifier weights and class means under balanced data and suitable training conditions. Figure 1 (left) illustrates the directional inconsistencies between naively merged representations and those obtained by joint training, motivating explicit alignment.

This perspective motivates us to ask: *to what extent does the failure of model merging stem from deviations from this ideal ETF geometry?* To answer this, we measure how closely the merged space align with the ETF structure and whether such deviations correlate with performance degradation. Specifically, we quantify this deviation as: $\Delta_{\text{etf}} = \Delta_{\text{Ours}} - \Delta_{\text{TSV}}, \Delta_{\text{Ours}} = \frac{1}{T}\sum_{t=1}^{T} \cos\big(W^t_{\text{Ours}}, W^t_{\text{ETF}}\big), \Delta_{\text{TSV}} = \frac{1}{T}\sum_{t=1}^{T} \cos\big(W^t_{\text{TSV}}, W^t_{\text{ETF}}\big), \cos(A,B) = \frac{\langle A,B\rangle_F}{\|A\|_F\|B\|_F}$, where W^t_{Ours}, W^t_{TSV}, and W^t_{ETF} denote the parameter representations produced by our method, TSV [Gargiulo et al., 2025], and the ETF reference for task t. Here T is the number of tasks, $\langle A, B\rangle_F = \text{Tr}(A^\top B)$, and $\|A\|_F = \sqrt{\langle A, A\rangle_F}$. Thus Δ_{Ours} and Δ_{TSV} measure cosine similarity to the reference, and Δ_{etf} measures their difference. Figure 1 (right) shows its correlation with the performance gap Δ_{diff} across eight task configurations on ViT-B/32 and ViT-L/14, supporting the importance of directional alignment.

Motivated by the above analysis, we propose a novel *direction alignment* framework that can operate flexibly in both parameter and feature spaces. Concretely, in parameter space, we first apply low-rank decomposition to task-specific parameter vectors and reconstruct a shared parameter space. The reconstructed parameters are aligned with an ETF-inspired reference to encourage geometric coherence. In feature space, we introduce a joint optimization framework that simultaneously learns fusion coefficients and task-specific rotation matrices. This process aligns the extracted features with the ETF structure, while balancing entropy minimization, alignment consistency, and rotation regularization. Together, these two components, parameter-space alignment and feature-space

alignment, enable the merged model to narrow the gap to mitigating the performance degradation. Our contributions are threefold:

1. We highlight the critical role of directional alignment, which becomes particularly significant in model merging, encompassing both parameter and feature spaces.

2. We theoretically demonstrate the structural advantages introduced by directional alignment, showing that it promotes more coherent geometric organization in both parameter and feature spaces and yields provable generalization benefits.

3. We validate our approach through extensive experiments across diverse benchmarks and task configurations, varying in class counts, model scales, and data availability. Our results demonstrate the effectiveness of the proposed method, showing that it consistently improves performance across diverse tasks and datasets. The source codes of our framework are made publicly available at `https://github.com/chenzk202212/MDA`.

2 Related Work

Data-free model merging Data-Free Model Merging methods [Ilharco et al., 2022, Chen et al., 2025, Yang et al., 2026] operate solely on model parameters, offering significant practical advantages by eliminating the need for any training or validation data. Task Arithmetic [Ilharco et al., 2022] introduces the fundamental concept of task vectors, the difference between fine-tuned and pre-trained weights, and merges them through simple arithmetic averaging. TIES-Merging [Yadav et al., 2023] enhances this approach by trimming low-magnitude parameters, resolving sign conflicts, and performing disjoint merging of sparse vectors. Fisher Merging [Matena and Raffel, 2022] employs the Fisher Information matrix to prioritize important parameters during weighted averaging. More recent advances include Consensus Merging [Wang et al., 2024], which identifies and removes selfish parameters that benefit only specific tasks, and Adaptive Weight Disentanglement (AWD) [Xiong et al., 2024] that explicitly promotes orthogonality among task vectors to reduce interference. Task Singular Vectors (TSV) [Gargiulo et al., 2025] represents a geometric approach by applying singular value decomposition to per-layer task matrices, enabling compression and interference reduction through whitening transformations. Similarly, DOGE [Wei et al., 2025] formulates merging as a constrained optimization problem solved via adaptive projective gradient descent within a shared subspace. Most recently, Isotropic Merging (ISO) [Marczak et al., 2025] proposes to flatten the singular value spectrum of task matrices to enhance subspace alignment between task-specific and merged models, achieving state-of-the-art performance through both common and task-specific subspace integration. Although previous data-free methods have achieved notable performance improvements, most of them rely on decomposing the parameter space into different subspaces. Such partitioning breaks the global geometric coherence of the parameter space among tasks. In contrast, our approach introduces a new paradigm that performs **global parameter-space alignment** from a unified geometric perspective rather than separating the space into different components.

Data-based model merging Data-Based Model Merging methods leverage additional data, typically unlabeled examples from the test distribution, to guide the merging process and achieve improved performance. AdaMerging [Yang et al., 2023a] treats merging coefficients as learnable parameters optimized by minimizing the prediction entropy in test data. Representation Surgery [Yang et al., 2024b] extends this approach by explicitly aligning the intermediate representations of the merged model with those of individual task-specific models, using test data to calibrate feature-level biases. These methods generally achieve higher accuracy by adapting to the target data distribution, but this advantage comes with significant costs: the requirement for data collection introduces practical constraints and potential privacy concerns, the additional computation during merging increases resource requirements, and the performance gains are often sensitive to the quality and representativeness of the provided data. Moreover, these approaches typically overlook the geometric implications of neural collapse in feature space, failing to preserve the optimal directional relationships between class prototypes that emerge in well-trained models. To address this, our framework extends the same directional alignment principle into the feature space, where we not only consider optimizing the fusion coefficients but also explicitly enforcing the directional coherence among task representations.

3 Preliminaries

3.1 Task Vector

We draw inspiration from prior work [Ilharco et al., 2022] that introduces the concept of task vectors and leverages task arithmetic to combine task-specific vectors from multiple tasks, thereby enabling the merged representation to inherit multi-task capabilities and achieve better generalization. For a given task k, its corresponding task vector $\boldsymbol{\tau}_k$ is computed as the vector by subtracting the pre-trained parameters θ_0 from the fine-tuned parameters θ_k: $\boldsymbol{\tau}_k = \theta_k - \theta_0$.

Furthermore, multiple task vectors $\{\boldsymbol{\tau}_k\}_{k=1}^K$ are aggregated and combined with the pre-trained model through a weighted summation, yielding the multi-task model: $\theta_{\text{MTL}} = \theta_0 + \lambda \sum_{k=1}^K \boldsymbol{\tau}_k$, where the merging coefficient λ denotes the task-wise importance coefficient for adaptive vector fusion.

To achieve finer-grained control over the fusion process, our method, inspired by Yang et al. [2023a], performs layer-wise adaptation of task vectors $\boldsymbol{\tau}_k^{(l)} = \theta_k^{(l)} - \theta_0^{(l)}$. Thus, the formulation can be generalized as follows: $\theta^{(l)} = \theta_0^{(l)} + \lambda^{(l)} \sum_{k=1}^K \boldsymbol{\tau}_k^{(l)}$, where the merging coefficient $\lambda^{(l)}$ denotes the layer-wise importance coefficient for vector fusion.

3.2 Neural Collapse

Neural collapse [Papyan et al., 2020] refers to a highly structured phenomenon that emerges in the terminal phase of training on balanced datasets (i.e., after achieving a near-zero training error). Specifically, the last-layer features and classifier weights converge towards a simplex ETF geometry, formally defined as follows:

Definition 1 (Simplex Equiangular Tight Frame)**.** *A simplex ETF is a matrix $W_{ETF} = [w_1, w_2, \ldots, w_C] \in \mathbb{R}^{d\times C}$ composed of C vectors and $w_i \in \mathbb{R}^d$ with $d \geq C-1$, satisfying:*

$$W_{ETF} = \sqrt{\frac{C}{C-1}}\mathbf{U}\left(\mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top\right), \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{d\times C}$ satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top$. For any $w_i, w_j \in W_{ETF}$, we have

$$w_i^\top w_j = \frac{C}{C-1}\delta_{i,j} - \frac{1}{C-1}, \quad \forall i,j \in [1,C], \tag{2}$$

with $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

Building on this definition, neural collapse can be characterized by four convergent properties:

NC1 (Variability collapse). Intra-class feature variability vanishes, and features from the same class collapse to their class mean: $\Sigma_W^c = \frac{1}{n_c}\sum_{i=1}^{n_c}(\mathbf{h}_{c,i} - \mathbf{h}_c)(\mathbf{h}_{c,i} - \mathbf{h}_c)^T \to 0$, where n_c is the number of samples in class c, $\mathbf{h}_{c,i}$ is the feature of the i-th sample, and $\mathbf{h}_c$ denotes the class mean of class c.

NC2 (Convergence to simplex ETF). Class means become maximally separated, forming a centered simplex ETF: $\hat{\mathbf{h}}_c = \frac{\mathbf{h}_c - \mathbf{h}_G}{\|\mathbf{h}_c - \mathbf{h}_G\|}, \mathbf{h}_G = \frac{1}{\sum_k n_c}\sum_{k=1}^K\sum_{i=1}^{n_c}\mathbf{h}_{c,i}$.

NC3 (Self-duality with classifier weights). The normalized class means align with the corresponding classifier weights: $\hat{\mathbf{h}}_c = \frac{w_c}{\|w_c\|}$, where w_c denotes the classifier weight vector for class c.

NC4 (Nearest class center decision rule). The classifier reduces to a nearest-class-mean classifier: $\arg\max_c \langle \mathbf{h}, w_c\rangle = \arg\min_c \|\mathbf{h} - \mathbf{h}_c\|$, where $\mathbf{h}$ is the feature of a test sample.

While standard Neural Collapse focuses on the final classifier layer, ETF-like structures also progressively emerge in intermediate layers [Parker et al., 2023]. Concurrently, the Neural Feature Ansatz (NFA) [Beaglehole et al., 2024a] reveals that weight matrices and feature covariances co-evolve during training, jointly aligning toward ETF-consistent directions to promote inter-class separation. Motivated by these insights, we operate under the assumption that the column space of each layer's weight matrix can be approximated by a near-ETF subspace. This theoretically and empirically grounded assumption captures the weight-feature co-alignment dynamics across depth, provides a tractable geometric foundation for our layer-wise regularization, and explains the empirical correlation between the directional deviation Δ_{ETF} and the performance gap Δ_{diff} (Figure 1). Consequently, it serves as the foundational premise for our subsequent derivations.

Algorithm 1 Parameter-Space Alignment

Input: Task vectors $\{\tau_t\}_{t=1}^T$, rank k
Output: ETF-guided task vectors $\tau_{\text{etf}} = \{\tau_t^{\text{etf}}\}_{t=1}^T$
for $t = 1$ to T **do**
  SVD: $\tau_t^{(l)}$: $\tau_t^{(l)} = U_t^{(l)}\Sigma_t^{(l)}(V_t^{(l)})^\top$
  Keep top-k components: $U_t^{(l)}[:, 1:k]$
end for
Concatenate $\{U_t^{(l,k)}\}_{t=1}^T$ to form $U_{\text{cat}}^{(l)}$
$U_{\text{cat}}^{(l)} = U_{\text{share}}^{(l)}\Sigma_{\text{share}}^{(l)}(V_{\text{share}}^{(l)})^\top$
Truncate $U_{\text{share}}^{(l)}$ to its first d_{out} columns to obtain $\tau_{\text{share}}^{(l)}$
Build frame W_{ETF} as in Appendix B.1; the ideal simplex satisfies:
$(W_{\text{ETF}}W_{\text{ETF}}^\top)_{ii} = ||w_i||^2 = 1$
$(W_{\text{ETF}}W_{\text{ETF}}^\top)_{ij} = w_i w_j^\top = -\frac{1}{C-1}, i \neq j$
$\tau_{\text{etf}} \leftarrow \tau_{\text{share}}W_{\text{ETF}}^\top W_{\text{ETF}}$

Algorithm 2 Feature-Space Alignment

Input: ETF-guided task vectors $\tau_{\text{etf}} = \{\tau_t^{\text{etf}}\}_{t=1}^T$, the total number of layers in the model is L, datasets $\{\mathcal{D}_t\}_{t=1}^T$, epochs E
Output: Fusion coefficients $\boldsymbol{\lambda}$, rotation matrices $\{\mathbf{R}^t\}_{t=1}^T$
Initialize $\boldsymbol{\lambda_0} = \frac{1}{T}\mathbf{1}_T$, $\mathbf{R}^t = \mathbf{I}_d$; construct ETF M^*
// Layer-wise Fusion
for $l = 1$ to L **do**
  Fusion: $\theta^{(l)} = \theta_0^{(l)} + \lambda^{(l)} \sum_{t=1}^T \tau_t^{\text{etf},(l)}$
end for
for $e = 1$ to E **do**
  Extract rotated features: $\tilde{\mathbf{h}}_{t,i} = \mathbf{R}^t\phi_{\boldsymbol{\theta}}(\mathbf{x}_{t,i})$
  Compute $\mathcal{L} = \mathcal{L}_{\text{entropy}} + \alpha\mathcal{L}_{\text{align}} + \beta\mathcal{L}_{\text{rotation}}$
  Update $\lambda^{(l)}$ and R^t jointly via the total loss $\mathcal{L}$
end for

4 Merging with Directional Alignment

Most prior works either partition the parameter space into different subspaces [Gargiulo et al., 2025, Wei et al., 2025, Marczak et al., 2025], or formulate model merging as an optimization problem over task-specific fusion coefficients [Yang et al., 2023a]. In contrast, our method provides a unified geometric perspective that views model merging through the lens of directional alignment. In the parameter space, we align the shared subspace with the simplex ETF directions to ensure globally consistent task representations. In the feature space, we leverage the ETF structure naturally formed between the final-layer features and classifier weights, and jointly optimize the fusion coefficients and rotation matrices to achieve feature-level directional consistency. By aligning task updates with the simplex ETF geometry, we bridge parameter-space and feature-space alignment within a single framework, and further provide a theoretical analysis.

4.1 Parameter-space Directional Alignment

We conduct parameter-space analysis under data-free conditions, focusing on optimizing the intrinsic geometric properties of multi-task parameters. Although the subsequent theoretical analysis could, in principle, apply to any τ_{share} satisfying $\tau_{\text{etf}} = \tau_{\text{share}}P_{\text{ETF}}$, we specifically adopt the SVD-based construction for consistency with our overall framework and in line with prior works [Gargiulo et al., 2025, Marczak et al., 2025, Wei et al., 2025]. This choice not only ensures comparability with existing model-merging approaches but also provides a statistically optimal low-rank approximation that captures the dominant shared components across tasks.

For tasks $\{T_1, T_2, \ldots, T_T\}$ and layers $l = 1, \ldots, L$, we first obtain their SVD decompositions:

$$\tau_t^{(l)} = U_t^{(l)}\Sigma_t^{(l)}(V_t^{(l)})^\top.$$

We retain the top-k principal components $U_t^{(l,k)} := U_t^{(l)}[:, 1:k]$ and construct a shared representation

$$\tau_{\text{share}}^{(l)} := U_{\text{share}}^{(l)}[:, 1:d_{\text{out}}] \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}},$$

where $U_{\text{share}}^{(l)}$ is obtained from the SVD of the concatenated basis

$$U_{\text{cat}}^{(l)} = \left[U_1^{(l,k)}, U_2^{(l,k)}, \cdots, U_T^{(l,k)}\right] \in \mathbb{R}^{d_{\text{in}} \times kT}, U_{\text{cat}}^{(l)} = U_{\text{share}}^{(l)}\Sigma_{\text{share}}^{(l)}(V_{\text{share}}^{(l)})^\top$$

Here d_{in} is the parameter input dimension at layer l, and d_{out} is the parameter output dimension. We set $k = d_{\text{out}}/T$ so that the concatenated dimension kT exactly matches the output dimension d_{out}, ensuring a balanced representation across all T tasks. The proposed method is illustrated through Algorithm 1 and Figure 3 in Appendix F.

The ETF-aligned task parameter space at each layer is obtained through the following direction alignment:

$$\tau_{\text{etf}}^{(l)} = \tau_{\text{share}}^{(l)} W_{\text{ETF}}^{\top} W_{\text{ETF}} = \tau_{\text{share}}^{(l)} \mathcal{P}_{\text{ETF}} \tag{3}$$

where $W_{\text{ETF}} \in \mathbb{R}^{C \times d_{\text{out}}}$ denotes an ETF-inspired frame, $C = \sum_t C_t$ classes, each task t corresponds to a subset of C_t classes, $\mathcal{P}_{\text{ETF}} = W_{\text{ETF}}^{\top} W_{\text{ETF}}$ is the direction alignment matrix, ETF-guided task vectors $\tau_{\text{etf}} = \{\tau_t^{\text{etf}}\}_{t=1}^{T}$ are obtained by projecting the shared parameter subspace τ_{share} onto the reference frame. Exact simplex identities require $d_{\text{out}} \geq C - 1$; numerical frames follow the stated compatibility conditions.

Theoretical Analysis. To understand the benefits of ETF-guided directional alignment, we provide theoretical analyses from two complementary perspectives: *ETF Structural Coherence* and *Generalization Advantage*. The former characterizes how directional alignment improves the approximation to jointly-trained parameters, while the latter examines its effect on statistical complexity. Detailed proofs are provided in Appendix B.2.

Theorem 1 (ETF Structural Coherence)**.** *Let $\tau_{\text{ideal}}^{(l)}$ denote the oracle jointly-trained parameters and $\tau_{\text{share}}^{(l)}$ the SVD-reconstructed parameters at layer l. Define $\mathcal{P}_{\text{ETF}} = W_{\text{ETF}}^{\top} W_{\text{ETF}}$ and let $\Pi_{\mathcal{S}}$ be the orthogonal projector onto $\mathcal{S} = \text{rowspan}(W_{\text{ETF}})$. Under the directional compatibility conditions in Appendix B.2.1, the aligned parameters $\tau_{\text{etf}}^{(l)} = \tau_{\text{share}}^{(l)} \mathcal{P}_{\text{ETF}}$ satisfy*

$$\|\tau_{\text{ideal}}^{(l)} - \tau_{\text{etf}}^{(l)}\|_F^2 \leq \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{share}}^{(l)}\|_F^2 - g \left\| \tau_{\text{share}}^{(l)} (I - \Pi_{\mathcal{S}}) \right\|_F^2, \tag{4}$$

where $g \in (0, 1]$ is the positive directional correction gain defined in the Appendix. For an exact simplex ETF, $\mathcal{P}_{\text{ETF}} = \frac{C}{C-1} \Pi_{\mathcal{S}}$, and the correction term is equivalently

$$\left\| \tau_{\text{share}}^{(l)} \left(I - \tfrac{C-1}{C} \mathcal{P}_{\text{ETF}} \right) \right\|_F^2. \tag{5}$$

Theorem 1 establishes a geometric connection between directional coherence and approximation to an oracle jointly-trained model. The residual $\|\tau_{\text{share}}^{(l)}(I - \Pi_{\mathcal{S}})\|_F^2$ measures the parameter energy outside the aligned subspace, where task-specific deviations and inconsistent update directions may accumulate. Under the stated compatibility conditions, ETF-guided alignment suppresses these components while controlling distortion within the retained subspace, yielding a smaller approximation error. Importantly, the correction depends on the magnitude of the off-subspace residual: larger directional discrepancies offer greater potential for improvement when the correction gain remains positive. Thus, SVD reconstruction and ETF alignment play complementary roles: the former integrates dominant task-specific spectral components, whereas the latter further regulates their directional organization.

Theorem 2 (Generalization Bound)**.** *Let $\mathcal{R}(\cdot)$ and $\hat{\mathcal{R}}(\cdot)$ denote population and empirical risks. Consider a fixed-feature linearized model $z_\tau(x) = z_0(x) + \tau^{\top} \phi(x)$, where $\|\phi(x)\|_2 \leq R$ and the loss is bounded and L-Lipschitz in the logits. Let $r = \dim(\mathcal{S}) \leq d_{\text{out}}$. Suppose $\tau_{\text{etf}} = \tau_{\text{share}} \mathcal{P}_{\text{ETF}}$ and both parameter norms are bounded by predetermined constants B_{share} and B_{etf}, respectively. If both models achieve the same empirical risk $\hat{\mathcal{R}}^*$, then, with probability at least $1 - \delta$, simultaneously,*

$$\begin{aligned} \mathcal{R}(\tau_{\text{share}}) &\leq \hat{\mathcal{R}}^* + \frac{2\sqrt{2} L R B_{\text{share}}}{\sqrt{n}} \sqrt{d_{\text{out}}} + c_\delta, \\ \mathcal{R}(\tau_{\text{etf}}) &\leq \hat{\mathcal{R}}^* + \frac{2\sqrt{2} L R B_{\text{etf}}}{\sqrt{n}} \sqrt{r} + c_\delta, \end{aligned} \tag{6}$$

where $c_\delta = 3\sqrt{\log(4/\delta)/(2n)}$. Consequently, the aligned model admits a tighter generalization upper bound when

$$B_{\text{etf}} \sqrt{r} < B_{\text{share}} \sqrt{d_{\text{out}}}. \tag{7}$$

For an exact simplex ETF, $r = C - 1$.

Theorem 2 complements the structural-coherence result by characterizing the inductive bias introduced by directional alignment. From a Rademacher-complexity perspective, restricting parameter

variation to a structured subspace reduces an upper bound on the effective hypothesis complexity, provided that the aligned parameter norm remains sufficiently controlled. The reduction depends on the dimension of the retained subspace rather than the full output dimension, yielding a potentially tighter generalization bound without requiring additional model capacity.

The two results also reveal a common geometric mechanism. The off-subspace residual

$$\frac{\|\tau_{\text{share}}(I - \Pi_{\mathcal{S}})\|_F^2}{\|\tau_{\text{share}}\|_F} \tag{8}$$

characterizes the magnitude of directional inconsistency before alignment. A larger residual indicates greater potential for geometric correction, while the restricted subspace limits unconstrained parameter variation. These properties help explain the observed relationship between ETF deviation and merging performance in Figure 1, although the theoretical bounds do not imply unconditional accuracy gains.

Taken together, Theorems 1 and 2 provide complementary explanations for the effectiveness of MDA. ETF-guided alignment acts as both a *directional correction mechanism*, which can reduce interference inherited from independently optimized tasks, and a *structural regularizer*, which constrains the merged representation toward a coherent and lower-complexity geometric organization. These effects provide theoretical support for the empirical improvements observed across different task configurations and model scales.

4.2 Feature-space Directional Alignment

We introduce task-specific directional alignment matrices and jointly optimize them with fusion coefficients to enhance feature-space consistency and mitigate task interference. Algorithms 1 and 2 are applied sequentially: Algorithm 1 first constructs the ETF-guided merged task vector τ_{etf} through parameter-space alignment, while Algorithm 2 further optimizes fusion coefficients and feature alignment matrices. The ETF reference is predefined. Prediction-based pseudo-labels guide alignment and prototype estimation without ground-truth labels.

Definition 2 (Task Vector Fusion with Feature Direction Alignment)**.** *The fusion model combines ETF-guided parameters and applies task-specific feature transformations:*

$$\begin{aligned} \theta^{(l)} &= \theta_0^{(l)} + \lambda^{(l)} \tau_{\text{etf}}^{(l)}, \\ \tilde{\mathbf{h}}_{t,i} &= \mathbf{R}^t \mathbf{h}_{t,i}, \quad \mathbf{R}^t \in \mathbb{R}^{d \times d}. \end{aligned} \tag{9}$$

Here, $\lambda^{(l)}$ is the trainable layer-wise fusion coefficient, $\mathbf{h}_{t,i}$ denotes the merged model's feature for sample i of task t, and $\mathbf{R}^t$ is the task-specific alignment matrix. In the orthogonal case $\mathbf{R}^t \in SO(d)$, feature norms and pairwise angles are preserved.

We jointly optimize fusion coefficients and alignment matrices through three complementary objectives:

$$\min_{\{\lambda^{(l)}\},\{\mathbf{R}^t\}} \mathcal{L} = \mathcal{L}_{\text{entropy}} + \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{rotation}}, \tag{10}$$

where α and β balance feature alignment and rotation regularization. In the implementation, β is the configured rotation multiplier times the current α.

Entropy Minimization. Following Yang et al. [2023a], we minimize prediction entropy on samples:

$$\mathcal{L}_{\text{entropy}} = -\sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{c=0}^{C_t-1} p_{t,i,c} \log p_{t,i,c}, \tag{11}$$

where $p_{t,i,c}$ denotes the merged model's predicted probability for class c. This encourages confident predictions.

ETF Feature Alignment. Inspired by Neural Collapse, we construct a fixed ETF reference $\mathcal{M}^* = \{\mathbf{m}_c^*\}_{c=1}^C$. For an exact simplex ETF, its normalized directions satisfy

$$\langle \mathbf{m}_c^*, \mathbf{m}_{c'}^* \rangle = \begin{cases} 1, & c = c', \\ -\frac{1}{C-1}, & c \neq c'. \end{cases}$$

When the feature dimension is insufficient for an exact simplex ETF, we adopt an ETF-inspired directional frame.

At each update, the current model predicts the hard pseudo-label $\hat{y}_{t,i} = o_t + \arg\max_{0 \le c < C_t} p_{t,i,c}$, where o_t is the task-specific class offset. The index is held fixed when selecting the ETF target. The alignment loss is

$$\mathcal{L}_{\text{align}} = \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \left\| \frac{\tilde{\mathbf{h}}_{t,i}}{\|\tilde{\mathbf{h}}_{t,i}\|_2} - \mathbf{m}^*_{\hat{y}_{t,i}} \right\|_2^2. \tag{12}$$

The entropy term encourages confident predictions, while the alignment term uses prediction-based pseudo-labels.

Directional Rotation Regularization. We further encourage each alignment matrix to remain close to an optimal Procrustes rotation:

$$\mathcal{L}_{\text{rotation}} = \frac{1}{d^2} \sum_{t=1}^{T} \left\| \mathbf{R}^t - \mathbf{R}^t_{\text{proc}} \right\|_F^2, \tag{13}$$

where the reference rotation aligns empirical feature prototypes with ETF targets:

$$\mathbf{R}^t_{\text{proc}} = \arg \min_{\mathbf{R} \in SO(d)} \left\| M^t \mathbf{R}^\top - M^{t*} \right\|_F^2. \tag{14}$$

For the classes present in the current task batch, define $\mathcal{I}_{t,c} = \{i : \hat{y}_{t,i} = c\}$ and the normalized pseudo-class means

$$\begin{aligned} \bar{\mathbf{h}}_{t,c} &= \frac{1}{|\mathcal{I}_{t,c}|} \sum_{i \in \mathcal{I}_{t,c}} \frac{\mathbf{h}_{t,i}}{\|\mathbf{h}_{t,i}\|_2}, \\ [M^t]_{c,:} &= \frac{\bar{\mathbf{h}}^\top_{t,c}}{\|\bar{\mathbf{h}}_{t,c}\|_2}. \end{aligned} \tag{15}$$

M^{t*} contains the corresponding ETF target rows. The Procrustes reference is computed from unrotated features and treated as a fixed target during each update. This regularizer promotes consistency between feature transformations and the reference geometry; setting $\beta = 0$ removes the Procrustes regularization.

Theorem 3 (Rademacher Complexity Bound)**.** *Consider the orthogonal specialization of the rotation-merged hypothesis class*

$$\mathcal{H}_{\text{rot}} = \left\{ x \mapsto g\left(\sum_{t=1}^{T} \theta_t^\top \mathbf{R}^t \phi(x) \right) : \mathbf{R}^t \in SO(d) \right\},$$

where ϕ and $\{\theta_t\}$ are fixed independently of the adaptation sample, $\|\phi(x)\|_2 \le R$, $\sum_t \|\theta_t\|_2^2 \le B^2$, and g is 1-Lipschitz. For a bounded loss that is L-Lipschitz in predictions, the expected Rademacher complexity satisfies

$$\mathfrak{R}_n(\mathcal{H}_{\text{rot}}) \le \frac{BR\sqrt{T}}{\sqrt{n}}. \tag{16}$$

With probability at least $1 - \delta$, the population risk satisfies

$$\mathcal{R}(\hat{h}) \le \hat{\mathcal{R}}(\hat{h}) + \frac{2LBR\sqrt{T}}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{17}$$

Furthermore, restricting the rotations to a fixed, sample-independent ETF-guided neighborhood $\sum_t \|\mathbf{R}^t - \mathbf{R}^t_*\|_F^2 \le \rho^2$ yields

$$\mathfrak{R}_n(\mathcal{H}_{\text{align}}(\rho)) \le \frac{BR}{\sqrt{n}} \min\{\sqrt{T}, \rho\}. \tag{18}$$

Thus, when $\rho < \sqrt{T}$, the restricted hypothesis class admits a tighter generalization upper bound.

Theorem 3 characterizes the generalization properties of feature-space alignment through Rademacher complexity. Orthogonal transformations preserve feature norms, while restricting their variation around a fixed geometric reference can reduce hypothesis complexity and tighten the generalization bound.

Under the stated assumptions, ETF-guided feature alignment acts as a structural regularizer that limits unconstrained directional variation. Together with parameter-space alignment, it provides complementary geometric constraints that can mitigate feature overlap and task interference, contributing to more coherent model merging. Detailed proofs are provided in Appendix B.3.1.

5 Experiments

To ensure a fair comparison, we strictly follow the experimental settings of prior work [Gargiulo et al., 2025, Wei et al., 2025, Fu et al., 2025]. For all baselines requiring scaling coefficients (e.g., TIES), we perform a grid search on the validation set to ensure they are evaluated at their optimal configurations.

Datasets and Models For vision tasks, we evaluate across three benchmark suites (8, 14, and 20 tasks) using CLIP [Radford et al., 2021] with varying ViT encoders (ViT-B/32, B/16, L/14) [Dosovitskiy et al., 2020] to assess model capacity effects. For standard NLP tasks, we employ Flan-T5-base [Chung et al., 2024] on eight GLUE datasets [Wang et al., 2019]. Furthermore, to test scalability to larger architectures and more challenging domains, we evaluate Qwen2-7B-Instruct [Yang et al., 2024a] on highly heterogeneous healthcare datasets. Comprehensive setup details, dataset descriptions, and extended NLP results (Table 15) are deferred to Appendices G, E.

Table 1: Average accuracy (%) when merging models across a larger number of tasks.

Method	ViT-B/32			ViT-B/16			ViT-L/14		
	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
Pre-trained	48.4	57.3	56.1	55.1	61.3	59.8	64.4	68.0	65.1
Weight averaging	66.5	64.4	61.1	72.1	69.3	65.3	79.4	76.6	71.5
Task Arithmetic	70.8	65.4	60.6	75.4	70.6	65.9	84.8	79.3	74.0
Ties-Merging	75.1	68.0	63.4	79.2	73.3	68.2	86.9	79.5	75.7
Consensus TA	75.0	70.4	65.4	79.0	74.5	69.7	86.2	82.2	78.9
Consensus TIES	74.8	67.7	63.2	78.2	75.3	67.1	86.9	81.5	76.8
TSV TA	85.7	79.8	76.4	89.0	84.5	80.5	92.9	88.2	87.1
ISO TA	84.9	78.9	73.1	89.1	84.6	79.5	93.3	88.0	87.0
ISO-CTS TA	83.8	79.7	76.6	88.5	85.8	82.2	92.9	89.0	**88.9**
DOGE TA	80.7	77.9	72.5	86.3	82.1	79.1	88.8	87.1	81.0
MDA TA	**87.4**	**81.9**	**77.3**	**90.6**	**86.1**	**82.7**	**93.6**	**89.4**	88.4

Table 2: Generalization results on two unseen tasks when merging ViT-B/32 models on six tasks.

Method	Seen Tasks							Unseen Tasks		
	SUN397	Cars	RESISC45	DTD	SVHN	GTSRB	Avg.	MNIST	EuroSAT	Avg.
Pre-trained	63.2	59.9	60.6	43.9	23.5	30.4	46.9	47.6	45.6	46.6
Task Arithmetic	64.3	63.0	73.2	54.9	84.7	79.5	69.9	75.5	42.6	59.1
Ties-Merging	68.3	65.5	76.9	54.9	75.4	72.0	68.9	73.1	47.3	60.2
AdaMerging	68.4	71.9	87.9	69.1	92.2	93.8	80.5	77.7	47.3	62.5
TSV TA	68.2	71.2	90.6	90.0	**95.8**	96.7	85.4	85.3	42.9	64.1
ISO TA	72.4	**74.2**	89.8	87.1	83.7	90.8	83.0	79.9	51.2	65.6
ISO-CTS TA	**72.6**	74.1	88.5	85.3	80.0	87.5	81.3	78.2	**52.3**	65.3
DOGE TA	69.8	72.6	86.6	67.6	90.8	91.6	79.8	81.3	48.2	64.8
MDA TA	71.1	72.9	**92.4**	**91.5**	**95.8**	**97.0**	**86.8**	**85.5**	47.7	**66.6**
MDA AM wo rotation	**71.5**	**73.2**	93.1	92.9	95.8	96.9	87.2	85.5	48.2	66.9
MDA AM	71.2	72.9	**93.3**	**94.3**	**96.1**	**97.3**	**87.5**	**86.1**	**48.8**	**67.5**

Baselines We compare with a broad range of existing methods, including both data-free and data-based approaches. Data-free methods include TA [Ilharco et al., 2022], Concrete TA [Tang et al., 2023], Ties-Merging [Yadav et al., 2023], Consensus Merging [Wang et al., 2024], AWD TA [Xiong et al., 2024], PCB-Merging [Du et al., 2024], TSV [Gargiulo et al., 2025], ISO [Marczak et al., 2025], and DOGE TA [Wei et al., 2025]. Data-based methods include AdaMerging [Yang et al., 2023a], Concrete AM [Tang et al., 2023], Representation Surgery [Yang et al., 2024b], AWD AM [Xiong et al., 2024], and DOGE AM [Wei et al., 2025]. We treat TA methods as data-free, since they operate purely in parameter space, while AM methods are data-based, leveraging data to optimize fusion coefficients. MDA TA follows Algorithm 1, and MDA AM further refines ETF-aligned parameters from Algorithm 1 using Algorithm 2.

Discussion on Parameter-space Alignment As shown in Table 1, the benefits of our parameter-space alignment amplify as the number of tasks increases. By enforcing parameter-space separation, our method constrains individual task adaptations to nearly orthogonal subspaces. This geometric separation effectively mitigates catastrophic interference, preserving task-specific functionality while enabling seamless multi-task integration.

Discussion on Feature-space Alignment As shown in Table 2 and Figure 2, incorporating feature direction alignment consistently improves performance. This suggests that feature-space alignment serves as an effective structural prior, guiding fusion beyond naive parameter averaging or standard optimization. Results in Tables 9 and 12 (Appendix G.2) further support these findings.

Discussion on Qwen2-7B-Instruct To validate scalability to Large Language Models, we evaluate Qwen2-7B-Instruct on heterogeneous healthcare datasets. Our method integrates multiple LoRA experts with zero inference overhead, consistently outperforming the baseline. Crucially, unlike

Table 3: The comparison of different methods on the healthcare domain using Qwen2-7B-Instruct.

	MedQA	CMExam	HealthcareMagic				cMedQA2			
Method	Acc.	Acc.	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Pretrained	51.92	74.65	1.05	25.90	3.27	12.65	9.77	40.30	19.05	20.94
Individual (oracle)	60.25	88.56	5.43	30.60	8.07	18.48	11.17	41.71	19.23	24.02
Weight Averaging	56.64	75.11	4.49	30.71	7.13	17.70	10.78	41.26	20.08	22.28
Task Arithmetic	57.97	**87.56**	5.00	30.33	7.72	18.11	5.55	32.70	12.02	19.15
Ties-Merging	58.05	87.47	4.23	**30.72**	7.03	17.58	11.30	42.06	20.60	22.85
TSV	56.64	85.92	5.12	30.61	7.71	18.10	12.55	**43.84**	**21.84**	24.40
ISO-C	58.05	86.68	4.52	29.58	7.12	17.60	9.78	39.60	17.93	23.65
ISO-CTS	57.82	86.74	4.54	29.73	7.20	17.77	11.97	42.84	20.28	25.26
MDA TA	**58.47**	87.12	**5.32**	30.47	**7.92**	**18.32**	**12.64**	43.62	21.54	**25.31**

Task Arithmetic which suffers from severe sequence disruption, our approach preserves the model's generative foundation for open-ended QA. By effectively resolving parameter conflicts, it achieves a superior balance across diverse languages (English/Chinese) and formats (MCQA/QA), demonstrating robust global optimization for merged LLMs. We evaluate MCQA via Accuracy [Devlin et al., 2019], and assess QA generation using BLEU-4 [Papineni et al., 2002] and ROUGE-1/2/L [Lin, 2004].

Discussion on Generalization Table 2 evaluates generalization on two unseen tasks after merging ViT-B/32 models trained on six seen tasks. Our MDA TA achieves the best average performance on unseen tasks, outperforming baseline methods while maintaining strong results on seen tasks. Similar gains are observed in feature-space generalization when combining directional alignment with fusion coefficient optimization. From a theoretical perspective, this alignment induces geometric properties that capture task-invariant structures, enabling more effective knowledge transfer to novel tasks.

Ablation Study We ablate the parameter- and feature-space alignments to isolate their individual contributions (Figure 2). For ViT-B/32, omitting parameter-space alignment notably degrades performance, particularly on the 14- and 20-task benchmarks, confirming its role in mitigating destructive interference. A similar degradation occurs in the AM setting when either alignment is removed, with the performance gap widening as the task count grows. Overall, both alignments are essential for scalable and generalizable model merging.

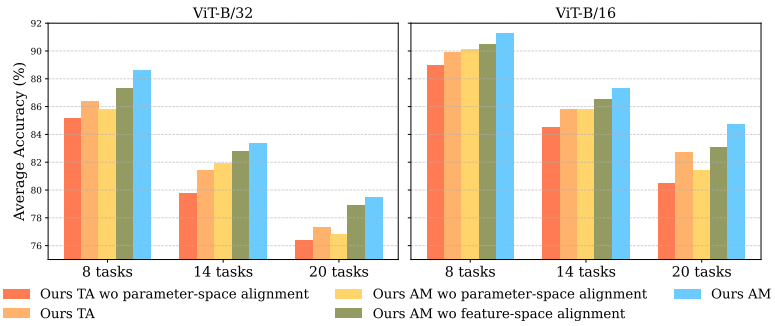


Figure 2: Ablation study of our method.

Additional analyses. Specifically, Appendix G.2.3 evaluates the loss hyperparameters α and β on ViT-B/32 across eight tasks, while Appendix G.2.6 examines the impact of the rank dimension k. Furthermore, to elucidate the relationship between initial task alignment and our method's effectiveness, Appendix G.2.2 provides an in-depth analysis of the correlation between geometric alignment and the resulting performance gains.

6 Conclusion

In this work, we proposed a unified framework for parameter- and feature-space alignment to preserve structural coherence and enhance generalization, validated both theoretically and empirically. Despite its effectiveness, certain limitations remain. First, optimizing rotation matrices introduces computational overhead that is non-trivial for extremely large models, though we introduce preliminary operations to mitigate this (see Table 17 in the Appendix). Second, merging models across different architectures remains an open challenge; future work will evaluate our framework on multimodal datasets. Finally, exploring lightweight strategies to reduce costs and studying alignment robustness under distribution shifts will further broaden our framework's applicability.

Acknowledgments and Disclosure of Funding

Tingting Zhu was supported by the Royal Academy of Engineering under the Research Fellowship scheme. MS was supported by JST ASPIRE Grant Number JPMJAP2405. Yu Zhang was supported by the National Natural Science Foundation of China under Grant no. 62136005, and the Shenzhen fundamental research program JCYJ20250604144724032. Sen Cui would like to acknowledge the financial support received from the Science and Technology Project of Beijing Municipal Science & Technology Commission (Grant No. Z251100008125030) and Shuimu Scholar program from Tsinghua University.

References

Daniel Beaglehole, Ioannis Mitliagkas, and Atish Agarwala. Feature learning as alignment: a structural property of gradient descent in non-linear neural networks. *arXiv preprint arXiv:2402.05271*, 2024a.

Daniel Beaglehole, Peter Súkeník, Marco Mondelli, and Mikhail Belkin. Average gradient outer product as a mechanism for deep neural collapse. *arXiv preprint arXiv:2402.13728*, 2024b.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 37:93671–93689, 2024.

Zhikang Chen, Abudukelimu Wuerkaixi, Sen Cui, Haoxuan Li, Ding Li, Jingfeng Zhang, Bo Han, Gang Niu, Houfang Liu, Yi Yang, et al. Learning without isolation: Pathway protection for continual learning. *arXiv preprint arXiv:2505.18568*, 2025.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yanqi Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. In *arXiv preprint arXiv:2210.11416*, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. *International Joint Conference on Neural Networks (IJCNN)*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.

Quinn Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse. *arXiv preprint arXiv:2402.06171*, 2024.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. Training-free llm merging for multi-task learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33111–33124, 2025.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

Li Guo, Keith Ross, Zifan Zhao, Andriopoulos George, Shuyang Ling, Yufeng Xu, and Zixuan Dong. Cross entropy versus label smoothing: A neural collapse perspective. *arXiv preprint arXiv:2402.03979*, 2024.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Hanglei Hu, Yingying Guo, Zhikang Chen, Sen Cui, Fei Wu, Kun Kuang, Min Zhang, and Bo Jiang. Advancing personalized learning with neural collapse for long-tail challenge. In *Forty-second International Conference on Machine Learning*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Vignesh Kothapalli, Tom Tirer, and Joan Bruna. A neural collapse perspective on feature evolution in graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023a.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5319–5329, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*, 2023.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. *arXiv preprint arXiv:2502.04959*, 2025.

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

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to route among specialized experts for zero-shot generalization. *arXiv preprint arXiv:2402.05859*, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Liam Parker, Emre Onal, Anton Stengel, and Jake Intrater. Neural collapse in the intermediate hidden layers of classification neural networks. *arXiv preprint arXiv:2308.02760*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

Peter Súkeník, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. *Advances in Neural Information Processing Systems*, 36, 2024.

Anke Tang, Li Shen, Yong Luo, Liang Ding, Han Hu, Bo Du, and Dacheng Tao. Concrete subspace learning based interference elimination for multi-task model fusion. *arXiv preprint arXiv:2312.06173*, 2023.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, pages 21478–21505. PMLR, 2022.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. *arXiv preprint arXiv:2405.07813*, 2024.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. *arXiv preprint arXiv:2501.01230*, 2025.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 527:60–70, 2023.

Feng Xiong, Runxi Cheng, Wang Chen, Zhanqiu Zhang, Yiwen Guo, Chun Yuan, and Ruifeng Xu. Multi-task model merging via adaptive weight disentanglement. *arXiv preprint arXiv:2411.18729*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Lilin Yan and Hongwei Zhao. One extra dimension suffices for the complete-graph squared-stress. *arXiv preprint arXiv:2610.06531*, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *eprint arXiv: 2407.10671*, 2024a.

Danni Yang, Zhikang Chen, Sen Cui, Mengyue Yang, Ding Li, Abudukelimu Wuerkaixi, Haoxuan Li, Jinke Ren, and Mingming Gong. Decentralized dynamic cooperation of personalized models for federated continual learning. *Advances in Neural Information Processing Systems*, 38:111897–111926, 2026.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023a.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *International Conference on Machine Learning*. PMLR, 2024b.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023b.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2883637.

Didi Zhu, Yinchuan Li, Min Zhang, Junkun Yuan, Jiashuo Liu, Kun Kuang, and Chao Wu. Bridging the gap: neural collapse inspired prompt tuning for generalization under class imbalance. *arXiv preprint arXiv:2306.15955*, 2023.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A Notation

We summarize the main symbols used in this paper as follows.

- T: the total number of tasks considered in model merging.
- C: the total number of classes across all tasks.
- θ_0: parameters of the pre-trained model.
- θ_k: parameters of the model fine-tuned on task k.
- $\tau_k = \theta_k - \theta_0$: task vector of task k, representing the update from pre-training to fine-tuning.
- $\tau_k^{(l)} = \theta_k^{(l)} - \theta_0^{(l)}$: layer-wise task vector at layer l.
- λ: merging coefficients; $\lambda^{(l)}$ denotes layer-wise coefficients.
- $W_{\mathrm{ETF}} \in \mathbb{R}^{C\times d}$: row-wise directional frame with C reference directions (the transpose of the column-wise convention in Definition 1).
- $\mathcal{S} = \mathrm{rowspan}(W_{\mathrm{ETF}}) \subset \mathbb{R}^d$: frame subspace, with dimension $C-1$ for an exact simplex ETF.
- $\Pi_{\mathcal{S}}$: orthogonal projector onto subspace $\mathcal{S}$.
- $\mathcal{P}_{\mathrm{ETF}} = W_{\mathrm{ETF}}^\top W_{\mathrm{ETF}}$: frame operator; $\mathcal{P}_{\mathrm{ETF}} = \frac{C}{C-1}\Pi_{\mathcal{S}}$ for an exact unit-norm simplex ETF.
- $\mathbf{h}_{c,i}$: feature vector of the i-th sample in class c.
- $\mathbf{h}_c = \frac{1}{n_c}\sum_i \mathbf{h}_{c,i}$: class mean feature of class c with n_c samples.
- $\mathbf{R}^t \in SO(d)$: task-specific rotation matrix for task t.
- $\tilde{\mathbf{h}}_{t,i} = \mathbf{R}^t\mathbf{h}_{t,i}$: rotated feature of sample i from task t.
- Σ_W^c: within-class covariance matrix for class c.
- $\mathcal{L} = \mathcal{L}_{\text{entropy}} + \alpha\mathcal{L}_{\text{align}} + \beta\mathcal{L}_{\text{rotation}}$: overall training loss combining entropy minimization, ETF alignment, and rotation regularization.
- $\mathcal{L}_{\text{entropy}}$: entropy minimization loss to calibrate predictions.
- $\mathcal{L}_{\text{align}}$: alignment loss encouraging ETF-inspired directional structure in features.
- $\mathcal{L}_{\text{rotation}}$: rotation regularization loss encouraging $\mathbf{R}^t$ close to the optimal Procrustes solution.
- Δ_{ETF}: deviation from the ETF geometry.
- Δ_{diff}: performance gap metric between methods.

B Theoretical Properties

B.1 ETF Construction Methods

Exact simplex reference (when $d_{\text{out}} \geq C-1$). Let $\mathbf{P}_C = \mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top$ and choose $\mathbf{V} \in \mathbb{R}^{C\times(C-1)}$ with $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_{C-1}$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{P}_C$. A thin QR decomposition of a Gaussian matrix in $\mathbb{R}^{d_{\text{out}}\times(C-1)}$ gives $\mathbf{Q} \in \mathbb{R}^{d_{\text{out}}\times(C-1)}$ with orthonormal columns. Then

$$\mathbf{W}_{\text{simplex}}^\top = \sqrt{\frac{C}{C-1}}\,\mathbf{Q}\mathbf{V}^\top \in \mathbb{R}^{d_{\text{out}}\times C}$$

has unit-norm class directions and off-diagonal inner products $-1/(C-1)$. This construction includes the boundary case $d_{\text{out}} = C-1$.

Numerical directional frames. The centering matrix $\mathbf{P}_C$ has rank $C-1$. For $d_{\text{out}} < C$, its truncated SVD supplies $\mathbf{W}^\top = \left(\sqrt{C/(C-1)}\,\mathbf{U}_{:,:d_{\text{out}}}\right)^\top$. For $d_{\text{out}} < C-1$, this is a lower-dimensional approximation and cannot retain the exact simplex Gram matrix. For $d_{\text{out}} \geq C$, QR decomposition of a Gaussian matrix in $\mathbb{R}^{d_{\text{out}}\times C}$ supplies an orthogonal directional reference $\mathbf{W} = \mathbf{Q}^\top$. The parameter-space implementation normalizes the columns of the row-wise frame before forming $\mathbf{W}^\top\mathbf{W}$. These numerical frames are ETF-inspired references; QR orthogonality or subsequent normalization alone

does not establish exact simplex identities. The frame-based analysis in Appendix B.2 accounts for the actual operator through its distortion and compatibility assumptions. Identities stated for an exact simplex ETF apply only to that ideal reference.

B.2 Parameter-space Direction Alignment

B.2.1 ETF Structural Coherence

Theorem 4 (ETF Structural Coherence)**.** *Let $\tau_{\text{ideal}}^{(l)} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ represent the ideal jointly-trained multi-class parameters at layer l, and let $\tau_{\text{share}}^{(l)}$ denote the SVD-reconstructed parameters obtained from task merging.*

Let $W_{\text{ETF}} \in \mathbb{R}^{C \times d_{\text{out}}}$ denote an ETF-inspired directional frame, and define

$$\mathcal{P}_{\text{ETF}} = W_{\text{ETF}}^{\top} W_{\text{ETF}}, \qquad \mathcal{S} = \text{rowspan}(W_{\text{ETF}}),$$

where $\Pi_{\mathcal{S}}$ denotes the orthogonal projector onto $\mathcal{S}$.

The aligned task vector is

$$\tau_{\text{etf}}^{(l)} = \tau_{\text{share}}^{(l)} \mathcal{P}_{\text{ETF}}.$$

For notational simplicity, write $A = \tau_{\text{share}}^{(l)}$ and $B = \tau_{\text{ideal}}^{(l)}$, and decompose

$$A_{\parallel} = A\Pi_{\mathcal{S}}, \qquad A_{\perp} = A(I - \Pi_{\mathcal{S}}),$$
$$B_{\parallel} = B\Pi_{\mathcal{S}}, \qquad B_{\perp} = B(I - \Pi_{\mathcal{S}}).$$

Define the frame distortion

$$\kappa = \|\mathcal{P}_{\text{ETF}} - \Pi_{\mathcal{S}}\|_2.$$

Assume that $\|A_{\perp}\|_F > 0$ and that the oracle parameters exhibit approximate ETF-subspace compatibility:

$$\|B_{\perp}\|_F \le \varepsilon \|A_{\perp}\|_F, \tag{19}$$
$$\|B_{\parallel} - A_{\parallel}\|_F \le \eta \|A_{\parallel}\|_F, \tag{20}$$

for some $\varepsilon \in [0, 1/2)$ and $\eta \ge 0$.

Define the directional correction gain

$$g = 1 - 2\varepsilon - (\kappa^2 + 2\kappa\eta) \frac{\|A_{\parallel}\|_F^2}{\|A_{\perp}\|_F^2}.$$

Whenever $g > 0$, the ETF-aligned task vector satisfies the following approximation bound:

$$\begin{aligned} \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{etf}}^{(l)}\|_F^2 \le \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{share}}^{(l)}\|_F^2 \\ - g\|\tau_{\text{share}}^{(l)}(I - \Pi_{\mathcal{S}})\|_F^2. \end{aligned} \tag{21}$$

In the ideal simplex ETF case, where

$$\mathcal{P}_{\text{ETF}} = \frac{C}{C-1} \Pi_{\mathcal{S}},$$

the bound equivalently becomes

$$\begin{aligned} \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{etf}}^{(l)}\|_F^2 \le \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{share}}^{(l)}\|_F^2 \\ - g \left\| \tau_{\text{share}}^{(l)} \left(I - \frac{C-1}{C} \mathcal{P}_{\text{ETF}} \right) \right\|_F^2. \end{aligned} \tag{22}$$

Here, C denotes the number of classes, and $\mathcal{P}_{\text{ETF}}$ represents the directional alignment operator induced by the ETF-inspired frame.

For an exact simplex ETF, the pairwise cosine similarity between distinct class directions is $-1/(C-1)$, and the factor $C/(C-1)$ arises from the corresponding frame operator.

Theorem 4 provides a geometric interpretation of how directional alignment can improve the approximation to an ideal jointly-trained representation. The improvement depends on the relationship between the oracle representation and the ETF-aligned subspace.

Specifically, the coefficient $g \in (0, 1]$ quantifies the effective directional correction under the stated compatibility conditions. A positive correction gain is obtained when the residual energy outside the aligned subspace is sufficiently large relative to the in-subspace distortion introduced by alignment.

In the ideal simplex case, the frame distortion reduces to $\kappa = 1/(C-1)$. Thus, increasing C reduces the scaling distortion associated with the ideal simplex frame operator. However, the resulting approximation improvement also depends on the selected subspace and its compatibility with the oracle parameters.

Consequently, ETF alignment is expected to be beneficial when it removes substantial incompatible parameter components while preserving the dominant directions associated with the jointly-trained representation. A complete derivation is provided below.

Frame structure and directional compatibility. **Setup and dimensions.** Consider first the ideal simplex ETF case. Let $W_{\mathrm{ETF}} \in \mathbb{R}^{C \times d_{\mathrm{out}}}$ be a row-wise unit-norm simplex ETF, with $d_{\mathrm{out}} \geq C-1$, satisfying

$$(W_{\mathrm{ETF}} W_{\mathrm{ETF}}^{\top})_{ii} = 1,$$

and

$$(W_{\mathrm{ETF}} W_{\mathrm{ETF}}^{\top})_{ij} = -\frac{1}{C-1}, \qquad i \neq j.$$

The corresponding Gram matrix is

$$W_{\mathrm{ETF}} W_{\mathrm{ETF}}^{\top} = \frac{C}{C-1}\left(I_C - \frac{1}{C}\mathbf{1}_C \mathbf{1}_C^{\top}\right).$$

Therefore, the frame operator satisfies

$$\mathcal{P}_{\mathrm{ETF}} = W_{\mathrm{ETF}}^{\top} W_{\mathrm{ETF}} = \frac{C}{C-1}\Pi_{\mathcal{S}},$$

where

$$\mathcal{S} = \mathrm{rowspan}(W_{\mathrm{ETF}}) \subseteq \mathbb{R}^{d_{\mathrm{out}}}.$$

The operator has eigenvalue $C/(C-1)$ with multiplicity $C-1$ on $\mathcal{S}$ and eigenvalue zero on $\mathcal{S}^{\perp}$.

For $\tau_{\mathrm{share}}^{(l)} \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{out}}}$, the aligned parameters are

$$\tau_{\mathrm{etf}}^{(l)} = \tau_{\mathrm{share}}^{(l)} \mathcal{P}_{\mathrm{ETF}} = \frac{C}{C-1}\tau_{\mathrm{share}}^{(l)} \Pi_{\mathcal{S}}.$$

Inner-product identity. By cyclicity of trace,

$$\begin{aligned}
&\left\langle \tau_{\mathrm{share}}^{(l)}, \tau_{\mathrm{etf}}^{(l)} \right\rangle_F \\
&= \mathrm{Tr}\left((\tau_{\mathrm{share}}^{(l)})^{\top} \tau_{\mathrm{share}}^{(l)} \mathcal{P}_{\mathrm{ETF}} \right) \qquad (23) \\
&= \frac{C}{C-1} \left\| \tau_{\mathrm{share}}^{(l)} \Pi_{\mathcal{S}} \right\|_F^2. \qquad (24)
\end{aligned}$$

This identity characterizes the energy of the reconstructed parameters preserved within the ETF-aligned subspace.

Implication for alignment bounds. Since

$$\|W_{\mathrm{ETF}}\|_F^2 = \mathrm{Tr}(W_{\mathrm{ETF}} W_{\mathrm{ETF}}^{\top}) = C,$$

we obtain

$$\begin{aligned}&\|\tau_{\text{share}}^{(l)} W_{\text{ETF}}^{\top}\|_F^2 \\ &= \frac{C}{C-1}\|\tau_{\text{share}}^{(l)}\Pi_{\mathcal{S}}\|_F^2.\end{aligned} \tag{25}$$

For nonzero $\tau_{\text{share}}^{(l)}$, define the normalized directional compatibility coefficient

$$\gamma^2 = \frac{\|\tau_{\text{share}}^{(l)} W_{\text{ETF}}^{\top}\|_F^2}{\|\tau_{\text{share}}^{(l)}\|_F^2 \|W_{\text{ETF}}\|_F^2}. \tag{26}$$

It follows that

$$\gamma^2 = \frac{1}{C-1}\frac{\|\tau_{\text{share}}^{(l)}\Pi_{\mathcal{S}}\|_F^2}{\|\tau_{\text{share}}^{(l)}\|_F^2} \leq \frac{1}{C-1}. \tag{27}$$

In the best-aligned case, where the row directions of $\tau_{\text{share}}^{(l)}$ lie entirely in $\mathcal{S}$, we obtain

$$\gamma = \frac{1}{\sqrt{C-1}}.$$

Thus, the coefficient γ measures the fraction of parameter energy captured by the ETF-aligned subspace, up to the normalization induced by the simplex frame geometry.

More generally, for a numerically constructed ETF-inspired frame, $\mathcal{P}_{\text{ETF}}$ remains symmetric positive semidefinite. Its deviation from the orthogonal projector is characterized by

$$\kappa = \|\mathcal{P}_{\text{ETF}} - \Pi_{\mathcal{S}}\|_2.$$

This quantity allows the approximation analysis to account for deviations from the ideal simplex frame operator. □

Approximation to Ideal Parameters.

Proof of the approximation bound. For simplicity, let

$$A = \tau_{\text{share}}^{(l)}, \qquad B = \tau_{\text{ideal}}^{(l)}, \qquad P = \mathcal{P}_{\text{ETF}}, \qquad \Pi = \Pi_{\mathcal{S}}.$$

Since $P = W_{\text{ETF}}^{\top} W_{\text{ETF}}$ is symmetric positive semidefinite, its range is $\mathcal{S}$, and it vanishes on $\mathcal{S}^{\perp}$.

We first decompose the reconstructed parameters and the oracle parameters into components parallel and orthogonal to $\mathcal{S}$:

$$A = A_{\parallel} + A_{\perp}, \qquad B = B_{\parallel} + B_{\perp},$$

where

$$A_{\parallel} = A\Pi, \qquad A_{\perp} = A(I - \Pi),$$

and similarly for B.

Because these components occupy orthogonal row subspaces, the approximation error before alignment is

$$\begin{aligned}\|B - A\|_F^2 = \|B_{\parallel} - A_{\parallel}\|_F^2 \\ + \|B_{\perp} - A_{\perp}\|_F^2.\end{aligned} \tag{28}$$

After applying directional alignment, the reconstructed parameters become

$$\tau_{\text{etf}}^{(l)} = AP = A_{\parallel}P.$$

Therefore,

$$\begin{aligned}\|B - AP\|_F^2 = \|B_{\parallel} - A_{\parallel}P\|_F^2 \\ + \|B_{\perp}\|_F^2.\end{aligned} \tag{29}$$

Using

$$B_\parallel - A_\parallel P = B_\parallel - A_\parallel + A_\parallel(I - P),$$

and expanding the squared norms, we obtain the exact identity

$$\begin{aligned} &\|B - A\|_F^2 - \|B - AP\|_F^2 \\ &= \|A_\perp\|_F^2 - 2\langle B_\perp, A_\perp\rangle_F \\ &\quad - \|A_\parallel(I - P)\|_F^2 \\ &\quad - 2\langle B_\parallel - A_\parallel, A_\parallel(I - P)\rangle_F. \end{aligned} \tag{30}$$

This decomposition separates the benefit of suppressing off-subspace components from the distortion introduced within the aligned subspace.

Since $A_\parallel = A_\parallel \Pi$, we have

$$\|A_\parallel(I - P)\|_F = \|A_\parallel(\Pi - P)\|_F \tag{31}$$

$$\leq \|\Pi - P\|_2 \|A_\parallel\|_F \tag{32}$$

$$= \kappa \|A_\parallel\|_F. \tag{33}$$

Furthermore, by the assumed approximate ETF compatibility,

$$\|B_\perp\|_F \leq \varepsilon \|A_\perp\|_F,$$

and

$$\|B_\parallel - A_\parallel\|_F \leq \eta \|A_\parallel\|_F.$$

Applying the Cauchy–Schwarz inequality to Eq. equation 30 gives

$$\begin{aligned} &\|B - A\|_F^2 - \|B - AP\|_F^2 \\ &\geq (1 - 2\varepsilon)\|A_\perp\|_F^2 \\ &\quad - (\kappa^2 + 2\kappa\eta)\|A_\parallel\|_F^2. \end{aligned} \tag{34}$$

By defining

$$g = 1 - 2\varepsilon - (\kappa^2 + 2\kappa\eta)\frac{\|A_\parallel\|_F^2}{\|A_\perp\|_F^2},$$

whenever $g > 0$, we obtain

$$\|B - A\|_F^2 - \|B - AP\|_F^2 \geq g\|A_\perp\|_F^2.$$

Equivalently,

$$\begin{aligned} \|B - AP\|_F^2 \leq \|B - A\|_F^2 \\ - g\|A(I - \Pi)\|_F^2. \end{aligned} \tag{35}$$

Substituting the original notation yields

$$\begin{aligned} &\|\tau_{\text{ideal}}^{(l)} - \tau_{\text{etf}}^{(l)}\|_F^2 \\ &\leq \|\tau_{\text{ideal}}^{(l)} - \tau_{\text{share}}^{(l)}\|_F^2 \\ &\quad - g\|\tau_{\text{share}}^{(l)}(I - \Pi_{\mathcal{S}})\|_F^2. \end{aligned} \tag{36}$$

For the ideal simplex ETF,

$$P = \frac{C}{C-1}\Pi, \qquad \kappa = \frac{1}{C-1},$$

and consequently

$$\Pi = \frac{C-1}{C}P.$$

The preceding result thus becomes

$$\begin{aligned}&\|\tau_{\text{ideal}}^{(l)} - \tau_{\text{etf}}^{(l)}\|_F^2 \\ \leq\ &\|\tau_{\text{ideal}}^{(l)} - \tau_{\text{share}}^{(l)}\|_F^2 \\ &- g\left\|\tau_{\text{share}}^{(l)}\left(I - \frac{C-1}{C}\mathcal{P}_{\text{ETF}}\right)\right\|_F^2.\end{aligned} \tag{37}$$

This establishes the claimed approximation bound.

Geometric interpretation. Although canonical Neural Collapse focuses on final-layer class features and classifier weights, subsequent studies suggest that structured feature geometry can also emerge in intermediate representations [Parker et al., 2023].

The Neural Feature Ansatz [Beaglehole et al., 2024a] further motivates the relationship between weight directions and feature covariance during learning.

These observations support the use of an ETF-inspired subspace as a geometric reference, while the compatibility conditions above explicitly characterize when this reference provides a beneficial correction to the reconstructed parameters.

Specifically, alignment improves the approximation when the energy removed from incompatible directions dominates the distortion introduced within the preserved subspace.

Thus, SVD reconstruction and ETF-guided directional alignment play complementary roles: the former identifies and combines dominant task-specific components, whereas the latter imposes additional geometric structure on the resulting parameter representation. □

Enhanced Direction Error Bound.

Proof. We analyze the directional residual

$$\tau_{\text{share}}^{(l)}(I - \mathcal{P}_{\text{ETF}}).$$

For simplicity, write $A = \tau_{\text{share}}^{(l)}$, $P = \mathcal{P}_{\text{ETF}}$, and $\Pi = \Pi_{\mathcal{S}}$.

Because $A_\perp = A(I - \Pi)$ is orthogonal to the aligned subspace, while $A_\parallel(I - P)$ remains within that subspace, we obtain

$$\begin{aligned}\|A(I-P)\|_F^2 = &\|A_\perp\|_F^2 \\ &+ \|A_\parallel(I-P)\|_F^2.\end{aligned} \tag{38}$$

Therefore,

$$\|A(I-P)\|_F^2 \geq \|A(I-\Pi)\|_F^2. \tag{39}$$

The aligned component satisfies the general trace identity

$$\|AP\|_F^2 = \text{Tr}(A^\top AP^2). \tag{40}$$

In the ideal simplex ETF case,

$$P = \frac{C}{C-1}\Pi,$$

so

$$\|AP\|_F^2 = \left(\frac{C}{C-1}\right)^2 \|A\Pi\|_F^2. \tag{41}$$

Furthermore,

$$\begin{aligned}\|A(I-P)\|_F^2 = &\|A(I-\Pi)\|_F^2 \\ &+ \frac{1}{(C-1)^2}\|A\Pi\|_F^2.\end{aligned} \tag{42}$$

This decomposition explicitly separates the off-subspace directional residual from the scaling effect of the ETF frame operator.

We next characterize the residual using the stable rank of A:

$$r_s = \frac{\|A\|_F^2}{\|A\|_2^2}.$$

Let

$$m = \min\left(d_{\text{in}}, \operatorname{rank}(P)\right).$$

Since $\operatorname{rank}(A\Pi) \le m$, we have

$$\|A\Pi\|_F^2 \le m\|A\Pi\|_2^2 \tag{43}$$

$$\le m\|A\|_2^2 \tag{44}$$

$$= \frac{m}{r_s}\|A\|_F^2. \tag{45}$$

Using the orthogonal decomposition

$$\|A\|_F^2 = \|A\Pi\|_F^2 + \|A(I-\Pi)\|_F^2,$$

we obtain

$$\|A(I-\Pi)\|_F^2 \ge \left(1 - \frac{m}{r_s}\right)_+ \|A\|_F^2, \tag{46}$$

where $(x)_+ = \max\{x, 0\}$.

Combining this inequality with Eq. equation 38 gives the general directional residual bound

$$\boxed{\|A(I-P)\|_F^2 \ge \left(1 - \frac{m}{r_s}\right)_+ \|A\|_F^2.} \tag{47}$$

For the exact simplex ETF, $\operatorname{rank}(P) = C-1$, and $m = \min(d_{\text{in}}, C-1)$. Using Eq. equation 42, a refined bound follows:

$$\begin{aligned}\|A(I-P)\|_F^2 \ge \Bigg[&\frac{1}{(C-1)^2} \\ &+ \left(1 - \frac{1}{(C-1)^2}\right)\left(1 - \frac{m}{r_s}\right)_+\Bigg]\|A\|_F^2.\end{aligned} \tag{48}$$

This result characterizes how the directional residual depends on both the rank structure of the reconstructed parameters and the geometry of the ETF-aligned subspace.

In particular, when the reconstructed parameters contain substantial energy outside the aligned subspace, the residual provides a quantitative measure of the directional components modified by alignment.

Together with Theorem 4, this establishes a geometric connection between directional correction, parameter structure, and approximation to the jointly-trained representation. □

B.2.2 Generalization Advantage

To characterize the statistical effect of directional alignment, we analyze a fixed-feature linearized layer. The result identifies when restricting parameter directions to the ETF-aligned subspace yields a tighter generalization bound, without assuming that alignment necessarily decreases the population risk of every individual model.

Theorem 5 (Generalization Bound). *Let $S_n = \{(x_i, y_i)\}_{i=1}^n$ be an i.i.d. sample and let $z_A(x) = z_0(x) + A^\top \phi(x)$, where z_0 and the feature map $\phi(x) \in \mathbb{R}^{d_{\text{in}}}$ are fixed independently of S_n, with*

$\|\phi(x)\|_2 \le R$. *Assume that* $0 \le \ell(z,y) \le 1$ *and that* ℓ *is L-Lipschitz in the logits under the Euclidean norm. Let* $W_{\mathrm{ETF}} \in \mathbb{R}^{C\times d_{\mathrm{out}}}$ *be a fixed ETF-inspired frame and define*

$$\mathcal{P}_{\mathrm{ETF}} = W_{\mathrm{ETF}}^{\top} W_{\mathrm{ETF}}, \quad \mathcal{S} = \mathrm{rowspan}(W_{\mathrm{ETF}}), \quad r = \dim(\mathcal{S}).$$

Let $\Pi_{\mathcal{S}}$ *be the orthogonal projector onto* $\mathcal{S}$*. Consider* $\tau_{\mathrm{etf}} = \tau_{\mathrm{share}}\mathcal{P}_{\mathrm{ETF}}$ *and suppose* $\|\tau_{\mathrm{share}}\|_F \le B_{\mathrm{share}}$ *and* $\|\tau_{\mathrm{etf}}\|_F \le B_{\mathrm{etf}}$*, where the positive radii are specified independently of* S_n*. Then, with probability at least* $1-\delta$*, the following bounds hold simultaneously:*

$$\mathcal{R}(\tau_{\mathrm{share}}) \le \widehat{\mathcal{R}}(\tau_{\mathrm{share}}) + \frac{2\sqrt{2}LRB_{\mathrm{share}}}{\sqrt{n}}\sqrt{d_{\mathrm{out}}} + 3\sqrt{\frac{\log(4/\delta)}{2n}}, \tag{49}$$

$$\mathcal{R}(\tau_{\mathrm{etf}}) \le \widehat{\mathcal{R}}(\tau_{\mathrm{etf}}) + \frac{2\sqrt{2}LRB_{\mathrm{etf}}}{\sqrt{n}}\sqrt{r} + 3\sqrt{\frac{\log(4/\delta)}{2n}}. \tag{50}$$

In particular, if the two models attain the same empirical risk, the aligned model has a tighter upper bound *whenever* $B_{\mathrm{etf}}\sqrt{r} < B_{\mathrm{share}}\sqrt{d_{\mathrm{out}}}$*. For an exact simplex ETF,* $r = C-1$ *and* $\mathcal{P}_{\mathrm{ETF}} = \frac{C}{C-1}\Pi_{\mathcal{S}}$*. Furthermore, for any such parameters,*

$$\left|\mathcal{R}(\tau_{\mathrm{share}}) - \mathcal{R}(\tau_{\mathrm{etf}})\right| \le LR\left\|\tau_{\mathrm{share}}(I - \mathcal{P}_{\mathrm{ETF}})\right\|_F. \tag{51}$$

Theorem 5 characterizes the inductive bias introduced by ETF-guided directional alignment. The bound depends on the dimension of the retained subspace rather than the full parameter-output dimension, provided that the effective Frobenius-norm budgets are comparable. Thus, alignment can reduce a Rademacher-complexity upper bound while preserving the empirical fit. This is a statement about the statistical capacity of the restricted hypothesis class, not an unconditional guarantee that every aligned solution has lower population risk. The residual-dependent bound in Eq. equation 51 separately quantifies how much the population risk can change under the alignment transformation.

Proof. **ETF Structure and Orthogonal Decomposition.** Because $\mathcal{P}_{\mathrm{ETF}}$ is symmetric positive semidefinite and has range $\mathcal{S}$, we have $\mathcal{P}_{\mathrm{ETF}} = \mathcal{P}_{\mathrm{ETF}}\Pi_{\mathcal{S}}$, and hence $\tau_{\mathrm{etf}} = \tau_{\mathrm{etf}}\Pi_{\mathcal{S}}$. For an exact simplex ETF,

$$\mathcal{P}_{\mathrm{ETF}} = \frac{C}{C-1}\Pi_{\mathcal{S}}, \qquad \tau_{\mathrm{etf}} = \frac{C}{C-1}\tau_{\Pi}, \qquad \tau_{\Pi} := \tau_{\mathrm{share}}\Pi_{\mathcal{S}}.$$

Frobenius Norm Reduction. Orthogonality gives

$$\|\tau_{\mathrm{share}}\|_F^2 = \|\tau_{\Pi}\|_F^2 + \|\tau_{\mathrm{share}}(I - \Pi_{\mathcal{S}})\|_F^2. \tag{52}$$

Consequently, when $\tau_{\mathrm{share}} \neq 0$,

$$\begin{aligned}\|\tau_{\mathrm{share}}\|_F - \|\tau_{\Pi}\|_F &= \frac{\|\tau_{\mathrm{share}}(I - \Pi_{\mathcal{S}})\|_F^2}{\|\tau_{\mathrm{share}}\|_F + \|\tau_{\Pi}\|_F} \\ &\ge \frac{\|\tau_{\mathrm{share}}(I - \Pi_{\mathcal{S}})\|_F^2}{2\|\tau_{\mathrm{share}}\|_F}.\end{aligned} \tag{53}$$

This reduction applies to the orthogonal projection τ_{Π}; the actual ETF-aligned parameters also depend on the frame scaling.

Spectral Characterization of the Residual. Let $\{v_j\}_{j=1}^{d_{\mathrm{out}}}$ be an orthonormal basis whose first r vectors span $\mathcal{S}$. Then

$$\|\tau_{\mathrm{share}}(I - \Pi_{\mathcal{S}})\|_F^2 = \sum_{j=r+1}^{d_{\mathrm{out}}} \|\tau_{\mathrm{share}}v_j\|_2^2. \tag{54}$$

This is precisely the parameter energy outside the aligned subspace. In the canonical simplex representation in $\mathbb{R}^C$, $\mathcal{S} = \mathbf{1}_C^{\perp}$ and $I - \Pi_{\mathcal{S}} = \mathbf{1}_C\mathbf{1}_C^{\top}/C$. For an exact simplex frame, the actual parameter change satisfies

$$\begin{aligned}\|\tau_{\mathrm{share}}(I - \mathcal{P}_{\mathrm{ETF}})\|_F^2 = \|\tau_{\mathrm{share}}(I - \Pi_{\mathcal{S}})\|_F^2 \\ + \frac{1}{(C-1)^2}\|\tau_{\Pi}\|_F^2.\end{aligned} \tag{55}$$

Generalization Gap via Rademacher Complexity. For a predetermined $B > 0$, consider the classes of logits

$$\mathcal{H}_{\text{full}}(B) = \{z_A : \|A\|_F \le B\},$$
$$\mathcal{H}_{\mathcal{S}}(B) = \{z_A : \|A\|_F \le B,\ A = A\Pi_{\mathcal{S}}\}.$$

Let $\epsilon_i \in \{\pm 1\}^{d_{\text{out}}}$ have independent Rademacher coordinates. The vector contraction inequality, followed by Cauchy–Schwarz, yields

$$\begin{aligned}\widehat{\mathfrak{R}}_{S_n}(\ell \circ \mathcal{H}_{\mathcal{S}}(B)) &\le \frac{\sqrt{2}LB}{n}\mathbb{E}_\epsilon\Big\|\sum_{i=1}^n \phi(x_i)\epsilon_i^\top \Pi_{\mathcal{S}}\Big\|_F \\ &\le \frac{\sqrt{2}LBR}{\sqrt{n}}\sqrt{r}. \end{aligned} \tag{56}$$

Indeed, independence of the Rademacher coordinates gives

$$\mathbb{E}_\epsilon\Big\|\sum_{i=1}^n \phi(x_i)\epsilon_i^\top \Pi_{\mathcal{S}}\Big\|_F^2 = r\sum_{i=1}^n \|\phi(x_i)\|_2^2 \le nrR^2.$$

Taking $\Pi_{\mathcal{S}} = I$ gives the corresponding bound with $r = d_{\text{out}}$ for $\mathcal{H}_{\text{full}}(B)$. The standard bounded-loss uniform generalization inequality, applied to both classes with a union bound, establishes Eqs. equation 49–equation 50.

Finally, the Lipschitz property implies

$$\begin{aligned}\big|\mathcal{R}(A) - \mathcal{R}(A\mathcal{P}_{\text{ETF}})\big| &\le L\,\mathbb{E}_x\big\|(A - A\mathcal{P}_{\text{ETF}})^\top \phi(x)\big\|_2 \\ &\le LR\|A(I - \mathcal{P}_{\text{ETF}})\|_F, \end{aligned} \tag{57}$$

which proves Eq. equation 51. □

Lemma 1 (Coupling Between Parameter-Space and Feature-Space Alignment)**.** *Let $W_0, \Delta W \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ denote the pretrained weight and the merged task update. In agreement with parameter-space alignment, let*

$$W = W_0 + \Delta W, \qquad \widetilde{W} = W_0 + \mathcal{P}_{\text{ETF}}\Delta W.$$

For class c, let the input have mean μ_c and covariance Σ_c, with $\|\mu_c\|_2 \le M$ and $\|\Sigma_c\|_2 \le \sigma^2$. Define $h_c = W\mu_c$ and $\widetilde{h}_c = \widetilde{W}\mu_c$. Then

$$\widetilde{h}_c = \mathcal{P}_{\text{ETF}}h_c + (I - \mathcal{P}_{\text{ETF}})W_0\mu_c, \tag{58}$$

and the aligned within-class covariance satisfies

$$\widetilde{\Sigma}_c = \widetilde{W}\Sigma_c\widetilde{W}^\top, \qquad \|\widetilde{\Sigma}_c\|_2 \le \sigma^2\|\widetilde{W}\|_2^2. \tag{59}$$

Moreover, let $a > 0$, $\zeta = \|\mathcal{P}_{\text{ETF}} - a\Pi_{\mathcal{S}}\|_2$, and $\rho = \min_c \|\Pi_{\mathcal{S}}h_c\|_2 > 0$. Define

$$b = M\big(\zeta\|W\|_2 + \|(I - \mathcal{P}_{\text{ETF}})W_0\|_2\big).$$

Suppose $b < a\rho$ and normalized projected class means satisfy

$$\max_c \left\|\frac{\Pi_{\mathcal{S}}h_c}{\|\Pi_{\mathcal{S}}h_c\|_2} - w_c^*\right\|_2 \le \epsilon,$$

for prescribed unit reference prototypes $w_c^* \in \mathcal{S}$ (simplex ETF prototypes when dimensionally feasible). Then, for any classes c, c',

$$\left|\frac{\langle \widetilde{h}_c, \widetilde{h}_{c'}\rangle}{\|\widetilde{h}_c\|_2\|\widetilde{h}_{c'}\|_2} - \langle w_c^*, w_{c'}^*\rangle\right| \le 2\epsilon + \frac{4b}{a\rho}. \tag{60}$$

For an exact simplex ETF, one may take $a = C/(C-1)$ and $\zeta = 0$.

Proof. The first identity follows from $\widetilde{W} = W_0 + \mathcal{P}_{\mathrm{ETF}}\Delta W$ and $W = W_0 + \Delta W$. The covariance identity follows from linear propagation of covariance; its operator-norm bound follows from $\|\Sigma_c\|_2 \le \sigma^2$. Moreover,

$$\begin{aligned}\|\widetilde{h}_c - a\Pi_{\mathcal{S}} h_c\|_2 &\le \|\mathcal{P}_{\mathrm{ETF}} - a\Pi_{\mathcal{S}}\|_2 \|h_c\|_2 + \|(I - \mathcal{P}_{\mathrm{ETF}})W_0\|_2 \|\mu_c\|_2 \\ &\le b. \end{aligned} \tag{61}$$

Since $\|a\Pi_{\mathcal{S}} h_c\|_2 \ge a\rho > b$, the aligned means are nonzero. For any nonzero vectors u, v, $\|u/\|u\|_2 - v/\|v\|_2\|_2 \le 2\|u - v\|_2/\|v\|_2$. Consequently,

$$\left\| \frac{\widetilde{h}_c}{\|\widetilde{h}_c\|_2} - \frac{\Pi_{\mathcal{S}} h_c}{\|\Pi_{\mathcal{S}} h_c\|_2} \right\|_2 \le \frac{2b}{a\rho}.$$

Applying the triangle inequality to the inner products of normalized means and using the assumed prototype deviation proves Eq. equation 60. Thus, parameter-space alignment controls the transformation of feature means, while feature-space alignment can further reduce their mismatch to the prescribed reference prototypes. The bound does not assume that projection alone automatically produces an exact ETF geometry. □

B.3 Feature-space Direction Alignment

Low-dimensional regime ($\mathrm{d} < C - 1$): In this case, the feature space lacks sufficient degrees of freedom to support class-level separation, leading to highly entangled representations. Optimizing fusion coefficients using standard entropy minimization loss yields suboptimal results compared to joint multi-task learning. However, by introducing an ETF-based alignment loss that enforces directional consistency with an equiangular tight frame (ETF), we observe substantial performance gains. This highlights the importance of directionality under low-dimensional constraints.

High-dimensional regime ($\mathrm{d} \ge C - 1$): Here, the feature space is sufficiently expressive to allow class separation. Optimizing only the fusion coefficients already achieves competitive performance, and additional directional alignment provides marginal improvements. This suggests that direction constraints are less crucial in high-dimensional settings.

B.3.1 Generalization Analysis

To analyze the generalization properties of feature-space directional alignment, we study the hypothesis complexity of rotation-merged models under prediction-guided adaptation. Our analysis characterizes how orthogonal transformations and ETF-guided directional constraints regulate model capacity and improve generalization bounds.

Theorem 6 (Rademacher Complexity Bound)**.** *Let $\mathcal{H}_{\mathrm{rot}}$ denote the hypothesis space of rotation-merged models:*

$$\mathcal{H}_{\mathrm{rot}} = \left\{ x \mapsto g\left(\sum_{t=1}^{T} \theta_t^\top R^t \phi(x) \right) : R^t \in SO(d) \right\}, \tag{62}$$

where the feature map ϕ and coefficients $\{\theta_t\}_{t=1}^T$ are fixed independently of the adaptation sample, $\|\phi(x)\|_2 \le R$, $\sum_{t=1}^T \|\theta_t\|_2^2 \le B^2$, and g is 1-Lipschitz.

Let ℓ be a bounded L-Lipschitz adaptation loss with values in $[0, 1]$. Given n i.i.d. adaptation samples x_i, define the population and empirical risks as

$$\mathcal{R}(h) = \mathbb{E}_x[\ell(h(x))], \qquad \hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i)).$$

Then the expected Rademacher complexity satisfies

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{rot}}) \le \frac{BR\sqrt{T}}{\sqrt{n}}. \tag{63}$$

For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, every $h \in \mathcal{H}_{\mathrm{rot}}$ satisfies

$$\mathcal{R}(h) \le \hat{\mathcal{R}}(h) + \frac{2LBR\sqrt{T}}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{64}$$

Furthermore, let $\{R_*^t\}_{t=1}^T$ be a fixed ETF-guided reference, chosen independently of the adaptation sample. Consider the restricted hypothesis space

$$\mathcal{H}_{\text{align}}(\rho) = \left\{ h \in \mathcal{H}_{\text{rot}} : \sum_{t=1}^{T} \|R^t - R_*^t\|_F^2 \le \rho^2 \right\}. \tag{65}$$

Its expected Rademacher complexity satisfies

$$\mathfrak{R}_n(\mathcal{H}_{\text{align}}(\rho)) \le \frac{BR}{\sqrt{n}} \min\{\sqrt{T}, \rho\}. \tag{66}$$

With probability at least $1-\delta$, simultaneously with Eq. equation 64, every $h \in \mathcal{H}_{\text{align}}(\rho)$ satisfies

$$\mathcal{R}(h) \le \hat{\mathcal{R}}(h) + \frac{2LBR}{\sqrt{n}} \min\{\sqrt{T}, \rho\} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{67}$$

Thus, when $\rho < \sqrt{T}$, ETF-guided directional constraints yield a tighter generalization upper bound.

Theorem 6 characterizes the statistical complexity of rotation-based feature alignment. By constraining task-specific transformations to orthogonal rotations, the hypothesis class preserves feature norms and admits a generalization bound with $\mathcal{O}(\sqrt{T/n})$ dependence under bounded coefficient norms.

Furthermore, ETF-guided alignment introduces an additional structural constraint on the admissible rotation matrices. Restricting their deviation from a fixed geometric reference reduces the effective hypothesis complexity, yielding a tighter bound when the alignment radius is sufficiently small. This provides a theoretical interpretation of feature-space alignment as a structural regularizer that limits unconstrained directional variation across tasks.

The bound controls generalization of the stated loss under its assumptions rather than directly guaranteeing improved classification accuracy. Prediction-guided directional alignment can contribute to more stable and coherent model merging.

Proof of Theorem 6. **Hypothesis Space Decomposition.** Consider the score function

$$f_{\mathbf{R}}(x) = \sum_{t=1}^{T} \theta_t^\top R^t \phi(x), \qquad R^t \in SO(d). \tag{68}$$

The hypothesis class is obtained by applying the 1-Lipschitz function g to the score. By the contraction principle, its Rademacher complexity is bounded by that of the underlying score class.

Rademacher Complexity Analysis. Let $\{\sigma_i\}_{i=1}^n$ denote independent Rademacher variables and define

$$s_\sigma = \sum_{i=1}^{n} \sigma_i \phi(x_i).$$

For a fixed sample, the empirical Rademacher complexity satisfies

$$\begin{aligned} \hat{\mathfrak{R}}_n(\mathcal{H}_{\text{rot}}) &\le \frac{1}{n} \mathbb{E}_\sigma \sup_{\{R^t\}} \sum_{i=1}^{n} \sigma_i f_{\mathbf{R}}(x_i) \\ &= \frac{1}{n} \mathbb{E}_\sigma \sup_{\{R^t\}} \sum_{t=1}^{T} \theta_t^\top R^t s_\sigma. \end{aligned} \tag{69}$$

Since $R^t \in SO(d)$, we have $\|R^t\|_2 = 1$. Applying the Cauchy–Schwarz inequality,

$$\begin{aligned} \hat{\mathfrak{R}}_n(\mathcal{H}_{\text{rot}}) &\le \frac{1}{n} \sum_{t=1}^{T} \|\theta_t\|_2 \mathbb{E}_\sigma \|s_\sigma\|_2 \\ &\le \frac{B\sqrt{T}}{n} \mathbb{E}_\sigma \|s_\sigma\|_2. \end{aligned} \tag{70}$$

Using independence of the Rademacher variables,

$$\begin{aligned}\mathbb{E}_\sigma \|s_\sigma\|_2^2 &= \sum_{i=1}^{n} \|\phi(x_i)\|_2^2 \\ &\le nR^2. \end{aligned} \tag{71}$$

Therefore,

$$\mathbb{E}_\sigma \|s_\sigma\|_2 \le R\sqrt{n},$$

which yields

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{rot}}) \le \frac{BR\sqrt{T}}{\sqrt{n}}. \tag{72}$$

ETF-Guided Alignment Constraint. Consider the restricted hypothesis space $\mathcal{H}_{\mathrm{align}}(\rho)$ and write

$$R^t = R_*^t + \Delta R^t,$$

where

$$\sum_{t=1}^{T} \|\Delta R^t\|_F^2 \le \rho^2.$$

Because the reference rotations are fixed independently of the sample, their contribution to the expected Rademacher sum vanishes.

Consequently,

$$\begin{aligned}&\hat{\mathfrak{R}}_n(\mathcal{H}_{\mathrm{align}}(\rho)) \\ &\le \frac{1}{n}\mathbb{E}_\sigma \sup_{\{\Delta R^t\}} \sum_{t=1}^{T} \theta_t^\top \Delta R^t s_\sigma \\ &\le \frac{\rho}{n}\mathbb{E}_\sigma \left(\sum_{t=1}^{T} \|\theta_t s_\sigma^\top\|_F^2 \right)^{1/2} \\ &\le \frac{\rho B}{n}\mathbb{E}_\sigma \|s_\sigma\|_2 \\ &\le \frac{\rho BR}{\sqrt{n}}. \end{aligned} \tag{73}$$

Furthermore, since $\mathcal{H}_{\mathrm{align}}(\rho) \subseteq \mathcal{H}_{\mathrm{rot}}$, we also have

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{align}}(\rho)) \le \mathfrak{R}_n(\mathcal{H}_{\mathrm{rot}}).$$

Combining the two bounds gives

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{align}}(\rho)) \le \frac{BR}{\sqrt{n}} \min\{\sqrt{T}, \rho\}. \tag{74}$$

Generalization Bound. For a bounded loss, the standard uniform generalization inequality based on expected Rademacher complexity states that, with probability at least $1-\delta$, every hypothesis h in a fixed hypothesis class $\mathcal{H}$ satisfies

$$\mathcal{R}(h) \le \hat{\mathcal{R}}(h) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{75}$$

By the Lipschitz contraction principle,

$$\mathfrak{R}_n(\ell \circ \mathcal{H}) \le L\mathfrak{R}_n(\mathcal{H}). \tag{76}$$

Applying these results to $\mathcal{H}_{\text{rot}}$ and $\mathcal{H}_{\text{align}}(\rho)$, and using a union bound with failure probability $\delta/2$ for each class, we obtain, simultaneously,

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \frac{2LBR\sqrt{T}}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}, \tag{77}$$

$$\mathcal{R}(h_{\text{align}}) \leq \hat{\mathcal{R}}(h_{\text{align}}) + \frac{2LBR}{\sqrt{n}} \min\{\sqrt{T}, \rho\} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{78}$$

This establishes the claimed generalization bounds. □

B.4 Computational Complexity

B.4.1 Notation

Table 4 summarizes the main symbols used in this section.

Table 4: Notation summary.

Symbol	Meaning
T	Number of tasks (models to be merged).
L	Number of network layers in each model.
n	Matrix dimension of each layer weight ($n \times n$ for simplicity).
m	Smaller dimension for non-square layers ($m = \min(p, q)$ for a $p \times q$ matrix).
$\tau_t^{(\ell)}$	Parameter matrix of task t at layer ℓ.
$U_t^{(\ell)}, \Sigma_t^{(\ell)}, V_t^{(\ell)}$	Singular value decomposition (SVD) components of $\tau_t^{(\ell)}$.
$U_{\text{cat}}^{(\ell)}, V_{\text{cat}}^{(\ell)}$	Concatenated matrices of top-k components across tasks (for shared subspace).
W_{ETF}	Equiangular Tight Frame (ETF) basis matrix.
C	Number of classes (ETF dimension).
$\mathcal{O}(\cdot)$	Big-O notation for asymptotic computational complexity.

B.4.2 Complexity of Algorithm 1

For each layer $\ell = 1, \ldots, L$, Algorithm 1 performs three main computational steps.

Per-task SVDs. For each task $t = 1, \ldots, T$, the SVD

$$\tau_t^{(\ell)} = U_t^{(\ell)} \Sigma_t^{(\ell)} {V_t^{(\ell)}}^\top$$

requires $\mathcal{O}(n^3)$ time. The total cost across all layers and tasks is $\mathcal{O}(TLn^3)$.

Concatenated SVDs. After concatenating the top-k components of all tasks into $U_{\text{cat}}^{(\ell)}$ and $V_{\text{cat}}^{(\ell)}$, two additional SVDs are performed. Each SVD costs $\mathcal{O}(n^3)$, yielding an additional $\mathcal{O}(2Ln^3)$.

ETF projection. The ETF projection,

$$\tau_{\text{etf}}^{(\ell)} = \tau_{\text{share}}^{(\ell)} W_{\text{ETF}}^\top W_{\text{ETF}},$$

involves two matrix multiplications of size $n \times C$ and $C \times n$, with per-layer cost $\mathcal{O}(n^2 C)$, or $\mathcal{O}(n^2 LC)$ in total.

Overall complexity. Combining all terms gives:

$$\boxed{\mathcal{O}\big((T+2)\, L\, n^3 + n^2 LC\big)}. \tag{79}$$

Since $C \ll n$ in most practical cases, the ETF term is negligible relative to the cubic SVD cost.

Table 5: Computational complexity comparison per layer across methods. All methods involve L layers with weight dimension $n \times n$.

Method	Main SVD operations per layer	Total Complexity
ISO-C [Marczak et al., 2025]	One SVD on Δ_{TA}	$\mathcal{O}(Ln^3)$
ISO-CTS [Marczak et al., 2025]	One SVD on Δ_{TA} + T on Δ_t + two orthogonalization SVDs	$\mathcal{O}((T+3)Ln^3)$
TSV-M [Gargiulo et al., 2025]	T task SVDs + two for reconstruction	$\mathcal{O}((T+2)Ln^3)$
Ours (Alg. 1)	T per-task SVDs + two concatenated SVDs + ETF projection	$\mathcal{O}((T+2)Ln^3 + n^2LC)$

B.4.3 Complexity Comparison with Baselines

Table 5 compares our method with existing baselines in terms of their main SVD operations and asymptotic complexity per layer. Complexity expressions follow the standard cubic-time assumption for dense SVD.

When $C \ll n$, our method scales similarly to TSV-M and Iso-CTS in asymptotic order, while introducing only a lightweight ETF projection step.

B.4.4 Data-based Alignment: Empirical Overhead

When optimizing rotation matrices and fusion coefficients with data, the additional wall-clock cost per epoch is marginal (about $1-3\%$). The dominant extra FLOPs come from the Procrustes step (solving for the optimal rotation via SVD).

Table 6: Per-epoch timing breakdown (averaged across architectures). The Procrustes step accounts for a small fraction ($\sim$1–3%) of total training time.

Model	Forward (s)	Backward (s)	Procrustes (s)	Epoch total (s)	Procrustes ratio
ViT-L/14	10.82	3.33	0.24	$\approx$14.47	1.7%
ViT-B/32	4.18	0.87	0.16	$\approx$5.27	3.0%
ViT-B/16	6.12	1.12	0.16	$\approx$7.48	2.2%

FLOPs analysis. Table 7 reports the floating-point operation counts. The counters measure computational cost rather than loss components.

Table 7: FLOPs of rotation and alignment components per epoch. The Procrustes step dominates the additional FLOPs cost.

Model	rot_forward	rot_total	procrustes
ViT-L/14	3.78×10^7	1.13×10^8	9.06×10^8
ViT-B/32	1.68×10^7	5.03×10^7	2.68×10^8
ViT-B/16	1.68×10^7	5.03×10^7	2.68×10^8

Interpretation. The alignment step, comprising rotation application and Procrustes optimization, adds negligible wall-clock time and a manageable FLOPs increase dominated by the SVD-based Procrustes computation.

C Related Work

Neural Collapse. Recently, several studies [Ji et al., 2021, Zhu et al., 2021, Tirer and Bruna, 2022, Zhu et al., 2023, Li et al., 2023b, Xie et al., 2023, Yang et al., 2023b, Beaglehole et al., 2024b, Fisher et al., 2024, Guo et al., 2024, Kothapalli et al., 2024, Súkeník et al., 2024, Chen et al., 2024, Hu et al., 2025] have utilized this phenomenon to guide the training process in imbalanced data sets. Among them, Yang et al. [2023b] introduce a continual learning framework that pre-allocates a fixed number of classes within a simplex ETF, thereby guiding the representation learning of minority classes in subsequent incremental steps. This design enforces intra-class feature convergence to predetermined positions while maximizing and uniformly separating inter-class features. To address class imbalance, Li et al. [2023b] leverage its geometric structure under distributed conditions to

align the representation directions across clients, while preserving client-specific characteristics through fine-tuning. The aforementioned methods only consider collapsing the same class onto a single point in a fixed simplex ETF without accounting for the impact of intra-class spurious correlations. Kothapalli [2022] theoretically analyze the generalization benefits induced by neural collapse and propose a generalization bound based on the Class Distance Normalized Variance (CDNV), which demonstrates that training-induced collapse can facilitate generalization, albeit requiring substantial data support. Related work in Euclidean distance geometry establishes zero-residual second-order stationary points for complete, exact, uniformly weighted squared-stress objectives when the embedding dimension exceeds the ground-truth dimension [Yan and Zhao, 2026]. To the best of our knowledge, we are the first to revisit the importance of directional alignment through the lens of neural collapse, conducting the analysis from both the parameter space and the feature space.

D Baselines

Data-Free Methods

Task Arithmetic (TA) [Ilharco et al., 2022] introduces a simple yet effective approach by treating task-specific knowledge as arithmetic operations in the parameter space. It computes task vectors as the difference between fine-tuned and pre-trained parameters, enabling direct manipulation through addition and negation operations to compose or forget specific capabilities.

Concrete Task Arithmetic [Tang et al., 2023] proposes a multi-task model merging method based on continuous relaxation of discrete (Concrete) subspace learning. The key idea is to identify common low-dimensional subspaces in the parameter space and exploit the shared information therein to effectively mitigate task interference in multi-task fusion, while preserving overall performance as much as possible. We formulate the problem as a bi-level optimization and introduce a meta-learning framework, where gradient-based techniques are employed to learn shared Concrete masks that guide model merging within the subspace. Experiments across multiple tasks in both vision and language domains validate the effectiveness of the proposed approach, showing substantial improvements over existing Task Arithmetic and AdaMerging methods.

Ties-Merging [Yadav et al., 2023], a novel method for merging multiple task-specific models into a single multitask model without additional training. Existing merging techniques suffer from performance degradation due to parameter interference caused by redundant values and sign conflicts across models. Ties-Merging addresses these issues through three key steps: trimming low-magnitude parameters, electing dominant signs, and disjointly merging only aligned values. Extensive experiments across NLP and vision domains, various model architectures, and fine-tuning settings demonstrate that Ties-Merging consistently outperforms prior methods, achieving significant gains in both in-domain and out-of-domain generalization. Our work highlights the critical role of resolving sign interference and provides a robust, hyperparameter-efficient solution for model merging.

Consensus Merging [Wang et al., 2024] consistently improves upon previous model merging techniques such as Task Arithmetic and TIES, achieving state-of-the-art performance. Additionally, our compression scheme using TALL-masks reduces storage requirements by over 85% while retaining 99.7% of the original performance, demonstrating a highly effective trade-off between model efficiency and multi-task capability.

AWD Task Arithmetic [Xiong et al., 2024] uses the following techniques by explicitly promoting orthogonality among task vectors. By decomposing task vectors into a shared redundant component and disentangled orthogonal components, AWD effectively reduces inter-task interference while preserving task-specific performance. When integrated with Task Arithmetic and AdaMerging, AWD achieves state-of-the-art results on multi-task benchmarks, demonstrating strong generalization and robustness across varying task numbers and scaling coefficients.

PCB-Merging [Du et al., 2024] proposes a novel, training-free model merging method, which effectively balances parameter-level competition across tasks via intra- and inter-task balancing mechanisms, leading to significant performance gains in cross-task, cross-domain, and out-of-domain generalization settings without requiring additional data or retraining.

Task Singular Vectors (TSV) [Gargiulo et al., 2025] proposes a novel model merging framework, which leverages singular value decomposition at the layer level to compress task-specific updates

and reduce interference through a geometrically-informed whitening transformation, achieving state-of-the-art performance without additional training or validation data.

Isotropic Model Merging (ISO) [Marczak et al., 2025] proposes a novel model merging framework based on the analysis of the correlation between subspace alignment and model merging performance, which enhances alignment between task-specific and merged model subspaces by flattening the singular value spectrum of the merged task matrix and incorporating both common and task-specific directions.

DOGE Task Arithmetic [Wei et al., 2025] proposes a novel adaptive projective gradient descent framework for multi-task model merging, which formulates the problem as a constrained optimization objective to minimize performance gaps with individual task-specific models while preserving shared knowledge through subspace-constrained gradient updates and training-free task-aware merging coefficients.

Data-Based Optimization Methods

AdaMerging [Yang et al., 2023a] adaptively learns merging coefficients through gradient-based optimization on a small validation set. It introduces both task-wise and layer-wise coefficient learning strategies, allowing fine-grained control over the merging process based on actual task performance.

Concrete AdaMerging [Tang et al., 2023] extends AdaMerging with concrete masks, combining the benefits of adaptive coefficient learning with parameter-level masking. This dual optimization approach achieves superior performance by simultaneously learning what and how much to merge.

Representation Surgery [Yang et al., 2024b] operates at the representation level rather than parameter space. It aligns and merges intermediate representations across models using surgical precision, employing optimal transport and feature matching techniques to ensure semantic consistency.

AWD AdaMerging [Xiong et al., 2024] integrates Adaptive Weight Disentanglement with AdaMerging's optimization framework. This combination leverages both the structural decomposition of AWD and the data-driven optimization of AdaMerging for enhanced performance.

DOGE AdaMerging [Wei et al., 2025] extends the DOGE framework to the data-based setting, utilizing validation data to guide the discrete optimization process. It employs advanced search strategies and ensemble techniques to discover optimal merging configurations that maximize multi-task performance.

These baselines represent the current state-of-the-art in model merging, spanning from simple arithmetic operations to sophisticated optimization frameworks, providing a comprehensive benchmark for evaluation.

E Datasets

For vision tasks, we evaluate our approaches over three benchmark suites comprising 8, 14, and 20 tasks, respectively. The first suite, introduced in [Radford et al., 2021], consists of eight datasets: Stanford Cars [Krause et al., 2013], DTD [Cimpoi et al., 2014], EuroSAT [Helber et al., 2019], GTSRB [Stallkamp et al., 2011], MNIST [LeCun et al., 2002], RESISC45 [Cheng et al., 2017], SUN397 [Xiao et al., 2016], and SVHN [Netzer et al., 2011].

The 14-task benchmark extends the above by incorporating six additional datasets: CIFAR-100 [Krizhevsky et al., 2009], STL10 [Coates et al., 2011], Flowers102 [Nilsback and Zisserman, 2008], Oxford-IIIT Pet [Parkhi et al., 2012], PCAM [Veeling et al., 2018], and FER2013 [Goodfellow et al., 2013].

Finally, the 20-task benchmark includes all the previous 14 tasks plus six more: EMNIST [Cohen et al., 2017], CIFAR-10 [Krizhevsky et al., 2009], Food101 [Bossard et al., 2014], Fashion-MNIST [Xiao et al., 2017], Rendered-SST2 [Socher et al., 2013], and KMNIST [Clanuwat et al., 2018]. To investigate the effect of model capacity, we evaluate our method using three CLIP [Radford et al., 2021] variants, each employing a different ViT [Dosovitskiy et al., 2020] visual encoder: ViT-B/32, ViT-B/16, and ViT-L/14.

SUN397: A large-scale scene recognition dataset containing 397 categories with over 100K images. The dataset covers a diverse range of indoor and outdoor scenes, such as churches, bedrooms,

highways, and landscapes, making it one of the most comprehensive scene benchmarks. It is widely used to evaluate the generalization ability of visual recognition models, especially in transfer learning and domain generalization studies.

Stanford Cars: A fine-grained object classification dataset with 196 car models and 16K images. Each image is annotated with the make, model, and year of the car, requiring models to distinguish between visually similar subcategories. It is a standard benchmark for fine-grained recognition and is often used to evaluate representation learning, attention mechanisms, and metric learning methods.

RESISC45: A remote sensing image scene classification dataset comprising 31.5K images across 45 categories. Images are collected from Google Earth, covering diverse scenes such as residential areas, airports, rivers, and forests. The dataset is designed to evaluate robustness under varying imaging conditions (viewpoints, illumination, resolutions), and is widely used in remote sensing and geospatial vision research.

EuroSAT: A remote sensing dataset based on Sentinel-2 satellite imagery, containing 27K images across 10 land-use and land-cover categories, such as industrial areas, forests, and highways. With multispectral bands available, it provides a valuable benchmark for both RGB and multispectral classification tasks. EuroSAT is widely used for studying domain adaptation, few-shot learning, and environmental monitoring applications.

SVHN: The Street View House Numbers dataset, with 73K training images and 10 digit categories, extracted from Google Street View. Unlike MNIST, SVHN features digits in natural scene images, often with cluttered backgrounds, varying scales, and illumination conditions. It is considered a more challenging real-world digit recognition benchmark and is widely used to test the robustness of deep learning models.

GTSRB: The German Traffic Sign Recognition Benchmark, containing 50K images across 43 traffic sign categories. Images vary significantly in lighting, perspective, and occlusion, making it highly representative of real-world autonomous driving conditions. GTSRB is a standard benchmark for evaluating models in traffic sign recognition, robustness under distribution shifts, and safety-critical perception tasks.

MNIST: A canonical digit classification dataset with 70K grayscale images of handwritten digits (0–9). As one of the earliest benchmarks in computer vision, MNIST remains a widely used baseline for algorithm prototyping and teaching, despite being relatively easy for modern deep learning models. It has inspired many extended datasets (e.g., EMNIST, Fashion-MNIST, KMNIST).

DTD: The Describable Textures Dataset, containing 5.6K images across 47 categories, each corresponding to a human-describable texture attribute (e.g., "bumpy," "striped," "zigzagged"). Unlike object or scene classification datasets, DTD emphasizes fine-grained texture perception, making it useful for evaluating mid-level representations and generalization across visual domains.

Flowers102: A fine-grained flower recognition dataset with 102 categories and 8K images. Each category corresponds to a specific flower species, with large intra-class variation due to differences in viewpoint, lighting, and scale. Flowers102 is frequently used for transfer learning and fine-grained classification, particularly in evaluating few-shot and zero-shot learning algorithms.

PCAM: The PatchCamelyon dataset for histopathology image classification, containing 327K image patches with binary labels (tumor vs. normal). Derived from the CAMELYON16 challenge, it is widely used in medical imaging benchmarks to evaluate weakly supervised learning, multiple-instance learning, and robustness in clinical decision support systems.

FER2013: A facial expression recognition dataset with 35.8K grayscale images across 7 emotion categories (e.g., happy, sad, angry, surprised). Images were collected in the wild from the ICML 2013 Kaggle competition, making the dataset challenging due to variations in pose, illumination, and occlusion. It is widely used for affective computing and human-computer interaction research.

Oxford-IIIT Pet: A fine-grained object dataset containing 37 pet breeds with 7K images. Each image includes both class labels and pixel-level segmentation masks, enabling tasks in both classification and semantic segmentation. The dataset is commonly used for fine-grained recognition and multi-task learning in vision.

STL-10: A dataset similar to CIFAR but with higher-resolution images (96×96), containing 10 categories and 13K labeled images. In addition to the labeled set, it includes a large set of 100K

unlabeled images, making it well-suited for semi-supervised and unsupervised representation learning. STL-10 is often used to evaluate scalable learning methods.

CIFAR-100: A widely used benchmark dataset with 100 fine-grained categories and 60K images. Each category has 500 training and 100 testing images, requiring models to handle limited data per class. CIFAR-100 is considered more challenging than CIFAR-10 and is a standard benchmark for evaluating the scalability of deep learning models.

CIFAR-10: A subset of CIFAR with 10 categories and 60K images. It remains one of the most widely used datasets for small-scale object classification, serving as a testbed for model architectures, optimization strategies, and regularization techniques.

Food-101: A large-scale food recognition dataset with 101 categories and 101K images. Each category represents a distinct food dish, with significant variation in presentation and context. It is a benchmark for fine-grained recognition, domain adaptation, and applications in food computing and lifestyle analysis.

Fashion-MNIST: A drop-in replacement for MNIST, containing 70K grayscale images from 10 clothing categories such as T-shirts, trousers, and coats. Compared to MNIST, Fashion-MNIST provides a more challenging benchmark while maintaining the same data format, making it a popular choice for evaluating lightweight models.

EMNIST: The Extended MNIST dataset, comprising 800K images across up to 62 categories (digits and uppercase/lowercase letters). It provides multiple splits (balanced, letters, digits, etc.), making it a versatile benchmark for handwritten character recognition and multi-class classification tasks.

KMNIST: A dataset of 70K grayscale images from 10 categories of handwritten Japanese characters (Kuzushiji). As a more challenging alternative to MNIST, KMNIST captures complex character structures and is used to evaluate transfer learning, few-shot learning, and robustness across different scripts.

Rendered SST-2: A rendered version of the Stanford Sentiment Treebank (SST-2), where binary sentiment labels (positive vs. negative) are transformed into visual classification tasks through rendered images. It bridges NLP and computer vision by enabling cross-modal evaluation, particularly in vision-language pretraining and multimodal learning research.

F Method

As illustrated in Figure 3, We provide a schematic illustration of our proposed method. In the parameter space, we leverage an Equiangular Tight Frame (ETF) to enforce directional alignment across the subspaces associated with different classes, thereby establishing a consistent geometric structure among task-specific parameters. In the feature space, each task is assigned an initialized alignment matrix that guides the orientation of its representations. During training, we jointly optimize both the alignment matrices and the fusion coefficients, which allows the model to adaptively refine the directions and magnitudes of the merged representations. This joint optimization yields a more coherent alignment across tasks and ultimately leads to improved performance and generalization.

G Performance

G.1 Implementation Details

To assess the role of directional structure in the parameter space, we remove residual scale effects and align the merged parameters with a simplex equiangular tight frame (ETF), thereby enforcing global directional consistency. Meanwhile, to evaluate the importance of feature-space alignment, we draw upon the theory of neural collapse. Specifically, we constrain the penultimate-layer features to align with a simplex ETF, while simultaneously aligning the classifier prototypes to the same simplex structure. This joint constraint, optimized with a learning rate of 1×10^{-3}, updates both the feature rotation matrix and the fusion coefficients, thereby enforcing a coherent geometric configuration that not only validates the importance of direction but also facilitates faster convergence and improved generalization.

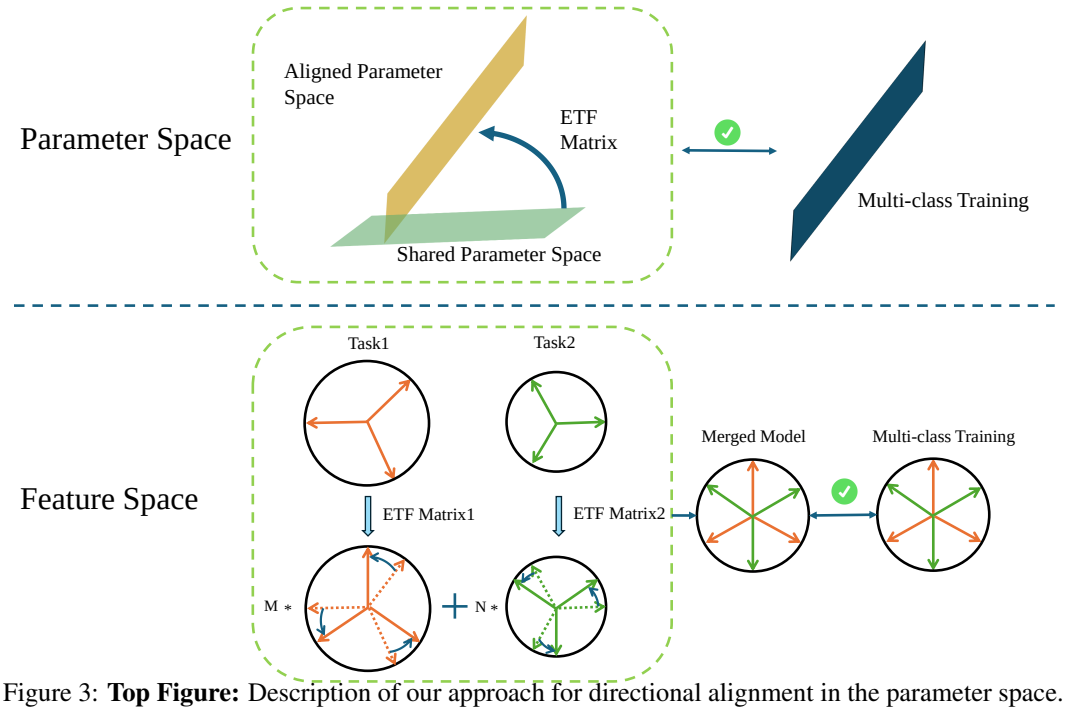


Figure 3: **Top Figure:** Description of our approach for directional alignment in the parameter space. **Bottom Figure:** Description of our approach for directional alignment in the feature space.

Following standard practice in model merging, we adopt the experimental setups used in TSV [Gargiulo et al., 2025] and AdaMerging [Yang et al., 2023a] for a fair comparison under both data-free and data-based regimes.

Data-free setting (Algorithm 1). In the data-free regime, our parameter-space alignment experiments employ exactly the same data splits as those used by TSV. Each task corresponds to an independently fine-tuned CLIP model on one of the datasets from the 8-, 14-, and 20-task. No additional data are used during merging; only the fine-tuned model weights are required.

Data-based setting (Algorithm 2). For adaptive (feature-space) alignment, we use the validation/test splits provided for each dataset to optimize the rotation matrices and fusion coefficients. The entropy term follows AdaMerging, while our alignment regularizer uses hard argmax pseudo-labels from the current model predictions. Ground-truth labels are used only for evaluation, not for adaptation.

The experiments use the same dataset partitions across baselines, with prediction-based adaptation as stated above. Specifically, the 8-, 14-, and 20-task benchmarks include datasets such as Cars, DTD, EuroSAT, RESISC45, SVHN, GTSRB, MNIST, SUN397, and progressively expanded subsets up to 20 tasks, matching the configurations reported in TSV. We emphasize that no labeled data are used in the data-free setting, while the data-based setup uses adaptation samples with model-generated pseudo-labels for feature-space optimization.

G.2 Results

G.2.1 Analysis of Qwen2-7B-Instruct on Healthcare Datasets

Datasets Details To evaluate our method, we utilize four healthcare datasets covering diverse languages (English and Chinese) and task formats. Specifically, MedQA is an English multiple-choice question answering (MCQA) dataset, consisting of 10,000 training, 400 validation, and 400 test samples. CMExam is a Chinese MCQA dataset containing 50,000 training, 4,000 validation, and 4,000 test samples. For open-ended question answering (QA) tasks, we employ the English HealthCareMagic and the Chinese cMedQA2 datasets. Both QA datasets share an identical data split scale, each comprising 30,000 training, 1,000 validation, and 1,000 test samples.

Training Details During the fine-tuning phase, we adopt Qwen2-7B-Instruct as the base model and utilize the LLaMA-Factory framework for training. Specifically, we apply Low-Rank Adaptation (LoRA) with strictly controlled hyperparameters: a rank (r) of 8, a scaling factor (α) of 16, and a dropout rate of 0.01. Furthermore, the fine-tuning process is conducted with a batch size of 64 and an initial learning rate of 1.0×10^{-4}, employing a cosine decay learning rate schedule integrated with a warm-up mechanism.

Generalization on the Qwen Architecture. As shown in Table 3, our proposed method demonstrates superior parameter integration capabilities on the Qwen2-7B-Instruct architecture. Healthcare data is highly specialized and heterogeneous, spanning from clinical licensing exams (CMExam [Liu et al., 2023], MedQA [Jin et al., 2021]) to real-world online medical consultations (HealthcareMagic [Li et al., 2023a], cMedQA2 [Zhang et al., 2018]). Experimental results indicate that our approach effectively unlocks Qwen's latent instruction-following abilities. Without incurring additional inference overhead, it successfully deconstructs and recombines the expert knowledge of four distinct LoRA modules. Compared to the pretrained baseline, our method achieves significant improvements across all 10 evaluation metrics, securing optimal performance in 5 of them, thereby proving its robustness in handling complex medical knowledge.

Mitigating Catastrophic Forgetting in Medical Generation. Maintaining the coherence and professionalism of generated text in open-ended medical QA tasks (e.g., HealthcareMagic and cMedQA2) is a major challenge. Traditional merging methods often disrupt Qwen's original language generation distribution. For instance, while Task Arithmetic performs strongly on Chinese MCQA (CMExam), its BLEU-4 score on cMedQA2 drops to 5.55, which is lower than the base model without any adapters (9.77). This demonstrates that simple weight addition interferes with Qwen's sequence generation logic. In contrast, our method successfully preserves the base LLM's generative foundation while integrating medical expert knowledge, achieving the highest BLEU-4 (12.64) and ROUGE-L (25.31) on cMedQA2, ensuring both the accuracy and naturalness of medical advice.

Cross-lingual and Multi-task Trade-offs. The cross-lingual nature (English and Chinese) of the utilized medical datasets places high demands on model merging. Comparisons show that existing state-of-the-art methods like TSV-M, despite performing excellently in specific Chinese generation metrics, experience a notable performance drop in English medical exams (MedQA). Through a refined conflict resolution mechanism, our approach achieves a better balance between different languages and task formats. It yields the highest accuracy on MedQA (58.47%) while remaining highly competitive on CMExam (87.12%). This demonstrates that our merging strategy effectively resolves weight conflicts among multiple medical LoRA experts, achieving global performance optimization.

G.2.2 Correlation Between Geometric Alignment and Performance Gains

To investigate the link between geometric alignment and model performance, we analyzed the correlation between the cosine similarity gap (Δ_{ETF}) and the performance difference (Δ_{diff}).

As detailed in Table 8, there is a strong positive correlation between these two metrics. A larger Δ_{ETF} indicates a weaker initial directional alignment among task-specific models. As this similarity gap widens, our method yields a consistently higher performance improvement (Δ_{diff}). For instance, in ViT-B/16, as the cosine similarity gap increases from 0.09 (8 tasks) to 0.15 (20 tasks), the performance gain correspondingly amplifies from 0.9% to 2.2%. While ViT-B/32 presents a minor fluctuation at the 20-task setting, the overarching trend clearly demonstrates that the degree of initial directional incoherence strongly predicts the structural benefits and performance gains delivered by our alignment mechanism.

G.2.3 Ablation Study on Loss Components

Interplay of Alignment and Rotation Constraints. To better understand the role of different components in our objective, we conduct an ablation study on the coefficients α and β. Specifically, α controls the neural collapse alignment loss ($\mathcal{L}_{\text{align}}$), while β balances the feature-space directional alignment loss ($\mathcal{L}_{\text{rotation}}$). As illustrated in the heatmap (Figure 5), optimizing these two constraints jointly forms a distinct performance surface, demonstrating that neither loss can be entirely omitted (e.g., accuracy drops to 87.3% at $\alpha = 0, \beta = 0$) or overly weighted.

Table 8: Correlation between cosine similarity gap (Δ_{ETF}) and performance gap (Δ_{diff}) across different task scales.

Model	8 Tasks	14 Tasks	20 Tasks
Δ_{ETF} **(Cosine Similarity Gap)**			
ViT-B/32	0.17	0.20	0.14
ViT-B/16	0.09	0.12	0.15
ViT-L/14	0.07	0.11	0.14
Δ_{diff} **(Performance Gap)**			
ViT-B/32	1.2%	1.6%	0.9%
ViT-B/16	0.9%	1.3%	2.2%
ViT-L/14	0.7%	1.0%	1.3%

Effect of Neural Collapse Alignment (α). Focusing on the horizontal axis, increasing α generally improves performance up to a sweet spot. For instance, when fixing $\beta = 0.4$, the accuracy rises from 87.2% ($\alpha = 0$) to the global peak of 88.9% ($\alpha = 0.6$). This highlights that feature alignment with the shared ETF space is crucial for mitigating task interference and enhancing generalization. However, continuing to increase α to 1.0 causes a slight drop, indicating that overly large α may over-constrain the model and restrict its capacity for task-specific representations.

Effect of Directional Alignment (β) and Optimal Balance. Similarly, observing the vertical axis reveals the impact of $\mathcal{L}_{\text{rotation}}$. Moderate values of β (e.g., 0.2 to 0.4) significantly improve performance by providing necessary structural consistency. Conversely, larger values ($\beta \geq 0.8$) gradually degrade accuracy, particularly when α is low (dropping to 86.6% at $\alpha = 0, \beta = 1.0$). This implies that excessive directional consistency hinders the model's adaptation to diverse task directions. Ultimately, the model achieves its best performance by striking a balance at $\alpha = 0.6$ and $\beta = 0.4$, confirming that a calibrated combination of ETF alignment and directional flexibility is essential.

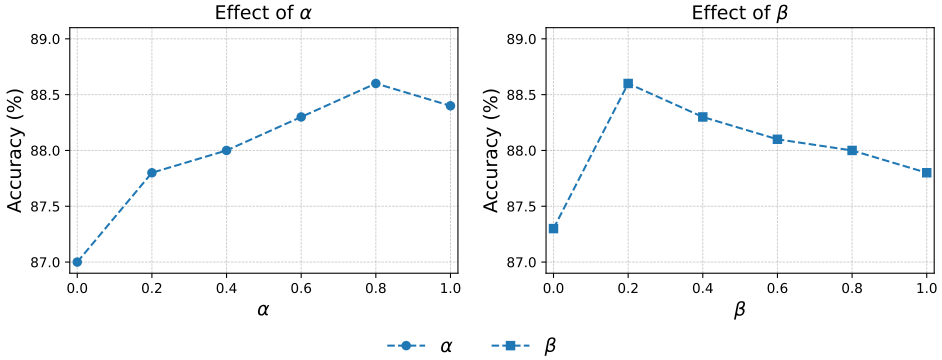


Figure 4: Ablation on the coefficients of different modules in the loss function.

G.2.4 Multi-task performance when merging ViT-B/32 models on 8-task vision benchmark

We conducted experiments on eight tasks using the ViT-B-32 architecture, based on the checkpoints provided by [Marczak et al., 2025, Gargiulo et al., 2025]. The results are summarized in Table 10.

Table 9 presents a comprehensive evaluation of multi-task performance when merging ViT-B/32 models across eight vision datasets. The results clearly demonstrate that incorporating feature-space alignment to optimize task-wise fusion coefficients yields consistent performance improvements across a variety of tasks. Our method, Ours AM, achieves the highest average accuracy of 88.6%, surpassing all baseline merging methods, including TSV AM (85.8%) and DOGE AM (85.9%).

Table 11 reports the ablation results on ViT-B/32 across eight vision benchmarks. Our rotation-aware alignment method (Ours AM) achieves the highest average accuracy of 88.6%, outperforming both the variant without rotation alignment (Ours w/o rotation AM, 87.3%) and the task-arithmetic baseline (Ours TA, 86.4%). The improvement is particularly pronounced on challenging datasets such as DTD (91.8% vs. 88.9%) and RESISC45 (89.1% vs. 87.7%), where feature distributions are highly

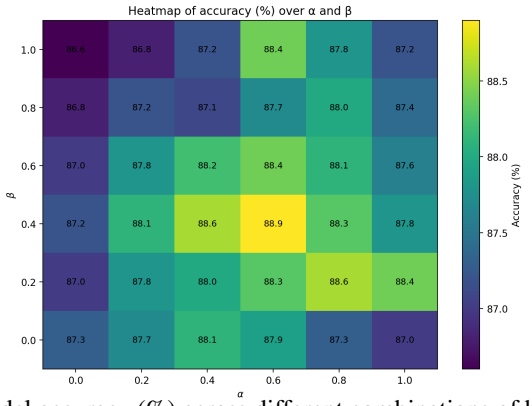


Figure 5: Heatmap of model accuracy (%) across different combinations of hyperparameters α and β. Both parameters are evaluated in the range of $[0.0, 1.0]$ with a step size of 0.2. The color gradient indicates the performance surface, with the optimal accuracy (88.9%) achieved at $\alpha = 0.6$ and $\beta = 0.4$.

Table 9: Multi-task performance when merging ViT-B/32 models on 8-task vision benchmark using checkpoints in [Tang et al., 2023].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
Non-merging Methods									
Pre-trained	62.3	59.7	60.7	45.5	31.4	32.6	48.5	43.8	48.0
Individual	79.2	77.7	96.1	99.7	97.5	98.7	99.7	79.4	91.0
Traditional MTL	73.9	74.4	93.9	98.2	95.8	98.9	99.5	77.9	89.1
Data-free Methods									
Task Arithmetic	55.2	54.9	66.7	78.9	80.2	69.7	97.3	50.4	69.1
Ties-Merging	59.8	58.6	70.7	79.7	86.2	72.1	98.3	54.2	72.4
Consensus Merging	65.7	63.6	76.5	77.2	81.7	70.3	97.0	57.1	73.6
AWD TA	63.5	61.9	72.6	84.9	85.1	79.1	98.1	56.7	75.2
PCB-Merging	66.7	65.5	78.5	79.3	86.4	77.1	98.2	59.1	76.3
Concrete TA	62.5	61.1	76.0	95.7	91.0	81.9	98.5	51.9	77.3
TSV TA	67.7	70.8	87.3	95.9	93.6	94.9	84.7	86.9	85.2
ISO TA	72.1	**73.8**	88.5	91.3	81.4	90.0	80.1	86.0	82.9
ISO-CTS TA	**72.2**	73.7	87.1	88.5	75.8	86.3	78.9	83.9	80.8
DOGE TA	67.7	70.1	82.0	90.3	86.3	86.8	98.3	64.0	80.7
MDA TA	70.4	71.9	**88.9**	**96.7**	**94.3**	**95.4**	**85.3**	**88.6**	**86.4**
Data-based Optimization Methods									
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
Representation Surgery	63.8	59.9	83.3	**97.9**	87.0	87.0	98.6	69.4	80.9
AWD AM	68.1	71.4	83.4	94.8	87.7	93.6	97.9	66.1	82.9
Concrete AM	67.8	70.0	87.5	96.0	91.6	**96.7**	98.7	63.8	84.0
TSV AM	68.1	72.1	87.7	96.4	92.3	93.3	**99.4**	88.9	87.3
DOGE AM	**70.5**	**74.8**	88.7	94.1	91.6	95.7	98.8	72.5	85.9
MDA AM	70.0	72.0	**89.1**	97.6	**93.7**	95.1	**99.4**	**91.8**	**88.6**

Table 10: Multi-task performance when merging ViT-B/32 models on 8-task vision benchmark using checkpoints in [Gargiulo et al., 2025].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
TSV TA	67.2	70.3	85.7	94.5	91.6	92.2	99.3	84.8	85.7
ISO TA	71.8	73.8	87.7	90.9	82.3	88.1	98.2	86.0	84.9
ISO-CTS TA	72.4	74.0	87.9	90.6	76.9	85.3	97.5	85.5	83.8
MDA TA	70.3	72.0	87.9	96.0	92.5	93.3	99.3	87.5	87.4

heterogeneous. This result highlights that incorporating rotation alignment is crucial for preserving feature geometry across tasks.

G.2.5 Multi-task performance when merging ViT-L/14 models on 8-task vision benchmark

Table 12 presents results for ViT-L/14. In the data-free regime, our method (Ours TA) achieves an average accuracy of 92.0%, outperforming strong baselines such as TSV (91.3%) and AWD TA (87.5%). In the optimization-based regime, Ours AM further improves performance to 92.7%,

Table 11: Ablation study on multi-task performance when merging ViT-B/32 models on 8-task vision benchmark using checkpoints in [Tang et al., 2023].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
Ours TA	**70.4**	71.9	88.9	96.7	**94.3**	**95.4**	85.3	88.6	86.4
Ours w/o rotation AM	68.1	**72.1**	87.7	96.4	92.3	93.3	**99.4**	88.9	87.3
Ours AM	70.0	72.0	**89.1**	**97.6**	93.7	95.1	**99.4**	**91.8**	**88.6**

Table 12: Multi-task performance when merging ViT-L/14 models on 8-task vision benchmark using checkpoints in [Tang et al., 2023].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
Non-merging Methods									
Pre-trained	66.8	77.7	71.0	59.9	58.4	50.5	76.3	55.3	64.5
Individual	82.3	92.4	97.4	100	98.1	99.2	99.7	84.1	94.2
Traditional MTL	80.8	90.6	96.3	96.3	97.6	99.1	99.6	84.4	93.1
Data-free Methods									
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	95.7	90.3	83.3	99.0	68.8	86.0
Consensus Merging	75.0	84.3	89.4	95.6	88.3	82.4	98.9	68.0	85.2
AWD TA	76.2	85.4	88.7	96.1	92.4	92.3	99.3	69.4	87.5
PCB-Merging	76.8	86.2	89.4	96.5	88.3	91.0	98.6	73.6	87.5
Concrete TA	**86.2**	66.9	**96.7**	93.4	**99.1**	89.0	74.6	**93.6**	87.4
TSV TA	78.3	90.0	94.1	98.5	95.2	96.1	99.5	78.9	91.3
ISO AM	79.9	90.9	94.8	98.3	91.4	95.5	99.2	79.2	91.1
DOGE TA	76.7	87.7	91.6	96.2	94.4	93.4	98.9	71.6	88.8
MDA TA	79.1	**90.6**	94.7	**98.7**	95.6	**97.0**	**99.6**	80.4	**92.0**
Data-based Optimization Methods									
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
Representation Surgery	75.7	84.4	93.1	98.8	91.3	93.4	99.1	76.1	89.0
AWD AM	**79.8**	90.6	91.8	97.0	93.9	**98.4**	99.2	81.1	91.5
Concrete AM	77.8	**91.2**	92.1	97.0	94.4	97.9	99.0	79.5	91.1
MDA AM	79.1	91.0	**94.7**	**98.9**	**95.9**	97.9	**99.7**	**84.0**	**92.7**

Table 13: Multi-task performance when merging ViT-L/14 models on 8-task vision benchmark using checkpoints in [Gargiulo et al., 2025].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
TSV TA	77.8	89.8	93.5	98.7	94.7	96.1	99.5	93.2	92.9
ISO TA	79.9	91.4	94.8	99.0	90.5	95.5	99.3	96.3	93.3
ISO-CTS TA	79.6	91.7	94.6	98.8	88.8	95.4	99.2	95.6	92.9
MDA TA	78.8	90.7	94.3	99.2	95.2	97.0	99.6	94.5	93.6

surpassing AdaMerging (90.8%) and AWD AM (91.5%). Notably, on EuroSAT, Ours AM achieves 98.9%, which is higher than TSV (98.5%) and AdaMerging (96.2%). These results indicate that our alignment mechanism generalizes well to large-scale backbones and heterogeneous tasks.

Table 14 further confirms the benefit of rotation alignment on ViT-L/14. Ours AM achieves the best average performance of 92.7%, compared to Ours w/o rotation AM (92.3%) and Ours TA (92.0%). Although the gains appear smaller than in ViT-B/32, the improvements are consistent across all tasks (e.g., DTD: 84.0% vs. 83.1%, EuroSAT: 98.9% vs. 98.7%), showing that rotation alignment stabilizes feature fusion and ensures reliable improvements even at larger scales.

We conducted experiments on eight tasks using the ViT-L-14 architecture, based on the checkpoints provided by [Marczak et al., 2025, Gargiulo et al., 2025]. The results are summarized in Table 13.

G.2.6 Ablation on the rank dimension k.

When $k = d_{\text{out}}/T$, the concatenated dimension kT naturally equals d_{out}, so the shared representation $\tau_{\text{share}}^{(l)}$ directly matches the layer's output dimension without requiring any truncation. More generally, when k varies across tasks or $kT \neq d_{\text{out}}$, the formulation still holds:

1. If $kT > d_{\text{out}}$, we simply truncate $U_{\text{share}}^{(l)}$ to its top-d_{out} singular components, preserving the directions that explain the highest shared variance.

Table 14: Ablation study on multi-task performance when merging ViT-L/14 models on 8-task vision benchmark using checkpoints in [Tang et al., 2023].

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg.
MDA TA	79.1	90.6	**94.7**	98.7	95.6	97.0	99.6	80.4	92.0
MDA w/o rotation AM	78.5	90.8	94.0	98.7	95.9	97.7	99.6	83.1	92.3
MDA AM	**79.1**	**91.0**	**94.7**	**98.9**	**95.9**	**97.9**	**99.7**	**84.0**	**92.7**

Table 15: Multi-task performance when merging Flan-T5-base (LoRA fine-tuned) models on all eight tasks.

Method	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
Non-merging Methods									
Individual	69.1	82.7	85.5	90.9	84.0	84.4	92.9	87.4	84.6
Data-free Methods									
Weight Averaging	**69.7**	59.7	78.9	90.1	83.8	80.5	91.2	72.0	78.2
Task Arithmetic	68.8	55.2	78.7	89.8	83.7	79.1	91.5	72.4	77.4
Ties-Merging	68.3	56.3	79.2	89.8	83.7	79.4	91.6	71.2	77.5
Concrete TA	69.1	58.1	78.4	89.9	83.5	79.4	91.6	73.4	78.0
TSV TA	69.3	**77.1**	80.4	90.0	83.6	79.1	92.5	82.5	81.8
ISO TA	69.1	57.4	76.7	88.6	82.7	80.1	91.3	63.3	76.2
DoGE TA	69.1	71.9	**80.9**	**90.3**	83.5	79.8	**92.5**	71.1	79.9
MDA TA	69.5	76.9	76.9	89.6	**83.9**	**82.4**	**92.5**	**84.7**	**82.1**
Data-based Optimization Methods									
AdaMerging++	69.1	60.3	78.4	90.0	83.6	79.1	91.6	74.1	78.3
Concrete AM	69.0	59.4	80.1	89.9	82.9	79.1	91.7	75.4	78.5

2. If $kT < d_{\text{out}}$, the missing dimensions can be completed either by zero-padding or implicitly through the subsequent ETF projection.

We investigate the impact of the subspace dimension k on model merging performance. Recall that k controls the number of principal components retained per task in the shared subspace construction. By default, we set $k = d_{\text{out}}/T$, where d_{out} is the output dimension of each layer and T is the number of tasks being merged. This choice ensures that the concatenated dimension kT naturally matches d_{out}, leading to a balanced contribution from each task without requiring additional normalization or scaling.

Table 16 reports the results of varying k across different ViT architectures and task counts. We observe that the performance remains stable near $k = d_{\text{out}}/T$, with slight fluctuations depending on the total number of tasks. Specifically,Smaller k values (e.g., $k = 0.01\, d_{\text{out}}$) underrepresent task-specific subspaces and lead to noticeable drops in accuracy, while excessively large k values (e.g., $k = 0.5\, d_{\text{out}}$ or $0.9\, d_{\text{out}}$) introduce redundancy and degrade generalization. Interestingly, when k is chosen close to d_{out}/T, the model achieves a favorable trade-off between shared and task-specific components, and in some cases even surpasses the performance of the setting with $k = d_{\text{out}}/T$. In most configurations, the setting $k = d_{\text{out}}/T$ achieves a favorable trade-off between efficiency and accuracy, and is therefore adopted as the default in all main results, consistent with prior work [Gargiulo et al., 2025]. MDA TA 0.01 means that we assign the top 1% proportion of feature dimensions d_{out} as the principal components for each task.

Table 16: Ablation study on the subspace dimension k for different ViT architectures and numbers of tasks.

Method	ViT-B/32			ViT-B/16			ViT-L/14		
	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
MDA TA $(1/T)$	86.4	81.4	**77.3**	89.9	85.8	**84.7**	92.0	**89.4**	**88.4**
MDA TA 0.01	77.9	76.3	72.6	82.0	80.5	78.2	87.8	86.6	85.4
MDA TA 0.06	86.1	**81.9**	77.1	90.1	**86.1**	82.5	91.9	89.3	88.2
MDA TA 0.1	**86.7**	81.2	75.9	**90.6**	85.5	81.1	**92.1**	89.3	87.1
MDA TA 0.5	77.8	70.6	66.7	83.4	76.3	73.0	87.9	82.9	79.5
MDA TA 0.9	70.5	66.8	64.3	78.3	72.9	70.5	84.2	79.8	76.9

G.2.7 Multi-task performance when merging Flan-T5-base (LoRA fine-tuned) models on all eight tasks.

Table 15 reports multi-task performance on Flan-T5-base (LoRA fine-tuned) across eight NLP benchmarks. Our method achieves the highest average score of 82.1% among data-free methods, outperforming TSV (81.8%) and DoGE (79.9%). For instance, on STSB, Ours TA obtains 84.7%, substantially higher than TSV (82.5%) and Ties-Merging (71.2%). Compared to optimization-based methods, Ours TA (82.1%) also surpasses AdaMerging++ (78.3%) and Concrete AM (78.5%) *without access to training data*, demonstrating the effectiveness of our approach in data-free language model merging.

Across both vision and NLP benchmarks, our results demonstrate that *directional alignment* is a key factor in successful model merging. Traditional methods often suffer from feature misalignment: when task-specific representations are merged without considering their geometric orientation, the resulting model exhibits degraded generalization, especially on heterogeneous datasets. Our method aligns task vectors via rotation, which brings two major benefits: (1) it preserves the relative geometry of task-specific features, leading to more coherent shared representations; (2) it reduces destructive interference between tasks, as misaligned directions in parameter space are corrected before fusion. This is empirically validated by the consistent improvements across backbones (ViT-B/32 and ViT-L/14) and modalities (vision and NLP). These findings suggest that *geometric alignment is not merely a technical detail but a fundamental principle* for effective model merging.

G.2.8 Effect of Class-to-Dimension Ratio and Tuning Paradigms

Vision: Dense class space and full parameter tuning. In our vision experiments, the benchmarks consist of 8, 14, or 20 tasks, accumulating to a total of 758 to over 1,300 categories. For models like ViT-B/32 ($d_{\text{out}} = 768$) and ViT-L/14 ($d_{\text{out}} = 1024$), the total class count can exceed the layer dimension, depending on the task configuration. In this dense regime, enforcing balanced angular separation is geometrically nontrivial, making directional alignment via ETF critical. Moreover, these vision models are trained via *full parameter fine-tuning*. This provides degrees of freedom for feature- and parameter-space directional alignment, without guaranteeing an exact simplex when the dimension is insufficient. Consequently, ETF alignment yields substantial structural and generalization gains.

NLP: Sparse class space and LoRA capacity bottleneck. Conversely, the NLP setting (eight GLUE tasks) involves very few labels, mostly binary or three-way classifications (e.g., SST-2, MRPC). The combined class count across all tasks remains under 20, which is drastically smaller than the output dimension of Flan-T5-base ($d_{\text{out}} = 2048$). In this sparse regime ($C \ll d_{\text{out}}$), class vectors naturally attain near-orthogonal separation without explicit regularization. Furthermore, unlike the vision tasks, our NLP models are adapted using Low-Rank Adaptation (LoRA). The intrinsically small rank r imposes a strict capacity bottleneck on the trainable parameters, restricting the model's ability to execute complex geometric realignments. Due to the combination of a naturally separated feature space and constrained optimization capacity, the performance improvements from our alignment mechanism are less pronounced in the NLP domain.

G.2.9 Why ETF Rather Than Orthogonal or Low-Rank Structures?

Theoretical motivation. For C class directions in $\mathbb{R}^{d_{\text{out}}}$, pairwise orthogonality requires $d_{\text{out}} \geq C$, whereas an exact simplex ETF requires $d_{\text{out}} \geq C - 1$. The ideal simplex has balanced norms and equal pairwise inner products $-1/(C-1)$, consistent with the Neural Collapse reference geometry. When $d_{\text{out}} < C - 1$, neither this exact simplex nor C orthogonal directions can be realized. We therefore use an ETF-inspired numerical frame as a geometric regularizer rather than claim exact equal-angle structure in that regime. Its empirical benefit is assessed by the comparisons below; the theoretical statements retain their stated frame and compatibility assumptions.

Empirical comparison. To further validate this design choice, we compared ETF-based alignment against **Low-Rank Alignment** and **Orthogonal Alignment** baselines, following the reviewer's suggestion. The rank in low-rank alignment was fixed to $d_{\text{out}}/4$. Although low-rank alignment preserved variance, it failed to maintain balanced angular separation across tasks, resulting in weaker class discrimination. Orthogonal alignment enforced stronger independence but sacrificed geometric

uniformity when $C > d_{\text{out}}$. As shown in Table 17, both baselines perform consistently worse than our ETF-based method, supporting the empirical benefit of ETF-inspired directional regularization.

Table 17: Comparison between ETF-based, low-rank, and orthogonal alignment strategies across different ViT architectures and task counts.

Method	ViT-B/32			ViT-B/16			ViT-L/14		
	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
MDA TA (ETF)	**86.4**	**81.4**	**77.3**	**89.9**	**85.8**	**84.7**	**92.0**	**89.4**	**88.4**
MDA TA (Low-Rank)	79.7	74.0	68.1	85.6	79.8	75.6	90.6	87.1	85.8
MDA TA (Orthogonal)	84.7	79.7	75.0	88.6	84.2	80.5	91.2	88.2	87.0

Conclusion. The ideal simplex geometry motivates our directional reference, and the empirical comparisons support its use as an alignment target. The numerical frame encourages balanced directional organization without requiring exact equal-angle relationships in dimension-limited settings.

H Impact Statement

This paper presents work whose goal is to advance the field of machine learning by improving the reliability and scalability of model merging through principled directional alignment in parameter and feature spaces. By enabling multiple task-specific models to be merged without joint retraining or access to task data, the proposed approach can reduce storage and deployment costs in multi-task and edge settings, potentially improving the accessibility and efficiency of machine learning systems. At the same time, more effective model merging may amplify both the strengths and limitations of the underlying models, including existing biases or errors inherited from task-specific fine-tuning. The proposed method does not introduce new data sources or application domains, and its ethical and societal implications are therefore largely aligned with those of existing model merging and multi-task learning techniques.

I Use of Large Language Models (LLMs)

In preparing this paper, we used large language models (LLMs) solely as an assistive tool for language polishing and minor writing improvements. The models were not involved in research ideation, experimental design, data analysis, or drawing scientific conclusions. All conceptual and technical contributions are the work of the authors. The authors take full responsibility for the contents of this paper. For the experimental evaluation, we used Qwen2-7B-Instruct as the base model to conduct all experiments.

J Ethics Statement

This work does not involve human subjects, sensitive personal data, or experiments that may directly cause harm to individuals, groups, or society. All datasets used in our study are publicly available, widely used in the machine learning community, and subject to their respective licenses. We have taken care to ensure fair use of these datasets and avoided introducing biases beyond what is inherent in the original data.

To promote reproducibility and transparency, we provide detailed experimental settings, algorithmic descriptions, and publicly available code. Our contributions are methodological in nature and aim to advance the understanding of the model merging field. We are committed to the principles outlined in the NeurIPS Code of Ethics and confirm that this work adheres to them in full.

K Reproducibility Statement

To facilitate community development, we provide detailed descriptions of the datasets used in our experiments (Section E), the compared baselines (Section D), the training configurations (Section G.2.1), as well as our proposed methods (Algorithms 1, 2 and Figure 3). In addition, we release publicly available code to reduce the difficulty of reproduction.

L Devices

In the experiments, we conduct all methods on a local Linux server equipped with one AMD EPYC 7742 64-Core Processor (128 logical threads). All methods are implemented using the PyTorch framework, and all models are trained on NVIDIA A800-SXM4-80G GPUs (80GB HBM2e memory).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims about the "Merging with Directional Alignment (MDA)" method, mapping directly to the three main theoretical and empirical contributions outlined at the end of Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are explicitly discussed at the end of the conclusion/impact section, acknowledging computational overhead from optimizing rotation matrices and the open challenge of merging models across different architectures.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides theoretical analyses of structural coherence, with full proofs and approximation error decompositions detailed in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix J contains a "Reproducibility statement" referencing detailed dataset descriptions, baseline configurations, training setups, algorithms, and anonymized source code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The authors provide an anonymized link to the source code repository in the Introduction and affirm all datasets used are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Extensive experimental details, including training configurations and dataset descriptions, are provided in Appendices E and G to ensure reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports average performance metrics across multiple tasks and architectures in extensive benchmarks, but explicit error bars or standard deviations are omitted due to the deterministic evaluation protocols standard in large-scale model merging.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper explicitly analyzes empirical overhead, providing theoretical asymptotic complexity comparisons (Table 5) and computational resource discussions in Appendix B.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper includes a dedicated Ethics Statement (Appendix I) explicitly confirming full adherence to the ICML/NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Impact Statement discusses the positive impacts on storage/deployment costs as well as the negative risk of amplifying biases inherited from base models.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The framework proposed in the paper uses widely available standard datasets and models, presenting no new release of high-risk assets or scraped data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The Ethics Statement (Appendix I) affirms that all datasets are subject to their respective licenses and have been properly used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper proposes an algorithmic improvement but does not introduce or release any standalone new datasets requiring novel documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The Ethics Statement explicitly confirms that the research does not involve human subjects or sensitive personal data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The research does not involve human subjects, rendering IRB approvals unnecessary.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We explicitly declare the usage of LLMs in our work. First, as noted in the appendix, LLMs were utilized as an assistive tool solely for language polishing and editing. Second, open-source LLMs, specifically Qwen2-7B-Instruct, were employed as the base models to conduct our experimental evaluations and test the proposed model merging framework, as detailed in the Experiments section.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.