NULL-SPACE FILTERING FOR DATA-FREE CONTINUAL MODEL MERGING: PRESERVING TRANSPARENCY, PROMOTING FIDELITY

Zihuan Qiu[1], **Lei Wang**[1], **Yang Cao**[3], **Runtong Zhang**[1], **Bing Su**[3], **Yi Xu**[2], **Fanman Meng**[1], **Linfeng Xu**[1], **Qingbo Wu**[1], **Hongliang Li**[1]
[1]University of Electronic Science and Technology of China,
[2]Dalian University of Technology, Dalian, China, [3]Jiigan Technology

ABSTRACT

Data-free continual model merging (DFCMM) aims to fuse independently fine-tuned models into a single backbone that evolves with incoming tasks without accessing task data. This paper formulate two fundamental desiderata for DFCMM: *transparency*, avoiding interference with earlier tasks, and *fidelity*, adapting faithfully to each new task. This poses a challenge that existing approaches fail to address: how to bridge data-level desiderata with parameter-space optimization to ensure transparency and fidelity in the absence of task data. To this end, we propose NUFILT (**NU**ll-space **FILT**ering), a data-free framework that directly links these desiderata to optimization. Our key observation is that task vectors approximately align with representation subspaces, providing structural surrogates for enforcing transparency and fidelity. Accordingly, we design a null-space projector that preserves prior responses by filtering out overlapping components of new task vectors, thereby ensuring transparency, and a lightweight LoRA adapter that injects complementary task-specific signals, enabling fidelity in adapting to new tasks. The adapter is trained with a projection-based surrogate loss to retain consistency with previous knowledge while introducing novel directions. This joint filtering–adaptation process allows the backbone to absorb new knowledge while retaining existing behaviors, and the updates are finally fused back in a layer-wise linear fashion without extra parameters or inference cost. Theoretically, we establish approximate subspace alignment guarantees that justify null-space filtering. Empirically, NUFILT achieves state-of-the-art performance with minimal forgetting on both vision and NLP benchmarks, improving average accuracy by 4–7% over OPCM and WUDI-Merging, while narrowing the gap to fine-tuning and reducing computation overhead.

1 INTRODUCTION

Modern machine learning systems are often deployed in dynamic environments where tasks arrive one after another. Training a new model from scratch for each task is not only computationally expensive but also requires retaining all historical training data, which is often impractical in real applications. At the same time, maintaining a separate checkpoint for every task quickly becomes infeasible due to memory and deployment constraints (McMahan et al., 2017; Fang et al., 2024; Qiu et al., 2024). A more appealing alternative is to reuse existing models—either drawn from a model library or trained for new tasks—and consolidate their knowledge into a single backbone that evolves over time (Zhou et al., 2024b; Yu et al., 2024b; Huang et al., 2024a; Tang et al., 2025). However, this consolidation is expected to be performed while retaining only the merged model for storage efficiency, and without accessing any data in order to safeguard privacy, which makes the problem particularly challenging. These constraints highlight the urgent need for approaches that integrates knowledge across sequential tasks directly in parameter space without extra storage or data costs.

Recent studies investigate data-free continual model merging (DFCMM) (Liu & Soatto, 2023; Porrello et al., 2025; Tang et al., 2025), where tasks arrive sequentially and, at each step, only the newly fine-tuned task model and the previously merged backbone are available (see Fig. 1). We formulate

two fundamental desiderata for DFCMM: *transparency*, avoiding interference with earlier tasks, and *fidelity*, adapting faithfully to each new task. These desiderata, not explicitly considered in prior work, highlight a new open challenge for the field. Since they are inherently defined in terms of data, existing approaches lack a principled translation of data-level desiderata into parameter-space objectives. As a result, simple averaging or arithmetic updates often cause interference that undermines transparency (Izmailov et al., 2018; Ilharco et al., 2023); projection methods fail to preserve fidelity when task vectors are correlated (Tang et al., 2025; Yadav et al., 2023); and adaptive strategies typically rely on auxiliary data, which violates the data-free constraint (Yang et al.; Tang et al., 2024b; Qiu et al., 2025). Achieving both transparency and fidelity under the strict requirements of DFCMM therefore remains an open and unresolved challenge.

In this paper, we introduce NUFILT (**NU**ll-space **FILT**ering), a novel framework for DFCMM that directly bridges data-level desiderata with parameter-space optimization. The key observation is that task vectors exhibit approximate alignment with representation subspaces, enabling structural surrogates for enforcing transparency and fidelity without data. To achieve transparency, a null-space projector filters out components of the new task vector overlapping with earlier subspaces, suppressing interfering activations. To achieve fidelity, a lightweight LoRA adapter performs projection-aware adaptation, optimized with a data-free surrogate loss: updates are constrained to remain consistent with previous tasks while introducing complementary signals along directions unique to the new task. Finally, the projector, task vector, and adapter are fused back into the backbone in a layer-wise linear fashion, so updates are absorbed without extra parameters or inference cost. Theoretically, we provide guarantees of approximate subspace alignment, establishing a rigorous foundation for null-space filtering. Empirically, we demonstrate that NUFILT achieves state-of-the-art accuracy with minimal forgetting across both vision and NLP tasks. Compared to recent methods such as OPCM (Tang et al., 2025) and WUDI-Merging (Cheng et al.), NUFILT improves average accuracy by 4–7%, while substantially narrowing the gap to individual fine-tuning and reducing computation overhead.

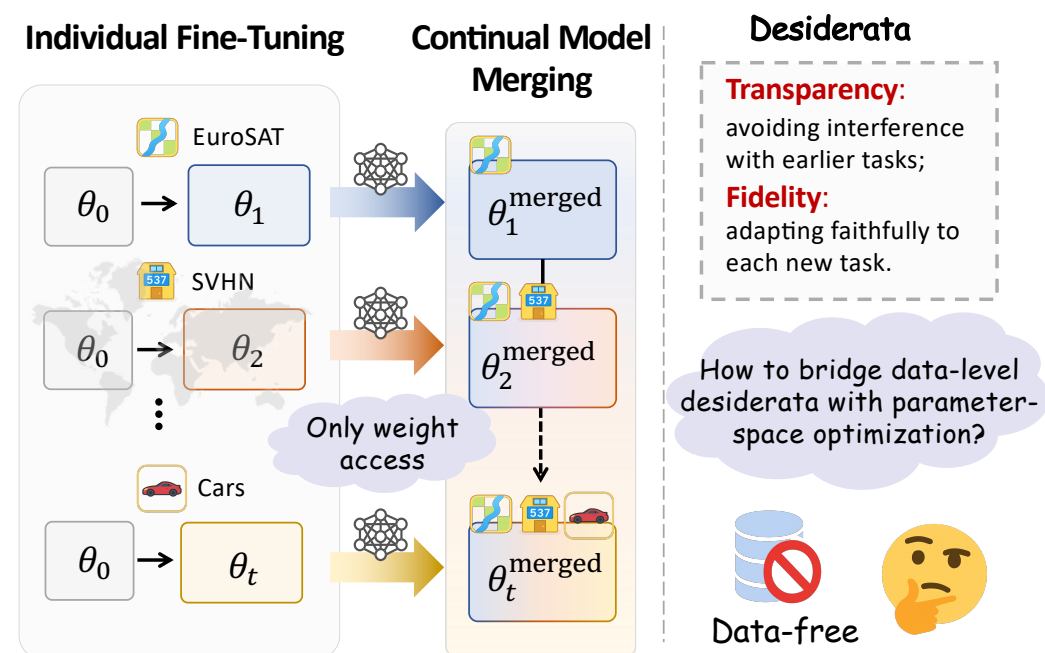

Figure 1: Illustration of **data-free continual model merging** (DFCMM). At each step, only the current task model and the previously merged model are accessible, and the merging process is performed without access to any data. The merged model is expected to preserve prior knowledge (transparency) while adapting efficiently to new tasks (fidelity).

To summarize, our main contributions are as follows:

- We formulate *transparency* and *fidelity* as two fundamental desiderata for data-free continual model merging, framing a new open challenge absent from prior work.
- We establish that task vectors exhibit approximate alignment with data representation subspaces, revealing a geometric property that explains how task vectors interact with representations.
- We propose NUFILT, a data-free continual model merging framework that combines null-space filtering with projection-aware adaptation to enforce both transparency and fidelity.
- Extensive experiments on vision and NLP benchmarks demonstrate that NUFILT achieves state-of-the-art performance, surpassing prior methods in accuracy while better resisting forgetting.

2 Related Work

Model Merging. Model merging provides an efficient alternative to multi-task or continual training by combining multiple fine-tuned models directly in parameter space, without revisiting their training data. Early efforts adopted simple weight averaging (Utans, 1996; Shoemake, 1985), which can connect models through linear mode connectivity (Entezari et al., 2021; Ainsworth et al., 2022) but often suffers from severe performance degradation due to parameter interference. Task Arithmetic (TA) (Ilharco et al., 2023) formalizes merging through task vectors, making merging operations more explicit. However, since standard fine-tuning rarely guarantees disentangled updates (Ortiz-Jimenez et al., 2023), TA can amplify interference, prompting structured finetuning strategies (Ortiz-Jimenez

et al., 2023; Liu & Soatto, 2023; Porrello et al., 2025). To mitigate conflicts, several methods introduce additional structure. Sparsity-based approaches such as TIES-Merging (Yadav et al., 2023) prune redundant or conflicting parameters, while reweighting schemes (Yu et al., 2024b) balance task contributions. Projection-based methods enforce orthogonality between task vectors to separate task directions (Tang et al., 2025; Cheng et al.), but struggled when tasks are inherently correlated. Adaptive strategies further adjust merging process with auxiliary calibration data (Yang et al.; Tang et al., 2025; Qiu et al., 2025), but this departs from the strictly data-free setting. Overall, these methods reduce interference to varying degrees but either rely on data or assume strong independence between tasks, leaving the core challenge of simultaneously achieving transparency and fidelity in the data-free continual setting largely unresolved.

Continual Learning. Continual learning tackles catastrophic forgetting (McCloskey & Cohen, 1989), where models lose performance on earlier tasks when adapting to new ones. Solutions include constraining parameter updates via importance weights (Kirkpatrick et al., 2016; Zenke et al., 2017; Aljundi et al., 2018), preserving knowledge through distillation (Hou et al., 2019; Douillard et al., 2020), or replaying exemplars and surrogates (Rebuffi et al., 2016; Liu et al., 2021). Other strategies expand capacity with dynamic architectures (Lee et al., 2017; Zhou et al., 2024a) or adopt parameter-efficient modules such as adapters and prompts (Yu et al., 2024a; Huang et al., 2024b). Beyond traditional data-driven approaches, a complementary direction explores continual learning through model merging. Early efforts merged fine-tuned checkpoints to alleviate forgetting (Mirzadeh et al., 2021; Wen et al., 2023; Marczak et al., 2024), but these typically relied on training data or sequential fine-tuning. More recently, continual model merging (Jin et al., 2023; Liu & Soatto, 2023; Porrello et al., 2025; Tang et al., 2025; Qiu et al., 2025) has emerged, aiming to fuse independently fine-tuned models directly in parameter space, avoiding both data access and checkpoint storage while improving scalability and privacy.

3 Background and motivation

We formalize data-free continual model merging in Sec. 3.1, review representative approaches in Sec. 3.2, outline key desiderata for data-free merging in Sec. 3.3, and analyze the geometric relation between task vectors and representations in Sec. 3.4.

3.1 Problem setting

We study data-free continual model merging, where a sequence of task-specific models $\{\theta_t\}_{t=1}^T$ are independently fine-tuned from a shared pre-trained model θ_0 on labeled datasets $\mathcal{D}_t$ with disjoint label sets $\mathcal{C}_t$. The goal is to obtain a single merged model θ_T^{merged} that generalizes to the union label space $\mathcal{C}_{1:T} = \bigcup_{t=1}^T \mathcal{C}_t$. Unlike conventional continual learning, we assume *no access* to raw training data. All knowledge integration must therefore occur directly in parameter space, via recursive merging of task-adapted checkpoints:

$$\theta_t^{\text{merged}} = \text{Merge}\Big(\theta_{t-1}^{\text{merged}}, \theta_t\Big), \quad (\theta_1^{\text{merged}} = \theta_1). \tag{1}$$

To expose the underlying structure, let $\tau_t = \theta_t - \theta_0$ denote the *task vector* (Ilharco et al., 2023), capturing the update induced by task t. Rather than merging two checkpoints directly, we reinterpret continual merging as learning a transformed update:

$$\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tilde{\tau}_t, \quad (\tilde{\tau}_t = F(\tau_t)). \tag{2}$$

Here, F transforms task vectors to stay compatible with previously merged parameters, ensuring seamless integration of new updates.

3.2 Representative Continual Model Merging Solutions

Several continual model merging approaches can be viewed as special cases of task vector transformation, each with different assumptions on $F(\tau_t)$. Below we outline representative solutions:

❶ **Weight Averaging (WA)** (Izmailov et al., 2018): Updates parameters by simple averaging, $\theta_t^{\text{merged}} = \frac{1}{t}\big[(t-1)\theta_{t-1}^{\text{merged}} + \theta_t\big]$. This stabilizes optimization but assumes task compatibility and

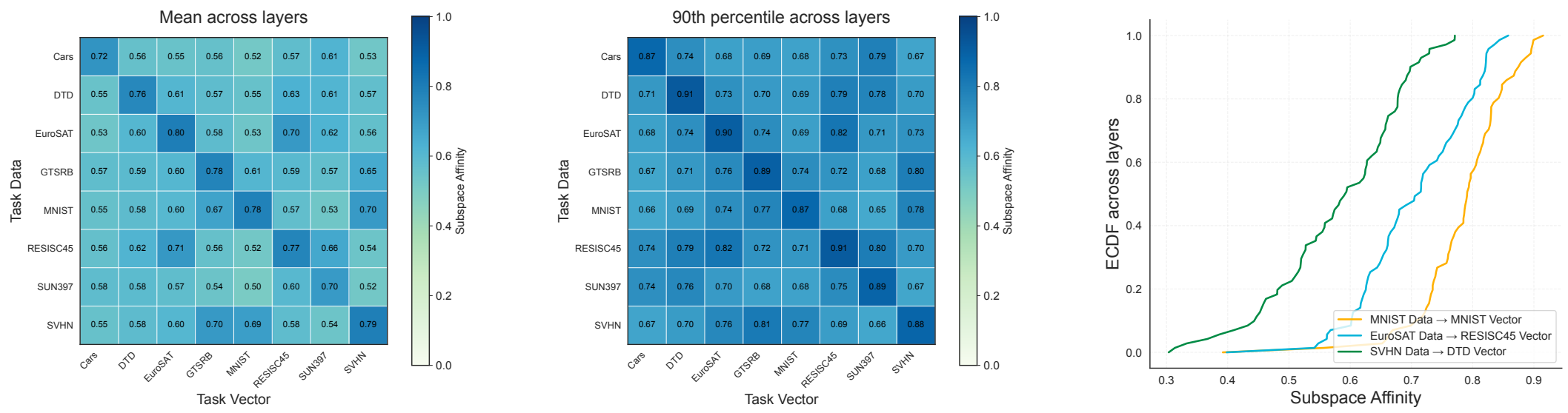

(a) Mean across layers (b) 90th percentile across layers (c) ECDF across layers

Figure 2: Subspace affinity between data and task vectors in ViT-B/16 across eight datasets. Heatmaps show diagonal dominance, and layer-wise ECDFs confirm higher affinities for matched pairs.

equal importance of checkpoints, making it sensitive to semantic conflicts. ❷ **Task Arithmetic (TA)** (Ilharco et al., 2023): Adds scaled task vectors, $\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \lambda\, \tau_t$, where λ is a tunable coefficient. It can outperform naive averaging but lacks structural constraints, leading to scale sensitivity and task interference. ❸ **Orthogonal Projection-based Continual Merging (OPCM)** (Tang et al., 2025): Projects each task vector onto the orthogonal complement of previous directions, $\theta_t^{\text{merged}} = \theta_0 + \frac{1}{\lambda_t}\left[\lambda_{t-1}\tau_{t-1}^{\text{merged}} + \mathcal{P}^{(t-1)}(\tau_t)\right]$, where $\mathcal{P}^{(t-1)}(\cdot)$ removes overlapping subspaces. This enforces geometric separation but struggles when task vectors are inherently entangled or non-orthogonal. ❹ **AdaMerging** (Yang et al.): Adapts coefficients using a small unlabeled test set, $\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \lambda_t(\mathcal{D}_t^{\text{test}})\, \tau_t$. By exploiting test-time signals, it improves task alignment and mitigates scaling issues, but requires auxiliary data, violating the data-free assumption.

3.3 Desiderata for Data-Free Continual Merging

We identify two desiderata for the transformed task vectors $\tilde{\tau}_i = F(\tau_i)$ when merging without data:

① **Transparency to previous tasks.** New task updates should not overwrite knowledge from earlier tasks. For any input X from $\mathcal{D}_{i\le t-1}$, the predictions after adding $\tilde{\tau}_t$ should remain consistent with those before:

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{X\sim\mathcal{D}_{i\le t-1}}\Big[\ell\Big((\theta_0 + \sum_{i=1}^{t}\tilde{\tau}_i)X^\top,\ (\theta_0 + \sum_{i=1}^{t-1}\tilde{\tau}_i)X^\top\Big)\Big], \tag{3}$$

where ℓ is a distance metric (*e.g.*, squared ℓ_2, KL divergence).

② **Fidelity to the current task.** The merged model should also reproduce the behavior of the task-specific model θ_t on its in-distribution inputs:

$$\mathcal{L}_{\text{fid}} = \mathbb{E}_{X\sim\mathcal{D}_t}\Big[\ell\Big((\theta_0 + \sum_{i=1}^{t}\tilde{\tau}_i)X^\top,\ (\theta_0 + \tau_t)X^\top\Big)\Big]. \tag{4}$$

In principle, optimizing $\mathcal{L}_{\text{trans}}$ and $\mathcal{L}_{\text{fid}}$ would yield task vectors that preserve prior knowledge while faithfully incorporating new tasks. Yet this is infeasible in the data-free setting, since $\mathcal{D}_{i\le t-1}$ and $\mathcal{D}_t$ are unavailable. This leads to the central challenge:

*How can continual merging achieve **transparency** and **fidelity** without data?*

3.4 Approximate Alignment Between Representation and Task Vector Subspaces

We examine the geometric relation between task vectors and data representations. We hypothesize that task vectors tend to align with principal directions in the representation space that capture task-relevant information. To validate this, we conduct a subspace alignment analysis and provide a theoretical guarantee that *task vector subspaces are approximately aligned with data subspaces*, which in turn motivates our null-space filtering approach.

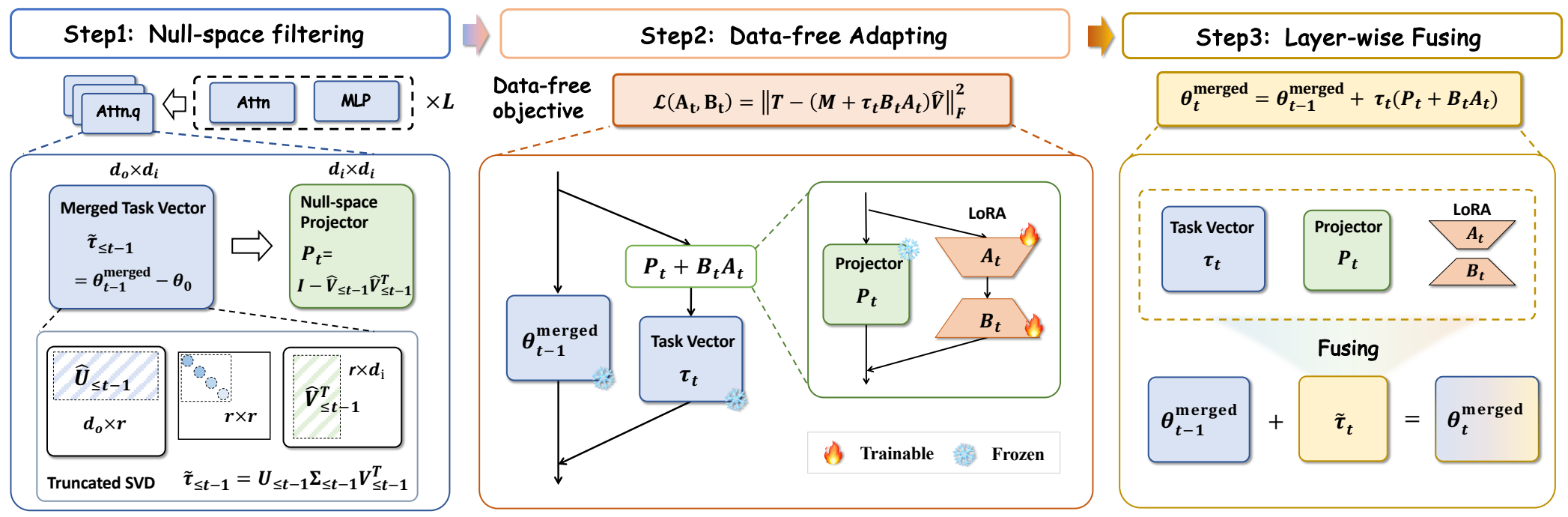

Figure 3: Overview of the NUFILT procedure. ❶ **Filtering**: the new task vector is processed through a null-space projector that suppresses activations from previous tasks, ensuring transparency to past knowledge. ❷ **Adapting**: within the filter, a lightweight LoRA adapter refines the update for the current task using a data-free objective. ❸ **Fusing**: the filter, task vector, and LoRA module are merged back into the backbone, keeping the parameter count and inference cost unchanged.

For each task and layer l, let $V_d^{(l)} \in \mathbb{R}^{d_i \times r_d}$ be the singular vectors of the representation covariance. For the task vector $\tau^{(l)}$, we compute its truncated SVD, $\tau^{(l)} \approx \hat{U}^{(l)}\hat{\Sigma}^{(l)}\hat{V}^{(l)\top}$, where $\hat{V}^{(l)} \in \mathbb{R}^{d_i \times r_v}$ is the top r_v right singular vectors. We define the subspace affinity:

$$\mathcal{A}(V_d^{(l)}, \hat{V}^{(l)}) = \frac{1}{r_d}\|\hat{V}^\top V_d\|_F^2, \quad (r_d \leq r_v \leq d_i) \tag{5}$$

which lies in $[0, 1]$ and measures the degree of alignment between the data and vector subspace.

Empirically, across ViT-B/16 models fine-tuned on eight datasets, affinity heatmaps (Fig. 2) show strong diagonal dominance: matched data–vector pairs exhibit much higher affinities than mismatched ones. Layer-wise ECDFs (Fig. 2c) further reveal a spectrum of overlaps, from high (MNIST–MNIST), moderate (EuroSAT–RESISC45), to low (SVHN–DTD). These results indicate that task vectors are consistently aligned with task-relevant representation directions, a trend observed in both vision and NLP tasks (additional results for more models provided in the Appendix C.2).

Theorem 1 (Approximate Subspace Alignment)**.** *Let $\tau^{(l)} \in \mathbb{R}^{d_o \times d_i}$ be the task vector at layer l, and $H \in \mathbb{R}^{N \times d_i}$ a representation matrix of rank r_d with right singular vectors $V_d \in \mathbb{R}^{d_i \times r_d}$. Suppose $\tau^{(l)}$ admits the decomposition $\tau^{(l)} = T_0 + E$ with $T_0 = BH$, $\mathrm{rank}(T_0) = r_d$, where each row E_k of E satisfies $\|E_k\|_2 \leq \Psi_k$. If $\sigma_{r_d}(T_0) > \sqrt{d_o}\,\max_k \Psi_k$, then for any $r_v \geq r_d$, the span of the top r_v right singular vectors $\hat{V}$ of $\tau^{(l)}$ is approximately aligned with $\mathrm{span}(V_d)$ in the sense that*

$$\mathcal{A}(V_d^{(l)}, \hat{V}^{(l)}) \;\leq\; \zeta^2, \tag{6}$$

where $\mathcal{A}(V_d^{(l)}, \hat{V}^{(l)}) = 1 - \frac{1}{r_d}\|\hat{V}^\top V_d\|_F^2$ and $\zeta^2 = \left(\frac{\sqrt{d_o}\,\max_k \Psi_k}{\sigma_{r_d}(T_0) - \sqrt{d_o}\,\max_k \Psi_k}\right)^2$.

This theorem shows that $\tau^{(l)}$ consists of a low-rank component aligned with the representation space plus a bounded perturbation, so its principal subspace is approximately aligned with the data subspace. Together with the empirical evidence, this provides both geometric intuition and theoretical justification for treating task vectors as carriers of task-relevant representation directions.

4 Method: NUFILT

Motivated by the above, we propose NUFILT (**NU**ll-space **FILT**ering), a data-free method that enforces structural constraints on τ_t. The key idea is to leverage the geometry of prior tasks to suppress activations tied to earlier ones, enabling continual merging while mitigating forgetting.

4.1 Steps in NUFILT: Filtering, Adapting & Fusing

The procedure of NUFILT unfolds in three steps, as illustrated in Fig. 3:

Step ❶: Filtering. To prevent interference with prior tasks, we introduce a null-space filter before applying each new task vector. For layer l, we compute the cumulative update

$$\tilde{\tau}_{\le t-1}^{(l)} = \theta_{t-1}^{\text{merged},(l)} - \theta_0^{(l)}. \tag{7}$$

and obtain its top-r_p right singular vectors $\hat{V}_{\le t-1}^{(l)}$. The null-space filter is defined as

$$P_t^{(l)} = I - \hat{V}_{\le t-1}^{(l)} \hat{V}_{\le t-1}^{(l)\top}, \tag{8}$$

which removes components aligned with earlier task directions. Thus, for old task features $h^{(l)} \in \operatorname{span}(\hat{V}_{\le t-1}^{(l)})$, we have $P_t^{(l)} h^{(l)} = 0$, ensuring transparency to prior knowledge.

Step ❷: Adapting. Within the filtered subspace, we insert a low-rank LoRA adapter

$$P_t^{(l)} \;\to\; P_t^{(l)} + B_t^{(l)} A_t^{(l)}, \tag{9}$$

where $A_t^{(l)} \in \mathbb{R}^{r_l \times d_i}$ and $B_t^{(l)} \in \mathbb{R}^{d_o \times r_l}$. This lightweight term restores task-specific flexibility and is trained with a data-free objective (Sec. 4.2), allowing the merged model to recover the performance of the fine-tuned model on task t without requiring raw data.

Step ❸: Fusing. Finally, the filter, task vector, and LoRA module are merged into the backbone:

$$\theta_t^{\text{merged},(l)} = \theta_{t-1}^{\text{merged},(l)} + \tau_t^{(l)} \big(P_t^{(l)} + B_t^{(l)} A_t^{(l)}\big). \tag{10}$$

Since all operations are linear, the composite update can be absorbed into the weight matrix, leaving the parameter count and inference cost identical to an individual model.

4.2 DATA-FREE OBJECTIVE OF NUFILT

From Theorem 1, we obtain a data-free upper bound on parameter–representation interactions.

Corollary 1 (Data-free upper bound)**.** *Let $X \in \mathbb{R}^{N \times d_i}$ with largest singular value $\sigma_1(X)$, rank r_d, and right singular vectors V_d. Let $\tau \in \mathbb{R}^{d_o \times d_i}$ with top-r_v right singular vectors $\hat{V}$ ($r_v \ge r_d$). If Theorem 1 ensures $1 - \frac{1}{r_d} \|\hat{V}^\top V_d\|_F^2 \;\le\; \zeta^2$, then for any $\rho \in \mathbb{R}^{d_o \times d_i}$,*

$$\big\|(\rho - \tau) X^\top\big\|_F^2 \;\le\; 2\,\sigma_1(X)^2 \Big(\big\|(\rho - \tau)\hat{V}\big\|_F^2 + r_d\,\zeta^2\,\|\rho - \tau\|_2^2\Big). \tag{11}$$

Subspace surrogate for losses. Under squared ℓ_2 loss, the data losses (Eq. 3–4) are

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{X \sim \mathcal{D}_{i \le t-1}} \|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) X^\top\|_F^2, \quad \mathcal{L}_{\text{fid}} = \mathbb{E}_{X \sim \mathcal{D}_t} \|(\tilde{\tau}_{\le t} - \tau_t) X^\top\|_F^2, \tag{12}$$

where $\tilde{\tau}_{\le t} = \sum_{i=1}^{t} \tilde{\tau}_i$. Let X_o and X_n denote feature matrices from old and new tasks. Applying Corollary 1 with $(\rho, \tau) = (\tilde{\tau}_{\le t}, \tilde{\tau}_{\le t-1})$ and $(\tilde{\tau}_{\le t}, \tau_t)$, and denoting the corresponding misalignments by ζ_o and ζ_n, we obtain the following *data-free upper bounds*:

$$\|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) X_o^\top\|_F^2 \le 2\,\sigma_1(X_o)^2 \Big(\|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) \hat{V}_{\le t-1}\|_F^2 + r_o \zeta_o^2 \|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1})\|_2^2\Big), \tag{13}$$

$$\|(\tilde{\tau}_{\le t} - \tau_t) X_n^\top\|_F^2 \le 2\,\sigma_1(X_n)^2 \Big(\|(\tilde{\tau}_{\le t} - \tau_t) \hat{V}_t\|_F^2 + r_n \zeta_n^2 \|(\tilde{\tau}_{\le t} - \tau_t)\|_2^2\Big). \tag{14}$$

Thus when the misalignment ζ is small, both losses are governed by the terms $\|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) \hat{V}_{\le t-1}\|_F^2$ and $\|(\tilde{\tau}_{\le t} - \tau_t)\hat{V}_t\|_F^2$, offering a projection-aware objective for effective data-free surrogates.

Projection-aware objective. At layer l, the merged task vector (subtracting $\theta_0^{(l)}$ from Eq. 10) is

$$\tilde{\tau}_{\le t}^{(l)} = \tilde{\tau}_{\le t-1}^{(l)} + \tau_t^{(l)} (P_t^{(l)} + B_t^{(l)} A_t^{(l)}). \tag{15}$$

Substituting Eq. 15 into the projection terms yields the compact objective:

$$\mathcal{L}(A_t^{(l)}, B_t^{(l)}) = \big\|T - (M + \tau_t^{(l)} B_t^{(l)} A_t^{(l)}) \hat{V}\big\|_F^2, \tag{16}$$

with

$$T = \begin{bmatrix} \tilde{\tau}_{\le t-1}^{(l)} \hat{V}_{\le t-1}^{(l)} \\ \tau_t^{(l)} \hat{V}_t^{(l)} \end{bmatrix}, \quad \hat{V} = \begin{bmatrix} \hat{V}_{\le t-1}^{(l)} \\ \hat{V}_t^{(l)} \end{bmatrix}, \quad M = \tilde{\tau}_{\le t-1}^{(l)} + \tau_t^{(l)} P_t^{(l)}. \tag{17}$$

Here the residual $\tau_t^{(l)} B_t^{(l)} A_t^{(l)}$ bridges $M\hat{V}$ and T, ensuring transparency to past tasks and fidelity to the new one. Since $\tau_t^{(l)}$ is generally non-square and low-rank, closed-form for $(A_t^{(l)}, B_t^{(l)})$ are unstable; we thus optimize Eq. 16 via gradient descent. The full procedure is summarized in Algo. 1.

Algorithm 1 NUFILT Procedure

Input: pre-trained model θ_0; fine-tuned models $\{\theta_t\}_{t=1}^T$; rank parameters $\{r_p, r_l, r_v\}$; learning rate η; and maximum number of iterations MaxIter.
Initialize: $\theta_1^{\text{merged}} = \theta_1$

for each task $t \in \{2, \ldots, T\}$ **do**
  for selected linear layer index $l \in \{1, \ldots, L\}$ **do**
    $P_t^{(l)} = I - \hat{V}_{\leq t-1}^{(l)} \hat{V}_{\leq t-1}^{(l)\top}$ ▷ null-space via $\hat{V}_{\leq t-1}^{(l)} = \text{TOPRIGHTSVD}(\tilde{\tau}_{\leq t-1}^{(l)}, r_p)$
    $A_t^{(l)} \leftarrow \mathbf{0}, \quad B_t^{(l)} \sim \mathcal{N}(0, \sigma_{init}^2)$ ▷ initialize LoRA components with low rank r_l
  for iter $= 1$ **to** MaxIter **do**
    compute $M = \tilde{\tau}_{\leq t-1}^{(l)} + \tau_t^{(l)} P_t^{(l)}$, T, and $\hat{V}$ via Eq. 17
    $\mathcal{L}(A_t^{(l)}, B_t^{(l)}) = \left\| T - \left(M + \tau_t^{(l)} B_t^{(l)} A_t^{(l)} \right) \hat{V} \right\|_F^2$ ▷ data-free objective (Eq. 16)
    $A_t^{(l)} \leftarrow A_t^{(l)} - \eta \nabla_{A_t^{(l)}} \mathcal{L}, \quad B_t^{(l)} \leftarrow B_t^{(l)} - \eta \nabla_{B_t^{(l)}} \mathcal{L}$ ▷ update LoRA components
  $\left\{ \theta_t^{\text{merged},(l)} \right\}_{l=1}^L = \left\{ \theta_{t-1}^{\text{merged},(l)} + \tau_t^{(l)} \left(P_t^{(l)} + B_t^{(l)} A_t^{(l)} \right) \right\}_{l=1}^L$ ▷ layer-wise fusion
Output: θ_T^{merged}

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUPS

Benchmarks and Protocols. We evaluate our approach on both vision and NLP tasks to assess scalability and generality. For vision, following Tang et al. (2025), we adopt CLIP-ViT backbones (Radford et al., 2021) and construct three groups of 8, 14, and 20 image classification tasks. We use publicly available ViT-B/32, ViT-B/16, and ViT-L/14 checkpoints, each fine-tuned on up to 20 datasets (Tang et al., 2024a). To ensure robustness to task order, all experiments are repeated over 10 random permutations (seeds 42–51). For NLP, we evaluate on eight GLUE tasks (Wang et al., 2019) using Flan-T5-base (Chung et al., 2024).

Implementation Details We insert the null-space filter in a cascaded fashion before selected linear layers of the backbone. All experiments share a single set of *global* hyper-parameters—used across both CLIP and Flan models, as well as all task orders and scales—without task-specific tuning: null-space rank $r_p = 128$, LoRA rank $r_l = 64$, and task projection rank $r_v = 8$. Each task is adapted for 50 iterations with Adam at a learning rate of 1×10^{-3}.

Metrics and Baselines. We evaluate performance with two standard metrics: average accuracy (ACC) and backward transfer (BWT) (Lin et al., 2022). ACC is the mean accuracy of the final merged model across all tasks, $\text{ACC} = \frac{1}{T} \sum_{i=1}^T a_i(\theta_T^{\text{merged}})$, where $a_i(\cdot)$ denotes accuracy on task i[1]. BWT quantifies the effect of merging on past tasks by comparing their performance before and after the final merge: $\text{BWT} = \frac{1}{T-1} \sum_{i=1}^{T-1} \left[a_i(\theta_T^{\text{merged}}) - a_i(\theta_i^{\text{merged}}) \right]$. In addition to the methods in Sec. 3.2, we compare against TIES-MERGING (Yadav et al., 2023), MAGMAX-IND (Marczak et al., 2024), WEMOE (Tang et al., 2024b), and WUDI-MERGING (Cheng et al.). Detailed descriptions are provided in the Appendix B.3.

5.2 MAIN RESULTS

Results of Vision Tasks The comparative results on vision tasks are summarized in Tab. 1, highlighting that our method, NUFILT, consistently outperforms prior merging techniques across various model architectures and task sequences, all without needing extra parameters or data access. Relative to baselines like OPCM and WUDI-Merging, NUFILT achieves marked improvements in average accuracy while reducing backward transfer. For example, on ViT-B/32 with 8 tasks, it reaches 83.6% accuracy, outperforming OPCM by 8.1% and WUDI-Merging by 8.9%; for 20 tasks, the gains are 5.3% over OPCM. On ViT-L/14, it surpasses OPCM by 4.6% on 8 tasks and 8.7% on 20 tasks, with backward transfer limited to -1.1% versus -2.6% for OPCM on 8 tasks. Overall, NUFILT closes

[1]For STSB (NLP), we report Spearman's ρ, denoted by $a_i(\cdot)$ for notational consistency.

Table 1: Comparative results of continual merging methods, reporting average accuracy and backward transfer over ten task orders (mean±std). EP and DA denote method assumptions: the need for extra parameters or access to data. Best results are in **bold**, and the second best are underlined.

	Method	Requirement EP / DA	ViT-B/32 8 tasks	ViT-B/32 14 tasks	ViT-B/32 20 tasks	ViT-B/16 8 tasks	ViT-B/16 14 tasks	ViT-B/16 20 tasks	ViT-L/14 8 tasks	ViT-L/14 14 tasks	ViT-L/14 20 tasks
	PRE-TRAINED	– / –	48.1	56.9	55.6	55.4	62.0	59.8	64.9	69.1	65.6
	INDIVIDUAL	– / –	90.4	89.3	89.8	92.4	91.3	91.6	94.3	93.4	93.5
	C. FINE-TUNED	– / –	79.8	67.4	62.6	82.9	72.2	68.2	90.0	70.9	77.7
ACC (%) ↑	WEIGHT AVERAGING	✗ / ✗	66.3 ±0.0	65.4 ±0.0	61.1 ±0.0	72.3 ±0.0	69.7 ±0.0	64.8 ±0.0	80.0 ±0.0	77.5 ±0.0	71.1 ±0.0
	TASK ARITHMETIC	✗ / ✗	67.5 ±0.0	66.5 ±0.0	60.0 ±0.0	77.1 ±0.0	70.9 ±0.6	64.2 ±0.0	82.1 ±0.0	77.9 ±0.0	70.3 ±0.0
	TIES-MERGING	✗ / ✗	49.0 ±10.2	66.2 ±0.6	59.9 ±0.7	66.8 ±3.7	70.5 ±0.8	63.0 ±1.6	64.3 ±7.0	78.0 ±0.6	68.3 ±0.9
	MAGMAX-IND	✗ / ✗	70.7 ±0.0	67.0 ±0.0	61.2 ±0.0	76.7 ±1.8	67.0 ±0.0	62.5 ±0.0	83.4 ±0.0	71.2 ±0.0	71.2 ±0.0
	LW ADAMERGING	✗ / ✓	53.4 ±3.2	59.8 ±1.6	59.7 ±7.4	59.9 ±2.3	64.3 ±1.2	61.5 ±1.1	68.8 ±2.9	73.1 ±5.7	66.9 ±1.1
	LORA-WEMOE	✓ / ✓	68.8 ±7.8	63.8 ±3.4	49.6 ±15.4	72.6 ±3.7	67.9 ±2.9	55.0 ±7.0	75.6 ±7.8	74.0 ±5.0	56.9 ±19.8
	WUDI-MERGING	✗ / ✗	74.7 ±6.6	67.0 ±6.9	63.7 ±3.8	81.0 ±4.7	75.0 ±4.1	69.6 ±4.7	87.5 ±3.3	84.2 ±3.7	78.1 ±2.8
	OPCM	✗ / ✗	75.5 ±0.5	71.9 ±0.3	65.7 ±0.2	81.8 ±0.3	77.1 ±0.5	70.3 ±0.2	87.0 ±0.4	83.5 ±0.2	76.0 ±0.2
	NUFILT (Ours)	✗ / ✗	**83.6** ±0.2	**78.0** ±0.2	**71.0** ±0.9	**87.3** ±0.1	**83.1** ±0.3	**78.1** ±0.9	**91.6** ±0.1	**89.2** ±0.1	**84.7** ±0.8
BWT (%) ↑	WEIGHT AVERAGING	✗ / ✗	-11.5 ±2.2	-8.0 ±1.3	-7.1 ±2.1	-9.7 ±1.5	-7.1 ±1.4	-7.3 ±1.7	-7.3 ±1.4	-5.8 ±1.0	-6.4 ±1.5
	TASK ARITHMETIC	✗ / ✗	-9.6 ±1.5	-1.3 ±1.6	-3.4 ±1.0	-4.2 ±1.0	-1.3 ±0.4	-3.6 ±0.4	-7.1 ±0.8	-1.8 ±0.3	-3.3 ±0.3
	TIES-MERGING	✗ / ✗	-15.3 ±8.0	**1.9** ±0.6	**-1.5** ±0.7	-5.5 ±0.4	**1.4** ±0.7	**-1.5** ±1.2	-13.0 ±5.7	**-1.1** ±0.4	**-2.9** ±1.0
	MAGMAX-IND	✗ / ✗	-8.3 ±1.3	-7.4 ±1.4	-7.2 ±1.6	-6.1 ±1.3	-7.4 ±2.0	-8.0 ±2.2	-5.0 ±0.8	-6.0 ±2.1	-6.5 ±2.1
	LW ADAMERGING	✗ / ✓	-32.5 ±3.6	-24.1 ±1.7	-22.7 ±4.3	-27.8 ±2.7	-22.1 ±1.4	-21.4 ±1.2	-24.3 ±3.3	-19.6 ±1.7	-21.7 ±1.1
	LORA-WEMOE	✓ / ✓	-20.4 ±9.0	-20.2 ±3.9	-24.5 ±10.0	-18.0 ±6.2	-18.8 ±3.4	-25.8 ±7.9	-17.8 ±5.9	-16.8 ±5.3	-27.9 ±17.2
	WUDI-MERGING	✗ / ✗	-17.0 ±7.5	-22.8 ±7.3	-26.0 ±4.1	-12.6 ±5.4	-16.9 ±4.4	-18.5 ±14.2	-7.3 ±3.7	-9.4 ±4.0	-15.8 ±2.9
	OPCM	✗ / ✗	-6.3 ±1.1	-6.0 ±1.0	-7.8 ±1.5	-4.8 ±0.7	-5.1 ±1.4	-6.3 ±2.2	-2.6 ±1.0	-4.3 ±0.7	-6.5 ±1.8
	NUFILT (Ours)	✗ / ✗	**-2.7** ±0.7	-5.7 ±0.9	-8.9 ±2.3	**-1.6** ±0.5	-3.5 ±0.6	-7.1 ±1.9	**-1.1** ±0.3	-2.0 ±0.3	-4.6 ±0.7

the gap to the individual fine-tuning benchmark, lagging by only 2.7% on ViT-L/14 for 8 tasks, underscoring its robustness in continual merging settings.

Results of NLP Tasks The results for NLP tasks using the Flan-T5-base model on 8 tasks are detailed in Tab. 2, showing that NUFILT excels all merging strategies. Against comparable baselines such as OPCM and WUDI-Merging, NUFILT delivers notable gains in overall performance. It achieves an average accuracy of 83.7%, exceeding WUDI-Merging by 1.5% and OPCM by 3.1%, while limiting backward transfer to -1.5% compared to -3.9% for WUDI-Merging and -2.5% for OPCM. NUFILT narrows the divide with individual fine-tuning, trailing by merely 2.7%, affirming its efficacy in continual merging for language models.

Table 2: Results of continual merging Flan-T5-base models on 8 tasks, ordered alphabetically.

Method	EP / DA	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	ACC ↑	BWT ↑
PRE-TRAINED	– / –	69.1	56.5	76.2	88.4	82.1	80.1	91.2	62.2	75.7	-
INDIVIDUAL	– / –	75.0	83.4	87.5	91.5	85.4	85.9	93.6	88.7	86.4	-
TASK ARITHMETIC	✗ / ✗	69.1	58.1	77.9	88.9	83.1	79.1	90.7	74.0	77.6	-4.6
TIES-MERGING	✗ / ✗	39.3	70.0	**82.4**	88.8	81.8	75.8	89.7	76.8	75.6	-6.1
LW ADAMERGING	✗ / ✓	69.1	58.1	77.9	88.9	83.1	79.1	90.7	74.2	77.6	-4.7
LORA-WEMOE	✓ / ✓	71.5	80.6	78.2	90.3	82.7	80.5	91.3	76.2	81.4	**0.1**
WUDI-MERGING	✗ / ✗	71.9	73.4	79.2	89.7	82.9	79.1	93.1	**88.2**	82.2	-3.9
OPCM	✗ / ✗	69.9	72.9	78.7	90.3	83.8	**83.0**	92.2	73.7	80.6	-2.5
NUFILT (Ours)	✗ / ✗	**72.5**	**83.3**	78.7	**91.1**	**84.2**	79.4	**93.4**	87.1	**83.7**	-1.5

5.3 ABLATION AND ANALYSIS RESULTS

Ablation on Each Component. We conduct an ablation study to disentangle the contributions of the null-space filter and the data-free objective. Results are summarized in Tab. 3. (1) Simply accumulating task vectors without any constraint leads to severe performance degradation. Both accuracy and backward transfer drop sharply as the number of tasks increases, confirming that direct parameter addition suffers from catastrophic interference. (2) Projecting task vectors onto the null-space of previous tasks significantly stabilizes merging. Compared to naive merging, this strategy improves accuracy by more than 20 points across all task counts and reduces forgetting to near zero. (3) Using the data-free objective alone (without null-space filtering) recovers part of the performance but still suffers from negative backward transfer, especially when scaling to more tasks. Combining both components yields the best balance between transparency and fidelity.

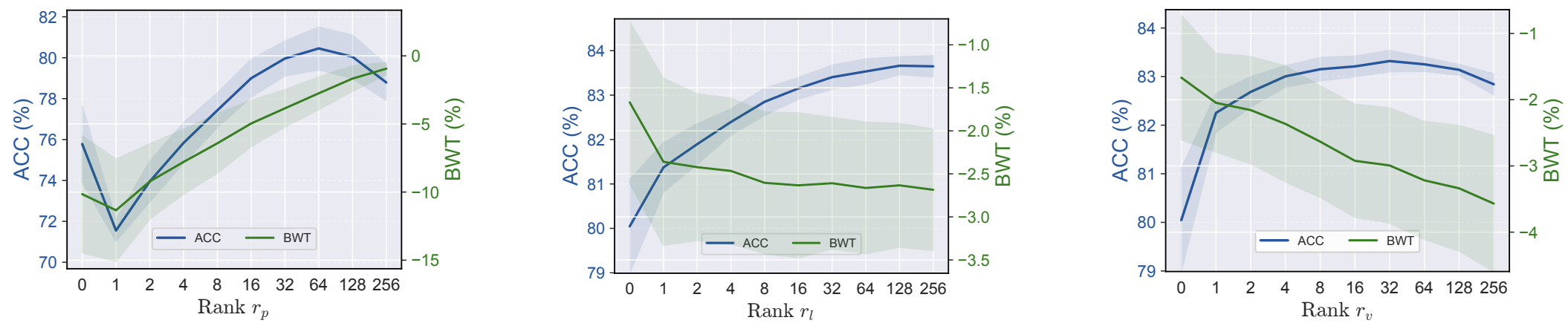

(a) Vary r_p, $r_l = r_v = 0$ (b) Vary r_l, $r_p = 128, r_v = 8$ (c) Vary r_v, $r_p = 128, r_l = 16$

Figure 4: Hyper-parameter sensitivity analysis on the 8-task continual merging protocol. Setting $r = 0$ corresponds to removing the associated component.

Table 3: Ablation study of NUFILT with CLIP ViT-B/32 over 8, 14, and 20 tasks.

Method	**Component**	**ACC(%)** ↑			**BWT(%)** ↑		
	Null-space / LoRA	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
(1) Naive Merging ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t$)	✗ / ✗	62.1±0.0	46.5±0.0	34.3±0.0	-18.5±6.2	-25.8±2.2	-24.7±5.1
(2) Only Null-space ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t P_t$)	✓ / ✗	80.0±1.1	76.7±1.0	67.0±0.7	-1.7±0.9	-4.2±0.9	-6.2±1.8
(3) Data-free Objective							
w/o null-space filter ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t B_t A_t$)	✗ / ✓	75.8±1.9	63.7±1.6	51.7±1.0	-10.2±4.3	-17.1±2.7	-20.6±4.9
full method ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t (P_t + B_t A_t)$)	✓ / ✓	83.6±0.2	78.0±0.2	71.0±0.9	-2.7±0.7	-5.7±0.9	-8.9±2.3

Hyper-parameter Sensitivity. We analyze the impact of rank hyper-parameters in Fig. 4. (a) Higher null-space rank r_p initially reduces forgetting and boosts performance, but excessively large ranks reduce projection selectivity, impairing new task adaptation and lowering accuracy. (b) Low-rank adaptation enhances adaptability while controlling forgetting. Increasing rank r_l improves accuracy steadily, with backward transfer remaining stable. (c) Projecting updates onto $\hat{V}_t^{(l)}$ improves new task adaptability. Moderate ranks yield consistent gains, while overly large r_v reduces projection specificity, approaching full-space projection and diminishing benefits.

Adaptation Overhead. Tab. 4 details the adaptation overhead for ViT-B/32 under 8-task continual merging. NUFILT outperforms baselines in both efficiency and accuracy. Compared to LW AdaMerging and LoRA-WEMoE, it uses similar or less GPU memory and runtime while achieving higher accuracy. WUDI-Merging is efficient but lags in accuracy. Remarkably, NUFILT reaches 83.3% accuracy with just 25 iterations and 9.3s, under one-third baseline runtimes. This efficiency underscores NUFILT's scalability and practicality for continual merging.

Table 4: Overhead of ViT-B/32 under 8-task continual merging: per-task iterations, total solving time, peak GPU memory, and final parameter count.

Model	**Iters.**	**Time**	**GPU Mem.**	**Param.**	**ACC(%)**
LW AdaMerging	50	47.1s	1.6GB	87.5M	53.4 ±3.2
LoRA-WEMoE	50	72.9s	1.8GB	103.7M	68.8 ±7.8
WUDI-Merging	50	28.9s	2.0GB	87.5M	74.7 ± 6.6
NUFILT	25	9.3s	1.8GB	87.5M	83.3 ±0.3
	100	35.9s	1.8GB	87.5M	83.6 ±0.2
	50	18.4s	1.8GB	87.5M	83.6 ±0.2

6 Conclusion

In this work, we addressed the problem of data-free continual model merging, where independently fine-tuned models must be consolidated sequentially into a single backbone without access to task data or earlier checkpoints. The central challenge lies in enforcing transparency and fidelity. To this end, we established that task vectors approximately align with representation subspaces, providing a geometric foundation for continual merging. Building on this insight, we proposed NUFILT, a novel framework that integrates null-space filtering to suppress interference with prior knowledge and projection-aware adaptation to recover task-specific fidelity via data-free surrogate objectives. Experiments across vision and NLP benchmarks demonstrated that NUFILT achieves state-of-the-art results, improving over recent methods by 4–7% on average while substantially reducing forgetting and narrowing the gap to individual fine-tuning. We believe this work provides a principled and practical step toward scalable, theoretically grounded solutions for continual model merging.

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

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

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 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*, pp. 3606–3613, 2014.

Tarin Clanuwat, Marc Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2018.

Adam Coates, Andrew Y 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*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. *European Conference on Computer Vision*, 2020.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Congyu Fang, Adam Dziedzic, Lin Zhang, Laura Oliva, Amol Verma, Fahad Razak, Nicolas Papernot, and Bo Wang. Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data. *EBioMedicine*, 101, 2024.

Ian J Goodfellow, Dumitru Erhan, Pierre-Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, William 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*, pp. 117–124. Springer, 2013.

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

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 831–839, 2019.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic loRA composition. In *First Conference on Language Modeling*, 2024a.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024b.

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

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

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

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016.

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

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 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, 1998.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. *Advances in Neural Information Processing Systems*, 35: 16165–16177, 2022.

Tian Yu Liu and Stefano Soatto. Tangent model composition for ensembling and continual fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18676–18686, 2023.

Yaoyao Liu, Bernt Schiele, and Qianru Sun. Rmm: Reinforced memory management for class-incremental learning. *Advances in Neural Information Processing Systems*, 34:3478–3490, 2021.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision (ECCV)*, 2024.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations*, 2021.

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

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

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

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C V Jawahar. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A second-order perspective on model compositionality and incremental learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zihuan Qiu, Yi Xu, Fanman Meng, Hongliang Li, Linfeng Xu, and Qingbo Wu. Dual-consistency model inversion for non-exemplar class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24025–24035, 2024.

Zihuan Qiu, Yi Xu, Chiyuan He, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Mingle: Mixtures of null-space gated low-rank experts for test-time continual model merging. *arXiv preprint arXiv:2505.11883*, 2025.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2016.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew 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*, pp. 1631–1642, 2013.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

Anke Tang, Li Shen, Yong Luo, Han Hu, Bo Du, and Dacheng Tao. FusionBench: A Comprehensive Benchmark of Deep Model Fusion, June 2024a.

Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. Merging multi-task models via weight-ensembling mixture of experts. In *International Conference on Machine Learning*, pp. 47778–47799. PMLR, 2024b.

Anke Tang, Enneng Yang, Li Shen, Yong Luo, Han Hu, Bo Du, and Dacheng Tao. Merging models on the fly without retraining: A sequential approach to scalable continual model merging. *arXiv preprint arXiv:2501.09522*, 2025.

Joachim Utans. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models. AAAI Press*, pp. 133–138. Citeseer, 1996.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco S Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. *arXiv preprint arXiv:1806.03962*, 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. *arXiv preprint arXiv:1804.07461*, 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 *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Haitao Wen, Haoyang Cheng, Heqian Qiu, Lanxiao Wang, Lili Pan, and Hongliang Li. Optimizing mode connectivity for class incremental learning. In *International Conference on Machine Learning*, pp. 36940–36957. PMLR, 2023.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

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

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.

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

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024a.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024b.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *International conference on machine learning*, pp. 3987–3995, 2017.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564, 2024a.

Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. Metagpt: Merging large language models using model exclusive task arithmetic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1724, 2024b.

APPENDIX

The appendix is organized into sections, each providing supplementary explanations and supporting details.

A PROOF OF THEORETICAL RESULTS

A.1 PROOF OF THEOREM 1

Theorem 1 extends Proposition A.1 (Cheng et al.) from individual task vectors to subspaces. While Proposition A.1 demonstrates that each task vector can be approximated as a linear combination of representation features, Theorem 1 further establishes approximate alignment between the singular subspaces of task vectors and task representations.

Theorem A.1 (Weyl inequality for singular values (Weyl, 1912)). *For matrices $A \in \mathbb{R}^{m\times n}$ and $E \in \mathbb{R}^{m\times n}$, with singular values $\sigma_j(A)$ and $\sigma_j(A+E)$ ordered decreasingly ($\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$), and $\|E\|_2$ the operator norm of E, the following holds for $j = 1, 2, \ldots, \min(m, n)$:*

$$|\sigma_j(A+E) - \sigma_j(A)| \leq \|E\|_2. \tag{18}$$

Theorem A.2 (Wedin's Sin-Theta Theorem (Wedin, 1972)). *Let $M \in \mathbb{R}^{m\times n}$ with SVD $M = U\Sigma V^\top, U = [U_0 \quad U_\perp], V = [V_0 \quad V_\perp]$, where $U_0 \in \mathbb{R}^{m\times r}$ and $V_0 \in \mathbb{R}^{n\times r}$ correspond to the top-r singular values. Let $\hat{M} = M + H$ with SVD $\hat{M} = \hat{U}\hat{\Sigma}\hat{V}^\top, \hat{U} = \begin{bmatrix}\hat{U}_0 & \hat{U}_\perp\end{bmatrix}, \hat{V} = \begin{bmatrix}\hat{V}_0 & \hat{V}_\perp\end{bmatrix}$. Suppose there exist $a \in \mathbb{R}$ and $\Delta > 0$ such that $\sigma_r(M) \geq a, \sigma_{r+1}(\hat{M}) \leq a - \Delta$. Then the canonical angles between the singular subspaces satisfy*

$$\max\Big\{\mathrm{dist}(\hat{U}_0, U_0),\ \mathrm{dist}(\hat{V}_0, V_0)\Big\} \leq \frac{\max\{\|HV_0\|_2,\ \|H^\top U_0\|_2\}}{\Delta} \leq \frac{\|H\|_2}{\Delta}, \tag{19}$$

where $\mathrm{dist}(\hat{U}_0, U_0) = \|\sin\Theta(\hat{U}_0, U_0)\|_2$ denotes the largest principal angle between subspaces.

Proposition A.1 (Approximate Linear Combination (Cheng et al.)). *Let $\tau_k^{(l)}$ denote the task vector of neuron k in linear layer l. Consider N input samples $\{x_n\}_{n=1}^N$, and let $x_{n,t}^{(l)}$ denote the corresponding input to layer l after t iteration for sample x_n during fine-tuning. Assume that the model before layer l is $\mathcal{C}_l$-Lipschitz continuous, and the gradient of loss is bounded in ℓ_2-norm by $\mathcal{G}_l$, and the gradient of the loss with respect to the product $\theta_{k,t-1}^{(l)} x_{n,t}^{(l)}$ is bounded by $\Gamma_k^{(l)}$. Then, the following inequality holds:*

$$\left\| \tau_k^{(l)} - \sum_{n=1}^{N} \beta_{k,n}^{(l)} (x_{n,T}^{(l)})^\top \right\|_2 \leq \Phi_k \cdot \left(\sum_{t=1}^{T} \sum_{i=t}^{T} \eta_t \eta_i \right). \tag{20}$$

where $\Phi_k = N \cdot C_l \cdot \mathcal{G}_l \cdot \Gamma_k^{(l)}$ and $\beta_{k,n}^{(l)} = \sum_{t=1}^{T} -\eta_t \frac{\partial L(\theta_t)}{\partial(\theta_{k,t-1}^{(l)} x_{n,t}^{(l)})}$.

Theorem A.3 (Approximate Subspace Alignment). *Let $\tau^{(l)} \in \mathbb{R}^{d_o \times d_i}$ be the task vector at layer l, and $H \in \mathbb{R}^{N \times d_i}$ a representation matrix of rank r_d with right singular vectors $V_d \in \mathbb{R}^{d_i \times r_d}$. Suppose $\tau^{(l)}$ admits the decomposition $\tau^{(l)} = T_0 + E$ with $T_0 = BH$, $\operatorname{rank}(T_0) = r_d$, where each row E_k of E satisfies $\|E_k\|_2 \leq \Psi_k$. If $\sigma_{r_d}(T_0) > \sqrt{d_o} \max_k \Psi_k$, then for any $r_v \geq r_d$, the span of the top r_v right singular vectors $\hat{V}$ of $\tau^{(l)}$ is approximately aligned with $\operatorname{span}(V_d)$ in the sense that*

$$\mathcal{A}(V_d^{(l)}, \hat{V}^{(l)}) \;\leq\; \zeta^2, \tag{21}$$

where $\mathcal{A}(V_d^{(l)}, \hat{V}^{(l)}) = 1 - \frac{1}{r_d} \|\hat{V}^\top V_d\|_F^2$ and $\zeta^2 = \left(\frac{\sqrt{d_o} \max_k \Psi_k}{\sigma_{r_d}(T_0) - \sqrt{d_o} \max_k \Psi_k} \right)^2$.

Proof. The proof of Theorem 1 requires three ingredients. First, Proposition A.1 shows that task vectors can be approximated by linear combinations of representation features, yielding a decomposition into a structured term plus bounded error. Second, Weyl's inequality (Theorem A.1) provides a perturbation bound for singular values, which ensures a nontrivial spectral gap under bounded error. Finally, Wedin's sin–Theta theorem (Theorem A.2) translates this spectral gap into a guarantee of approximate alignment between the true and perturbed subspaces.

From Proposition A.1 (Approximate Linear Combination), each row $\tau_k^{(l)}$ of the task vector matrix satisfies

$$\|\tau_k^{(l)} - \sum_{n=1}^{N} \beta_{k,n}^{(l)} (x_{n,T}^{(l)})^\top \|_2 \leq \Psi_k, \tag{22}$$

where $\Psi_k = \Phi_k \cdot \left(\sum_{t=1}^{T} \sum_{i=t}^{T} \eta_t \eta_i \right)$ is a per-neuron error bound, and Φ_k depends on the number of samples, Lipschitz constant, gradient bounds, and learning rates. This yields the decomposition $\tau^{(l)} = T_0 + E$, with $T_0 = BH$, $B_{k,n} = \beta_{k,n}^{(l)}$, and $\|E_k\|_2 \leq \Psi_k$ for each row E_k, so $\|E\|_2 \leq \sqrt{d_o} \max_k \Psi_k$.

(1) Exact case ($E = 0$). In this case, $\tau^{(l)} = T_0 = BH$. Since H has rank r_d, the right singular space of H is exactly $\operatorname{span}(V_d)$. Under the rank assumption $\operatorname{rank}(T_0) = r_d$ (which holds when $d_o \geq r_d$ and B has sufficient rank, as motivated by the NTK regime), T_0 and H share the same r_d-dimensional right singular subspace. Hence $\operatorname{span}(\hat{V})$ and $\mathcal{A}(V_d, \hat{V}) = 1$.

(2) Perturbed case ($E \neq 0$). Apply Wedin's $\sin \Theta$ theorem A.2. Let V_0 be the top r_d right singular vectors of T_0, and set $a = \sigma_{r_d}(T_0)$. Define $\Delta = \sigma_{r_d}(T_0) - \|E\|_2$. The stability assumption $\sigma_{r_d}(T_0) > \sqrt{d_o} \max_k \Psi_k \geq \|E\|_2$ ensures $\Delta > 0$. With $\operatorname{rank}(T_0) = r_d$ and $\tau^{(l)} = T_0 + E$, Weyl's inequality A.1 gives $|\sigma_{r_d+1}(\tau^{(l)}) - \sigma_{r_d+1}(T_0)| \leq \|E\|_2$. Since $\sigma_{r_d+1}(T_0) = 0$, this yields $\sigma_{r_d+1}(\tau^{(l)}) \leq \|E\|_2 = a - \Delta$.

By Wedin's theorem, the canonical angles Θ between $\mathrm{span}(V_0)$ and $\mathrm{span}(\hat{V}_{r_d})$ satisfy

$$\|\sin\Theta(V_0, \hat{V}_{r_d})\|_2 \le \frac{\|E\|_2}{\Delta}. \tag{23}$$

We have $\mathrm{span}(V_0) = \mathrm{span}(V_d)$ and $\|E\|_2 \le \sqrt{d_o}\, \max_k \Psi_k$, hence

$$\|\sin\Theta(V_d, \hat{V}_{r_d})\|_2 \le \frac{\sqrt{d_o}\, \max_k \Psi_k}{\sigma_{r_d}(T_0) - \sqrt{d_o}\, \max_k \Psi_k}. \tag{24}$$

The subspace affinity then obeys

$$\mathcal{A}(V_d, \hat{V}_{r_d}) = \sqrt{\frac{1}{r_d}\sum_{j=1}^{r_d} \cos^2\theta_j} \;\ge\; \cos\theta_{\max} = \sqrt{1 - \sin^2\theta_{\max}}. \tag{25}$$

Using $1 - \sqrt{1-u} \le u$ for $u \in [0,1]$, we obtain

$$1 - \mathcal{A}(V_d, \tilde{V}_{r_d}) \le \sin^2\theta_{\max} \le \left(\frac{\sqrt{d_o}\, \max_k \Psi_k}{\sigma_{r_d}(T_0) - \sqrt{d_o}\, \max_k \Psi_k}\right)^2. \tag{26}$$

Finally, since enlarging the subspace only increases the overlap, for any $r_v \ge r_d$,

$$\mathcal{A}(V_d, \hat{V}) \ge \mathcal{A}(V_d, \hat{V}_{r_d}), \tag{27}$$

so the same bound holds. □

A.2 Proof of Corollary 1

Corollary A.1 (Data-free upper bound). *Let $X \in \mathbb{R}^{N \times d_i}$ with largest singular value $\sigma_1(X)$, rank r_d, and right singular vectors V_d. Let $\tau \in \mathbb{R}^{d_o \times d_i}$ with top-r_v right singular vectors $\hat{V}$ ($r_v \ge r_d$). If Theorem 1 ensures $1 - \frac{1}{r_d}\|\hat{V}^\top V_d\|_F^2 \;\le\; \zeta^2$, then for any $\rho \in \mathbb{R}^{d_o \times d_i}$,*

$$\big\|(\rho - \tau)X^\top\big\|_F^2 \;\le\; 2\,\sigma_1(X)^2\Big(\big\|(\rho-\tau)\hat{V}\big\|_F^2 + r_d\,\zeta^2\,\|\rho - \tau\|_2^2\Big). \tag{28}$$

Proof. Let $P = \hat{V}\hat{V}^\top$ be the projection onto $\mathrm{span}(\hat{V})$. Decompose

$$X^\top = PX^\top + (I-P)X^\top. \tag{29}$$

By $\|a+b\|_F^2 \le 2\|a\|_F^2 + 2\|b\|_F^2$,

$$\|(\rho-\tau)X^\top\|_F^2 \;\le\; 2\|(\rho-\tau)PX^\top\|_F^2 + 2\|(\rho-\tau)(I-P)X^\top\|_F^2. \tag{30}$$

(1) Projection term.

$$\begin{aligned}\|(\rho-\tau)PX^\top\|_F^2 &= \mathrm{trace}\big((\rho-\tau)PX^\top XP(\rho-\tau)^\top\big)\\ &= \mathrm{trace}\big(P(\rho-\tau)^\top(\rho-\tau)P\,X^\top X\big).\end{aligned}$$

Let $B = P(\rho-\tau)^\top(\rho-\tau)P \succeq 0$ and $C = X^\top X \succeq 0$. By von Neumann's trace inequality, $\mathrm{trace}(BC) \le \|C\|_2\,\mathrm{trace}(B)$. Since $\|C\|_2 = \|X^\top X\|_2 = \sigma_1^2(X)$ and

$$\mathrm{trace}(B) = \mathrm{trace}\big(P(\rho-\tau)^\top(\rho-\tau)P\big) = \|(\rho-\tau)P\|_F^2 = \|(\rho-\tau)\hat{V}\|_F^2, \tag{31}$$

we obtain

$$\|(\rho-\tau)PX^\top\|_F^2 \;\le\; \sigma_1^2(X)\,\|(\rho-\tau)\hat{V}\|_F^2. \tag{32}$$

(2) Residual term. By $\|AB\|_F \le \|A\|_2\|B\|_F$,

$$\|(\rho-\tau)(I-P)X^\top\|_F^2 \;\le\; \|\rho-\tau\|_2^2\,\|X(I-P)\|_F^2. \tag{33}$$

Using the SVD $X = U_d \Sigma_d V_d^\top$,

$$\|X(I-P)\|_F^2 = \operatorname{trace}\big(\Sigma_d^2 V_d^\top (I-P) V_d\big)\,. \tag{34}$$

Since $\Sigma_d^2 \preceq \sigma_1^2(X) I$,

$$\|X(I-P)\|_F^2 \;\le\; \sigma_1^2(X)\operatorname{trace}\big(V_d^\top (I-P) V_d\big)\,. \tag{35}$$

Noting that

$$\operatorname{trace}(V_d^\top (I-P) V_d) = r_d - \|V_d^\top \hat{V}\|_F^2 \;\le\; r_d \zeta^2, \tag{36}$$

we conclude

$$\|X(I-P)\|_F^2 \;\le\; \sigma_1^2(X)\, r_d \zeta^2. \tag{37}$$

Hence

$$\|(\rho-\tau)(I-P)X^\top\|_F^2 \;\le\; \sigma_1^2(X)\, r_d \zeta^2\, \|\rho-\tau\|_2^2. \tag{38}$$

(3) Combine. Summing the two parts yields

$$\|(\rho-\tau)X^\top\|_F^2 \;\le\; 2\sigma_1^2(X)\,\|(\rho-\tau)\hat{V}\|_F^2 + 2\sigma_1^2(X)\, r_d \zeta^2\, \|\rho-\tau\|_2^2, \tag{39}$$

as claimed. □

B ADDITIONAL DESCRIPTIONS

B.1 DETAILS OF DATASET AND TASK SETTINGS

Overview of Vision Tasks To comprehensively examine continual model merging, we curated a collection of 20 diverse image classification benchmarks, spanning natural objects, remote sensing, medical imagery, and text-rendered datasets. This selection largely follows prior practice in multi-domain evaluation (Tang et al., 2025), while ensuring balanced inclusion of datasets with varying granularity (from binary recognition to hundreds of categories). Specifically, the benchmarks include: SUN397 (Xiao et al., 2010), Stanford Cars (Krause et al., 2013), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), SVHN (Netzer et al., 2011), GTSRB (Stallkamp et al., 2012), MNIST (LeCun et al., 1998), DTD (Cimpoi et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), PCAM (Veeling et al., 2018), FER2013 (Goodfellow et al., 2013), Oxford-IIIT Pet (Parkhi et al., 2012), STL-10 (Coates et al., 2011), CIFAR-100 and CIFAR-10 (Krizhevsky & Hinton, 2009), Food-101 (Bossard et al., 2014), Fashion-MNIST (Xiao et al., 2017), EMNIST (Cohen et al., 2017), KMNIST (Clanuwat et al., 2018), and Rendered SST-2 (Socher et al., 2013).

Overview of NLP Tasks In addition to vision benchmarks, we also consider widely-used natural language understanding datasets. Specifically, we evaluate on the GLUE benchmark (Wang et al., 2018), which covers tasks ranging from single-sentence acceptability judgments to pairwise entailment and semantic similarity. For classification-oriented datasets, we report *exact match accuracy*, including CoLA (linguistic acceptability), MNLI (natural language inference), MRPC (paraphrase detection), QNLI (question–answer entailment), QQP (duplicate question detection), RTE (recognizing textual entailment), and SST-2 (sentiment classification). For STS-B, which measures semantic textual similarity, performance is reported using *Spearman's ρ correlation coefficient*.

Task Grouping We evaluate continual merging under three progressively larger task sets. For each setting, models are assessed using *average accuracy (ACC)* and *backward transfer (BWT)*. To ensure robustness, we generate 10 random task orders for every group (Tab. 5), reporting the mean and standard deviation across runs.

- **8-Task Group**: (1) SUN397, (2) Stanford Cars, (3) RESISC45, (4) EuroSAT, (5) SVHN, (6) GTSRB, (7) MNIST, (8) DTD.
- **14-Task Group**: Extends the 8-task set with (9) Flowers102, (10) PCAM, (11) FER2013, (12) Oxford-IIIT Pet, (13) STL-10, (14) CIFAR-100.
- **20-Task Group**: Builds upon the 14-task set by adding (15) CIFAR-10, (16) Food-101, (17) Fashion-MNIST, (18) EMNIST, (19) KMNIST, (20) Rendered SST-2.

For the NLP benchmarks, tasks are organized in **alphabetical order**: CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST-2, and STS-B.

Table 5: Dataset orderings used for vision experiments in each task group.

	Order	Dataset Order (by ID)
8 tasks	1	(04 → 05 → 07 → 08 → 03 → 06 → 01 → 02)
	2	(07 → 08 → 05 → 04 → 02 → 06 → 03 → 01)
	3	(03 → 06 → 04 → 02 → 01 → 08 → 05 → 07)
	4	(06 → 08 → 02 → 01 → 03 → 07 → 04 → 05)
	5	(07 → 06 → 03 → 08 → 05 → 01 → 04 → 02)
	6	(07 → 02 → 03 → 08 → 05 → 04 → 01 → 06)
	7	(07 → 01 → 04 → 03 → 08 → 05 → 02 → 06)
	8	(08 → 05 → 06 → 07 → 01 → 04 → 03 → 02)
	9	(01 → 04 → 05 → 02 → 06 → 03 → 07 → 08)
	10	(08 → 03 → 01 → 02 → 06 → 05 → 07 → 04)
14 tasks	1	(09 → 13 → 08 → 07 → 14 → 12 → 06 → 03 → 10 → 04 → 05 → 01 → 02 → 11)
	2	(09 → 10 → 11 → 14 → 07 → 13 → 04 → 02 → 06 → 08 → 03 → 12 → 05 → 01)
	3	(05 → 08 → 12 → 06 → 11 → 01 → 10 → 04 → 14 → 03 → 02 → 13 → 09 → 07)
	4	(03 → 10 → 09 → 12 → 04 → 13 → 01 → 06 → 11 → 02 → 14 → 08 → 07 → 05)
	5	(08 → 14 → 09 → 06 → 12 → 13 → 05 → 03 → 04 → 11 → 10 → 01 → 07 → 02)
	6	(03 → 12 → 13 → 01 → 11 → 04 → 10 → 05 → 14 → 08 → 09 → 07 → 02 → 06)
	7	(07 → 01 → 12 → 10 → 02 → 08 → 13 → 04 → 05 → 11 → 14 → 03 → 06 → 09)
	8	(05 → 12 → 04 → 11 → 03 → 08 → 10 → 01 → 09 → 13 → 14 → 07 → 06 → 02)
	9	(10 → 07 → 09 → 02 → 03 → 13 → 01 → 12 → 14 → 04 → 11 → 06 → 05 → 08)
	10	(01 → 02 → 11 → 06 → 08 → 12 → 07 → 05 → 10 → 14 → 03 → 13 → 09 → 04)
20 tasks	1	(20 → 06 → 15 → 05 → 10 → 14 → 16 → 19 → 07 → 13 → 18 → 11 → 02 → 12 → 03 → 17 → 08 → 09 → 01 → 04)
	2	(09 → 14 → 06 → 03 → 07 → 04 → 18 → 01 → 17 → 19 → 08 → 20 → 13 → 16 → 11 → 12 → 15 → 05 → 10 → 02)
	3	(09 → 15 → 16 → 11 → 03 → 13 → 08 → 10 → 12 → 02 → 20 → 01 → 05 → 19 → 07 → 06 → 04 → 18 → 17 → 14)
	4	(17 → 04 → 11 → 19 → 18 → 10 → 07 → 15 → 12 → 13 → 08 → 02 → 01 → 06 → 05 → 03 → 20 → 16 → 14 → 09)
	5	(14 → 16 → 04 → 20 → 15 → 17 → 07 → 11 → 06 → 18 → 12 → 01 → 19 → 09 → 10 → 05 → 08 → 02 → 13 → 03)
	6	(02 → 06 → 17 → 04 → 19 → 18 → 08 → 16 → 20 → 01 → 10 → 13 → 07 → 09 → 05 → 11 → 15 → 14 → 03 → 12)
	7	(19 → 01 → 09 → 14 → 06 → 20 → 17 → 04 → 08 → 02 → 15 → 03 → 16 → 13 → 12 → 07 → 10 → 05 → 11 → 18)
	8	(15 → 07 → 08 → 02 → 10 → 06 → 17 → 20 → 05 → 19 → 16 → 01 → 18 → 09 → 13 → 11 → 04 → 14 → 12 → 03)
	9	(10 → 05 → 07 → 11 → 01 → 03 → 17 → 15 → 18 → 04 → 14 → 19 → 02 → 06 → 13 → 20 → 08 → 12 → 09 → 16)
	10	(01 → 11 → 02 → 15 → 03 → 10 → 12 → 19 → 16 → 13 → 07 → 05 → 09 → 04 → 14 → 20 → 06 → 18 → 17 → 08)

B.2 Details of Downstream Models

In this section, we describe the evaluation protocol for both pre-trained and fine-tuned models across vision and natural language processing (NLP) downstream tasks.

For vision tasks (Tab. 6), we report the zero-shot performance of pre-trained CLIP-ViT models as well as the test accuracy of task-specific fine-tuned models. Fine-tuned checkpoints are obtained from Hugging Face (`https://huggingface.co/tanganke`), where each model is trained on its respective dataset using a standard recipe. During fine-tuning, the visual encoder is updated while the text encoder remains fixed. Training follows a consistent setup: cross-entropy loss, Adam optimizer, cosine annealing schedule, learning rate of 1×10^{-5}, batch size of 128, and 4000 training steps.

For NLP tasks, we adopt the 8-task GLUE benchmark using Flan-T5-base. As summarized in Tab. 7, we compare pre-trained and fine-tuned Flan-T5-base models across all GLUE tasks. Fine-tuning is conducted with learning rate 4×10^{-5}, batch size 16, and 2000 training steps for each task.

B.3 Details of Baselines

In addition to the methods presented in Sec. 3.2, we introduce the following additional baselines.

- **Ties-Merging** An extension of Task Arithmetic that alleviates parameter redundancy and sign conflicts during model merging (Yadav et al., 2023). For task t, we first compute the task-specific difference vector $\tau_t = \theta_t - \theta_0$, which is then trimmed and sign-normalized. The update is defined as $\tau_t^{\text{Ties}} = \text{Ties}\big(\tau_{t-1}^{\text{Ties}},\, \tau_t\big)$, and the merged model is updated by $\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \lambda \tau_t^{\text{Ties}}$.

Table 6: Performance of the CLIP pre-trained model and individually fine-tuned models on different vision downstream tasks.

	Model	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Flowers102	PCAM
Pre-trained	CLIP-ViT-B/32	63.2	59.6	60.3	45.0	31.6	32.5	48.3	44.2	66.4	60.6
	CLIP-ViT-B/16	65.5	64.7	66.4	54.1	52.0	43.5	51.7	45.0	71.3	54.0
	CLIP-ViT-L/14	68.2	77.9	71.3	61.2	58.4	50.5	76.3	55.5	79.2	51.2
Fine-tuned	CLIP-ViT-B/32	74.9	78.5	95.1	99.1	97.3	98.9	99.6	79.7	88.6	88.0
	CLIP-ViT-B/16	78.9	85.9	96.6	99.0	97.6	99.0	99.7	82.3	94.9	90.6
	CLIP-ViT-L/14	82.8	92.8	97.4	99.1	97.9	99.2	99.8	85.5	97.7	91.1
	Model	FER2013	OxfordIIITPet	STL10	CIFAR100	CIFAR10	Food101	FashionMNIST	EMNIST	KMNIST	RenderedSST2
Pre-trained	CLIP-ViT-B/32	41.3	83.3	97.1	63.7	89.8	82.4	63.0	12.0	10.0	58.6
	CLIP-ViT-B/16	46.4	88.4	98.3	66.3	90.8	87.0	67.3	12.4	11.2	60.6
	CLIP-ViT-L/14	50.0	93.2	99.4	75.1	95.6	91.2	67.0	12.3	9.7	68.9
Fine-tuned	CLIP-ViT-B/32	71.6	92.5	97.5	88.4	97.6	88.4	94.7	95.6	98.2	71.3
	CLIP-ViT-B/16	72.8	94.5	98.2	88.8	98.3	91.9	94.5	95.3	98.1	75.7
	CLIP-ViT-L/14	75.9	95.7	99.2	93.0	99.1	94.8	95.3	95.4	98.3	80.5

Table 7: Performance of pre-trained and fine-tuned Flan-T5-base models on the 8-task NLP GLUE benchmark.

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB
Flan-T5-base (Pre-trained)	69.1	56.5	76.2	88.4	82.1	80.1	91.2	62.2
Flan-T5-base (Fine-tuned)	75.0	83.4	87.5	91.5	85.4	85.9	93.6	88.7

- **Maximum Magnitude Selection (MAGMAX).** An extension of Task Arithmetic that selects, for each parameter dimension, the update with the larger absolute value (Marczak et al., 2024). Formally, $\tau_t^{\text{MagMax}} = \text{MagMax}(\tau_{t-1}^{\text{MagMax}}, \tau_t)$, and the merged model is updated as $\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \lambda\tau_t^{\text{MagMax}}$. In our experiments, we apply MAGMAX to merge individually fine-tuned models, denoted as MAGMAX-IND.
- **Weight-Ensembling MoE (WEMOE).** A mixture-of-experts based merging strategy (Tang et al., 2024b), where task-specific MLP layers serve as experts and are aggregated via a gating function. In the continual merging setting, however, WEMOE does not converge: the expert MLPs introduce excessive parameters, which makes learning the gating function unreliable with only a few unlabeled test samples. To mitigate this issue, we compress the MLP experts using LoRA, yielding the variant LORA-WEMOE. For test-time adaptation, we fine-tune the gating module using 5 randomly sampled instances per class from the test set of each new task.
- **WUDI-Merging.** A data-free merging method that reduces task interference by enforcing orthogonality between task vectors and their residual components (Cheng et al.). In the continual setting, each linear layer l maintains two task vectors: the cumulative merged vector from previous tasks $\tau_{t-1}^{\text{merged},(l)} = \theta_{t-1}^{\text{merged},(l)} - \theta_0^{(l)}$ and the current task vector $\tau_t^{(l)} = \theta_t^{(l)} - \theta_0^{(l)}$. The merged vector $\tau_m^{(l)}$ is obtained by optimizing $\mathcal{L} = \sum_i \frac{1}{\|\tau_{i,l}\|_F^2} \left\|(\tau_{m,l} - \tau_{i,l})(\tau_{i,l})^\top\right\|_F^2$. The update follows gradient descent, and the final merged parameters are $\theta_t^{\text{merged}} = \theta_0 + \tau_m$.

Details of Baseline Hyper-parameters. As summarized in Tab. 8, we report the key hyperparameter settings for all baseline methods and task configurations. Notably, our approach employs a *single*

fixed configuration applied uniformly across models (CLIP, Flan-T5), task scales (8, 14, 20), and all 10 task orders. This design highlights the robustness and generality of our method, ensuring that performance improvements do not stem from task-specific hyperparameter tuning.

Table 8: Hyperparameter settings for all baselines across different task configurations.

Method	Tasks	Scale Factor (λ)	α	Top-k (%)	LR	Steps	r_p	r_l	r_v
TASK ARITHMETIC	8	0.3	-	-	-	-	-	-	-
	14/20	0.1	-	-	-	-	-	-	-
TIES-MERGING	8	0.3	-	20	-	-	-	-	-
	14/20	0.1	-	20	-	-	-	-	-
MAGMAX-IND	8/14/20	0.5	-	-	-	-	-	-	-
LW. ADAMERGING	8/14/20	0.3	-	-	1e-4	50	-	-	-
LORA-WEMOE	8/14/20	0.3	-	-	1e-4	50	-	64	-
WUDI-MERGING	8/14/20	-	-	-	1e-5	50	-	-	-
OPCM	8/14/20	-	0.5	-	-	-	-	-	-
NUFILT (Ours)	8/14/20	-	-	-	1e-3	50	128	64	8

C Additional Results

In this section, we provide additional experimental results to support the findings reported in the main paper. Specifically, we include: (1) detailed overall performance results (C.1); (2) extended visualizations on subspace alignment (C.2).

C.1 Detailed Overall Performance Results

Tab. 9 expands on the average results in Tab. 1 by reporting per-task average accuracy after continually merging 20 tasks. We compare six methods, SWA, Task Arithmetic, Ties-Merging, MagMax-IND, OPCM, and our proposed NUFILT across three CLIP-ViT backbones (B/32, B/16, L/14). NUFILT achieves the highest accuracy on most tasks. These fine-grained results reinforce the main paper's findings, highlighting NUFILT's ability to improve performance on continual model merging.

C.2 Extended Visualizations on Subspace Alignment

This section provides extended subspace alignment visualizations for three additional models (ViT-B/32, ViT-L/14, Flan-T5-base; see Fig. 5, 6, and 7) to further validate the universal alignment of task vectors with task-relevant representations across architectures and layers. All visualizations follow a unified 3×3 grid structure where columns represent model components (left: full model; middle: attention layers; right: MLP layers) and rows show visualization types (top: layer-wise mean affinity heatmaps; middle: layer-wise 90th percentile affinity heatmaps; bottom: layer-wise ECDF curves). The layer-wise ECDFs reveal consistent spectral overlaps: for ViT models (Fig. 5 and 6), high (MNIST Data →MNIST Vector), moderate (EuroSAT Data →RESISC45 Vector), and low (SVHN Data →DTD Vector); for Flan-T5-base (Fig. 7), high (RTE Data →RTE Vector), moderate (RTE Data →QNLI Vector), and low (RTE Data →SST-2 Vector). These results across both vision and NLP domains confirm the consistent alignment phenomenon originally observed in Fig. 2, with all affinity heatmaps showing strong diagonal dominance where matched pairs exhibit significantly higher affinity than mismatched ones.

Table 9: Test set accuracy comparisons on different downstream tasks.

	Model	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Flowers102	PCAM
ViT-B/32	C. FINE-TUNED	53.9	38.2	64.7	98.7	45.4	34.4	86.7	58.4	57.5	67.7
	WA	64.2	59.6	64.8	60.9	47.3	43.1	71.8	46.4	66.5	63.9
	C.TA	62.0	53.7	60.9	58.1	48.5	48.9	79.4	46.1	61.1	73.4
	C.TIES	62.5	49.1	55.8	50.9	54.6	49.3	82.0	46.7	58.5	69.9
	MAGMAX-IND	63.6	53.1	59.7	49.1	53.8	53.1	79.8	43.2	56.9	75.1
	LW ADAMERGING	63.1	60.0	63.5	60.1	35.6	32.1	51.8	45.4	66.6	60.2
	LORA-WEMOE	51.4	45.8	63.3	43.5	42.9	34.6	58.9	46.5	47.5	60.1
	WUDI-MERGING	59.4	51.6	63.8	63.2	63.6	61.9	87.7	48.1	53.5	72.1
	OPCM	64.4	51.1	66.0	71.7	66.1	56.0	90.2	40.4	64.9	80.2
	NUFILT (Ours)	60.3	53.6	67.0	76.6	87.4	85.6	98.1	51.8	55.5	82.4
ViT-B/16	C. FINE-TUNED	62.7	58.0	67.6	99.1	46.0	29.2	93.9	61.9	64.1	75.2
	WA	67.1	64.6	69.3	63.4	62.4	52.7	80.7	46.6	71.8	63.1
	TA	65.8	57.5	63.8	59.5	64.7	54.0	88.0	45.3	67.5	67.1
	TIES	64.2	52.9	60.9	53.0	62.8	48.8	88.4	45.0	61.3	68.5
	MAGMAX-IND	65.8	51.8	57.8	42.6	54.4	43.7	83.0	42.8	60.4	69.8
	LW ADAMERGING	65.5	65.7	69.8	59.4	50.1	44.2	61.1	47.1	71.8	57.9
	LORA-WEMOE	62.7	60.2	69.4	37.7	52.1	39.9	63.1	45.3	64.3	51.7
	WUDI-MERGING	64.6	55.1	70.1	65.3	69.1	60.4	90.0	48.4	68.8	79.3
	OPCM	67.9	55.9	73.7	77.5	74.4	63.2	94.1	49.2	72.3	79.6
	NUFILT (Ours)	64.9	54.5	77.8	83.7	88.9	83.7	98.1	51.8	70.0	86.1
ViT-L/14	C. FINE-TUNED	69.5	73.6	78.3	99.2	59.3	49.3	98.6	69.7	83.2	78.3
	WA	70.7	77.7	76.4	75.3	69.5	62.1	93.7	57.7	80.0	73.6
	TA	70.4	74.1	73.9	66.3	69.9	65.6	95.1	56.6	78.6	70.4
	TIES	69.7	70.3	65.3	47.9	76.1	63.6	94.7	54.4	77.9	72.3
	MAGMAX-IND	73.1	73.7	75.6	64.6	73.7	68.8	94.6	56.1	78.0	71.7
	LW ADAMERGING	68.8	78.6	75.9	65.7	58.3	51.6	79.9	57.4	80.6	52.4
	LORA-WEMOE	62.1	68.1	68.7	53.2	47.5	49.4	69.8	49.1	66.2	54.2
	WUDI-MERGING	74.1	88.0	83.8	77.0	78.2	79.7	95.8	63.7	91.4	80.1
	OPCM	73.1	78.3	82.4	80.2	80.8	80.4	97.4	61.6	84.8	76.3
	NUFILT (Ours)	74.8	82.3	87.6	90.1	93.6	94.4	98.6	66.2	94.5	86.4

	Model	FER2013	OxfordIIITPet	STL10	CIFAR100	CIFAR10	Food101	FashionMNIST	EMNIST	KMNIST	RenderedSST2
ViT-B/32	C. FINE-TUNED	58.3	68.5	86.7	40.2	70.5	50.0	90.7	72.4	54.5	54.5
	WA	50.2	84.1	97.0	69.8	92.7	80.4	71.3	15.0	11.5	61.8
	TA	51.4	82.3	94.9	64.6	91.4	71.9	73.9	17.8	12.2	59.9
	TIES	49.5	81.3	95.2	63.7	91.2	70.2	73.7	17.8	16.9	59.8
	MAGMAX-IND	56.5	79.9	94.6	68.7	91.9	73.8	74.3	18.3	15.4	63.9
	LW ADAMERGING	43.2	83.7	96.8	67.0	89.9	81.6	63.7	16.8	10.7	59.1
	LORA-WEMOE	44.6	72.5	86.1	40.1	63.8	63.8	48.1	10.3	12.8	55.7
	WUDI-MERGING	60.7	80.3	93.1	60.9	85.4	62.8	76.1	38.2	21.9	70.2
	OPCM	55.8	82.9	95.9	67.6	92.8	74.0	76.3	22.4	18.3	64.6
	NUFILT (Ours)	62.5	81.6	95.1	63.1	91.1	66.0	85.4	42.6	43.5	71.6
ViT-B/16	C. FINE-TUNED	60.5	84.5	90.5	38.8	73.6	61.9	89.7	83.3	51.5	72.8
	WA	50.9	89.6	98.0	72.9	94.2	85.9	73.3	15.6	12.4	62.5
	TA	50.7	89.3	97.0	68.0	93.1	80.3	75.7	18.1	16.7	61.8
	TIES	50.4	87.9	96.3	63.1	91.7	78.0	75.0	23.4	24.9	61.5
	MAGMAX-IND	57.7	88.8	97.5	71.5	94.4	81.3	77.2	24.5	25.0	59.4
	LW ADAMERGING	46.8	88.9	98.1	69.2	91.4	86.6	67.2	17.2	11.0	59.2
	LORA-WEMOE	45.6	91.2	92.3	41.3	64.3	78.1	48.0	23.5	16.6	52.7
	WUDI-MERGING	64.7	91.5	95.9	67.5	90.2	78.3	81.4	50.8	30.5	70.0
	OPCM	59.5	91.8	97.7	73.2	94.7	83.1	81.3	26.5	23.4	66.8
	NUFILT (Ours)	66.0	92.3	97.0	70.5	94.0	80.8	88.3	69.9	71.0	72.1
ViT-L/14	C. FINE-TUNED	68.0	92.1	94.5	60.5	85.7	74.8	93.1	89.0	59.2	78.8
	WA	52.7	94.2	99.2	81.7	97.0	90.7	77.4	16.1	10.4	66.1
	TA	55.7	94.2	98.6	79.1	91.6	87.6	80.8	17.6	10.6	63.6
	TIES	57.6	93.5	97.8	74.0	95.6	84.7	79.7	20.2	12.6	58.4
	MAGMAX-IND	52.9	93.9	98.7	82.1	97.3	89.5	81.6	19.2	11.1	68.4
	LW ADAMERGING	49.2	93.5	99.3	77.2	95.8	91.1	68.2	18.6	9.8	66.6
	LORA-WEMOE	46.3	84.5	87.6	52.1	70.5	73.3	50.0	18.7	10.9	56.5
	WUDI-MERGING	66.2	95.6	98.6	79.5	95.5	99.1	84.0	46.1	23.9	78.3
	OPCM	61.8	95.4	99.2	83.0	97.8	90.9	86.0	26.4	14.7	71.0
	NUFILT (Ours)	67.0	95.8	98.8	79.6	96.5	90.4	91.9	51.0	74.3	77.2

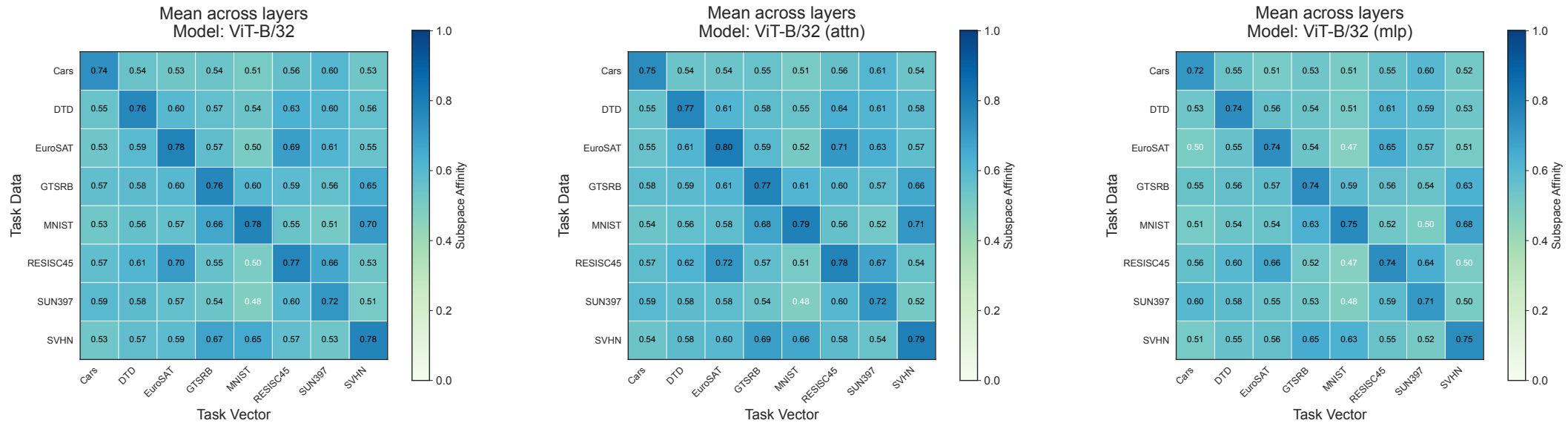

(a) Mean heatmaps: full model (left), attention layers (middle), and MLP layers (right).

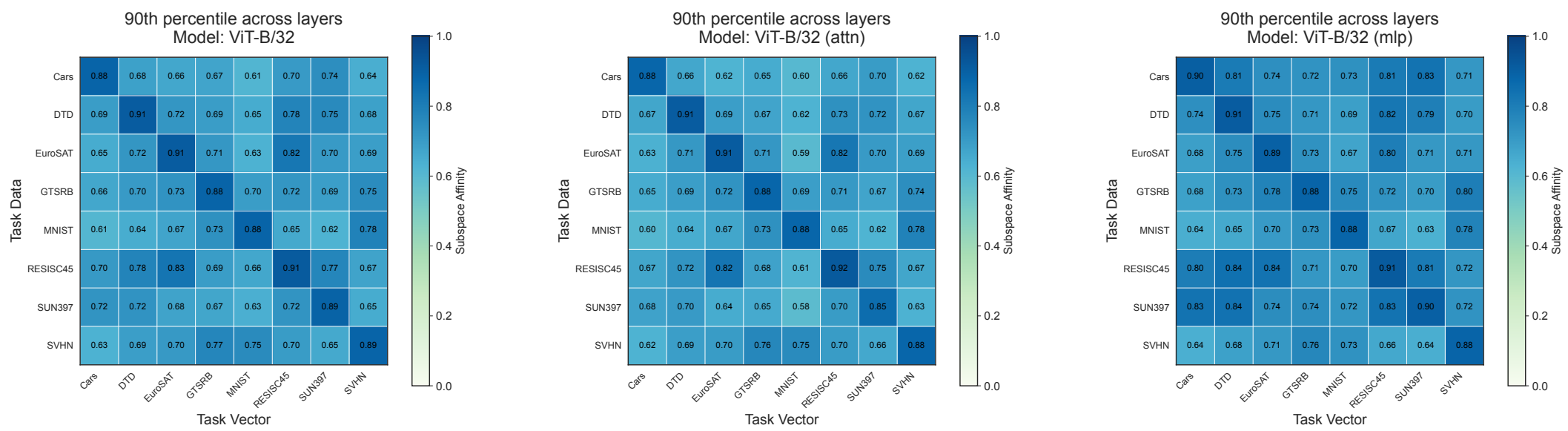

(b) 90th percentile heatmaps: full model (left), attention layers (middle), and MLP layers (right).

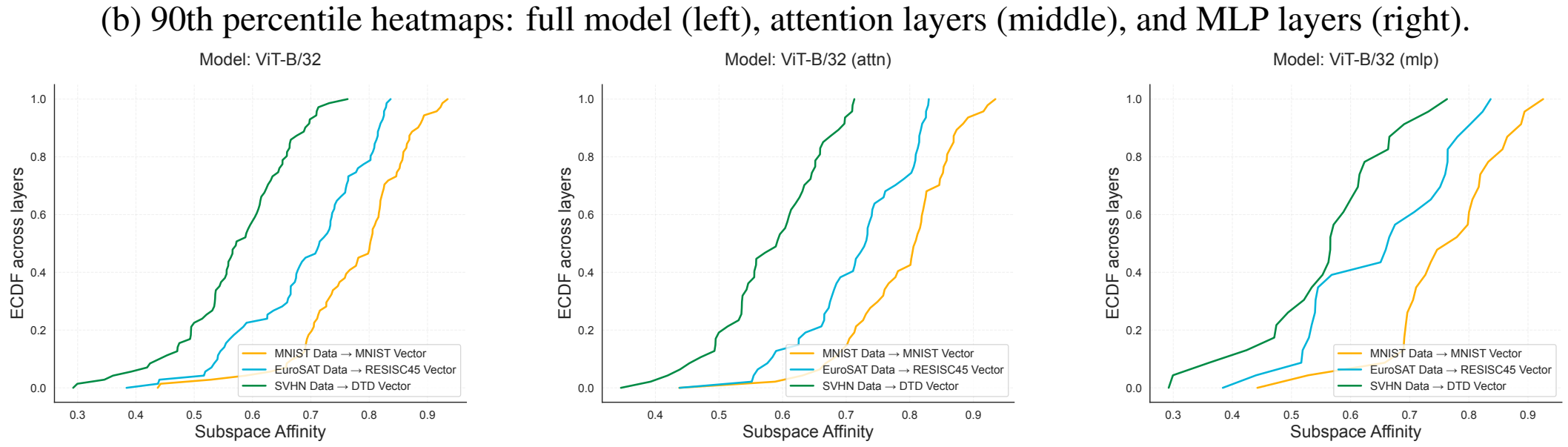

(c) ECDF distributions: full model (left), attention layers (middle), and MLP layers (right).

Figure 5: Subspace affinity between data and task vectors in ViT-B/32 across eight datasets.

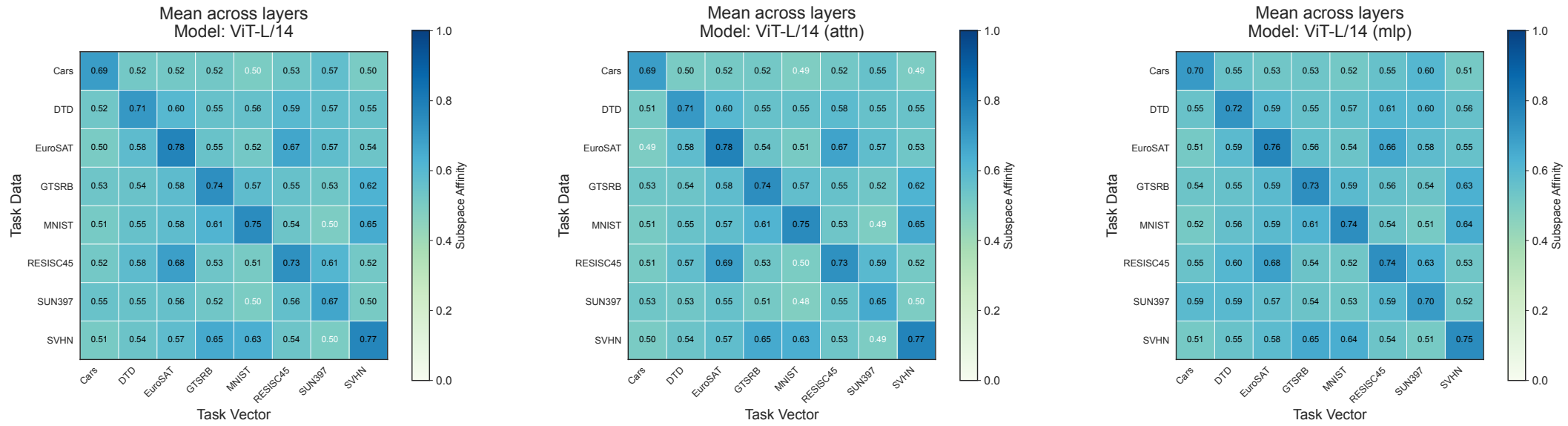

(a) Mean heatmaps: full model (left), attention layers (middle), and MLP layers (right).

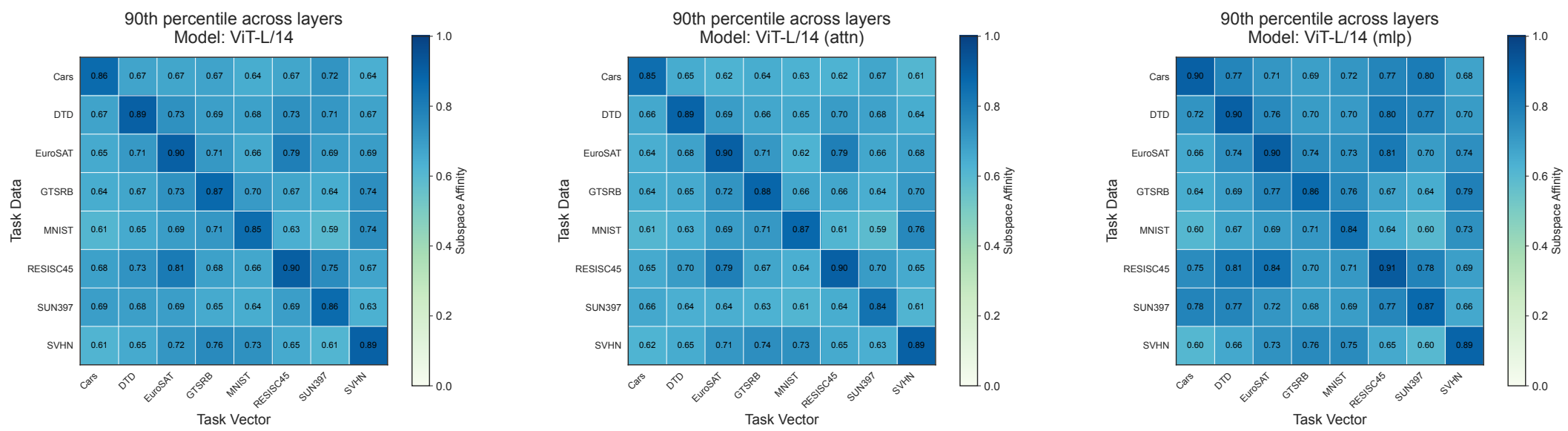

(b) 90th percentile heatmaps: full model (left), attention layers (middle), and MLP layers (right).

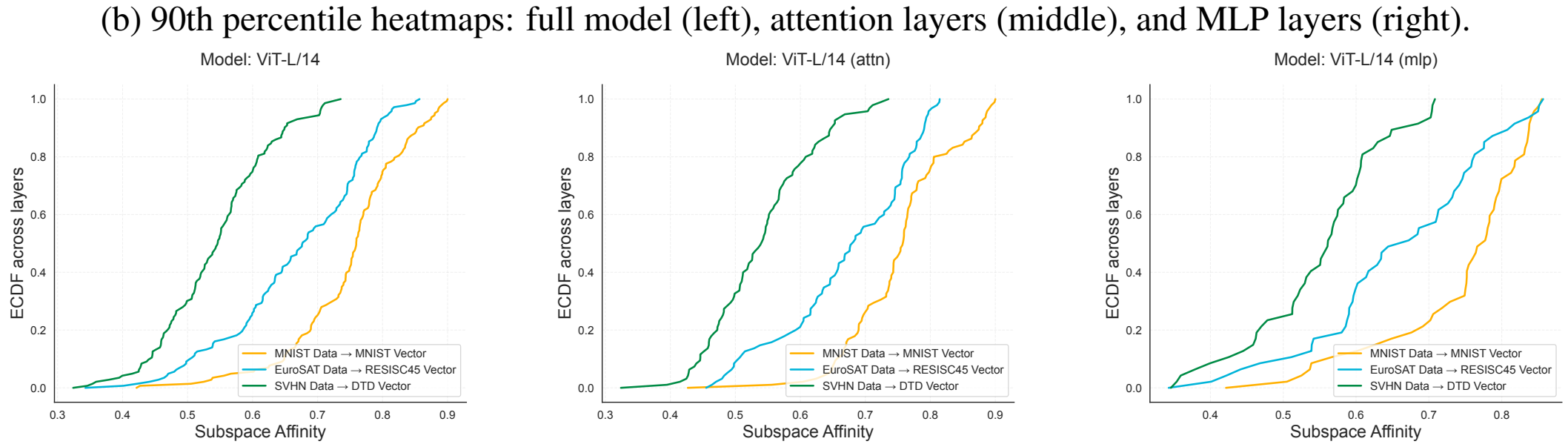

(c) ECDF distributions: full model (left), attention layers (middle), and MLP layers (right).

Figure 6: Subspace affinity between data and task vectors in ViT-L/14 across eight datasets.

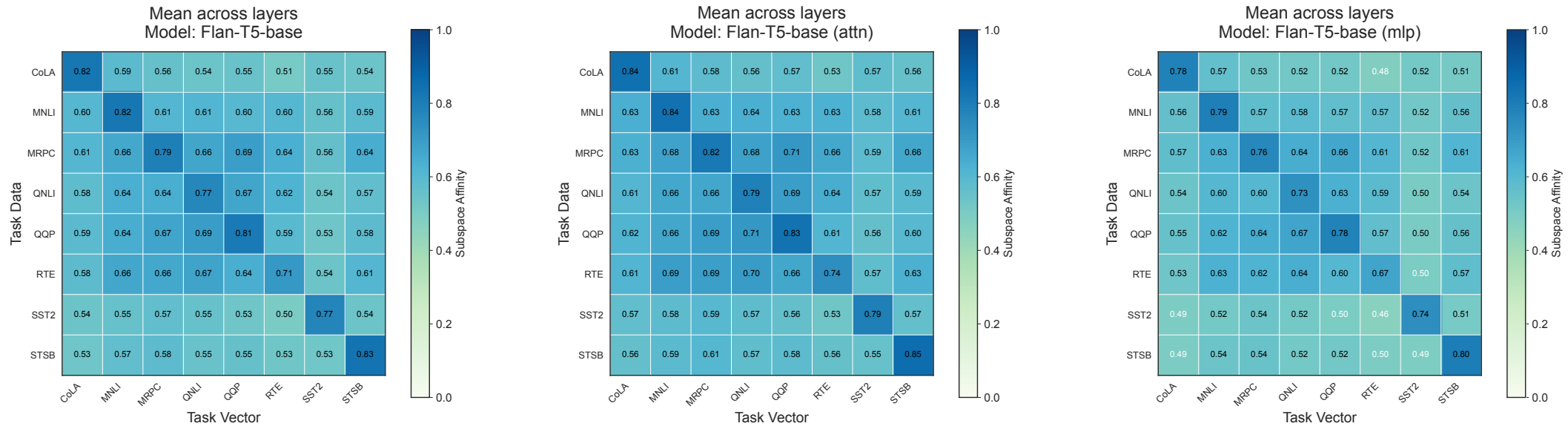

(a) Mean heatmaps: full model (left), attention layers (middle), and MLP layers (right).

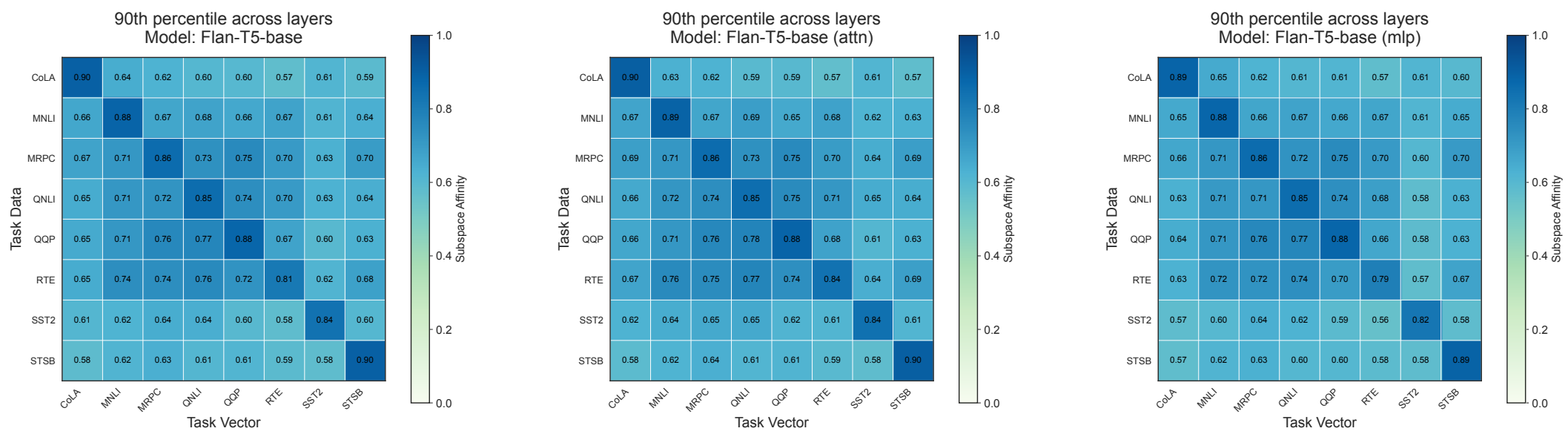

(b) 90th percentile heatmaps: full model (left), attention layers (middle), and MLP layers (right).

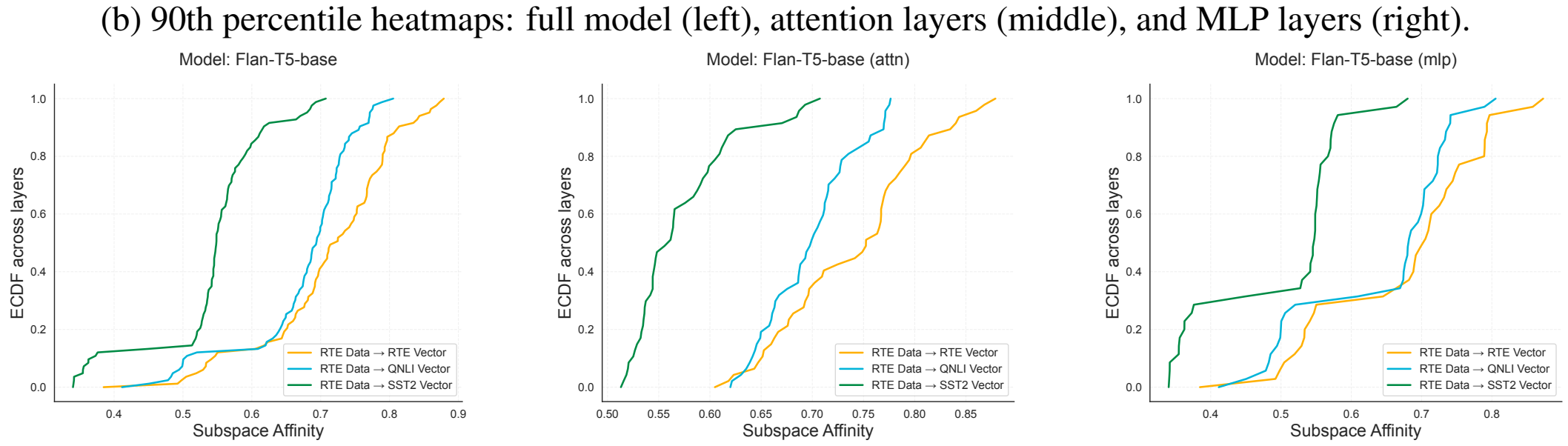

(c) ECDF distributions: full model (left), attention layers (middle), and MLP layers (right).

Figure 7: Subspace affinity between data and task vectors in Flan-T5-base across eight datasets.

D Discussions

D.1 Limitations

Similar to prior approaches in model merging, our NUFILT framework is built on the assumption that all task-specific fine-tuned models $\{\theta_t\}_{t=1}^{T}$ originate from a common pre-trained initialization θ_0. The implications of this assumption—such as its effect on the alignment of task vectors in the underlying subspace—remain insufficiently explored and merit further investigation. Our current experiments are restricted to models sharing the same backbone (*e.g.*, CLIP ViT variants, Flan-T5-base); extending the framework to heterogeneous initializations or architectures represents an interesting avenue for future work. Moreover, due to computational limitations, our study is confined to models with fewer than one billion parameters.

D.2 Broader Impacts

This paper aims to advance the Machine Learning field. Our work has potential societal impacts, but none require specific highlighting here.

D.3 LLM Usage

In preparing this submission, we used large language models (LLMs) solely as an assistive tool for sentence-level editing, including grammar correction, spelling adjustments, and minor word-choice refinements. The LLM was not involved in research ideation, methodological design, experimental analysis, or content generation beyond language editing. All substantive scientific contributions are solely those of the authors.