LoRA-DV: Spectral Rethinking for Reducing Task Interference via Difference Vector in Model Merging

Anonymous ACL submission

Abstract

Model merging integrates multi-task capabilities into Large Language Models without retraining, with spectral space offering better disentanglement of task interference than traditional parameter-space methods. However, current spectral approaches rely on static and one-off operations like truncation, which lack the granularity to resolve residual feature conflicts, resulting in a suboptimal merged model with unsatisfied multi-task performances. To bridge this gap, we propose LoRA-DV, a universal post-hoc framework that continuously refines merged weights via a Spectral Rethinking Mechanism. By employing iterative anisotropic scaling to modulate Difference Vectors (DVs)—defined as parameter displacements from the pre-trained state that encapsulate historical optimization knowledge—LoRA-DV acts as a high-precision spectral equalizer to suppress noise and amplify task signals with minimal learnable parameters. Experiments show that LoRA-DV significantly enhances existing baselines, effectively reducing spectral interference and boosting multi-task performance through fine-grained calibration.

1 Introduction

Large Language Models (LLMs) have achieved remarkable proficiency across a broad spectrum of natural language processing tasks. The accessibility of high-performance open-source foundations has catalyzed the proliferation of specialized expert models, ranging from automated code generation to clinical medical analysis. This shift from monolithic general-purpose systems toward a diverse ecosystem of domain-centric intelligence represents a significant paradigm shift. This naturally motivates the question of whether multiple task-specific expert LLMs can be unified into a single general-purpose model with multi-task capabilities (Ilharco et al., 2023; Yadav et al., 2023).

Currently, model merging has emerged as an efficient paradigm for integrating multiple specialized models into a unified framework without the prohibitive cost of joint training. A cornerstone of this field is Task Arithmetic (TA) (Ilharco et al., 2023), which defines a Task Vector as the element-wise difference between a fine-tuned model and its pre-trained base. By linearly combining these vectors, TA enables highly efficient capability transfer. Building on the efficacy of TA, more advanced methodologies have explored sophisticated strategies for learning task vector coefficients to improve multi-task performance. These include one-off learning frameworks, such as Ties-Merging (Yadav et al., 2023), Model Breadcrumbs (Davari and Belilovsky, 2024), and Consensus Merging (Wang et al., 2024), as well as inference-time dynamic learning exemplified by Twin-Merging (Lu et al., 2024) and EMR-Merging (Huang et al., 2024).

Traditional methods have predominantly operated in the parameter space. However, recent evidence suggests that traditional algorithms systematically exacerbate task interference during merging (Yadav et al., 2023; Yang et al., 2024b), causing a suboptimal trade-off among tasks where gains in one area come at the expense of others. This phenomenon is primarily caused by treating neural networks as flattened parameter vectors, thereby overlooking the critical algebraic structure inherent within weight matrices. Consequently, applying TA methods to such flattened updates frequently precipitates severe task interference, as it fails to account for the complex geometric interactions between the subspaces of task-specific weights.

To resolve this issue, current studies have transitioned the merging paradigm toward the spectral domain (Stoica et al., 2025; Gargiulo et al., 2025; Marczak et al., 2025). Singular Value Decomposition (SVD) offers a clear physical interpretation by decomposing weights into singular vectors and values, representing knowledge directions and their

corresponding importance. Through this geometrically informed lens, spectral methodologies can explicitly quantify the scaling of task vectors along critical parameter directions, facilitating a more granular analysis of inter-task interactions during the merging process.

Despite their potential, current spectral methodologies still face limitations. Most rely on static, one-off operations during initial aggregation, such as heuristic rank truncation (Gargiulo et al., 2025) or isotropic alignment (Marczak et al., 2025). These strategies lack sufficient granularity; they fail to adaptively calibrate task weights in the spectral space, leaving residual feature conflicts unresolved. Consequently, the merged model fails to achieve an optimal trade-off given the historical knowledge already acquired. Specifically, static strategies lack the capacity to adaptively balance task-specific representations, often leading to suboptimal compromises where dominant tasks suppress minority or conflicting ones. This rigidity prevents the model from searching for solutions that minimize interference, limiting its ability to exploit shared structures while preserving task-specific nuances.

Building on this, we argue that effective model merging requires a spectral rethinking mechanism. This mechanism leverages Difference Vectors (DVs) (Wang et al., 2025)—which encode historical optimization knowledge—to perform post-hoc reflection in the spectral space. Such reflection enables dynamic adjustments that discover better task trade-offs and reduce interference adaptively. To make this practical at the LLM scale, we adopt LoRA (Hu et al., 2022) as the problem-aligned parameterization. By exploiting the inherent low-rank structure of LoRA, iterative spectral analysis becomes computationally efficient, allowing us to learn a minimal set of parameters to suppress inter-task interference.

Thus, in this paper, we introduce LoRA-DV, a universal correction framework designed to dynamically modulate the spectral distribution captured in DVs. Rather than introducing new task knowledge, LoRA-DV recalibrates the existing spectral distribution to correct imbalanced or conflicting modes that emerge after merging. Functioning as a high-precision spectral equalizer, LoRA-DV employs iterative SVD-based anisotropic scaling. It introduces a negligible number of learnable parameters—approximately 0.1% of the original LoRA parameter count—to selectively suppress noisy modes and amplify constructive signals under sparse data guidance. Through extensive experiments on the 8-task GLUE benchmark and diverse generative scenarios, we demonstrate that LoRA-DV significantly enhances existing merging baselines. Our results show consistent performance gains across models ranging from Flan-T5 to Qwen-14B, effectively reducing task interference.

2 Related Work

2.1 Model Merging

Model merging is an established paradigm for endowing pre-trained models with multi-task capabilities without the prohibitive costs of joint training . Early approaches, such as Model Soups (Wortsman et al., 2022), demonstrated that averaging the weights of multiple fine-tuned models can enhance generalization. Building on this, Task Arithmetic (TA) (Ilharco et al., 2023) introduced task vectors—defined as the element-wise difference between fine-tuned and pre-trained weights—enabling the efficient addition or removal of specific capabilities. However, naive weight averaging often induces significant task interference.

To mitigate this, subsequent research has focused on refining the merging process within the parameter space. Ties-Merging (Yadav et al., 2023) resolves interference by trimming redundant parameters and aligning sign conflicts. Similarly, DARE (Yu et al., 2024) employs random dropouts to sparsify dense task vectors, thereby reducing inter-task overlap. Other methodologies, such as TALL-Mask and Consensus Merging (Wang et al., 2024), utilize task-specific masks to localize and retain only critical information. While effective, these methods predominantly operate in a flattened parameter space. They often neglect the underlying geometric structures and spectral properties of weight matrices, which we argue are essential for fine-grained interference resolution. LoRA-DV bridges this gap by leveraging these spectral attributes to resolve task interference.

2.2 SVD for Model Merging

The structural insights provided by Singular Value Decomposition (SVD) have recently been applied to model merging. KnOTS (Stoica et al., 2025) utilizes SVD by concatenating task-specific LoRA matrices and averaging their right-singular vectors to reconstruct the final merged weights. Similarly, TSV (Gargiulo et al., 2025) analyzes the interactions between the singular vectors of different tasks

to quantify interference and guide the merging process. Furthermore, Isotropic Merging (Iso) (Marczak et al., 2025) demonstrates that flattening the singular value spectrum can significantly enhance the alignment between merged and task-specific subspaces.

While sharing the motivation for spectral analysis, we contend that these prior methods rely on static operations that lack the granularity to resolve residual feature conflicts. In contrast, LoRA-DV dynamically modulates the spectral distribution to effectively reduce task interference and optimize multi-task performance.

3 Preliminaries

Low-Rank Adaptation (LoRA) Low-Rank Adaptation (LoRA) (Hu et al., 2022) is a widely used parameter-efficient finetuning paradigm that freezes the pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$and learns a low-rank parameter updated ΔW via a rank decomposition. The LoRA mechanism can be described as follows:

$$W = W_0 + \Delta W,\ \Delta W = BA, \tag{1}$$

where

$$B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k). \tag{2}$$

With only A, B are trainable, LoRA significantly reduces the trainable parameters and lower the hardware demand while maintaining comparable or better model performance.

From Task Vectors to Difference Vectors Task Vector (TV) (Ilharco et al., 2023) represents the element-wise difference between the model parameters of a pre-trained model and a finetuned model on a specific target. For task t, given pre-trained model weights as θ_{pre}, and the fine-tuned model weights θ^t_{ft}, the task vector is:

$$\tau_t = \theta^t_{\text{ft}} - \theta_{\text{pre}}. \tag{3}$$

Through simple vector arithmetic(addition, negation), task arithmetic enables effective model merging with high efficiency.

Difference Vectors (DVs) (Wang et al., 2025), as the extension of the TVs can be defined as the displacement from the pre-trained point at any optimization state. For an arbitrary parameter state $\hat{\theta}$, the DV can be calculated as:

$$\delta = \hat{\theta} - \theta_{\text{pre}}. \tag{4}$$

As the cumulative outcome of optimization, DVs encode the historical information from pre-training to the current point. With perturbing along the direction of the DVs, model weights are able to jump out of the local optima and keep searching without any extra components.

4 Methodology

We argue that spectral merging should not end at the initial decomposition. Thus, treating the merge as a one-off terminal operation is insufficient that spectral magnitudes can remain uncalibrated, leaving residual conflicts and producing suboptimal multi-task synergy. This motivates our central mechanism of spectral rethinking which rectifies the merged weight distribution after the initial model merging state.

4.1 Rethinking Mechanism

Our rethinking mechanism aims to adjust the parameter distribution of a merged model, however, due to the large scale of LLM tasks, the rethinking mechanism requires a framework that can efficiently adapt and regulate the distribution of model parameters. Therefore, we anchor the mechanism on LoRA difference vectors which is the natural carrier in the PEFT regime. Meanwhile its intrinsic low-rank form renders iterative spectral analysis computationally practical.

Formally, let θ_{pre} be the pre-trained model parameters and $\{\tau_i\}_{i=1}^N$ as task vectors form N LoRA modules where each LoRA task vector is $\tau_i = B_iA_i$. Thus, the standard initial merged state is:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \sum_{i=1}^{N} \alpha_i\tau_i. \tag{5}$$

Here, α_i is an isotropic scaling term. Most existing approaches simply tune α_i while ignoring the internal task interference. To dynamically adjust the weight distribution to suppress the task interference, we leverage the difference vectors of the merged model where:

$$\delta = \theta_{\text{merged}} - \theta_{\text{pre}}. \tag{6}$$

Since the difference vector δ contains the historical knowledge model learned form the pretrained to the merged state. Thus, we can realize our spectral rethinking mechanism through anisotropic scaling in singular-value space of the difference vectors.

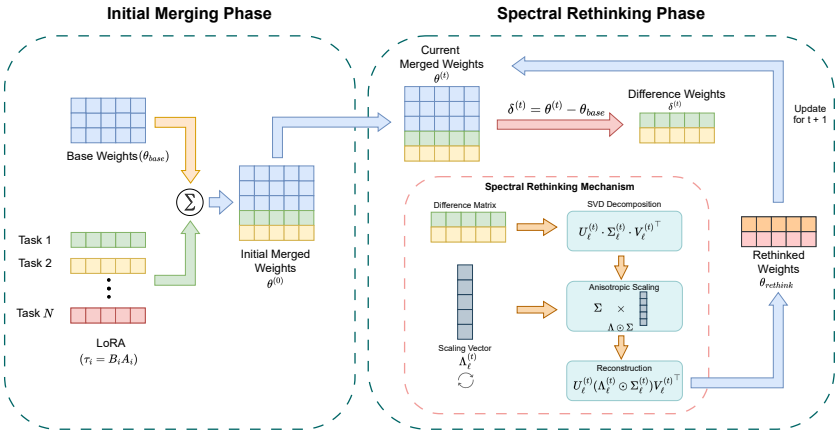


Figure 1: Overall pipeline of LoRA-DV, After the initial aggregation of base and LoRA weights, the Spectral Rethinking Mechanism is employed. It iteratively recalibrates the spectral distribution of difference vectors via SVD-based anisotropic scaling to reduce task interference effectively.

Given a layer-wise difference matrix $\Delta W \in \delta$, we perform the SVD decomposition:

$$\Delta W = U \Sigma V^\top, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r). \quad (7)$$

The singular vectors can now be seen as the knowledge directions while singular values are the strength. To recalibrate strengths without destroying learned geometry, we freeze U and V and introduce a learnable anisotropic scaling vector $\Lambda \in R^r$ that initialized to ones to dynamically re-weights the singular values:

$$\widetilde{\Delta W} = U(\Lambda \odot \Sigma)V^\top. \quad (8)$$

Equation 8 is the core rethinking action, instead of globally shrinking or truncating an update, we apply fine-grained gains per spectral mode, suppressing conflict noise modes while amplifying constructive task modes.

4.2 Iterative Rethinking Strategy

However, task interference is often layered, a single step correction may not adjust the parameter distribution into an optimal state. Thus, we introduce the iterative rethinking strategy. Formally, define $\{\theta^{(t)}\}_{t=0}^{T}$ with $\theta^{(0)}$is the initial merged model weights. At iteration $t+1$, we can extract the difference vector:

$$\delta^{(t)} = \theta^{(t)} - \theta_{\mathrm{pre}} \quad (9)$$

Then we apply the layer-wise rethinking operation. For each layer-wise difference matrix $\Delta W^{(t)} \in \delta^{(t)}$, we perform SVD:

$$\Delta W^{(t)} = U^{(t)} \Sigma^{(t)} V^{(t)\top}, \quad \Sigma^{(t)} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r). \quad (10)$$

and introduce a learnable anisotropic scaling vector $\Lambda^{(t)} \in R^r$(initialized to ones). With freezing $U^{(t)}$ and $V^{(t)}$, the spectrally reweighted update is reconstructed as:

$$\widetilde{\Delta W}^{(t)} = U^{(t)}(\Lambda^{(t)} \odot \Sigma^{(t)})V^{(t)\top}. \quad (11)$$

4.3 General Learning Objective

To make our rethinking mechanism well-defined at the model level and propose the general learning objective, we learn a global set of anisotropic scaling coefficients across all layers simultaneously. Formally, introduce a unified layer index set $\mathcal{L}$. For each $\ell \in \mathcal{L}$, let $\Delta W_\ell^{(t)} \in \delta^{(t)}$ denote the corresponding layer-wise difference matrix. The spectral rethinking SVD is $\Delta W_\ell^{(t)} = U_\ell^{(t)} \Sigma_\ell^{(t)} V^{(t)\top}_\ell$ followed by a learnable anisotropic scaling vector $\Lambda_\ell^{(t)}$. We define a global transform $F_\Lambda(\cdot)$ by applying the rethinking operation to all layers at iteration $t+1$ where:

$$F_{\Lambda^{(t)}}(\delta^{(t)}) := \left\{ U_\ell^{(t)} \left(\Lambda_\ell^{(t)} \odot \Sigma_\ell^{(t)} \right) V_\ell^\top \right\}_{\ell \in \mathcal{L}},$$
$$\Lambda^{(t)} := \{\Lambda_\ell^{(t)}\}_{\ell \in \mathcal{L}}. \quad (12)$$

Under a small calibration set $\mathcal{D}_{\mathrm{calib}}$, the *overall objective* is to jointly optimize the entire collection $\Lambda^{(t)} = \{\Lambda_\ell^{(t)}\}_{\ell \in \mathcal{L}}$ by minimizing the calibration

loss computed on the full model:

$$\Lambda^{*(t)} = \arg\min_{\Lambda^{(t)}} \mathcal{L}\Big(\theta_{\text{pre}} + F_{\Lambda^{(t)}}(\delta^{(t)}); \mathcal{D}_{\text{calib}}\Big). \tag{13}$$

4.4 Loss Functions

Notably, our spectral rethinking mechanism is available on both supervised and unsupervised training scenarios with different loss functions. For supervised training, we adopt the traditional cross-entropy loss. For unsupervised training loss $\mathcal{L}_{uns}$, following (Yang et al., 2024b), we use Shannon entropy of the predicted output of the neural network:

$$\mathcal{L}_{uns} = -\sum_{c=1}^{C} p(\hat{y}_{i,c}) \log p(\hat{y}_{i,c}). \tag{14}$$

Here, $p(\hat{y}_{i,c})$ denotes the prediction probability of input sample x_i is class c. According to (Yang et al., 2024b), the entropy loss in eq 14 is highly correlated to the actual prediction loss, thus, using the above unsupervised loss on the unlabeled calibration sets, our spectral rethinking mechanism is able to work effectively on unsupervised settings.

5 Experiment

5.1 Experimental Settings

Benchmarks and Baselines. For discriminative language tasks, following the standard setup (Yang et al., 2024b; Wang et al., 2024), we employ Flan-T5-base and Flan-T5-large (Chung et al., 2024) as the backbone and evaluate on the 8-task GLUE benchmark(Wang et al., 2019) to investigate the performances across fundamental linguistic understanding tasks. For language generative tasks, following the setting in (Lu et al., 2024), we evaluate Qwen-14B (Bai et al., 2023) on four scenarios: general knowledge (MMLU (Hendrycks et al., 2021)), factualness (TruthfulQA (Lin et al., 2022)), safety (BBQ (Parrish et al., 2022)), and summarization (CNN-DailyMail (Nallapati et al., 2016)).

Since LoRA-DV functions as a universal post-hoc rethinking module rather than a standalone merging method, we evaluate it atop representative general merging frameworks to demonstrate its effectiveness, including Model Soups (Wortsman et al., 2022), Task Arithmetic (Ilharco et al., 2023), Ties-Merging (Yadav et al., 2023), EMR-Merging (Huang et al., 2024), and DARE (Yu et al., 2024).

Implementation Details. We first obtain task-specific experts via LoRA fine-tuning. For Flan-T5-base and Flan-T5-large, we adopt the pre-trained LoRA adapters directly from the official FusionBench (Tang et al., 2025). For Qwen-14B, we set a higher LoRA rank of $r = 32$. The model is trained for 3 epochs using the Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 128. For our method LoRA-DV, we perform Singular Value Decomposition (SVD) on the task vectors of each LoRA adapter. To learn the anisotropic scaling factors for the singular values, we randomly sample a tiny calibration set of 32 samples from the training sets of the merged tasks. We optimize the scaling factors using the Adam optimizer with a learning rate of 0.005. The optimization process consists of 3 iterations, each comprising 10 epochs. More details in the Appendix.

5.2 Main Results

Results of Medium-Sized Models. For medium-sized language models, we select Flan-T5-base and Flan-T5-large (Chung et al., 2024) as base models and use eight discriminative tasks from the GLUE benchmark as target tasks. As detailed in Table 1 and Table 2, our method substantially outperforms baselines on average. For Flan-T5-base, LoRA-DV achieves consistent improvements across all methods. Specifically, the fully calibrated LoRA-DV improves WA, TA, TIES, and EMR by 2.7%, 2.4%, 2.6%, and 2.3%. Specifically, our approach achieves significant gains, such as a 4.4% improvement on MNLI under TA and a 6.7% improvement on STSB under EMR. Remarkably, even with unlabeled data (denoted as Yes(unlabeled)), our method still consistently surpasses the standard baselines, yielding average improvements of 1.2%, 1.1%, 1.4%, and 0.9% for WA, TA, TIES, and EMR, respectively. This validates the effectiveness of our spectral rethinking mechanism.

Similarly, for Flan-T5-large, our method yields average gains of 1.9% for WA, 1.7% for TA, 1.9% for TIES, and 1.6% for EMR. LoRA-DV maintains a significant advantage even with unlabeled data, outperforming the corresponding baselines by 0.9%, 0.9%, 1.3%, and 0.7% on average. Although the baselines marginally outperform our method on QQP in isolated cases, the performance gap is negligible and does not diminish the overall superiority of our approach.

Table 1: Performances of Flan-T5-base in 8 GLUE tasks. "Individual" refers to the metrics of each fine-tuned model on the dataset on which it was trained. **Bold** means the best performance across different methods. "No" indicates that LoRA-DV is not applied; "Yes (unlabeled)" indicates LoRA-DV uses entropy minimization (Eq. 14) on unlabeled data; and "Yes" indicates the use of the full LoRA-DV.

Method	LoRA-DV	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
Individual	-	69.1	82.7	85.5	90.9	84.0	84.5	92.9	87.4	84.6
WA	No	69.7	59.7	78.9	90.1	83.8	80.5	91.2	72.0	78.2
	Yes(unlabeled)	70.1	61.4	79.2	90.6	83.4	81.9	92.6	75.8	79.4 (+1.2%)
	Yes	**70.7**	**64.4**	**80.5**	**91.7**	**83.9**	**83.5**	**93.3**	**79.2**	**80.9 (+2.7%)**
TA	No	68.8	55.2	78.7	89.8	83.7	79.1	91.5	72.4	77.4
	Yes(unlabeled)	69.0	57.7	79.2	90.3	83.8	81.7	91.7	74.3	78.5 (+1.1%)
	Yes	**69.1**	**59.6**	**79.9**	**90.7**	**84.5**	**82.9**	**92.8**	**78.9**	**79.8 (+2.4%)**
TIES	No	68.3	56.3	79.4	89.8	**83.7**	79.4	91.6	71.2	77.5
	Yes(unlabeled)	68.4	59.2	79.9	90.3	82.8	81.7	92.3	76.3	78.9 (+1.4%)
	Yes	**68.6**	**61.7**	**80.5**	**91.2**	83.4	**83.1**	**93.1**	**78.9**	**80.1 (+2.6%)**
EMR	No	71.8	62.4	**79.7**	92.3	84.3	80.1	92.7	75.4	79.8
	Yes(unlabeled)	71.7	63.1	79.0	92.8	84.1	83.1	93.4	78.1	80.7 (+0.9%)
	Yes	**71.9**	**66.3**	79.4	**93.7**	**84.5**	**84.4**	**94.7**	**82.1**	**82.1 (+2.3%)**

Table 2: Performances of Flan-T5-large in 8 GLUE tasks. "Individual" refers to the metrics of each fine-tuned model on the dataset on which it was trained. **Bold** means the best performance across different methods. "No" indicates that LoRA-DV is not applied; "Yes (unlabeled)" indicates LoRA-DV uses entropy minimization (Eq. 14) on unlabeled data; and "Yes" indicates the use of the full LoRA-DV.

Method	LoRA-DV	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	Avg.
Individual	-	80.2	88.5	89.2	94.4	87.2	91.7	95.2	90.9	89.6
WA	No	74.6	84.3	84.1	92.8	86.3	87.4	94.8	88.0	86.5
	Yes(unlabeled)	76.2	85.6	85.3	93.3	86.2	88.1	95.1	89.2	87.4 (+0.9%)
	Yes	**76.8**	**87.0**	**86.5**	**93.8**	**87.0**	**89.8**	**95.8**	**90.8**	**88.4 (+1.9%)**
TA	No	76.9	85.4	85.3	93.9	85.8	88.1	95.2	87.8	87.3
	Yes(unlabeled)	77.3	86.2	86.9	94.6	86.3	90.1	95.2	89.3	88.2 (+0.9%)
	Yes	**78.5**	**87.5**	**87.5**	**94.8**	**86.4**	**90.5**	**95.8**	**91.0**	**89.0 (+1.7%)**
TIES	No	77.1	85.1	86.3	93.9	86.0	87.7	95.1	88.0	87.4
	Yes(unlabeled)	78.6	87.2	87.3	94.3	86.1	89.3	95.8	90.6	88.7 (+1.3%)
	Yes	**79.2**	**87.9**	**88.2**	**94.8**	**86.6**	**90.5**	**96.0**	**91.2**	**89.3 (+1.9%)**
EMR	No	78.8	86.7	85.9	94.1	86.1	87.8	95.0	88.4	87.9
	Yes(unlabeled)	79.5	87.8	87.5	94.3	86.2	88.9	95.3	89.1	88.6 (+0.7%)
	Yes	**80.5**	**89.0**	**88.0**	**95.2**	**86.8**	**90.5**	**96.0**	**90.0**	**89.5 (+1.6%)**

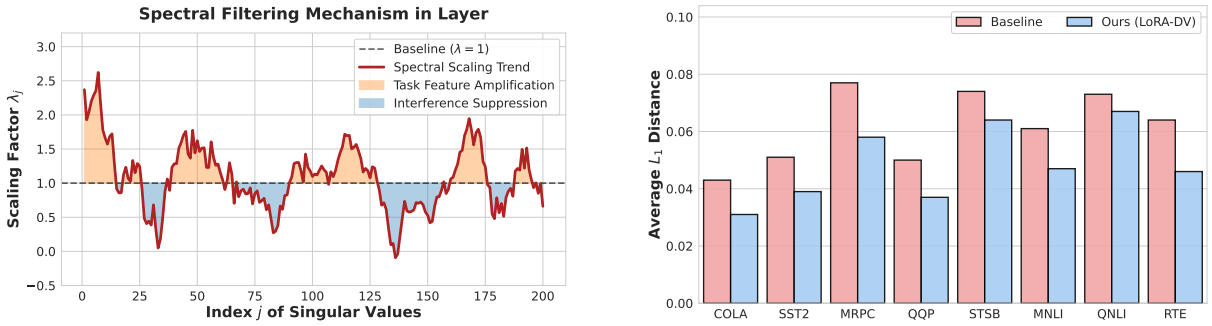
Results of Large Language Models. To further demonstrate the effectiveness of LoRA-DV in large language models, we supplement experiments on Qwen-14B. The experimental results of Qwen-14B are presented in Table 3, respectively. For Qwen-14B, our method achieves superior performance compared to both standard baselines and their DARE-enhanced variants. Specifically, under the TA framework, LoRA-DV improves the average score by 1.44%, surpassing the DARE-enhanced baseline. Similarly, under TIES, our approach yields an average improvement of 1.13%, with substantial gains observed on BBQ and MMLU benchmarks. Notably, LoRA-DV maintains a competitive performance even with unlabeled calibration data. LoRA-DV with unlabeled data (denoted as *(unlabeled)*) yields average improvements of 0.96% and 0.55% over the standard TA and TIES baselines, respectively. It is worth noting that under the TIES framework, our label-free method even surpasses the DARE-enhanced baseline (55.24 vs. 54.63). Although it performs slightly lower than TA w/ DARE (55.84 vs. 56.08), it still exhibits highly competitive performance given the unlabeled calibration data. These results indicate that our method remains robust and effective even when scaling to larger models and diverse task distributions, regardless of data availability.

5.3 Mechanism Analysis

In this section, we conduct a qualitative analysis to investigate the mechanism of LoRA-DV. Figure 2a illustrates the distribution of the learned scaling factors λ, visualizing how the model modulates feature importance. Specifically, we observe that the model selectively amplifies singular values corresponding to core task-specific directions ($\lambda > 1.0$), while explicitly suppressing components associ-

Table 3: Performances of Qwen-14B on general knowledge, factualness, safety, and summarization tasks. "Individual" refers to the metrics of each fine-tuned model on the dataset on which it was trained. **Bold** means the best performance across different methods. "No" indicates that LoRA-DV is not applied; "No (w/ DARE)" indicates that DARE is applied instead of LoRA-DV; "Yes (unlabeled)" indicates LoRA-DV uses entropy minimization (Eq. 14) on unlabeled data; and "Yes" indicates the use of the full LoRA-DV.

Method	LoRA-DV	MMLU	TruthfulQA	BBQ	CNN	Avg.
Individual	-	69.07	53.33	93.53	19.46	58.85
TA	No	68.33	52.39	78.24	20.55	54.88
	No (w/ DARE)	68.53	51.68	**82.83**	21.29	56.08
	Yes (unlabeled)	68.67	51.98	81.29	21.41	55.84 (+0.96%)
	Yes	**69.17**	**52.59**	81.71	**21.79**	**56.32** (+1.44%)
TIES	No	68.27	49.08	84.11	17.29	54.69
	No (w/ DARE)	69.31	**52.08**	81.19	15.92	54.63
	Yes (unlabeled)	69.04	49.31	84.70	17.92	55.24 (+0.55%)
	Yes	**69.78**	49.95	**85.37**	**18.17**	**55.82** (+1.13%)



(a) Spectral filtering mechanism

(b) Reduction in representation bias

Figure 2: Mechanism analysis of LoRA-DV. (a) Visualization of learned anisotropic scaling factors λ_j across singular value indices, illustrating the selective amplification of task-specific components and suppression of interference. (b) Comparison of representation bias (average L_1 distance) between Task Arithmetic and LoRA-DV across GLUE tasks, demonstrating consistent mitigation of representation drift.

Table 4: Ablation study of LoRA-DV

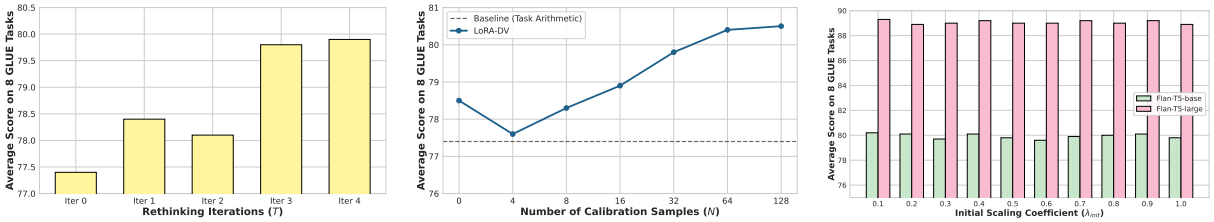
Method	**Flan-T5-base**	**Flan-T5-large**
Baseline	77.4	87.3
Random perturbation	23.5	29.0
Isotropic (scalar) scaling	78.2	87.5
LoRA-DV	79.8	89.2

ated with task interference or ($\lambda < 1.0$). This phenomenon implies that LoRA-DV effectively empowers the model to rethink, thereby decoupling critical task signals from interference.

To empirically validate LoRA-DV reduces task interference, we employ the metric of representation bias (Yang et al., 2024a), defined as the average L_1 distance between the hidden states of the merged model and those of the original single-task experts. This metric reflects task interference by quantifying the discrepancy between merged features and task-specific optimal features, with lower values indicating better mitigation of task interference. As shown in Figure 2b, compared to the baseline, LoRA-DV consistently yields significantly lower representation bias across all evaluated tasks. This reduction provides strong evidence that our anisotropic spectral calibration effectively mitigates representation drift, significantly alleviating task interference during the merging process.

5.4 Ablation Study

To demonstrate the effectiveness of our approach, we conducted ablation studies for LoRA-DV, summarized in Table 4. Replacing the learned spectral scaling with random perturbation leads to a drastic performance loss, highlighting the critical need for precise spectral calibration. Blind stochastic regularization impairs the model's capabilities, confirming that the directionality of the spectral update is essential. Replacing the fine-grained anisotropic vector with isotropic (scalar) scaling also results in a clear drop in performance, demonstrating the value of spectral-wise control. This is likely because isotropic scaling applies a uniform gain across all singular values, limiting its ability to disentangle constructive task features from conflicting noise within the same layer. In contrast,



(a) Impact of number of iterations (b) Impact of calibration set size (c) Robustness to initialization

Figure 3: Further analysis of LoRA-DV on GLUE benchmark. (a) Performance of Flan-T5-base with varying numbers of rethinking iterations T. (b) Performance of Flan-T5-base with respect to the size of the calibration dataset N. (c) Performance stability of Flan-T5-base and Flan-T5-large across varying initial scaling coefficients, demonstrating robustness to hyperparameter initialization.

LoRA-DV's anisotropic capability allows for the precise suppression of specific interference modes that a global scalar cannot address.

5.5 Further Analysis

Impact of Rethinking Iterations. In this section, we investigate the impact of LoRA-DV to the iteration depth T, Figure 3 (a) illustrates the progress of average GLUE performance as the iteration depth T increases. Compared to the non-iterative baseline, the initial rethinking step ($T = 1$) yields a substantial gain, boosting the score to 78.4. Although a transient performance dip is observed at $T = 2$ (78.1)—possibly attributable to unstable spectral calibration in intermediate steps—the performance recovers robustly and reaches 79.8 at $T = 3$. Further extending the iterations to $T = 4$ results in a marginal improvement to 79.9. Given the negligible performance gain beyond the third iteration, this demonstrates the rapid convergence of our approach. Accordingly, we adopt $T = 3$ as the default setting.

Impact of Calibration Set Size. We analyze the dependency on the calibration set size N. In Figure 3 (b), the method exhibits distinct behaviors across different sizes. Notably, in the absence of labeled data ($N = 0$), LoRA-DV, which employs entropy minimization (Eq. 14), already yields a performance of 78.5, superior to the baseline. However, introducing a tiny labeled set ($N = 4$) causes a temporary performance dip to 77.6, which still remains above the baseline. Performance rapidly recovers as N increases to 8 and 16, where the calibration set becomes representative enough to capture global conflict patterns. The score reaches a significant high at $N = 32$ (79.8) and shows further improvements at $N = 64$, finally saturating at $N = 128$ (80.5). The performances demonstrate that LoRA-DV demonstrates robust performance not only with calibration data but also remains remarkably effective in label-free scenarios.

Robustness to Initialization. Traditional merging methods often require meticulous tuning of the global scaling factor λ. To assess the robustness of our approach, we evaluate the performance of both Flan-T5-base and Flan-T5-large across a wide range of initial scaling coefficients $\lambda_{init} \in [0.1, 1.0]$. As visualized in Figure 3 (c), LoRA-DV demonstrates remarkable stability. Regardless of the initial value, the average scores for Flan-T5-base fluctuate minimally between 79.6 and 80.2, while Flan-T5-large remains steady around 89.0. This indicates that LoRA-DV can dynamically rectify the weight distribution to an optimal state, effectively eliminating the need for the manual hyperparameter grid search required by traditional scalar scaling methods.

6 Conclusion

In this paper, we demonstrate that treating model merging as a one-off operation fails to effectively reduce task interference under the current accumulated historical knowledge model learned, whereas the spectral properties of Difference Vectors offer a granular view for disentangling such conflicts. We argue that the task interference among expert models can be further reduced by dynamically modulating the difference vector in the spectral space. Thus, we introduce LoRA-DV, a universal and effective post-hoc rethinking framework that resolves interferences via anisotropic scaling, capable of operating with unlabeled data. Extensive experiments demonstrate the effectiveness of our method over state-of-the-art baselines. We believe that this work is one step toward shifting from static parameter aggregation to dynamic spectral rethinking.

Limitations

Task vectors depend on the particular pre-trained model, it is not yet feasible to compose and transfer knowledge across different architectures. With a suitable projection mechanism, this may become feasible in future work.

Ethical Considerations

Data Usage and Privacy. Our research utilizes publicly available datasets, including GLUE, MMLU, GSM8K, and specific domain datasets such as MedQA and HealthcareMagic. We have strictly adhered to the licenses and usage terms of these datasets (e.g., MIT, Apache-2.0, and CC-BY) as detailed in Appendix A.1. No private or personally identifiable information is collected or used in this study.

Potential Risks in Critical Domains. We evaluate our method on medical datasets (MedQA and HealthcareMagic) to demonstrate the effectiveness of model merging in heterogeneous tasks. However, we emphasize that the merged models are intended solely for research purposes. They have not undergone rigorous clinical validation and should not be used for providing medical advice or diagnosis in real-world scenarios.

Bias and Safety. Model merging inherits the capabilities but also potentially the biases of the pre-trained base models and fine-tuned experts. To monitor this, we included safety and bias benchmarks (BBQ and TruthfulQA) in our evaluation. While our method shows competitive performance on these benchmarks, users should exercise caution and conduct comprehensive safety testing before deploying merged models in sensitive applications.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

MohammadReza Davari and Eugene Belilovsky. 2024. Model breadcrumbs: Scaling multi-task model merging withnbsp;sparse masks. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXV*, page 270–287, Berlin, Heidelberg. Springer-Verlag.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. 2024. Della-merging: Reducing interference in model merging through magnitude-based sampling. *Preprint*, arXiv:2406.11617.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. 2025. 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, Vienna, Austria. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. 2024. Emr-merging: Tuning-free high-performance model merging. In *Advances in Neural Information Processing Systems*,

volume 37, pages 122741–122769. Curran Associates, Inc.

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

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

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *North American Chapter of the Association for Computational Linguistics*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. 2025. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. 2025. Model merging with SVD to tie the knots. In *The Thirteenth International Conference on Learning Representations*.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. 2025. Fusionbench: A unified library and comprehensive benchmark for deep model fusion. *Preprint*, arXiv:2406.03280.

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

Jinping Wang, Zhiqiang Gao, Dinggen Zhang, and Zhiwu Xie. 2025. Escaping optimization stagnation: Taking steps beyond task arithmetic via difference vectors. *arXiv preprint arXiv:2511.17987*.

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

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2024b. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*.

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

A Appendix

A.1 Experiment Details

Here we detailed illustrate the setting of our experiments

A.1.1 Benchmark.

Discriminative Tasks. we conduct experiments on the GLUE benchmark (Wang et al., 2019) with eight discriminative tasks, which is designed for classification tasks except for STS-B for the regression task. The detail of eight dataset can be found in the paper of Wang et al. (Wang et al., 2019). Consistent with prior research (Yu et al., 2024) we still split 10% of the training set as a validation set and employ the original validation data as the test set.

The licenses of QNLI, COLA, and STS-B are licensed under CC-BY-SA. QQP is licensed under MIT. SST-2 and MRPC are licensed under Apache 2.0. MNLI is licensed under OANC. RTE is licensed under CC BY 4.0. Thus, these datasets in GLUE are available for non-commercial research purposes.

Generative Tasks. Following the paper of Lu et al (Lu et al., 2024), we conduct experiments on four benchmarks:

- **MMLU** (Hendrycks et al., 2021): Designed to evaluate general and STEM knowledge, this benchmark encompasses 57 distinct subjects ranging from elementary to professional difficulty levels. We employ Exact-Match as the primary evaluation metric.
- **TruthfulQA** (Lin et al., 2022): This benchmark assesses the factualness and truthfulness of language models through a set of 817 questions. It covers 38 diverse categories, including health, law, finance, and politics, with performance measured using the Exact-Match metric.
- **BBQ** (Parrish et al., 2022): Utilized for safety evaluation, this dataset identifies social biases against protected classes across nine social dimensions within U.S. English-speaking contexts. We utilize Exact-Match as the metric for this assessment.
- **CNN-DailyMail** (Nallapati et al., 2016): This dataset serves as our benchmark for text summarization tasks. It requires the model to generate concise summaries from news articles, and we evaluate the generation quality using ROUGE-2 scores (Lin and Hovy, 2003).

We evaluated these tasks using the HELM benchmark (Liang et al., 2023) in a few-shot setting. For MMLU and TruthfulQA, which lack official training sets, we used the Dolly-15k dataset (Conover et al., 2023) for MMLU and the BigBench-sampled dataset for TruthfulQA.

The MMLU dataset is under the MIT License. TruthfulQA and CNN-DailyMail are under the Apache-2.0 License. BBQ is under the CC-BY 4.0 License.

Heterogeneous Tasks. Following Hi-Merging (Fu et al., 2025), we conduct experiments on three benchmarks:

- **GSM8K** (Cobbe et al., 2021): A dataset consisting of 8.5K high-quality grade school math word problems. It is designed to evaluate a model's ability to perform multi-step mathematical reasoning and basic arithmetic calculations to reach the correct solution. We report the accuracy as the evaluation metric.
- **MedQA** (Jin et al., 2021): A large-scale open-domain question answering benchmark collected from professional medical board examinations. It assesses the model's capacity to apply extensive professional medical knowledge and clinical reasoning to solve complex multiple-choice questions. We employ Accuracy as the evaluation metric.
- **HealthcareMagic** (Li et al., 2023): A dataset comprising approximately 100,000 authentic patient-physician conversations sourced from an online medical consultation platform. It evaluates the model's ability to understand patient inquiries and generate helpful medical responses. We assess performance using BLEU-4 (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003).

The GSM8K and MedQA datasets are under the MIT License. HealthcareMagic is under Apache-2.0 License.

A.1.2 Language Model Backbone.

For discriminative tasks, we use Flan-T5-base and Flan-T5-large (Chung et al., 2024) as our pre-trained backbone and select the official LoRA module from FusionBench (Tang et al., 2025). For

generative tasks, we utilized Qwen-14B (Bai et al., 2023) as the base model, fine-tuned via LoRA (Hu et al., 2022). The training setup consisted of a rank of 32, a batch size of 128, and a learning rate of $2e^{-4}$ over 3 epochs.

A.1.3 Baselines Details.

In this section, we provide a detailed description of the baseline models adopted for our comparative analysis.

- **Individual:** This setting deploys a dedicated fine-tuned model for each distinct task. While it eliminates inter-task interference, it lacks the capability to handle multiple tasks concurrently. We treat this as the *upper bound* of performance for single-task evaluation.
- **Weight Averaging (Wortsman et al., 2022) :** Representing the most rudimentary merging strategy, this method calculates the element-wise mean of parameters across different models. It is generally regarded as the lower bound for model merging performance.
- **Task Arithmetic (Ilharco et al., 2023):** This method leverages the concept of "task vectors"—derived by subtracting the pre-trained weights from the fine-tuned ones. These vectors are then aggregated and added to the base model to enable multi-task learning.
- **Ties-Merging (Yadav et al., 2023):** Designed to mitigate task interference, Ties-Merging focuses on removing redundant parameter updates. The algorithm proceeds through three distinct phases: *Trim*, *Elect Sign*, and *Disjoint Merge*.
- **Task Arithmetic (w/ DARE) (Yu et al., 2024):** This variant integrates the DARE technique, which applies Bernoulli resampling to sparsify the delta parameters (with a 70% drop rate) prior to executing Task Arithmetic (Ilharco et al., 2023).
- **Ties-Merging (w/ DARE) (Yu et al., 2024):** Analogous to the approach above, this method employs DARE to induce sparsity (70% random drop) before applying the standard Ties-Merging (Yadav et al., 2023) protocol.
- **EMR-Merging (Huang et al., 2024):** This tuning-free method introduces a unified model paired with lightweight task-specific modulators. It employs a three-stage process which is elect, mask, and rescale to align parameter direction and magnitude for each specific task without additional training.
- **EMR-Merging (Huang et al., 2024):** This tuning-free method introduces a unified model paired with lightweight task-specific modulators. It employs a three-stage process which is elect, mask, and rescale to align parameter direction and magnitude for each specific task without additional training.
- **DELLA-Merging (Deep et al., 2024):** This approach extends DARE by refining the parameter pruning process. Specifically, it determines the drop probability for each delta vector according to its absolute magnitude, which effectively mitigates instability during the merging process.

To determine the optimal coefficients for Task Arithmetic and Ties-Merging, we perform a small-scale grid search on the validation datasets. For DARE Merging, we adopt a constant coefficient of 0.7, consistent with the settings in prior works (Yu et al., 2024).

A.1.4 Hyperparameter.

For Flan-T5-base and Flan-T5-large (Chung et al., 2024), we select a learning rate of 0.005 and optimize the anisotropic scaling factors using the Adam optimizer. The optimization process consists of 3 rethinking iterations, with each iteration comprising 10 epochs. We utilize a calibration set constructed by randomly sampling 32 examples from the validation set of each task involved in the merger. For the Llama-3-8B (Grattafiori et al., 2024) and Qwen-14B (Bai et al., 2023), we maintain the same calibration data size and the number of iterations, we adjust the learning rate to 0.0001 and reduce the training duration to 5 epochs per iteration. Finally, all experimental results reported in this paper are averaged over three independent runs.

A.1.5 Further Experiment.

To further evaluate the generalization capability of LoRA-DV on other large-scale architectures and its efficacy in handling highly heterogeneous task interference, we conducted additional experiments using Llama-3-8B (Grattafiori et al., 2024). Specifically, we constructed a challenging cross-domain

Table 5: Performances of Llama-3-8B on heterogeneous merging scenarios, combining mathematical reasoning (GSM8K) and medical expertise (MedQA and HealthcareMagic). TA, TIES, DARE, and DELLA represent different merging frameworks, while Yes denotes the application of LoRA-DV. B-4, R-1, R-2, and R-L refer to BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L, respectively. **Bold** means the best performance across different methods.

Method	LoRA-DV	MedQA	HealthcareMagic				GSM8K	Avg.
		Acc.	B-4	R-1	R-2	R-L	Acc.	
Individual	-	63.59	31.42	30.11	9.74	19.11	69.75	37.29
TA	No	61.43	31.56	27.06	5.03	15.99	73.16	35.71
	Yes	**61.98**	**31.63**	**27.12**	**5.08**	**16.09**	**74.15**	**36.01** (+0.30%)
TIES	No	60.96	**35.38**	**28.83**	**6.24**	**17.95**	70.28	36.61
	Yes	**62.22**	35.31	28.76	6.18	17.87	**71.57**	**36.99** (+0.38%)
DARE	No	60.33	31.62	27.18	5.04	15.99	73.31	35.58
	Yes	**61.27**	**32.17**	**27.49**	**5.22**	**16.34**	**74.07**	**36.09** (+0.51%)
DELLA	No	61.51	20.33	23.46	3.82	12.57	74.98	32.78
	Yes	**63.32**	**20.42**	**23.53**	**3.91**	**12.65**	**75.82**	**33.28** (+0.50%)
TIES (w/ DARE)	No	61.35	32.23	27.25	5.12	16.16	72.18	35.72
	Yes	**63.39**	**32.24**	**27.60**	**5.20**	**16.41**	**73.92**	**36.46** (+0.74%)
TIES (w/ DELLA)	No	61.39	21.06	23.71	3.86	12.79	74.01	32.80
	Yes	**62.45**	**21.14**	**23.80**	**3.98**	**12.88**	**74.83**	**33.18** (+0.38%)

merging scenario that combines mathematical reasoning (GSM8K (Cobbe et al., 2021)) with medical expertise (MedQA (Jin et al., 2021) and HealthcareMagic (Li et al., 2023)).

The quantitative results are presented in Table 5. Our method demonstrates consistent improvements across all evaluated merging frameworks, including advanced composite methods such as TIES (w/ DARE) and TIES (w/ DELLA). On average, LoRA-DV outperforms the baselines in all settings. Notably, in this scenario involving the fusion of disparate domains, our approach significantly enhances domain-specific performance. For instance, under the TIES (w/ DARE) setting, LoRA-DV improves MedQA accuracy by approximately 2.04% compared to the baseline. Furthermore, our method consistently enhances the average performance on GSM8K across all settings, demonstrating its effectiveness in preserving critical reasoning capabilities during the heterogeneous merging process

A.1.6 Compute Resources Used and Runtimes.

The LoRA fine-tuning process is conducted on Nvidia A100 GPUs with 80GB VRAM. Specifically, training single-task LoRA models for Qwen-14B across the four generative tasks took approximately 1-2 hours per task, training single-task LoRA for Llama-3-8B need 1-2 hours on single GPUs. The other part of experiment is conduct on 2 Nvidia GeForce RTX 3090s, 1 iteration for medium-size model (Flan-T5-base and Flan-T5-large) costs 30 seconds - 1 minute, for large language model (Llama3-8B and Qwen-14B) costs 3 minutes - 5 minutes.