Model Merging on Loss Landscapes: A Geometry Perspective

Supplementary Material

A. Notations

To improve the readability for the general audience, Table 5 lists all the notation used in this paper.

B. Additional Details for Methodology

B.1. Proof for Proposition 1

The following provides a proof for proposition 1.

Proof. Let the global multi-task objective function be the weighted sum of the individual task losses, given by

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \lambda_t \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t} \left[\mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta})\right]. \tag{17}$$

Our objective is to find the optimal parameter $\boldsymbol{\theta}_m$ that minimizes the objective above. To achieve this, we first consider the Taylor expansion of the expected loss for a specific task t around its fine-tuned optimum $\boldsymbol{\theta}_t$. Denote $L_t(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t} \left[\mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta})\right]$ for simplicity, the second-order Taylor expansion at $\boldsymbol{\theta}_t$ gives

$$L_t(\boldsymbol{\theta}) \approx$$
$$L_t(\boldsymbol{\theta}_t) + (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \nabla_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t), \tag{18}$$

Based on our assumption 2 in the main paper, each fine-tuned model resides at a local minimum, which implies the first-order derivative vanishes, i.e., $\nabla_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta}_t) \approx \mathbf{0}$. Therefore, the expansion reduces to

$$L_t(\boldsymbol{\theta}) \approx L_t(\boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t). \tag{19}$$

Substituting the above approximation back into the global objective function, we have

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) &\approx \sum_{t=1}^{T} \lambda_t \left(L_t(\boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t) \right) \\ &= \frac{1}{2} \sum_{t=1}^{T} \lambda_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \text{const}.\end{aligned} \tag{20}$$

Herein, since the constant term does not affect the optimization, minimizing the global objective $\mathcal{J}(\boldsymbol{\theta})$ is equivalent to minimizing the quadratic form.

Recall that the Fréchet Mean objective defined in equation 6 is essentially finding a point $\boldsymbol{\theta}_m$ that minimizes the weighted sum of squared geodesic distances to all fine-tuned models. In a small neighborhood around each $\boldsymbol{\theta}_t$, the squared geodesic distance $d^2_{\mathbf{G}_t}(\theta, \theta_t)$ on a Riemannian manifold with metric tensor $\mathbf{G}_t = \boldsymbol{H}_t$ can be approximated by the quadratic form

$$d^2_{\mathbf{G}_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t) \tag{21}$$

Substituting this approximation into the Fréchet Mean objective yields

$$\boldsymbol{\theta}_m \approx \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \lambda_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_t). \tag{22}$$

Comparing the two approximations, it is obvious that the objective functions are identical (up to a constant scaling factor of $1/2$). Thus, minimizing the multi-task loss is locally equivalent to finding the Fréchet Mean on the manifold endowed with the Hessian metric. □

B.2. Closed-Form Fréchet Mean in equation 8

Denote the objective in equation 7 by $\mathcal{J}(\boldsymbol{\delta}_m)$. To minimize the objective, we take the first- and second-order derivatives with respect to $\boldsymbol{\delta}_m$,

$$\frac{\partial}{\partial \boldsymbol{\delta}_m} \mathcal{J}(\boldsymbol{\delta}_m) = 2 \sum_{t=1}^{T} \lambda_t \boldsymbol{H}_t (\boldsymbol{\delta}_m - \boldsymbol{\delta}_t), \tag{23}$$

$$\frac{\partial^2}{\partial \boldsymbol{\delta}_m^2} \mathcal{J}(\boldsymbol{\delta}_m) = 2 \sum_{t=1}^{T} \lambda_t \boldsymbol{H}_t. \tag{24}$$

Based on Assumption 2, the Hessian evaluated at each local minimum is positive-definite, $\boldsymbol{H}_t \succ 0$. With non-negative scalar factors $\lambda_t \geq 0$, the Hessian of $\mathcal{J}$ is positive-definite, making it strictly convex. Therefore, we can set equation 23 to zero and solve for the optimal $\boldsymbol{\delta}_m^*$ via

$$\sum_{t=1}^{T} \lambda_t \boldsymbol{H}_t \boldsymbol{\delta}_m^* = \sum_{t=1}^{T} \lambda_t \boldsymbol{H}_t \boldsymbol{\delta}_t, \tag{25}$$

which yields the closed-form solution in equation 8.

B.3. Empirical Fisher Diagonal in equation 9

Our goal is to approximate the expected Hessian $\boldsymbol{H}_t$ at the converged parameters $\boldsymbol{\theta}_t$. For models trained with a supervised log-likelihood objective, the expected Hessian coincides with the Fisher information matrix to second order, which equals the expected outer product of the per-sample gradient that writes

$$\boldsymbol{H}_t \approx \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t} \left[\nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta}_t) \nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta}_t)^\top\right]. \tag{26}$$

Table 5. Table of notation.

Notation	Explanation	Notation	Explanation
n	Dimension of the input data space.	T	Number of fine-tuning tasks.
m	Dimension of the parameter space.	L	Number of matrix-shaped layers in the backbone.
d_ℓ	Dimension of layer ℓ (square layers).	k	Per-task SVD truncation rank (also denoted r).
p	Subspace dimension, $p = LkT$.	Θ	Parameter space, with $\Theta \subseteq \mathbb{R}^m$.
λ_t	Merging weight of task t, with $\sum_t \lambda_t = 1$.	α	Global Frobenius rescaling of the merged delta.
η	Curvature heterogeneity diagnostic $\boldsymbol{c}^\top \bar{\boldsymbol{H}}^{-1}\boldsymbol{c}$.	τ	Geodesic parameter in $[0, 1]$.
$\mathbf{0}$	All-zero vector or matrix.	$\boldsymbol{I}_p$	Identity matrix of size p (likewise $\boldsymbol{I}_m$).
$\boldsymbol{x}$	Input data sample in $\mathbb{R}^n$.	$\boldsymbol{\theta}_0$	Pre-trained parameters.
$\boldsymbol{\theta}_t$	Fine-tuned parameters for task $1 \le t \le T$.	$\boldsymbol{\theta}_m$	Merged parameters.
$\boldsymbol{\delta}_t$	Task vector $\boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ for task t.	$\boldsymbol{\delta}_m$	Merged displacement $\boldsymbol{\theta}_m - \boldsymbol{\theta}_0$.
$\boldsymbol{\delta}_t^\perp$	Residual of $\boldsymbol{\delta}_t$ orthogonal to the subspace, $(\boldsymbol{I}_m - \boldsymbol{S}\boldsymbol{S}^\top)\boldsymbol{\delta}_t$.	$\tilde{\boldsymbol{\delta}}_t$	Projected task vector $\boldsymbol{S}^\top\boldsymbol{\delta}_t$ in the subspace.
$\tilde{\boldsymbol{\delta}}_m$	Merged displacement in the subspace, $\boldsymbol{S}^\top\boldsymbol{\delta}_m$.	$\bar{\boldsymbol{\delta}}$	Weighted average $\sum_t \lambda_t \tilde{\boldsymbol{\delta}}_t$.
$\tilde{\boldsymbol{\delta}}_I$	Flat-geometry (identity-metric) subspace solution, $\tilde{\boldsymbol{\delta}}_I = \bar{\boldsymbol{\delta}}$.	$\tilde{\boldsymbol{\delta}}_H$	Curvature-aware subspace Fréchet mean, $\bar{\boldsymbol{H}}^{-1}\sum_t \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t$.
$\boldsymbol{S}$	Subspace matrix $\boldsymbol{S} \in \mathbb{R}^{m\times p}$ with $\boldsymbol{S}^\top\boldsymbol{S} = \boldsymbol{I}_p$.	$\boldsymbol{S}^{(\ell)}$	Layer-ℓ block of $\boldsymbol{S}$ in $\mathbb{R}^{d_\ell^2 \times kT}$.
$\boldsymbol{U}_\perp^{(\ell)}$	Procrustes-whitened left factors in $\mathbb{R}^{d_\ell \times kT}$.	$\boldsymbol{V}_\perp^{(\ell)}$	Procrustes-whitened right factors in $\mathbb{R}^{d_\ell \times kT}$.
$\boldsymbol{U}_{\perp,i}^{(\ell)}$	i-th column of $\boldsymbol{U}_\perp^{(\ell)}$ (an atom of the tagged basis).	$\boldsymbol{V}_{\perp,i}^{(\ell)}$	i-th column of $\boldsymbol{V}_\perp^{(\ell)}$ (an atom of the tagged basis).
$\boldsymbol{F}_t$	Fisher information matrix of task t.	$\boldsymbol{H}_t$	Expected Hessian of task t evaluated at $\boldsymbol{\theta}_t$.
$\tilde{\boldsymbol{H}}_t$	Projected per-task Hessian $\boldsymbol{S}^\top\boldsymbol{H}_t\boldsymbol{S} \in \mathbb{R}^{p\times p}$.	$\tilde{\boldsymbol{H}}_t^{(\ell)}$	Per-layer projected Hessian $\boldsymbol{S}^{(\ell)\top}\boldsymbol{H}_t^{(\ell)}\boldsymbol{S}^{(\ell)} \in \mathbb{R}^{kT\times kT}$.
$\bar{\boldsymbol{H}}$	Weighted average of projected Hessians $\sum_{t=1}^T \lambda_t \tilde{\boldsymbol{H}}_t$.	$\boldsymbol{c}$	Curvature-task correlation $\sum_t \lambda_t(\tilde{\boldsymbol{H}}_t - \bar{\boldsymbol{H}})(\tilde{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}})$.
$\boldsymbol{u}$	First-moment vector in the optimizer state.	$\boldsymbol{v}$	Second-moment vector in the optimizer state.
$\boldsymbol{v}_t$	Empirical Fisher diagonal of task t, $\mathbb{E}[(\nabla\mathcal{L}_t)^2]$.		
$\mathcal{M}$	Differentiable manifold representing the parameter space.	$\mathbf{G}(\cdot)$	Riemannian metric tensor field on $\mathcal{M}$.
$\gamma(\cdot)$	Geodesic path $\gamma : [0, 1] \to \mathcal{M}$.		
$\mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta})$	Loss of task t at input $\boldsymbol{x}$ with parameter $\boldsymbol{\theta}$.	$\mathcal{J}(\boldsymbol{\theta})$	Multi-task objective $\sum_t \lambda_t \mathbb{E}[\mathcal{L}_t]$.
$\mathcal{F}(\tilde{\boldsymbol{\delta}})$	Subspace Fréchet objective, a quadratic in $\tilde{\boldsymbol{\delta}}$.		
$\mathcal{D}$	Collection of all task-specific datasets $\{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$.	$\mathcal{D}_t$	Task-t training distribution.
$\mathcal{B}^{(\ell)}$	Per-task tagged basis of rank-1 atoms at layer ℓ.	$\mathcal{V}_S$	Subspace Fréchet variance (irreducible task conflict).
$\mathcal{R}_S$	Residual energy $\sum_t \lambda_t(\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp$ lost by projection.		
$\mathbb{C}^2$	Class of twice continuously differentiable functions.	$\mathbb{R}$	Real numbers.

The off-diagonal entries of this $m \times m$ matrix are intractable to store at modern model scales. Hence, we propose to approximate it by its diagonal,

$$\boldsymbol{H}_t \;\approx\; \mathrm{diag}(\boldsymbol{v}_t), \quad \boldsymbol{v}_t \;=\; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}\big[(\nabla_{\boldsymbol{\theta}}\mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\theta}_t))^2\big], \tag{27}$$

which is exactly the empirical Fisher diagonal. The following Appendix D.4 verifies that $\boldsymbol{v}_t$ is robust to estimation from a tiny ($\le 0.5\%$) subsample of the training set, so this proxy is essentially data-free in practice. The diagonal proxy, by itself, is degenerate in the full m-dimensional space due to overparameterization in large neural networks, and each per-task curvature shares the same coordinate eigenbasis. Nevertheless, equation 12 shows that projecting it onto the per-task tagged basis $\boldsymbol{S}$ produces a *dense* $p \times p$ matrix that recovers the off-diagonal curvature signal that the ambient diagonal discards.

B.4. Proofs for the EpiMer Theory

This subsection proves Theorem 1, Theorems 2 and 3, and Proposition 2 stated in Sec. 3.3. Throughout, we work in the p-dimensional subspace defined by the orthonormal basis $\boldsymbol{S} \in \mathbb{R}^{m\times p}$ ($\boldsymbol{S}^\top\boldsymbol{S} = \boldsymbol{I}_p$) and write $\tilde{\boldsymbol{\delta}}_t = \boldsymbol{S}^\top\boldsymbol{\delta}_t$, $\boldsymbol{\delta}_t^\perp = (\boldsymbol{I}_m - \boldsymbol{S}\boldsymbol{S}^\top)\boldsymbol{\delta}_t$, and $\bar{\boldsymbol{H}} = \sum_{t=1}^T \lambda_t \tilde{\boldsymbol{H}}_t$. We assume the merging coefficients are normalized so that $\sum_{t=1}^T \lambda_t = 1$, matching the convention used by every method we compare against (including our own experiments, which use $\lambda_t = 1/T$).

B.4.1. Proof of Theorem 1 (Existence and Uniqueness)

Proof. Expanding the merging objective gives a quadratic form in $\tilde{\boldsymbol{\delta}}$:

$$\mathcal{F}(\tilde{\boldsymbol{\delta}}) = \tilde{\boldsymbol{\delta}}^\top \bar{\boldsymbol{H}}\, \tilde{\boldsymbol{\delta}} \;-\; 2\,\tilde{\boldsymbol{\delta}}^\top \sum_{t=1}^T \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t \;+\; \underbrace{\sum_{t=1}^T \lambda_t \tilde{\boldsymbol{\delta}}_t^\top \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t}_{\text{constant in } \tilde{\boldsymbol{\delta}}}. \tag{28}$$

Differentiating twice yields $\nabla^2\mathcal{F}(\tilde{\boldsymbol{\delta}}) = 2\bar{\boldsymbol{H}}$. By assumption $\bar{\boldsymbol{H}} \succ 0$, so the Hessian is symmetric positive-definite and $\mathcal{F}$ is strictly convex on $\mathbb{R}^p$. Consequently, $\mathcal{F}$ admits a unique global minimizer, and that minimizer is the unique

stationary point. Setting the gradient to zero,

$$\nabla \mathcal{F}(\tilde{\boldsymbol{\delta}}_m^*) = 2\bar{\boldsymbol{H}}\,\tilde{\boldsymbol{\delta}}_m^* - 2\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t = \mathbf{0}, \tag{29}$$

and using invertibility of $\bar{\boldsymbol{H}}$,

$$\tilde{\boldsymbol{\delta}}_m^* = \bar{\boldsymbol{H}}^{-1} \sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t = \left(\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t\right)^{-1} \left[\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t\right], \tag{30}$$

which is exactly equation 13. □

Remark. In our empirical-Fisher implementation we replace $\boldsymbol{H}_t$ by $\mathrm{diag}(\boldsymbol{v}_t) + \epsilon \boldsymbol{I}_m$ with a small jitter $\epsilon > 0$ before projecting (see `ensure_psd` in `hessian.py`). This guarantees $\tilde{\boldsymbol{H}}_t \succ 0$ for every task and hence $\bar{\boldsymbol{H}} \succ 0$, so the existence/uniqueness hypothesis holds unconditionally in practice.

B.4.2. Proof of Theorem 2 (Merging Error Bound)

Proof. By Assumption 1 the loss $\mathcal{L}_t$ is C^2, so a second-order Taylor expansion of $\mathcal{L}_t$ around $\boldsymbol{\theta}_t$ gives

$$\mathcal{L}_t(\boldsymbol{\theta}_m) - \mathcal{L}_t(\boldsymbol{\theta}_t) = \nabla \mathcal{L}_t(\boldsymbol{\theta}_t)^\top (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t) + \tfrac{1}{2}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t) + R_3^{(t)}, \tag{31}$$

with cubic remainder $R_3^{(t)} = \mathcal{O}(\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_t\|^3)$. Assumption 2 kills the linear term, leaving

$$\mathcal{L}_t(\boldsymbol{\theta}_m) - \mathcal{L}_t(\boldsymbol{\theta}_t) = \tfrac{1}{2}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t) + R_3^{(t)}. \tag{32}$$

Then, we can decompose $\boldsymbol{\theta}_m - \boldsymbol{\theta}_t$ along the subspace $\boldsymbol{S}$ and its orthogonal complement. Since $\boldsymbol{\theta}_m = \boldsymbol{\theta}_0 + \boldsymbol{S}\tilde{\boldsymbol{\delta}}_m^*$ and $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \boldsymbol{\delta}_t = \boldsymbol{\theta}_0 + \boldsymbol{S}\tilde{\boldsymbol{\delta}}_t + \boldsymbol{\delta}_t^\perp$, we have

$$\boldsymbol{\theta}_m - \boldsymbol{\theta}_t = \boldsymbol{S}\big(\tilde{\boldsymbol{\delta}}_m^* - \tilde{\boldsymbol{\delta}}_t\big) - \boldsymbol{\delta}_t^\perp. \tag{33}$$

Let $\Delta_t \triangleq \tilde{\boldsymbol{\delta}}_m^* - \tilde{\boldsymbol{\delta}}_t$. Substituting equation 33 into the quadratic form and using $\boldsymbol{S}^\top \boldsymbol{H}_t \boldsymbol{S} = \tilde{\boldsymbol{H}}_t$,

$$\begin{aligned}&(\boldsymbol{\theta}_m - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t) \\ &= \Delta_t^\top \tilde{\boldsymbol{H}}_t \Delta_t - 2\,\Delta_t^\top \boldsymbol{S}^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp + (\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp.\end{aligned} \tag{34}$$

The first and third terms are exactly the per-task contributions to $\mathcal{V}_S$ and $\mathcal{R}_S$. The cross term we bound via two applications of Cauchy–Schwarz. First, the $\boldsymbol{H}_t$-Cauchy–Schwarz inequality (valid because $\boldsymbol{H}_t \succeq 0$),

$$\begin{aligned}&\big|\Delta_t^\top \boldsymbol{S}^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp\big| = \\ &\big|(\boldsymbol{S}\Delta_t)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp\big| \le \sqrt{\Delta_t^\top \tilde{\boldsymbol{H}}_t \Delta_t}\,\sqrt{(\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp}.\end{aligned} \tag{35}$$

Multiplying by λ_t, summing over t, and applying the standard Cauchy–Schwarz inequality on the resulting sum,

$$\begin{aligned}\sum_{t=1}^{T} \lambda_t \big|\Delta_t^\top \boldsymbol{S}^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp\big| &\le \sum_{t=1}^{T} \sqrt{\lambda_t\,\Delta_t^\top \tilde{\boldsymbol{H}}_t \Delta_t}\,\sqrt{\lambda_t\,(\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp} \\ &\le \sqrt{\sum_{t=1}^{T} \lambda_t\,\Delta_t^\top \tilde{\boldsymbol{H}}_t \Delta_t}\,\sqrt{\sum_{t=1}^{T} \lambda_t\,(\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp} \\ &= \sqrt{\mathcal{V}_S}\,\sqrt{\mathcal{R}_S}.\end{aligned} \tag{36}$$

Combining all three terms above, we have

$$\begin{aligned}\sum_{t=1}^{T} \lambda_t (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t (\boldsymbol{\theta}_m - \boldsymbol{\theta}_t) &\le \mathcal{V}_S + 2\sqrt{\mathcal{V}_S}\sqrt{\mathcal{R}_S} + \mathcal{R}_S \\ &= \big(\sqrt{\mathcal{V}_S} + \sqrt{\mathcal{R}_S}\big)^2.\end{aligned} \tag{37}$$

Plugging this into the weighted sum of equation 32 and absorbing the per-task remainders into $R_3 = \sum_t \lambda_t R_3^{(t)} = \mathcal{O}(\max_t \|\boldsymbol{\theta}_m - \boldsymbol{\theta}_t\|^3)$ yields

$$\sum_{t=1}^{T} \lambda_t \big[\mathcal{L}_t(\boldsymbol{\theta}_m) - \mathcal{L}_t(\boldsymbol{\theta}_t)\big] \le \tfrac{1}{2}\big(\sqrt{\mathcal{V}_S} + \sqrt{\mathcal{R}_S}\big)^2 + R_3, \tag{38}$$

which is equation 15. □

Remark. Equality in the bound is attained when, for every t with $\lambda_t > 0$, the parallel error $\boldsymbol{S}\Delta_t$ and the residual $\boldsymbol{\delta}_t^\perp$ are co-linear under the metric $\boldsymbol{H}_t$ and the per-task ratios $\sqrt{\Delta_t^\top \tilde{\boldsymbol{H}}_t \Delta_t}/\sqrt{(\boldsymbol{\delta}_t^\perp)^\top \boldsymbol{H}_t \boldsymbol{\delta}_t^\perp}$ are constant. The bound is therefore tight rather than merely an order-of-magnitude estimate.

B.4.3. Proof of Theorem 3 (Curvature Advantage)

Proof. Let $\boldsymbol{b} \triangleq \sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t$, so the curvature-aware Fréchet mean of Theorem 1 is $\tilde{\boldsymbol{\delta}}_H = \bar{\boldsymbol{H}}^{-1}\boldsymbol{b}$. From the quadratic expansion in equation 28,

$$\mathcal{F}(\tilde{\boldsymbol{\delta}}_H) = \boldsymbol{b}^\top \bar{\boldsymbol{H}}^{-1} \bar{\boldsymbol{H}}\, \bar{\boldsymbol{H}}^{-1}\boldsymbol{b} - 2\boldsymbol{b}^\top \bar{\boldsymbol{H}}^{-1}\boldsymbol{b} + C = -\boldsymbol{b}^\top \bar{\boldsymbol{H}}^{-1}\boldsymbol{b} + C, \tag{39}$$

where $C \triangleq \sum_t \lambda_t \tilde{\boldsymbol{\delta}}_t^\top \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t$ is independent of $\tilde{\boldsymbol{\delta}}$. Substituting the flat-geometry choice $\tilde{\boldsymbol{\delta}}_I = \bar{\boldsymbol{\delta}} = \sum_t \lambda_t \tilde{\boldsymbol{\delta}}_t$,

$$\mathcal{F}(\tilde{\boldsymbol{\delta}}_I) = \bar{\boldsymbol{\delta}}^\top \bar{\boldsymbol{H}}\, \bar{\boldsymbol{\delta}} - 2\bar{\boldsymbol{\delta}}^\top \boldsymbol{b} + C. \tag{40}$$

Subtracting equation 39 from equation 40,

$$\begin{aligned}\mathcal{F}(\tilde{\boldsymbol{\delta}}_I) - \mathcal{F}(\tilde{\boldsymbol{\delta}}_H) &= \bar{\boldsymbol{\delta}}^\top \bar{\boldsymbol{H}}\, \bar{\boldsymbol{\delta}} - 2\bar{\boldsymbol{\delta}}^\top \boldsymbol{b} + \boldsymbol{b}^\top \bar{\boldsymbol{H}}^{-1}\boldsymbol{b} \\ &= \big(\bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} - \boldsymbol{b}\big)^\top \bar{\boldsymbol{H}}^{-1} \big(\bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} - \boldsymbol{b}\big),\end{aligned} \tag{41}$$

where the second equality is the identity $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$ valid for any $\boldsymbol{A} \succ 0$.

It remains to show that $\bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} - \boldsymbol{b} = -\boldsymbol{c}$ where $\boldsymbol{c} = \sum_t \lambda_t(\tilde{\boldsymbol{H}}_t - \bar{\boldsymbol{H}})(\tilde{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}})$. Expanding $\boldsymbol{c}$ and using $\sum_t \lambda_t = 1$,

$$\begin{aligned}\boldsymbol{c} &= \sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t - \Big(\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t\Big)\bar{\boldsymbol{\delta}} - \bar{\boldsymbol{H}} \sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{\delta}}_t + \bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} \sum_{t=1}^{T} \lambda_t \\ &= \boldsymbol{b} - \bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} - \bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} + \bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} = \boldsymbol{b} - \bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}}.\end{aligned} \tag{42}$$

Hence $\bar{\boldsymbol{H}}\,\bar{\boldsymbol{\delta}} - \boldsymbol{b} = -\boldsymbol{c}$, and the quadratic form is invariant under sign,

$$\mathcal{F}(\tilde{\boldsymbol{\delta}}_I) - \mathcal{F}(\tilde{\boldsymbol{\delta}}_H) = \boldsymbol{c}^\top \bar{\boldsymbol{H}}^{-1} \boldsymbol{c} \;\geq\; 0, \tag{43}$$

where non-negativity follows from $\bar{\boldsymbol{H}}^{-1} \succ 0$. Equality holds if and only if $\boldsymbol{c} = \boldsymbol{0}$, *i.e. iff the curvature deviations and task-vector deviations are uncorrelated under the weighting* $\{\lambda_t \mid t = 1, \ldots, T\}$. □

Remark. Three sufficient conditions for $\boldsymbol{c} = \boldsymbol{0}$ are immediate: *(a)* all $\tilde{\boldsymbol{H}}_t$ are equal (so $\tilde{\boldsymbol{H}}_t - \bar{\boldsymbol{H}} = \boldsymbol{0}$); *(b)* all $\tilde{\boldsymbol{\delta}}_t$ are equal (so $\tilde{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}} = \boldsymbol{0}$); *(c)* the first moment of the deviations $(\tilde{\boldsymbol{H}}_t - \bar{\boldsymbol{H}})(\tilde{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}})$ vanishes by symmetry. Cases (a) and (b) are degenerate (homogeneous curvature, identical task vectors); case (c) is the empirically interesting one and is precisely what the diagnostic $\eta = \boldsymbol{c}^\top \bar{\boldsymbol{H}}^{-1} \boldsymbol{c}$ measures, evaluated in the projected $p \times p$ space at $\mathcal{O}(p^3)$ cost.

B.4.4. Proof of Proposition 2 (Subsumptions)

Proof. We verify each of the five cases by direct substitution into the closed-form Fréchet mean

$$\tilde{\boldsymbol{\delta}}^*_m = \Big(\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t\Big)^{-1} \Big[\sum_{t=1}^{T} \lambda_t \tilde{\boldsymbol{H}}_t \tilde{\boldsymbol{\delta}}_t\Big], \quad \boldsymbol{\theta}_m = \boldsymbol{\theta}_0 + \boldsymbol{S}\,\tilde{\boldsymbol{\delta}}^*_m, \tag{44}$$

under the convention $\sum_{t=1}^{T} \lambda_t = 1$.

Case 1 (Task Arithmetic). Set $\boldsymbol{S} = \boldsymbol{I}_m$ and $\tilde{\boldsymbol{H}}_t = \boldsymbol{I}_m$. Then $\tilde{\boldsymbol{\delta}}_t = \boldsymbol{\delta}_t$ and equation 44 reduces to

$$\tilde{\boldsymbol{\delta}}^*_m = \Big(\sum_t \lambda_t\Big)^{-1} \sum_t \lambda_t \boldsymbol{\delta}_t = \sum_t \lambda_t \boldsymbol{\delta}_t, \tag{45}$$

so $\boldsymbol{\theta}_m = \boldsymbol{\theta}_0 + \sum_t \lambda_t \boldsymbol{\delta}_t$, which is exactly the Task Arithmetic update [7].

Case 2 (Fisher Averaging). Set $\boldsymbol{S} = \boldsymbol{I}_m$ and $\tilde{\boldsymbol{H}}_t = \mathrm{diag}(\boldsymbol{F}_t)$. Both $\sum_t \lambda_t \mathrm{diag}(\boldsymbol{F}_t)$ and $\sum_t \lambda_t \mathrm{diag}(\boldsymbol{F}_t)\boldsymbol{\delta}_t$ are coordinate-wise diagonal, so equation 44 gives

$$(\tilde{\boldsymbol{\delta}}^*_m)_i = \frac{\sum_{t=1}^{T} \lambda_t F_{t,i}\, \delta_{t,i}}{\sum_{t=1}^{T} \lambda_t F_{t,i}}, \qquad i = 1, \ldots, m, \tag{46}$$

which is the per-coordinate Fisher-weighted average of the task vectors [11].

Case 3 (Daheim et al.). Set $\boldsymbol{S} = \boldsymbol{I}_m$ and $\tilde{\boldsymbol{H}}_t = \boldsymbol{H}_t$ (full Hessian, possibly approximated). Then

$$\tilde{\boldsymbol{\delta}}^*_m = \Big(\sum_t \lambda_t \boldsymbol{H}_t\Big)^{-1} \sum_t \lambda_t \boldsymbol{H}_t \boldsymbol{\delta}_t, \tag{47}$$

which yields the full-space curvature-aware merge of [3].

Case 4 (TSV-M, projection scaffold). Let $\boldsymbol{S}$ be the orthonormal basis for the top-k left singular subspace of the stacked task vectors $[\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_T]$ (the joint-SVD construction of [5]), and set $\tilde{\boldsymbol{H}}_t = \boldsymbol{I}_p$. Then equation 44 simplifies to $\tilde{\boldsymbol{\delta}}^*_m = \sum_t \lambda_t \tilde{\boldsymbol{\delta}}_t = \sum_t \lambda_t \boldsymbol{S}^\top \boldsymbol{\delta}_t$, and lifting back to ambient space yields

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_0 + \boldsymbol{S}\boldsymbol{S}^\top \Big(\sum_t \lambda_t \boldsymbol{\delta}_t\Big), \tag{48}$$

that is, Task Arithmetic projected onto the rank-k task-singular subspace. This captures the *subspace-projection scaffold* underlying TSV-M. Algorithm 1 of [5] additionally performs a per-layer orthogonal Procrustes whitening of the stacked per-task singular vectors. This decorrelation is nonlinear in $\{\boldsymbol{\delta}_t \mid t = 1, \ldots, T\}$ and therefore lies outside the quadratic Fréchet template of equation 44. Accordingly, our hierarchy subsumes TSV-M's curvature-agnostic subspace-projection choice but leaves the Procrustes decorrelation as an orthogonal enhancement. Empirically, we still compare against Gargiulo's complete algorithm in Sec. 4.

Case 5 (EpiMer). EpiMer adopts the same per-task factored basis as TSV-M (equation 10), but replaces the identity metric $\tilde{\boldsymbol{H}}_t = \boldsymbol{I}_p$ of Case 4 with the projected curvature $\tilde{\boldsymbol{H}}_t = \boldsymbol{S}^\top \boldsymbol{H}_t \boldsymbol{S}$, where $\boldsymbol{S}$ collects the vec'd rank-1 atoms $\{\boldsymbol{U}_{\perp,i}\boldsymbol{V}_{\perp,i}^\top\}$ as columns. Substituting this curvature-aware metric into the quadratic Fréchet template equation 44 recovers equation 13 verbatim. On the per-task tagged basis, however, the Fréchet *mean* can suffer from a magnitude collapse pathology because each rank-1 atom is owned by exactly one task. In practice, our proposed EpiMer in Sec. 3.2.3 uses the matrix-weighted sum aggregator of equation 14, which is exactly αT times the Fréchet mean and reshapes the per-task contributions through the same curvature solve. In the homogeneous-curvature limit $\tilde{\boldsymbol{H}}_t \equiv \tilde{\boldsymbol{H}}$, equation 14 collapses to $\alpha \sum_t \tilde{\boldsymbol{\delta}}_t$, recovering TSV-M aggregation. Nevertheless, whenever the per-task curvatures differ, our matrix solve can yield a strictly different and empirically better merge. See Appendix D.2 for a more detailed comparison.

The above shows that in each case, the existing method is recovered by an appropriate $(\boldsymbol{S}, \tilde{\boldsymbol{H}}_t)$. Proof completes. □

Table 6. Per-task top-1 accuracy on the eight image classification tasks, expanded from Tab. 2. Best among merging methods is **bold**, second best is underlined. All baselines use $\lambda_t = 1/T$, so AM and TA collapse to the same merged delta and are listed jointly. TSV-M and EpiMer are reported at $k = 32$ and the per-method optimal global rescaling α from Sec. 4.3: TSV-M uses $\alpha = 0.70$ on ViT-B/32 and $\alpha = 1.0$ on ViT-B/16 and ViT-L/14; EpiMer uses $\alpha = 1.0$ on all three backbones. On ViT-L/14, the average accuracy of TSV-M and EpiMer differ by 0.06 percentage points (EpiMer 0.9065 vs TSV-M 0.9059); both display as .906 at the table's three-decimal precision.

Backbone	Methods	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN	Avg.
Baselines										
ViT-B/32	Linear-Probing	.597	.436	.451	.326	.482	.602	.631	.316	.480
	Fine-tuning	.806	.791	.999	.990	.997	.961	.755	.975	.909
ViT-B/16	Linear-Probing	.647	.441	.544	.434	.517	.663	.655	.520	.553
	Fine-tuning	.891	.822	.999	.992	.998	.968	.785	.978	.929
ViT-L/14	Linear-Probing	.779	.549	.613	.507	.763	.713	.682	.584	.649
	Fine-tuning	.934	.839	.999	.993	.997	.973	.824	.981	.943
Flat-geometry methods										
ViT-B/32	AM [15] / TA [7]	.575	.506	.729	.529	.871	.715	.650	.647	.653
	TIES [17]	.572	.532	.871	.741	.983	.696	.569	.835	.725
	TSV-M [5]	.675	.657	.959	.881	.990	.835	.674	.903	<u>.822</u>
ViT-B/16	AM [15] / TA [7]	.672	.504	.760	.599	.941	.759	.677	.765	.710
	TIES [17]	.693	.559	.838	.804	.989	.778	.637	.896	.774
	TSV-M [5]	.802	.711	.977	.937	.994	.860	.686	.955	<u>.865</u>
ViT-L/14	AM [15] / TA [7]	.797	.619	.901	.713	.967	.826	.715	.786	.791
	TIES [17]	.845	.675	.950	.907	.991	.873	.737	.890	.859
	TSV-M [5]	.899	.773	.985	.965	.995	.909	.762	.960	<u>.906</u>
Curvature-aware methods										
ViT-B/32	Fisher [11]	.751	.598	.409	.338	.542	.620	.707	.351	.539
	EpiMer	.672	.694	.965	.914	.993	.834	.665	.926	**.833**
ViT-B/16	Fisher [11]	.824	.610	.503	.408	.678	.688	.740	.547	.625
	EpiMer	.804	.722	.977	.944	.993	.869	.710	.941	**.870**
ViT-L/14	Fisher [11]	.874	.760	.646	.507	.755	.760	.788	.671	.720
	EpiMer	.901	.771	.984	.973	.995	.914	.763	.949	**.906**

C. Implementation Details for Experiments

C.1. Datasets

This section provides further details on the datasets used in our comparative experiments. Following prior works [7, 14, 15], we evaluated model merging on eight image classification tasks:

- **Stanford Cars (Cars) [8]** is an image classification dataset with 16,185 images of 196 classes of cars. The dataset is split into 8,144 training images and 8,041 testing images. Each class appears with the exact frequency in the training and testing sets.
- **Describable Texture Dataset (DTD) [2]** comprises of 3,760 training images and 1,880 testing image. Each image is labeled with one of 47 describable textures.
- **EuroSAT [6]** is a dataset for land use classification with 27,000 labeled Sentinel-2 satellite images. The dataset is split into 21,000 training and 6,000 testing images.
- **German Traffic Sign Recognition Benchmark Dataset (GTSRB) [13]** contains 39,270 images with 43 classes of traffic signs. The dataset is split into 26,640 training images and 12,630 testing images.
- **MNIST [9]** is a dataset of 10 handwritten digits, containing 60,000 training images and 10,000 testing images, with balanced presence of each digit in both splits.
- **Remote Sensing Image Scene Classification Dataset (RESISC45) [1]** is a remote sensing image classification dataset with 25,200 images of 45 classes of scenarios. The dataset is split into 18,900 training images and 6,300 testing images, with around 700 images per class.
- **Scene Understanding Benchmark (SUN397) [16]** contains 397 categories of scenes with 108,754 images in total. The number of images varies across categories, but each category contains at least 100 images.
- **Street View House Numbers (SVHN) [12]** is a dataset of 10 classes of housing number digits extracted from the

Google Street View images. The dataset contains $73,257$ training images and $26,032$ testing images.

C.2. Model Checkpoints

We use the publicly released, per-task fine-tuned CLIP-ViT-B/16, CLIP-ViT-B/32, and CLIP-ViT-L/14 checkpoints from [7], which were obtained by linear probing followed by full fine-tuning on each of the eight datasets in Appendix C.1. The checkpoints ship only the converged parameters $\boldsymbol{\theta}_t$, not the optimizer states; we therefore reconstruct the empirical Fisher diagonal $\boldsymbol{v}_t$ used by equation 9 via a single forward+backward pass over each task's training set at $\boldsymbol{\theta}_t$. Refer to Appendix D.4 for more details.

C.3. Baselines Implementation

For the baseline methods, we implement Arithmetic Mean (AM) [15], Task Arithmetic [7], Fisher-weighted Averaging [11], and TIES-Merging [17]. We follow the original papers for the implementation details of these methods. In particular, we use the same merging coefficient $\lambda_t = 0.125$ across all tasks for Task Arithmetic and TIES-Merging. We retain the top-20% of parameters and select signs by frequency in TIES-Merging. To facilitate a fair comparison, we leverage the square of the first momentum to construct a diagonal Fisher Information Matrix $\boldsymbol{F}_t \leftarrow \text{diag}(\boldsymbol{u}^2)$ in Fisher-weighted Averaging, following the definitions in the original paper.

D. Additional Results

D.1. Per-Task Accuracy Breakdown

Table 6 expands Tab. 2 from the main text with the per-task top-1 accuracy on every dataset, for all baselines and EpiMer, across the three CLIP-ViT backbones. Averages in the last column match the main-text table exactly.

D.2. Global Rescaling α Sensitivity

This section reports the full α sweep at $k \in \{4, 16, 32\}$ on each backbone and supplements the $k = 32$ slice that appears in Sec. 4.3. We sweep $\alpha \in \{0.20, 0.30, 0.40, 0.50, 0.70, 1.00\}$ on TSV-M [5] and the matrix-weighted aggregator of equation 14; Tabs. 7 to 9 report average accuracy across the eight tasks.

Beyond the $k = 32$ findings reported in Sec. 4.3, the lower-rank slices in Tabs. 7 to 9 establish that EpiMer dominates TSV-M at every rank tested at the per-method optimum: by $+0.79$, $+0.20$, $+1.10$ percentage points at $k = 4, 16, 32$ on ViT-B/32; by $+0.52$, $+0.66$, $+0.48$ percentage points on ViT-B/16; and by $+0.49$, $+0.25$, $+0.06$ percentage points on ViT-L/14. The L/14 margin shrinks at high rank because the per-task curvatures become more uniform on the larger backbone, pushing equation 14 toward its homogeneous-curvature limit (TSV-M); even there, EpiMer remains strictly ahead.

Table 7. ViT-B/32 α-sensitivity sweep. Best per column is **bold**.

	TSV-M			EpiMer		
α	$k=4$	$k=16$	$k=32$	$k=4$	$k=16$	$k=32$
0.20	.573	.624	.630	.573	.609	.601
0.30	.627	.689	.699	.633	.677	.670
0.40	.667	.737	.750	.674	.727	.724
0.50	.693	.770	.787	.703	.763	.764
0.70	**.720**	**.800**	**.822**	.728	.800	.812
1.00	.706	.788	.820	**.709**	**.802**	**.833**

Table 8. ViT-B/16 α-sensitivity sweep. Best per column is **bold**.

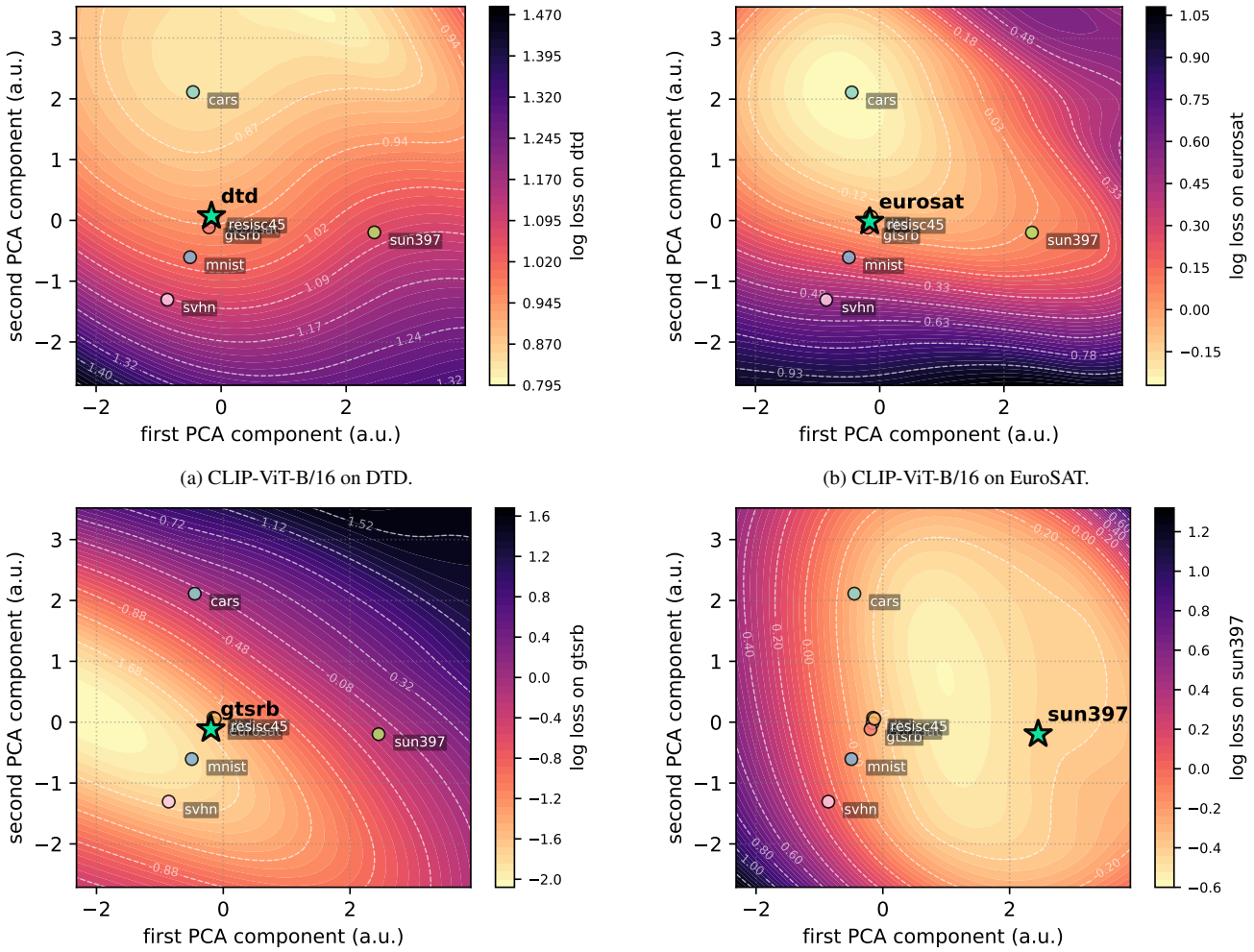
	TSV-M			EpiMer		
α	$k=4$	$k=16$	$k=32$	$k=4$	$k=16$	$k=32$
0.20	.636	.681	.688	.638	.672	.666
0.30	.676	.733	.747	.680	.725	.724
0.40	.703	.774	.792	.709	.766	.771
0.50	.725	.803	.822	.730	.796	.804
0.70	**.755**	.834	.857	**.760**	.832	.846
1.00	.754	**.836**	**.865**	.751	**.843**	**.870**

Table 9. ViT-L/14 α-sensitivity sweep. Best per column is **bold**.

	TSV-M			EpiMer		
α	$k=4$	$k=16$	$k=32$	$k=4$	$k=16$	$k=32$
0.20	.730	.763	.772	.733	.761	.766
0.30	.759	.801	.816	.764	.801	.808
0.40	.783	.830	.849	.789	.830	.841
0.50	.801	.850	.870	.809	.849	.863
0.70	.826	.875	.895	.832	.875	.890
1.00	**.837**	**.887**	**.906**	**.842**	**.890**	**.906**

D.3. Loss Landscape Visualizations

To support the connected-basin assumption (Assumption 3) used by EpiMer, we visualize the loss landscape of fine-tuned CLIP-ViT models around their fine-tuned positions. Following prior works [4, 10], we apply PCA on the matrix $\boldsymbol{\Theta}$ that stacks the parameter vectors of all fine-tuned models for a given backbone, take the top two principal directions δ, η, and evaluate the per-task loss on the grid $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \alpha\delta + \beta\eta$, where $\bar{\boldsymbol{\theta}}$ is the mean of all fine-tuned models. Figure 4 reports this visualization for CLIP-ViT-B/16 on DTD, EuroSAT, GTSRB, and SUN397, all of which exhibit a low-change-of-loss basin around the fine-tuned model with high-loss barriers between the per-task minima—qualitatively consistent with Assumption 3.



(a) CLIP-ViT-B/16 on DTD.

(b) CLIP-ViT-B/16 on EuroSAT.

(c) CLIP-ViT-B/16 on GTSRB.

(d) CLIP-ViT-B/16 on SUN397.

Figure 4. Loss landscape visualization for CLIP-ViT-B/16 fine-tuned on (a) DTD, (b) EuroSAT, (c) GTSRB, and (d) SUN397 datasets.

D.4. Data-Efficiency for Empirical Fisher Diagonal

This section reports the full setup used by the empirical Fisher robustness study summarized in Sec. 4.5; the headline figure is Fig. 3 in the main paper.

Setup. For each backbone and each subsample fraction $f \in \{0.5\%, 1\%, 5\%, 10\%, 25\%, 50\%, 100\%\}$ we shuffle the training set with a fixed seed, take the prefix of $\lfloor f \cdot N_t/B \rfloor$ batches with batch size $B = 64$, and accumulate the per-sample squared gradients on that prefix to obtain $\boldsymbol{v}_t^{(f)}$. The per-sample gradients are materialized in chunks of 8 samples via `jax.vmap` so that peak memory remains $\mathcal{O}(\text{chunk} \cdot |\boldsymbol{\theta}|)$ rather than $\mathcal{O}(B \cdot |\boldsymbol{\theta}|)$. We then plug $\{\boldsymbol{v}_t^{(f)}\}_{t=1}^{8}$ into EpiMer at the optimum with $k = 32$, $\alpha = 1.0$, and `frechet_sum` aggregator from Appendix D.2 to evaluate the merged backbone on all eight tasks. The same per-task classifier heads are used as in Tab. 2.

D.5. Computational Cost

Table 10 reports the end-to-end merge wall-clock time at the headline setting $k = 32$. The non-spectral baselines (AM/TA, TIES, Fisher Averaging) complete in under seven seconds on every backbone because they operate in the full parameter space with simple per-coordinate aggregation. TSV-M and EpiMer instead pay a fixed per-backbone JIT-compilation cost for the per-leaf SVD kernel that builds the per-task tagged basis $\mathcal{B}^{(\ell)}$ from equation 10—$\sim$34 seconds on ViT-B/32 and ViT-B/16, and $\sim$109 seconds on ViT-L/14, consistent with L/14's $\sim 3\times$ larger parameter count. On top of that shared floor, EpiMer's per-leaf $p \times p$ matrix solve in equation 14 with $p = kT = 256$ at $k = 32, T = 8$ adds only 3.7, 4.8, and 11.1 seconds over TSV-M on ViT-B/32, ViT-B/16, and ViT-L/14, respectively—at most 14%

Table 10. End-to-end merge wall-clock time at the headline setting $k = 32$ on a single RTX 3090 Ti, reported as the median of at least four repeat runs on the eight-task CLIP-ViT benchmark. AM/TA, TIES, and Fisher-weighted Averaging operate in the full parameter space and do not trigger the spectral JIT path. TSV-M and EpiMer both construct the per-task tagged basis of Eq. (10) and therefore share the JIT-compilation cost of the per-leaf SVD kernel; EpiMer's per-leaf $p \times p$ matrix solve with $p = kT = 256$ adds only 3.7–11.1 seconds on top of that shared baseline. The per-task empirical Fisher diagonal used by Fisher Averaging and EpiMer is accumulated once per backbone and amortized across every subsequent merge attempt.

Method	**ViT-B/32**	**ViT-B/16**	**ViT-L/14**
AM / Task Arithmetic	0.9 s	0.9 s	1.3 s
TIES-Merging	3.3 s	3.4 s	6.1 s
Fisher-weighted Averaging	1.7 s	1.8 s	3.8 s
TSV-M	34.2 s	34.2 s	108.7 s
EpiMer	37.9 s	39.0 s	119.8 s

of the corresponding TSV-M baseline. The per-task empirical Fisher diagonal $\boldsymbol{v}_t$ needs one forward+backward pass per batch, is robust to a tiny data fraction (Sec. 4.5), and is shared with Fisher-weighted Averaging, so it is amortized once per backbone across all downstream merge attempts.

References

[1] 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.

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching, 2024.

[4] Ruiqi Ding, Tao Li, and Xiaolin Huang. Better loss landscape visualization for deep neural networks with trajectory information. In *Proceedings of the 15th Asian Conference on Machine Learning*, pages 311–326. PMLR, 2024.

[5] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging, 2025.

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019.

[7] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic, 2023.

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[9] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[10] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

[11] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, pages 17703–17716. Curran Associates, Inc., 2022.

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

[13] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks*, pages 1453–1460, 2011.

[14] 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. 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*, pages 23965–23998. PMLR, 2022.

[15] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, 2022.

[16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[17] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, pages 7093–7115. Curran Associates, Inc., 2023.