Rebuttal Tables & Figures

Decentralized Instruction Tuning: Conflict-Aware Splitting and Weight Merging

Table R.1: Controlled 3B comparison (Qwen2.5-VL-3B-Instruct, 136 Vision-FLAN tasks). Avg. = mean over 8 benchmarks. Joint (1 ep) and MERIT-3D are averaged over 5 seeds; others over 3 seeds.

Method	**General MCQA**		**Preference**		**Text-Rich VQA**		**Reasoning**		**Avg.**
	Seed	MMB	LLaVA-W	MMVet	TextVQA	AI2D	MathV	MMMU	
Base model	66.8	79.7	53.2	34.0	61.2	63.8	29.6	41.2	53.7
Joint (0.5 ep)	67.4	79.4	40.2	33.1	67.2	60.9	31.4	41.6	52.7
Joint (1 ep)	69.2	80.5	41.9	36.4	68.0	62.6	34.2	41.9	54.3
Joint (2 ep)	70.0	81.4	42.8	37.6	63.4	62.5	36.5	43.0	54.7
Random (2 groups)	69.4	80.1	44.5	34.7	70.4	62.7	34.0	41.2	54.6
Random (4 groups)	70.4	81.0	40.6	34.7	70.4	63.1	34.0	40.8	54.4
Random (8 groups)	69.5	79.9	42.2	35.0	73.7	61.7	33.5	40.5	54.5
Conflict-ind. (2)	70.7	80.6	42.6	35.4	70.0	62.9	34.4	42.3	54.9
Uniform soup (2)	70.2	81.1	45.0	35.3	68.9	63.4	36.8	42.2	55.4
Uniform soup (3)	70.1	81.1	42.3	36.3	68.8	63.1	35.8	42.5	55.0
Uniform soup (4)	70.2	81.1	41.8	36.3	68.4	63.4	35.9	42.2	54.9
MERIT-1D (2 gr.)	71.0	80.0	43.1	35.0	72.4	62.1	36.5	41.4	55.2
MERIT-2D (4 gr.)	70.8	78.4	47.4	36.6	74.1	61.5	36.0	40.7	55.7
MERIT-3D (8 gr.)	70.5	80.1	52.0	37.7	75.2	62.5	35.4	42.7	**57.0**

Table R.2: PCA-based MERIT vs. K-means clustering. All methods use identical training budgets, hyperparameters, and token-weighted averaging. Evaluated on 11 benchmarks.

Method	K	MMB	MME	Seed	MathV	Hall	AI2D	MMVet	MIA	MMMU	TVQA	DocVQA	Avg.
MERIT-1D	2	80.0	82.6	71.0	36.5	54.4	62.1	35.0	34.4	41.4	72.4	50.0	56.3
MERIT-2D	4	78.4	83.0	70.8	36.0	53.0	61.5	36.6	37.3	40.7	74.1	50.8	56.6
MERIT-3D	8	80.1	81.4	70.5	35.4	54.2	62.5	37.7	39.2	42.7	75.2	50.4	**57.2**
K-means	2	80.1	79.8	70.2	34.3	55.1	62.1	36.9	34.4	38.6	71.2	49.3	55.8
K-means	4	79.7	80.6	70.9	35.2	56.2	61.9	34.4	35.4	42.3	73.7	50.2	56.4
K-means	8	79.4	79.7	70.4	35.1	53.6	62.1	35.1	39.4	41.1	75.0	50.4	56.5

Table R.3: Per-seed results for MERIT-3D (8 groups) over 5 independent runs. All runs use the same 200-sample calibration-based PCA partition; only the training seed differs.

Run	Seed	MMB	LLaVA-W	MMVet	TVQA	AI2D	MathV	MMMU	Avg.
TEST1	70.4	80.3	52.2	38.1	75.1	63.1	35.7	42.1	57.1
TEST2	70.6	80.1	53.7	38.8	75.0	62.1	35.9	43.3	57.4
TEST3	70.5	79.5	49.9	36.5	75.1	62.8	34.3	42.8	56.4
TEST4	70.5	80.2	52.7	38.0	75.3	61.5	34.7	42.3	56.9
TEST5	70.3	80.5	51.7	37.1	75.4	63.2	36.6	43.2	57.3
Mean	70.5	80.1	52.0	37.7	75.2	62.5	35.4	42.7	**57.0**
Std	0.1	0.4	1.4	0.9	0.2	0.7	0.9	0.5	0.4

Table R.4: Per-seed results for Joint training (1 epoch) over 5 independent runs. MERIT-3D outperforms in all 5 pairs ($p = 0.03125$, one-sided Wilcoxon signed-rank test).

Run	Seed	MMB	LLaVA-W	MMVet	TVQA	AI2D	MathV	MMMU	Avg.
TEST1	68.6	80.1	41.8	38.5	68.3	62.4	35.0	40.9	54.5
TEST2	69.6	80.6	41.7	36.0	67.9	62.4	34.0	41.9	54.3
TEST3	69.1	80.4	41.5	34.9	68.0	62.2	34.2	42.1	54.1
TEST4	69.3	80.7	42.3	36.8	67.9	63.6	33.8	42.5	54.6
TEST5	69.5	80.5	42.1	35.8	67.7	62.6	34.0	41.9	54.3
Mean	69.2	80.5	41.9	36.4	68.0	62.6	34.2	41.9	54.3
Std	0.4	0.2	0.3	1.3	0.2	0.5	0.5	0.6	0.2

Table R.5: Gradient calibration sensitivity across 100 Vision-FLAN tasks. Each cell reports the cosine similarity between the calibration-subset gradient and the full 1000-sample reference gradient, averaged over 100 tasks. At n=200 (used in all experiments), fidelity is already 0.847 with diminishing returns beyond (right panel).

n	Mean cos	Std
50	0.683	0.176
100	0.762	0.147
150	0.814	0.121
200	0.847	0.106
250	0.871	0.095
300	0.891	0.087
400	0.920	0.076
500	0.940	0.070
700	0.968	0.062
1000	1.000	0.000

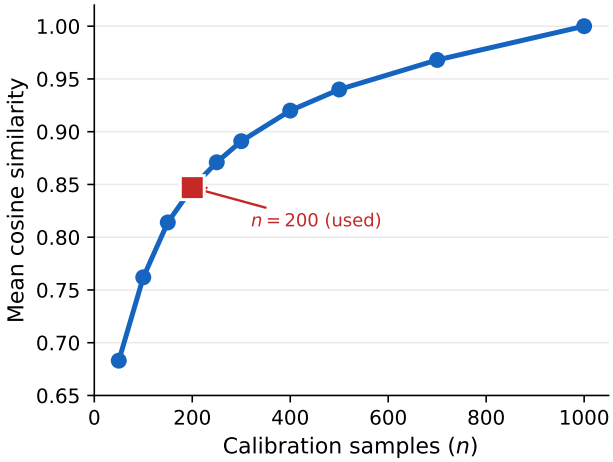

Table R.6: SGD learning-rate sweep (8-benchmark average). Joint training batch size selected via grid search over $\{4, 8, 16, 32, 64, 128\}$ (best: 32); MERIT branches use batch size 8. LR and scheduler are identical across all methods. At practical LRs ($\eta \geq 8\times10^{-5}$), all MERIT variants match or exceed Joint. Joint performance plateaus and slightly drops at the largest η, while MERIT continues to improve (right panel). MERIT-3D degrades at very small η, consistent with $\mathcal{G}_{\text{var}} \propto \eta^2 \to 0$.

	3e-5	4e-5	8e-5	2e-4	4e-4
Joint (1 ep)	17.7	30.8	51.7	53.4	53.3
MERIT (1D)	47.6	51.3	52.9	53.1	**53.7**
MERIT (2D)	18.2	32.3	51.5	53.2	**53.8**
MERIT (3D)	17.5	17.3	34.0	52.6	**53.7**

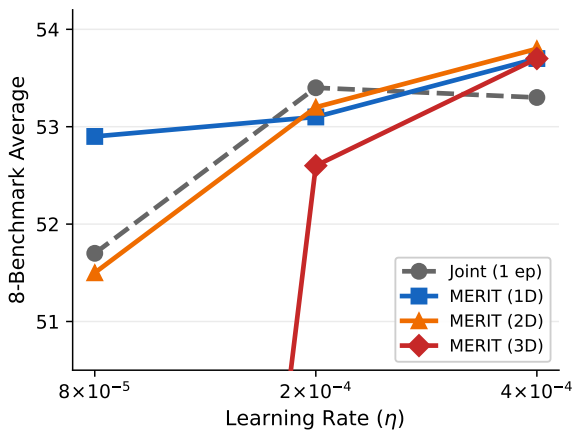

Table R.7: 7B per-benchmark results (Qwen2.5-7B-Instruct, 176 tasks, 1.6M examples). MERIT-2D gains are concentrated on open-ended reasoning benchmarks (MMVet, MathVista).

Benchmark	Joint FFT	MERIT-2D	Δ
SeedBench	70.1	70.6	+0.5
MMBench	76.2	75.6	−0.6
LLaVA-W	58.8	59.2	+0.4
MMVet	32.5	35.1	+2.6
TextVQA	73.8	72.4	−1.4
AI2D	58.6	58.4	−0.2
MathVista	32.9	35.4	+2.5
MMMU	36.4	36.1	−0.3

Table R.8: Text-only instruction tuning (Qwen2.5-3B, 66 FLAN tasks). MERIT-2D achieves the best average, demonstrating generality beyond multimodal settings.

Method	**Knowledge**		**Commonsense**		**Inference**		**Problem Solving**		Avg.
	MMLU	GPQA	HSwag	WGrande	BoolQ	XNLI	ARC-C	HEval	
Base	65.5	29.5	74.5	70.1	77.3	42.9	55.6	37.8	56.7
Joint (1 ep)	65.8	30.1	74.0	69.9	83.9	41.9	55.5	39.7	57.6
Joint (2 ep)	66.1	30.6	76.3	69.2	82.9	41.9	56.5	42.7	58.3
Random (2)	65.8	29.5	74.4	70.3	85.3	42.4	55.1	42.4	58.1
Random (4)	66.0	30.3	74.7	70.3	85.7	42.5	54.9	42.1	58.2
Conflict-ind. (2)	65.8	28.7	74.0	69.4	85.3	42.2	53.8	42.1	57.7
Uniform soup (2)	65.9	30.6	74.0	70.2	83.9	41.9	55.2	40.2	57.7
Uniform soup (3)	65.7	30.8	74.0	70.2	84.2	42.1	55.7	39.0	57.7
MERIT-1D (2)	66.0	30.6	74.4	69.9	86.3	42.1	54.2	41.2	58.1
MERIT-2D (4)	66.1	30.0	74.7	70.8	86.1	42.4	55.3	41.5	**58.4**

Table R.9: Raw-gradient PCA vs. cosine-similarity PCA. Both produce comparable performance, confirming that cosine-similarity PCA preserves the dominant conflict structure (Davis–Kahan bound; gradient norm CV = 0.17).

Method	Seed	MMB	LLaVA-W	MMVet	TVQA	AI2D	MathV	MMMU	Avg.
Raw PCA, 1D	70.9	79.9	41.5	35.0	72.5	62.7	36.2	41.3	55.0
Raw PCA, 2D	71.0	81.4	44.8	35.4	73.4	62.6	36.4	42.5	55.9
Raw PCA, 3D	70.6	81.2	40.0	35.4	75.4	62.2	33.6	41.0	54.9
Cos-sim PCA, 1D	71.0	80.0	43.1	35.0	72.4	62.1	36.5	41.4	55.2
Cos-sim PCA, 2D	70.8	78.4	47.4	36.6	74.1	61.5	36.0	40.7	55.7
Cos-sim PCA, 3D	70.5	80.1	52.0	37.7	75.2	62.5	35.4	42.7	**57.0**

Table R.10: Parameter displacement from shared initialization $\theta^{(0)}$ over training epochs. The merged model remains 2–3× closer to $\theta^{(0)}$ throughout, with the ratio widening monotonically.

Epoch	$\|\theta_{\text{joint}} - \theta^{(0)}\|_2$	$\|\bar{\theta}_{\text{merge}} - \theta^{(0)}\|_2$	Ratio
0.1	4.24	2.90	1.46×
0.5	13.73	5.65	2.43×
1.0	19.73	7.50	2.63×
2.0	28.15	10.11	2.78×
3.0	32.50	11.36	2.86×
4.0	34.18	11.79	2.90×
5.0	34.61	11.87	2.92×
6.0	34.65	11.88	2.92×

Table R.11: Training loss on the full mixture over epochs. The merged model has substantially higher training loss but achieves superior held-out performance (57.0 vs. 54.3), consistent with implicit regularization from merging.

Epoch	L_{joint}	L_{merged}	Gap
0.5	0.709	1.198	+0.489
1.0	0.560	1.172	+0.611
2.0	0.370	1.167	+0.797
3.0	0.205	1.202	+0.998
6.0	0.064	1.33	+1.27

Table R.12: Wall-clock time comparison. MERIT's end-to-end time is comparable to joint training even when branches run sequentially. In parallel deployment, MERIT is strictly faster.

Stage	3B (V100×8)	7B (A100×8)
Gradient extraction	1h 38m	1h 59m
Cosine similarity	28m	38m
PCA decomposition	<1s	<1s
Joint training (1 ep)	4h 22m	43h 18m
Joint training (2 ep)	8h 40m	–
MERIT-3D (sequential)	5h 24m	43h 39m
MERIT overhead	**+23%**	**+0.8%**

Table R.13: Computational comparison of conflict-resolution methods at T=136 tasks on a 3B model. PCGrad and GradNorm are infeasible at this scale; MERIT resolves conflicts in a one-time preprocessing step.

Method	When resolved	Per-step overhead	Communication
PCGrad	Per-step	$O(T^2)\times d$; ∼816 GB	All-reduce + T backpasses
GradNorm	Per-step	$O(T)$ forward	All-reduce + per-task loss
MERIT	One-time (∼2h)	Zero	Zero (one-shot merge)

Table R.14: MERIT + TIES-Merging on 7B checkpoints (density k=20%, 3 runs). TIES consistently degrades performance.

Method	**General MCQA**		**Preference**		**Text-Rich VQA**		**Reasoning**		**Avg.**
	Seed	MMB	LLaVA-W	MMVet	TextVQA	AI2D	MathV	MMMU	
Joint FFT	70.1	76.2	58.8	32.5	73.8	58.6	32.9	36.4	54.9
MERIT + Simple Avg	70.6	75.7	59.2	35.1	72.4	58.4	35.4	36.1	55.4
MERIT + TIES (run 1)	25.8	74.5	54.7	35.2	72.1	56.5	36.7	34.7	48.8
MERIT + TIES (run 2)	43.5	73.6	52.5	34.0	72.0	55.7	33.8	34.7	50.0
MERIT + TIES (run 3)	56.7	74.2	50.8	33.5	75.2	56.6	35.4	37.1	52.4

Table R.15: 7B results over three independent training seeds (Qwen2.5-7B-Instruct, 176 tasks). MERIT outperforms joint training in all three runs.

Seed	Method	**General MCQA**		**Preference**		**Text-Rich VQA**		**Reasoning**		**Avg.**
		Seed	MMB	LLaVA-W	MMVet	TextVQA	AI2D	MathV	MMMU	
Run 1	Joint FFT	70.1	76.2	58.8	32.5	73.8	58.6	32.9	36.4	54.9
	MERIT + Simple Avg	70.6	75.7	59.2	35.1	72.4	58.4	35.4	36.1	**55.4**
Run 2	Joint FFT	70.3	75.9	60.9	34.0	74.1	58.5	31.8	36.2	55.2
	MERIT + Simple Avg	70.8	75.5	63.0	34.1	72.0	58.6	31.9	36.2	**55.3**
Run 3	Joint FFT	70.1	75.9	60.8	34.1	77.1	58.6	31.7	35.9	55.5
	MERIT + Simple Avg	70.4	75.3	60.1	35.6	76.3	58.8	33.0	36.4	**55.7**

Table R.16: Linear Mode Connectivity (2D split, K=4). Barrier = $\max_\alpha[L(\theta(\alpha)) - ((1-\alpha)L(0) + \alpha L(1))]$ along 21-point interpolation ($\alpha \in [0, 1]$, step 0.05). All 10 barriers are exactly 0, confirming that branches remain in a shared flat basin with no loss barrier.

Path	$L(\alpha$=0)	$\min_\alpha L(\alpha)$	$L(\alpha$=1)	Barrier
E1: Branch ↔ Branch (pairwise)				
b0 ↔ b1	2.306	1.901 (α=0.70)	1.993	0.0
b0 ↔ b2	2.306	2.244 (α=0.35)	2.468	0.0
b0 ↔ b3	2.306	1.959 (α=0.65)	2.046	0.0
b1 ↔ b2	1.993	1.972 (α=0.15)	2.468	0.0
b1 ↔ b3	1.993	1.792 (α=0.45)	2.046	0.0
b2 ↔ b3	2.468	2.040 (α=0.95)	2.046	0.0
E3: Branch → Merged				
b0 → merged	2.306	1.973 (α=1.0)	1.973	0.0
b1 → merged	1.993	1.874 (α=0.50)	1.973	0.0
b2 → merged	2.468	1.973 (α=1.0)	1.973	0.0
b3 → merged	2.046	1.944 (α=0.60)	1.973	0.0

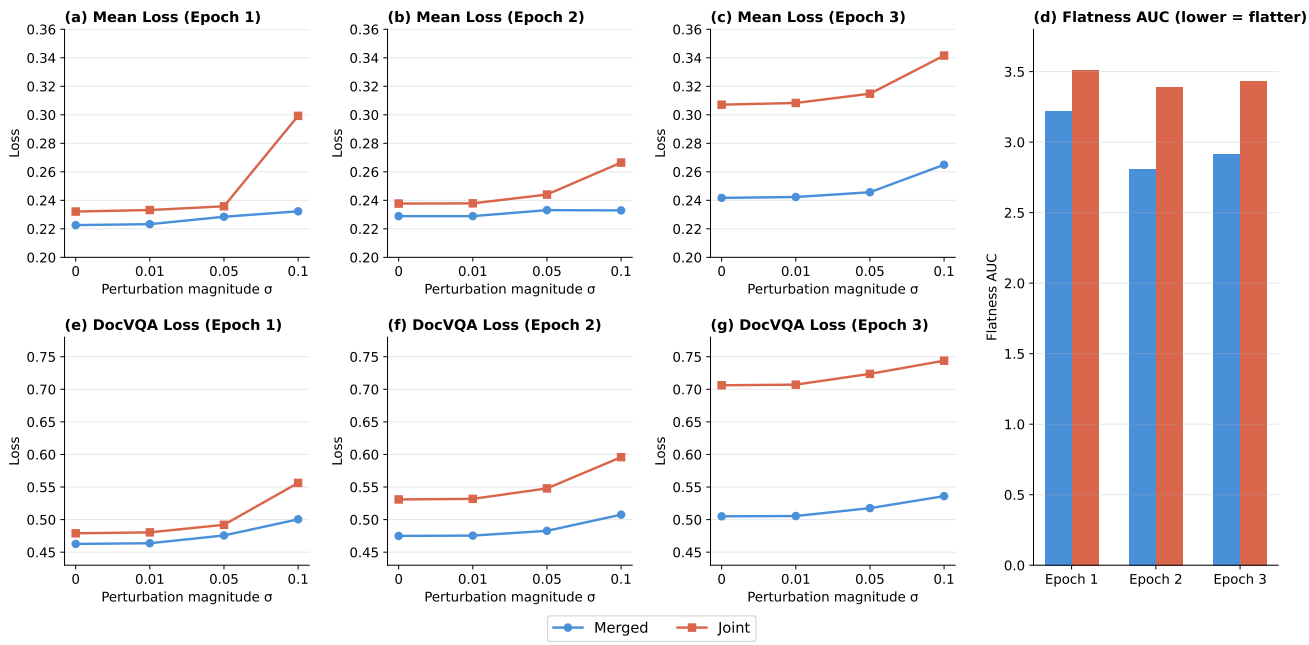

Figure R.1: Weight perturbation robustness analysis. (a–c) Mean loss under isotropic Gaussian perturbation ($\sigma \in \{0, 0.01, 0.05, 0.1\}$) at epochs 1–3. The merged model (blue) exhibits consistently lower loss sensitivity than the jointly trained model (red) at all perturbation levels and all epochs. (d) Flatness AUC (area under the ΔL–σ curve; lower = flatter basin): the merged model resides in a flatter region throughout training. (e–g) Per-benchmark breakdown on DocVQA shows the same pattern. This provides direct evidence that independently fine-tuned branches remain within a shared flat basin, supporting the merge-ready assumption.