Meta-Merging by Checkpoint Nowcasting

Albert M. Orozco Camacho [1 2] **Boris Knyazev** [3 4 2] **Eugene Belilovsky** [1 2] **Guy Wolf** [4 2]

Abstract

Model merging—the direct combination of parameters from independently fine-tuned networks—offers a way to compose task-specific capabilities without retraining or ensemble inference. Existing merge methods are often built from hand-crafted arithmetic or sparsification heuristics, leaving open whether general learned weight-space operators can be repurposed for merging directly. We study this question with NiNo, a pre-trained checkpoint-nowcasting meta-network originally designed to predict near-future training states from short checkpoint histories. We show that pre-trained NiNo can be reused as a data-free pairwise meta-merge operator for independently fine-tuned models. On an 8-task CLIP ViT-B/16 benchmark, NiNo is competitive with strong arithmetic baselines and consistently lands in the same functional region as weight averaging, Task Arithmetic, and TIES. Moreover, NiNo is best on HumanEval in a Qwen3 language extension among the compared merge methods, while extending meta-merge beyond pairs remains an open challenge. These results position learned checkpoint nowcasting as a practical starting point for data-free model merging and motivate future weight-space learners trained for merging explicitly.

1. Introduction

Large-scale foundation models now dominate vision and language pipelines, yet each new downstream task still demands an expensive fine-tune and extra GPU memory at deployment. Model merging tackles this bottleneck: instead of storing K separate checkpoints, it combines their parameters into a single network that behaves as if it had been fine-tuned on all K tasks. Merging is attractive for on-device inference, continual learning, and federated settings where data cannot be shared.

Recent work has shown paths toward making model parameters compatible with each other. This is a challenging open problem, to which clever algorithms have been proposed to tackle specific settings. Most successful techniques share a two-step recipe:

1. **Layer-wise alignment**: via pruning low-magnitude deltas (Davari & Belilovsky, 2024), or apply sparse binary masks (Yadav et al., 2023) so that corresponding layers become "compatible".

2. **Arithmetic aggregation**: Average or linearly interpolate the aligned weights, occasionally with a small post-hoc re-fine-tune ("stitching").

Knyazev et al. (2025) introduced NiNo, a meta-network algorithm capable of nowcasting parameters to speed up training optimization. Given a series of past checkpoints from a single training run as input, a graph neural network-based model predicts future model parameters. NiNo is *meta-trained* on a dataset of optimization trajectories; since it takes models as inputs and returns the updated model (with the same order of neurons as in the inputs), an unexplored question is: can a pre-trained meta-network, trained to extrapolate future parameters, also serve as a generic operator for merging independently fine-tuned models? We refer to this reuse as *meta-merge*: applying a learned update rule, trained on optimization trajectories, directly to task-specific checkpoints at test time.

This paper studies that question in the pairwise setting. Starting from the open-source CLIP ViT-B/16 visual encoder (Radford et al., 2021; Dosovitskiy et al., 2020), we fine-tune the same pre-trained checkpoint on eight vision tasks and treat each pair of task-specific checkpoints as a pseudo-trajectory rooted at their common initialization. Because the pre-trained NiNo model we use expects a five-checkpoint context, we choose a pairwise input sequence and later validate that choice by ablation. We then aggregate the resulting predicted checkpoints into a default merged model. We compare this merging-by-nowcasting recipe against weight averaging, Task Arithmetic (TA) (Il-

[1]Concordia University [2]Mila – Québec AI Institute [3]Samsung AI Lab, Montréal [4]Université de Montréal. Correspondence to: Albert M. Orozco Camacho <alorozco53@mila.quebec>, Boris Knyazev <b.knyazev@samsung.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

harco et al., 2023), TIES (Yadav et al., 2023), and Breadcrumbs (Davari & Belilovsky, 2024) on all 28 source-task pairs, using both top-1 accuracy and final-block Centered Kernel Alignment (CKA) (Kornblith et al., 2019; Davari et al., 2022).

Across the 28-pair CLIP grid, the default NiNo merge is competitive with the strongest arithmetic baselines: TIES remains best on average, while NiNo `avg 1:5` stays close in top-1 accuracy and consistently lands in the same activation-space neighborhood as weight averaging, TA, and TIES. The horizon trajectory is also informative: activation-space similarity peaks around k=3, while the horizon average `1:5` slightly improves accuracy and therefore serves as our default merged checkpoint. A compact Qwen3 extension at 0.6B shows a similar but weaker transfer pattern, and exploratory 4-task and 8-task compositions suggest that pairwise meta-merge is currently more robust than many-way scaling.

Contributions

- We propose merging-by-nowcasting as a data-free pairwise meta-merge operator for independently fine-tuned checkpoints.
- We characterize NiNo's horizon trajectory in both accuracy and activation space, showing that its default horizon average lands near a strong arithmetic merge cluster.
- We test this meta-merge picture beyond the main pairwise vision setting via a Qwen3 language extension and exploratory many-way compositions.

Taken together, the experiments uncover an unexpected capability of NiNo: a model trained to nowcast optimization trajectories can be repurposed as a practical pairwise meta-merge operator for independently fine-tuned checkpoints. This meta-merge perspective connects trajectory extrapolation and model merging, and suggests that learned update rules may offer an alternative route to composing task-specific models without task data or hand-crafted sparsity heuristics (Johnston et al., 1999; Hunter, 1986; Jolicoeur-Martineau et al., 2023; Izmailov et al., 2018; Morales-Brotons et al., 2024; Knyazev et al., 2025).

2. Methodology

We compare standard weight-space merge operators with a learned pairwise meta-merge alternative. Section 2.1 defines the arithmetic baselines used throughout the paper, while Section 2.2 describes how we repurpose pre-trained NiNo as a data-free pairwise meta-merge operator for fine-tuned checkpoints.

2.1. Weight-Space Merge Baselines

Problem set-up. Let $W_0 \in \mathbb{R}^P$ denote the parameters of a pre-trained foundation model and let $\{W_t\}_{t=1}^T$ be the checkpoints obtained after fine-tuning W_0 on T disjoint tasks. A *task vector* is the weight difference $\Delta_t = W_t - W_0$. Weight-space merging seeks a single parameter set $\widehat{W}$ that approximates the performance of $\{W_t\}$ *without* storing multiple models or resorting to ensemble inference. Most methods apply a layer-wise *alignment* operator $\mathcal{A}(\cdot)$—permutations, masks, sign filtering—followed by an arithmetic aggregation:

$$\widehat{W} \;=\; W_0 \;+\; \sum_{t=1}^{T} \alpha_t \, \mathcal{A}(\Delta_t), \tag{1}$$

where the coefficients α_t (scalars or per-layer vectors) are chosen by grid-search on a small validation set. This contrasts with ensembling, whose cost grows linearly in T at inference time.

Weight averaging and Task Arithmetic. Plain weight averaging applies Eq. equation 1 with $\mathcal{A}(\Delta_t) = \Delta_t$ and uniform coefficients, yielding

$$\widehat{W}_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} W_t = W_0 + \frac{1}{T}\sum_{t=1}^{T} \Delta_t.$$

Task Arithmetic (TA) (Ilharco et al., 2023) uses the same direction but introduces a global scaling coefficient λ:

$$\widehat{W}_{\text{TA}} = W_0 + \lambda \, \frac{1}{T}\sum_{t=1}^{T} \Delta_t.$$

This simple linear recipe is often surprisingly effective when the tasks are sufficiently "orthogonal," a *linear* combination of task vectors already yields a competent merged model, but it becomes increasingly sensitive to interference as more task vectors are added.

TIES and Breadcrumbs. Trim, Elect, Sign & Merge (TIES-Merging) (Yadav et al., 2023) mitigates interference by *masking* low-magnitude weights in each Δ_t, resolving sign conflicts among the surviving entries, and only then averaging the resulting task vectors. Breadcrumbs (Davari & Belilovsky, 2024) instead removes both very small and very large magnitudes within each layer, leaving a sparse, layer-balanced update that is added back to W_0 with strength α. Both methods aim to reduce task interference without retraining, but do so through different sparsification heuristics.

2.2. NiNo as a Meta-Merge Operator

NiNo (Knyazev et al., 2025) is a pre-trained meta-network for checkpoint nowcasting: given a short context of training checkpoints, it predicts future checkpoints along that

trajectory. In our setting, we treat independently fine-tuned models as a *pseudo-trajectory* rooted at the common pre-trained checkpoint W_0, and use NiNo directly as a pairwise meta-merge operator without additional training.

The pre-trained NiNo model used here expects a context of five checkpoints. This fixes the input length, but not the ordering of the task-specific checkpoints. For a pair of task-specific models W_{t_1} and W_{t_2}, we choose the context order

$$C_{01221} = [W_0, W_{t_1}, W_{t_2}, W_{t_2}, W_{t_1}],$$

which we refer to as the `01221` recipe. For each horizon $k \in \{1, \ldots, 5\}$, we denote the corresponding NiNo prediction by

$$\widehat{W}^{(k)}_{NiNo} = \text{NiNo}_k(C_{01221}). \tag{2}$$

Our default pairwise meta-merge is then the full horizon average

$$\widehat{W}_{NiNo} = \frac{1}{5}\sum_{k=1}^{5} \widehat{W}^{(k)}_{NiNo}, \tag{3}$$

which we denote as `avg 1:5`.

All pairwise NiNo results use the chosen `01221+avg 1:5` recipe by default. Appendix A.2 reports the context-order and horizon ablations that motivate this choice.

3. Experiments

Experimental set-up. We start from the open-source CLIP ViT-B/16 (Radford et al., 2021; Dosovitskiy et al., 2020) encoder (W_0) and fine-tune it on eight disjoint vision tasks from the benchmark of Ilharco et al. (2023): MNIST (*hand-written digits*), DTD (*textures*), EuroSAT (*satellite imagery*), Cars (*fine-grained vehicles*), RESISC-45 (*remote-sensing scenes*), SVHN (*street numbers*), GTSRB (*traffic signs*), and SUN397 (*scene recognition*). For each source-task pair $\{d_1, d_2\}$ we create two single-task checkpoints W_{d_1}, W_{d_2} and merge them with the five methods under study: plain weight averaging, Task Arithmetic (TA), TIES, Model Breadcrumbs, and NiNo. For NiNo, we use the pairwise recipe chosen in Section 2.2, namely `01221` followed by the full-horizon average `avg 1:5`. This yields all $\binom{8}{2} = 28$ benchmark pairs. Unless otherwise stated, we evaluate each merged model only on its two source datasets and report top-1 accuracy averaged over the 28 source-task pairs. A compact language-domain extension with Qwen3 is presented later in this section; full per-benchmark results are deferred to Appendix A.7.

Similarity Measurement. To quantify representational similarity, we use *linear Centered Kernel Alignment (CKA)* (Kornblith et al., 2019; Davari et al., 2022). For two activation matrices, $X, Y \in \mathbb{R}^{n\times d}$, it is defined as $\text{CKA}(X, Y) = \|X^\top Y\|_F^2 / \left(\|X^\top X\|_F \|Y^\top Y\|_F\right)$, where $\|\cdot\|_F$ denotes the Fröbenius norm. Linear CKA is invariant to orthogonal transformations and isotropic scaling, and takes values in $[0, 1]$, with 1 indicating identical representations. All CKA scores are computed on the last vision-block activations, after feeding the test sets of the corresponding source datasets.

Table 1. **Pairwise merge performance on the 28-pair CLIP ViT-B/16 benchmark.** Mean top-1 is averaged over all 28 source-task pairs and their two source datasets; standard error is over pairs. The CKA column reports mean final-block linear CKA to the arithmetic cluster {weight_avg, TA, TIES}, averaged over the same evaluations. Best overall and best NiNo candidate are bold. All NiNo rows use the paper's chosen `01221` context order for each pair.

Method	Mean top-1	CKA to arithmetic cluster
weight_avg	0.896 ± 0.012	0.984
TA	0.905 ± 0.011	0.985
TIES	$\mathbf{0.908} \pm \mathbf{0.011}$	0.989
breadcrumbs	0.831 ± 0.014	0.830
NiNo k=1	0.873 ± 0.011	0.909
NiNo k=2	0.882 ± 0.011	0.940
NiNo k=3	0.890 ± 0.012	*0.961*
NiNo k=4	0.874 ± 0.013	0.942
NiNo k=5	0.845 ± 0.015	0.873
NiNo avg 1:5	$\mathbf{0.891} \pm \mathbf{0.011}$	0.958

3.1. Pairwise Accuracy and Representational Alignment

To study NiNo's pairwise meta-merge behavior relative to the standard arithmetic and sparsified baselines, we compare both predictive performance and representational similarity on the full 28-pair grid. Table 1 summarizes the main pairwise vision results. TIES is the strongest method on average, reaching 0.908 top-1 accuracy, with TA close behind at 0.905. NiNo's horizon sweep under our chosen `01221` context order shows a clear best aggregate candidate: `avg 1:5` reaches 0.891 top-1 accuracy, narrowly above the best single horizon k=3 at 0.890. Breadcrumbs trails the other methods on this pairwise benchmark.

Figure 1 makes this pair-level geometry concrete for the nearest baseline, weight averaging. Across most task pairs, the CKA panel remains above 0.95, with the weakest alignment concentrated on EuroSAT-involving merges, while the accuracy panel shows only a small residual gap on average. The rightmost column of Table 1 summarizes activation-space proximity to the arithmetic merge cluster {weight_avg, TA, TIES}. This proximity is non-monotone across NiNo horizons: it rises from 0.909 at k=1 to a peak of 0.961 at k=3, then declines at later horizons. The default `avg 1:5` candidate retains nearly all of k=3's alignment (0.958 vs. 0.961) while slightly improving mean top-1 accuracy (0.891 vs. 0.890), which is why it remains our default merged checkpoint. Appendix A.3.2 and Appendix A.3.4

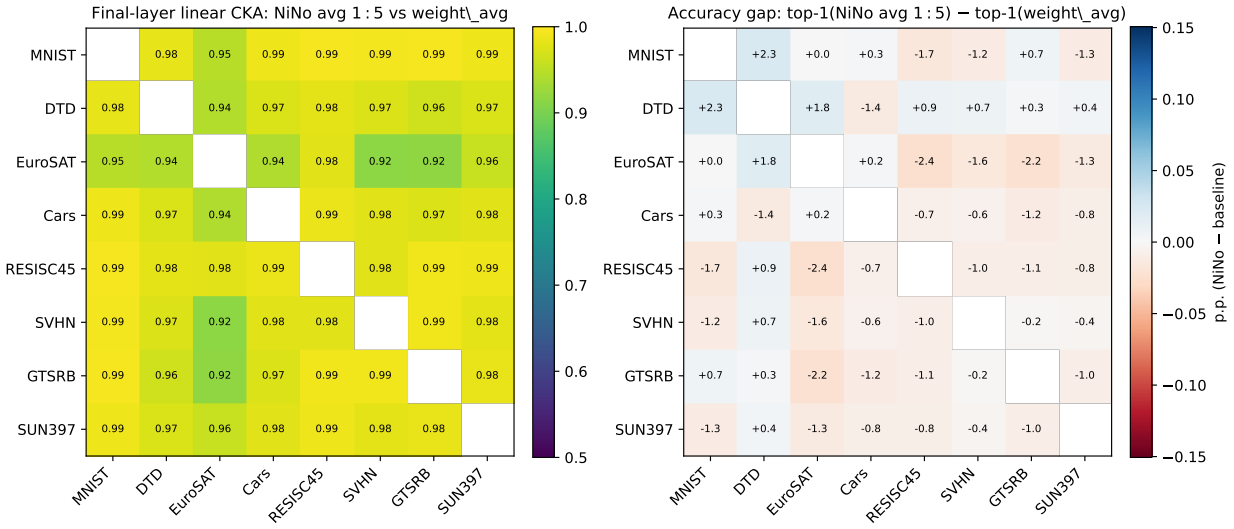

Figure 1. **Per-pair 8×8 comparison of NiNo avg 1:5 against weight averaging.** The left panel reports final-block linear CKA, averaged over the two in-pair evaluation datasets. The right panel reports the signed accuracy gap in percentage points. This is the closest baseline to NiNo in activation space on the 28-pair grid.

Table 2. **Qwen3 language-domain extension (0.6B model).** All numeric columns report the single 3-way merge (Math+Fr+Code). The last column is the macro-average over the four benchmark columns.

Method	GSM8K	French	GSM8K-fr	HE	Avg
zero-shot	19.5	25.1	21.1	38.4	26.0
weight avg	49.8	27.1	33.1	42.7	38.2
TA (λ=0.3)	**50.6**	**27.3**	**35.6**	42.7	**39.1**
NiNo	39.1	24.4	23.5	**46.3**	33.3

provide the full horizon trajectory, additional baseline comparisons, and the remaining per-pair heatmaps.

Beyond Pairwise Merging. We also explored 4-task and 8-task compositions of the same pre-trained NiNo operator. Because NiNo accepts only five context states, these many-way meta-merge experiments require heuristic trees, folds, or soups of saved pairwise meta-merges, and they are less robust than pairwise meta-merge. Among these heuristics, compositions built from saved pairwise meta-merges are more stable than direct hierarchical recipes. We therefore treat multi-task scaling as an informative limitation of zero-shot NiNo and defer the detailed ablations to Appendix A.4 and Appendix A.5.

Language-Domain Extension: Qwen3. We repeat the comparison on Qwen3 (Yang et al., 2025), using the released 0.6B model with three task fine-tunes (math, French, code). We summarize the resulting 3-way merge on four benchmarks here, and defer the full pairwise and benchmark-level results to Appendix A.7. All three merge operators improve substantially over the zero-shot model. TA remains strongest on the macro-average, while NiNo is best on HumanEval at 46.3, suggesting preserved code-specialized behavior despite weaker overall language-domain merge quality. This makes Qwen3 a useful stress test: aggregate language merging is harder here, yet the code-oriented capability remains highly competitive.

4. Discussion and Conclusion

In this paper, we showed that a pre-trained NiNo nowcaster can be repurposed as a practical data-free pairwise meta-merge operator. On the 28-pair CLIP ViT-B/16 benchmark, NiNo's chosen `01221+avg 1:5` recipe is competitive with strong arithmetic baselines, while consistently landing in the same activation-space neighborhood as weight averaging, Task Arithmetic, and TIES. The horizon analysis further shows that this behavior is structured rather than accidental.

Beyond the main vision setting, NiNo is strongest on code-oriented behavior in Qwen3, while its 4-task and 8-task compositions remain less robust than the pairwise operator itself. Taken together, these results make pairwise meta-merge the clearest current success case for repurposing learned weight-space dynamics for explicit model composition. More broadly, they suggest that merge-relevant structure already exists in general-purpose weight-space learners even before any explicit merge training. Future work should therefore target stronger many-way composition rules and explicit unsupervised objectives for learned weight-space merge operators.

References

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Davari, M. and Belilovsky, E. Model breadcrumbs: Scaling multi-task model merging with sparse masks, 2024. URL https://arxiv.org/abs/2312.06795.

Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., and Belilovsky, E. Reliability of cka as a similarity measure in deep learning, 2022. URL https://arxiv.org/abs/2210.16156.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Hunter, J. S. The exponentially weighted moving average. *Journal of quality technology*, 18(4):203–210, 1986.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Johnston, F., Boyland, J. E., Meadows, M., and Shale, E. Some properties of a simple moving average when applied to forecasting a time series. *Journal of the Operational Research Society*, 50(12):1267–1271, 1999.

Jolicoeur-Martineau, A., Gervais, E., Fatras, K., Zhang, Y., and Lacoste-Julien, S. Population parameter averaging (papa). *arXiv preprint arXiv:2304.03094*, 2023.

Knyazev, B., Moudgil, A., Lajoie, G., Belilovsky, E., and Lacoste-Julien, S. Accelerating training with neuron interaction and nowcasting networks. In *International Conference on Learning Representations*, 2025.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. E. Similarity of neural network representations revisited. *CoRR*, abs/1905.00414, 2019. URL http://arxiv.org/abs/1905.00414.

Lasbordes, M. and Gad, S. Luth: Efficient french specialization for small language models and cross-lingual transfer, 2025. URL https://arxiv.org/abs/2510.05846. arXiv:2510.05846.

Luo, Z., Xu, C., Zhao, P., Sun, Q., Geng, X., Hu, W., Tao, C., Ma, J., Lin, Q., and Jiang, D. Wizardcoder: Empowering code large language models with evol-instruct, 2023.

Morales-Brotons, D., Vogels, T., and Hendrikx, H. Exponential moving average of weights in deep learning: Dynamics and benefits. *arXiv preprint arXiv:2411.18704*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. Ties-merging: Resolving interference when merging models, 2023. URL https://arxiv.org/abs/2306.01708.

Yang, A. et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A. Appendix

A.1. Complementary Results

A.2. Context-Order and Horizon Ablation for NiNo

Because the pre-trained NiNo model expects a trajectory-like input, pairwise meta-merge depends on both the 5-state context order and on which predicted horizon(s) are retained. To justify the chosen pairwise recipe, we run an ablation on 10 source-task pairs sampled from the 8-task CLIP pool. For each pair, we evaluate the five single-horizon outputs and the prefix averages 1:2, 1:3, 1:4, and 1:5 across six candidate context orders.

Figure 2 summarizes the selection rule used in the paper. Panel A aggregates all six orders and shows the horizon trend: k=3 is the strongest single horizon, while the full average 1:5 remains competitive because it retains the early-horizon gains. Panel B reports the final 1:5 candidate by context sequence and highlights the chosen order $01221 = [W_0, W_1, W_2, W_2, W_1]$. Panel C repeats the horizon analysis after restricting to this chosen order, confirming that the same trend holds once the ordering is fixed. Error bars are standard errors over the 10 sampled pairs.

Figure 2 therefore supports the same rule used in the main text: k=3 is the strongest single horizon, 1:5 is the best robust default, and 01221 is the strongest tested full-average order.

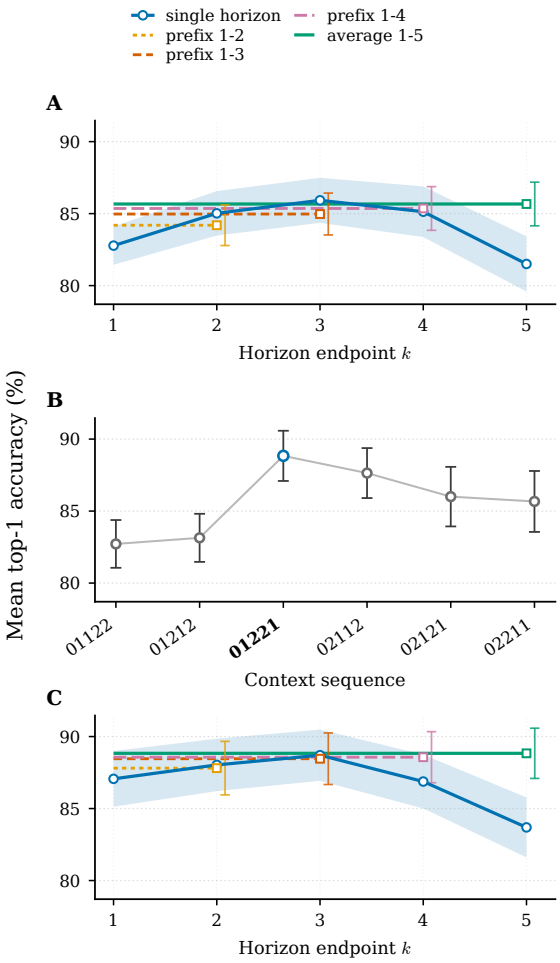

Figure 2. Context-order and horizon ablation for NiNo on two-task merges. Error bars indicate standard error over sampled task pairs.

A.3. Full 28-Pair Grid: Accuracy and Geometry

This section extends the main pairwise analysis to all 28 source-task pairs in the 8-task CLIP ViT-B/16 benchmark {MNIST, DTD, EuroSAT, Cars, RESISC45, SVHN, GTSRB, SUN397}. For each pair, we report NiNo's five pairwise meta-merge horizons and the average over 1:5, together with weight averaging, Task Arithmetic, TIES, breadcrumbs, and the zero-shot reference W_0. Pairwise NiNo always uses the paper's chosen `01221` recipe. All evaluations here are in-domain: each merged model is tested only on its two source datasets.

Appendix A.3.1 reports aggregate and per-dataset accuracy. Appendix A.3.2 places NiNo relative to the arithmetic baselines in weight and activation space. Appendix A.3.3 contrasts directional proximity with final accuracy. Appendix A.3.4 and Appendix A.3.5 provide the per-pair heatmaps and the midpoint ablation.

A.3.1. Accuracy across the 28-pair grid

Table 3 reports mean per-pair top-1 accuracy over all 28 source-task pairs. NiNo is competitive with weight averaging and Task Arithmetic at every horizon $k \geq 2$; the best single horizon k=3 (0.890) and the full average 1:5 (0.891) trail TA by 1.5 and 1.4 p.p., and TIES by 1.8 and 1.7 p.p. Breadcrumbs is clearly lower on this grid.

Table 3. **Top-1 accuracy of merged CLIP ViT-B/16 encoders, averaged over all 28 source-task pairs.** Standard error is over pairs. Best result in bold.

Method	Mean top-1
W_0 (zero-shot)	0.555 ± 0.011
weight_avg	0.896 ± 0.012
TA	0.905 ± 0.011
TIES	**0.908 ± 0.011**
breadcrumbs	0.831 ± 0.014
NiNo k=1	0.873 ± 0.011
NiNo k=2	0.882 ± 0.011
NiNo k=3	0.890 ± 0.012
NiNo k=4	0.874 ± 0.013
NiNo k=5	0.845 ± 0.015
NiNo avg 1:5	0.891 ± 0.011

The aggregate hides a useful per-dataset asymmetry. Table 4 shows that NiNo k=1 beats every baseline on DTD and Cars, but falls behind on RESISC45, SVHN, and GTSRB; those deficits are what pull its 28-pair mean below TIES. MNIST is nearly saturated for every merge method except W_0, consistent with NiNo's pair-level wins concentrating on MNIST-containing pairs.

A.3.2. NiNo's location in weight and activation space

Section 3.1 showed that NiNo lands near the arithmetic merge cluster in activation space. Here we extend that picture to the full 28-pair grid along two axes: (i) cosine between task-vector deltas in weight space, and (ii) linear CKA on the final visual-encoder block. Throughout, $\delta = W_{\text{merged}} - W_0$ denotes the candidate's task-vector direction.

Weight-space proximity. Table 5 shows two stable patterns. In weight space, each NiNo horizon is closest to weight averaging and Task Arithmetic, then to TIES, and least to breadcrumbs. The cosine also decays almost monotonically with horizon: k=1 is nearest to the raw summed-task-vector direction, while later horizons drift away from every baseline.

Table 4. **Top-1 accuracy of merged encoders, by evaluation dataset.** Each entry is the mean top-1 of the listed candidate when one of its two source datasets is the column dataset, averaged over the 7 partner pairs. Best result per column in bold.

Method	MNIST	DTD	EuroSAT	Cars	RESISC45	SVHN	GTSRB	SUN397	avg
W_0 (zero-shot)	0.517	0.447	0.555	0.647	0.664	0.520	0.434	0.655	0.555
weight_avg	0.994	0.740	0.988	0.824	0.934	0.962	0.968	0.756	0.896
TA	**0.995**	0.777	0.990	0.829	0.943	0.970	**0.978**	0.757	0.905
TIES	**0.995**	0.778	**0.992**	0.845	**0.947**	**0.971**	0.968	**0.772**	**0.908**
breadcrumbs	0.982	0.617	0.968	0.781	0.891	0.860	0.818	0.729	0.831
NiNo k=1	0.988	**0.790**	0.983	**0.862**	0.857	0.875	0.913	0.712	0.873
NiNo k=2	0.992	0.767	0.986	0.834	0.898	0.912	0.941	0.729	0.882
NiNo k=3	0.994	0.763	0.986	0.822	0.917	0.945	0.960	0.736	0.890
NiNo k=4	0.993	0.716	0.976	0.769	0.914	0.947	0.943	0.734	0.874
NiNo k=5	0.994	0.654	0.962	0.736	0.899	0.954	0.887	0.678	0.845
NiNo avg 1:5	0.994	0.770	0.987	0.829	0.914	0.940	0.958	0.736	0.891

Table 5. **Mean weight-space cosine between NiNo deltas and baseline deltas, over 28 pairs.** The pre-trained reference W_0 is omitted (delta is zero). Best per column in bold.

NiNo candidate	weight_avg	TA	TIES	breadcrumbs
NiNo k=1	**0.691**	**0.691**	**0.589**	**0.489**
NiNo k=2	0.513	0.513	0.442	0.372
NiNo k=3	0.463	0.463	0.399	0.336
NiNo k=4	0.319	0.319	0.278	0.237
NiNo k=5	0.282	0.282	0.243	0.206
NiNo avg 1:5	0.495	0.495	0.426	0.359

For calibration, the baselines themselves form a tighter one-dimensional chain: TA and weight averaging are co-linear, TIES is close to both, and breadcrumbs is farther away. NiNo preserves the same ordering, but sits outside that tighter weight-space cluster.

Table 6. **Mean weight-space cosine between baseline deltas, over 28 pairs.** Baselines lie on a one-dimensional cluster from breadcrumbs through TIES to TA/weight_avg.

	weight_avg	TA	TIES	breadcrumbs
weight_avg	1.000	1.000	0.866	0.718
TA	1.000	1.000	0.866	0.718
TIES	0.866	0.866	1.000	0.743
breadcrumbs	0.718	0.718	0.743	1.000

Activation-space proximity. Table 7 shows a different geometry in activation space. NiNo k=3 and avg 1:5 both reach CKA above 0.95 with weight averaging, TA, and TIES, placing them inside the same activation-space neighborhood despite their larger weight-space offsets. CKA to breadcrumbs is lower, and CKA to W_0 stays around 0.58–0.60 across horizons.

Table 7. **Mean final-block linear CKA between NiNo and the baselines, over 28 pairs $\times$ 2 in-pair datasets.** Best per column in bold.

NiNo candidate	W_0	weight_avg	TA	TIES	breadcrumbs
NiNo k=1	0.578	0.917	0.902	0.907	0.838
NiNo k=2	**0.600**	0.958	0.926	0.937	**0.877**
NiNo k=3	0.580	**0.973**	**0.950**	**0.959**	0.859
NiNo k=4	0.594	0.960	0.928	0.938	0.865
NiNo k=5	0.589	0.887	0.861	0.871	0.826
NiNo avg 1:5	0.586	0.972	0.946	0.956	0.866

The arithmetic baselines form an even tighter activation-space cluster: {weight_avg, TA, TIES} all lie at CKA ≥ 0.97, whereas breadcrumbs is closer to W_0 than to any of them.

Table 8. **Mean final-block linear CKA between baselines, over 28 pairs $\times$ 2 in-pair datasets.** Note that breadcrumbs is closer to W_0 in CKA (0.769) than to weight_avg, TA, or TIES, indicating a separate activation cluster.

	W_0	weight_avg	TA	TIES	breadcrumbs
W_0 (zero-shot)	1.000	0.585	0.521	0.544	0.769
weight_avg	0.585	1.000	0.970	0.983	0.874
TA	0.521	0.970	1.000	0.985	0.792
TIES	0.544	0.983	0.985	1.000	0.825
breadcrumbs	0.769	0.874	0.792	0.825	1.000

Horizon trajectory. Figure 3 combines the previous tables. Weight-space cosine decays almost monotonically with k, while activation-space CKA peaks at k=3 and stays nearly unchanged for avg 1:5. This is the same pattern summarized in the main text: early horizons align with the arithmetic direction, middle horizons align with the arithmetic representation, and the full average retains that representation while slightly improving accuracy.

Per-dataset breakdown. Table 9 shows that the ordering weight_avg $>$ TIES $\gtrsim$ TA $\gg$ breadcrumbs is stable across

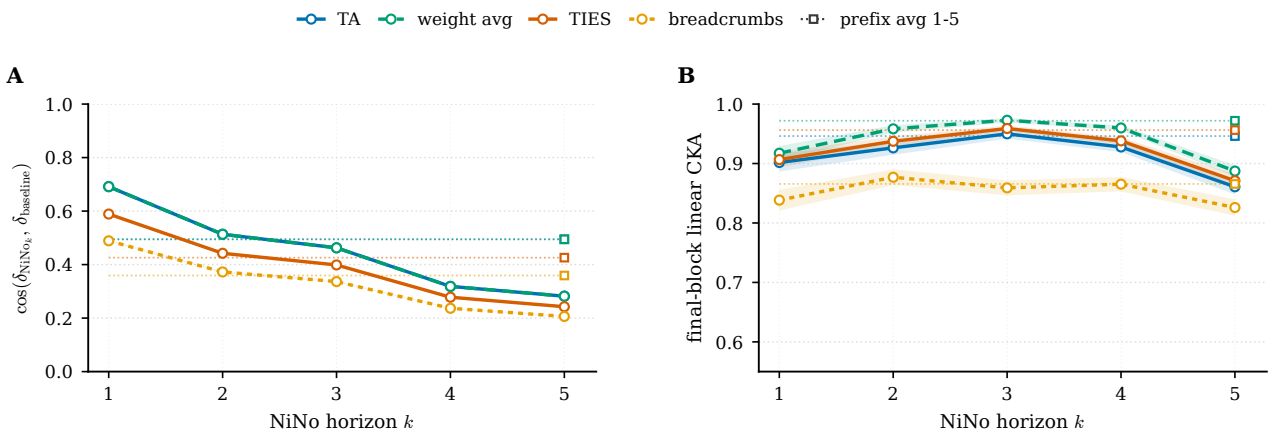

Figure 3. **NiNo horizon trajectory across the 28-pair grid. (A)** weight-space cosine between NiNo's task-vector delta and each baseline's. **(B)** final-block linear CKA between NiNo and each baseline. Solid/dashed lines with circle markers show the single-horizon means at $k \in \{1, \ldots, 5\}$; shaded bands are pair-level standard error. Dotted horizontal segments with square markers report the prefix average 1:5 as a constant reference per baseline. TA and weight averaging coincide in panel A by construction (their deltas are co-linear).

datasets. EuroSAT is the clearest outlier, producing the largest representation shift; RESISC45 and SUN397 stay highly aligned across all merge baselines. The avg 1:5 values are within ± 0.01 of k=3 throughout and are omitted for brevity.

A.3.3. Direction-similar $\neq$ accuracy-best

Tables 5 and 7 suggest that NiNo lives near the weight-averaging / Task-Arithmetic axis. Using either NiNo k=3 or avg 1:5 as the anchor gives the same picture: weight averaging is the nearest baseline in CKA (0.973 for k=3, 0.972 for avg 1:5), and TA/weight averaging are also the nearest in cosine (0.463 for k=3, 0.495 for avg 1:5), but TIES is the most accurate baseline overall.

So directional proximity explains NiNo's closeness to the arithmetic family, not which baseline is strongest. TIES gains accuracy by departing from the raw summed-task-vector direction through sparsification and sign resolution, even though that makes it slightly farther from NiNo in cosine and CKA.

A.3.4. Per-pair 8×8 heatmaps

Figures in this section give pair-level versions of the aggregate CKA and accuracy results. Each off-diagonal cell corresponds to one source-task pair and averages the two source-dataset evaluations. The left panel reports final-block CKA between NiNo and the baseline; the right panel reports the signed top-1 gap in percentage points. Pairwise NiNo always uses the paper's chosen `01221` recipe.

NiNo vs Task Arithmetic. Figures 4 and 5 show the headline comparison against TA. CKA stays high on most pairs, with the weakest cells on EuroSAT pairs, and the accuracy gaps are usually small and negative. The k=3 and avg 1:5 panels are nearly indistinguishable, consistent with their near-coincidence in activation space.

NiNo vs the other baselines. Figures 6–8 extend the same view to weight averaging, TIES, and breadcrumbs. The CKA panels follow the same aggregate ordering: weight averaging is nearest, then TA/TIES, then breadcrumbs. The accuracy panels show the complementary nuance: NiNo stays closest to weight averaging in representation space, but trails TIES more in accuracy because TIES's sparsification step changes the direction enough to help performance.

A.3.5. Midpoints input recipe

The chosen pairwise recipe, $[W_0, W_{t_1}, W_{t_2}, W_{t_2}, W_{t_1}]$, repeats the second task checkpoint as an intermediate state. A natural alternative is to insert synthetic midpoints, mirroring `merge_vit.py`'s `--midpoints` flag:

$$[W_0,\ \tfrac{W_0+W_{t_1}}{2},\ W_{t_1},\ \tfrac{W_{t_1}+W_{t_2}}{2},\ W_{t_2}].$$

The pre-trained NiNo meta-network is unchanged; only the input sequence is different. We evaluate the midpoint recipe on the same 28-pair grid and compare it to the paper's chosen `01221` recipe at every matched horizon.

The midpoint recipe is uniformly worse on this grid. For $k \in \{2, 3, 4\}$ and avg 1:5, it loses 1.4–2.0 p.p.; even its best matched horizon wins only 8 of the 28 pairs, and the average 1:5 variant wins 7 of 28.

The informative failure is k=5: midpoints collapses by 23.2 p.p. and wins no pair. Its weight-space cosine to the chosen k=5 output remains high (0.911), but the activa-

Table 9. **Final-block linear CKA between NiNo k=3 and each baseline, by evaluation dataset.** Each entry is the mean over the 7 partner pairs containing the column dataset. Best per column in bold.

Baseline	MNIST	DTD	EuroSAT	Cars	RESISC45	SVHN	GTSRB	SUN397
W_0 (zero-shot)	0.492	0.727	0.282	0.757	0.877	0.213	0.393	0.899
weight_avg	**0.997**	**0.967**	**0.912**	**0.976**	**0.991**	**0.982**	**0.981**	**0.977**
TA	0.991	0.940	0.868	0.970	0.984	0.935	0.942	0.969
TIES	0.988	0.962	0.904	0.971	0.985	0.930	0.964	0.966
breadcrumbs	0.928	0.879	0.621	0.907	0.969	0.833	0.785	0.951

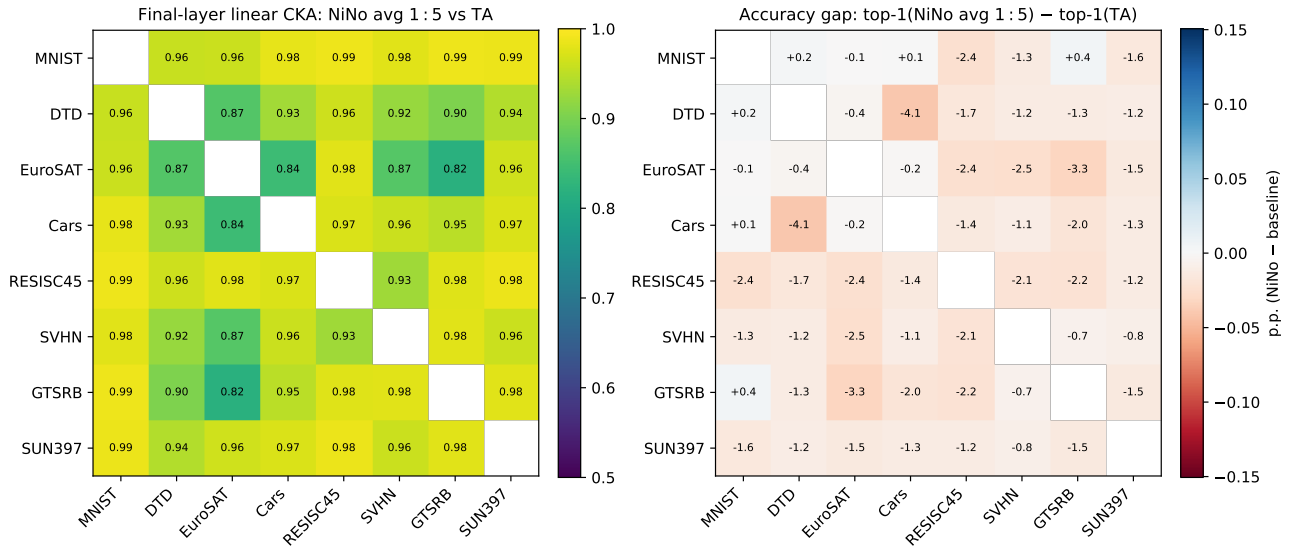

Figure 4. **Per-pair 8×8 comparison of NiNo avg 1:5 against Task Arithmetic.** Left: final-block linear CKA, averaged over the two in-pair evaluation datasets. Right: accuracy gap (NiNo − TA) in percentage points.

tion CKA drops to 0.791, far below the 0.92–0.98 range seen elsewhere. So the midpoint recipe points in roughly the same direction, but lands outside the useful activation cluster.

This makes the midpoint experiment a clean negative control for the paper's chosen `01221` ordering. Constructing synthetic intermediate states does not improve merging in this regime, and the late-horizon k=5 midpoint output is not useful in practice.

A.4. Multi-Subset 4-Task Benchmark

To test whether the pairwise picture extends beyond two tasks, we evaluate 10 strategies across five 4-task subsets sampled from the same 8-dataset pool. The baselines are Task Arithmetic, weight averaging, TIES, and breadcrumbs. NiNo contributes three composition families—direct, progressive-fold, and binary tree—each paired with both GNN and MLP backbones. All NiNo variants average horizons 1:5 and are evaluated on the four component validation splits. The five subsets cover different mixes of digit/symbol, texture, spatial, and fine-grained natural tasks. Table 10 summarizes the mean and standard deviation across these five subsets.

The same-run arithmetic baselines win every subset, with TIES best throughout. The strongest NiNo composition, `nino_tree_mlp`, reaches 80.06 average accuracy, trailing TIES by 5.9 p.p. and weight averaging by 1.5 p.p. It is also the most stable NiNo variant (σ=2.48), whereas `nino_direct_mlp` is the least stable (σ=8.18), collapsing on harder components such as DTD and SVHN. Across recipes, direct favors the GNN backbone, tree favors the MLP backbone, and progressive-fold is the weakest family.

A.5. 8-Task Same-Run Multi-Method Benchmark

We then scale to a single 8-task merge. Because NiNo only accepts five context states, same-run many-way meta-merge must be wrapped in either a binary tree or a progressive-fold topology. We also include an all-pair soup diagnostic that averages the 28 saved pairwise merges for each method.

Table 11 summarizes the representative same-run baselines, the most informative pair-soup diagnostics, and the strongest same-run NiNo wrappers. The same-run arith-

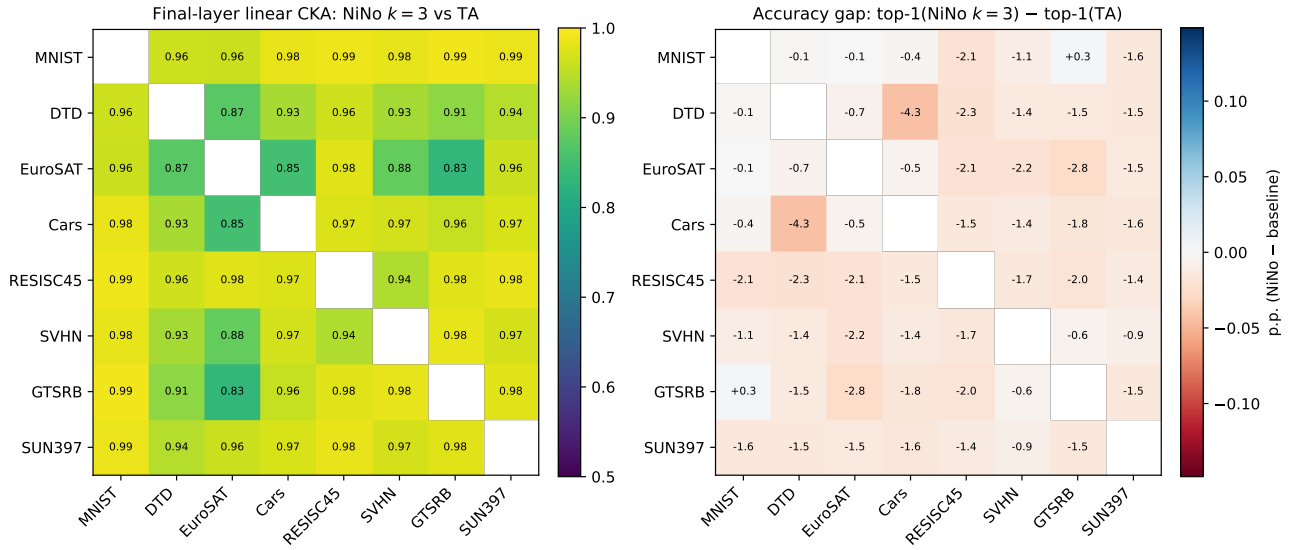

Figure 5. **Per-pair 8×8 comparison of NiNo k=3 against Task Arithmetic.** Same construction as Figure 4, with the single best NiNo horizon (k=3) as the anchor.

Table 10. **Top-1 accuracy of merged ViT-B/16 encoders across five 4-task subsets.** Columns report the mean and standard deviation over the five sampled subsets. Best mean in bold.

Method	Mean	Std
TIES	**85.96**	4.01
TA (λ=0.4)	84.32	4.04
breadcrumbs	83.44	4.76
weight_avg	81.55	4.49
NiNo tree (MLP)	80.06	2.48
NiNo direct (GNN)	78.53	5.21
NiNo prog-fold (GNN)	76.48	4.60
NiNo tree (GNN)	76.00	3.38
NiNo prog-fold (MLP)	72.66	3.44
NiNo direct (MLP)	62.52	8.18

Table 11. **Representative 8-task ViT-B/16 merge results, full validation, seed 42.** We report the strongest same-run baselines, the most informative all-pair soups over the 28-pair grid, and the strongest same-run NiNo wrappers. Best in each column in bold.

Strategy	Avg	Min	HM
breadcrumbs	**79.83**	**57.82**	**77.89**
TIES	76.93	57.18	74.93
TA (λ=0.2)	74.99	52.50	72.87
weight_avg	71.67	50.05	69.47
soup-28-pair: TA	73.83	52.77	71.87
soup-28-pair: weight_avg	71.67	50.05	69.46
soup-28-pair: NiNo-avg-1:5	71.09	51.44	69.07
NiNo tree (MLP)	70.86	49.84	68.43
NiNo prog-fold (GNN, α=0.75)	68.07	46.33	65.78

metic baselines still dominate: breadcrumbs and TIES lead, followed by TA and weight averaging. The best same-run NiNo variant is the binary tree with the MLP backbone (`tree_default_mlp`, 70.86), close to the soup of pairwise NiNo merges (71.09) and to souped weight averaging (71.67), but still below every same-run arithmetic baseline.

Per-dataset winners. The per-dataset winners reinforce the leaderboard. Breadcrumbs wins six of the eight tasks, TIES wins SVHN, and the only NiNo-derived per-dataset win is SUN397 via progressive-fold with fixed α=0.5. Since SUN397 is the final task in the canonical progressive-fold order, this is best read as late-task retention rather than broad many-way strength.

All-pair soup diagnostic. The all-pair-soup rows in Table 11 average the 28 saved per-pair merged checkpoints for a method. TA and weight averaging are nearly invariant under this construction: TA drops from 74.99 to 73.83, and weight averaging is unchanged at 71.67. TIES loses 3.51 p.p., and breadcrumbs loses 13.36 p.p., indicating that their per-pair sparsification rules do not transfer cleanly to the many-way setting.

Backbone/topology interaction. The GNN/MLP preference flips with topology. On the binary tree, the MLP backbone beats the GNN by 15.4 p.p.; on progressive-fold, the GNN beats the MLP by 4.5 p.p. The simplest reading is that MLP-NiNo prefers small, symmetric deltas, whereas GNN-NiNo is more tolerant of spread-out contexts. The negative-control tree rows with alternative anchors support

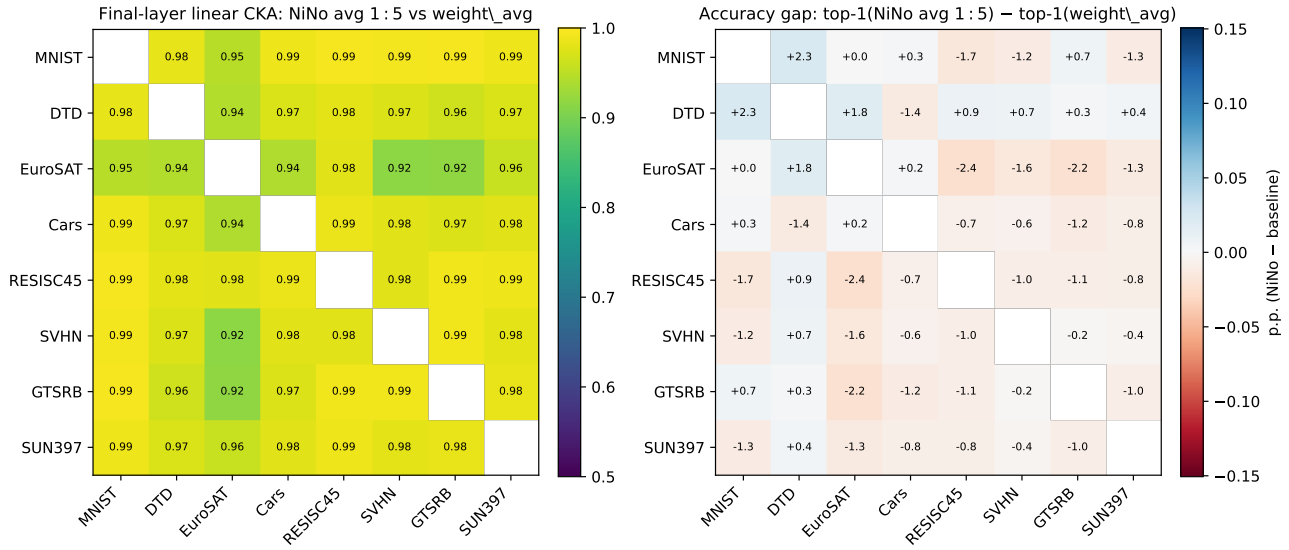

Figure 6. **NiNo avg** 1:5 **against weight averaging.** The CKA panel saturates above 0.95 on most pairs, confirming weight averaging as NiNo's nearest baseline in activation space.

the same point: changing the anchor does not rescue the default GNN tree.

Progressive-fold stage diagnostics. A stage-wise diagnostic of the best progressive-fold family on the full 8-task validation set, not only on the tasks absorbed so far, shows that stage 3 is stronger than the final stage 4 checkpoint for all four runs.

For all four runs, the stage-3 checkpoint has a higher full-dataset average than the final stage-4 checkpoint. The last fold into GTSRB and SUN397 therefore helps some late tasks but hurts overall balance. Smaller α protects earlier tasks better, while larger α shifts capacity toward later tasks; in this batch, decay α_0=0.75 gives the best final average and harmonic mean, and fixed α=0.375 gives the best final minimum.

A.6. 8-Task Tree-Merge Order Ablation

Because the 8-task tree builds level-1 pair merges from a fixed input order, changing the permutation changes which tasks meet at each tree level and how strongly they are preserved through the 7-step hierarchy. We test four orderings while keeping the tree recipe itself fixed.

Table 12 shows that the mean spread across orderings is only 1.5 p.p. (53.98–55.45), and even the best ordering remains far behind every same-run arithmetic baseline. Reordering the leaves alone does not close the 8-task gap.

The small mean spread hides much larger task-level swings. Table 13 shows that MNIST and SVHN move by about 12 p.p. across orderings, while DTD and SUN397 are nearly invariant. The consistent effect is positional: tasks placed in the leftmost level-1 pair are preserved better, reflecting the asymmetric role of the a and b slots under the fixed `[zs,a,b,b,a]` tree recipe.

Table 13. **Per-task ordering sensitivity.** Range of top-1 accuracy across the four 8-task tree orderings, sorted descending. "Best ord. (Worst)" lists the ordering achieving the column max (resp. min).

Task	Min	Max	Range (p.p.)	Best ord. (Worst)
SVHN	44.89	56.92	12.03	default (resisc_first)
MNIST	82.97	94.52	11.55	default (resisc_first)
RESISC45	53.44	61.33	7.89	resisc_first (reverse)
GTSRB	37.88	43.56	5.68	reverse (resisc_first)
EuroSAT	52.47	57.19	4.72	resisc_first (reverse)
Cars	49.99	54.43	4.44	hard_first (reverse)
SUN397	54.15	56.08	1.93	reverse (default)
DTD	39.52	40.90	1.38	resisc_first (default)

Fixing the canonical 8-task tree by input permutation alone therefore cannot close the baseline gap. A better many-way recipe would need either a flatter merge structure or a more order-symmetric construction.

A.7. Language-Domain Extension: Qwen3

To test whether the meta-merge picture survives outside vision, we run a small Qwen3 extension. Starting from `Qwen/Qwen3-0.6B` (Yang et al., 2025), we use three public fine-tunes (math, French, and code) and evaluate both the 3-way merge and all three pairwise merges on GSM8K (Cobbe et al., 2021), French-Bench (Lasbordes & Gad, 2025), GSM8K-fr, and HumanEval-Instruct (Luo et al., 2023). Because this is a small shape-matched exten-

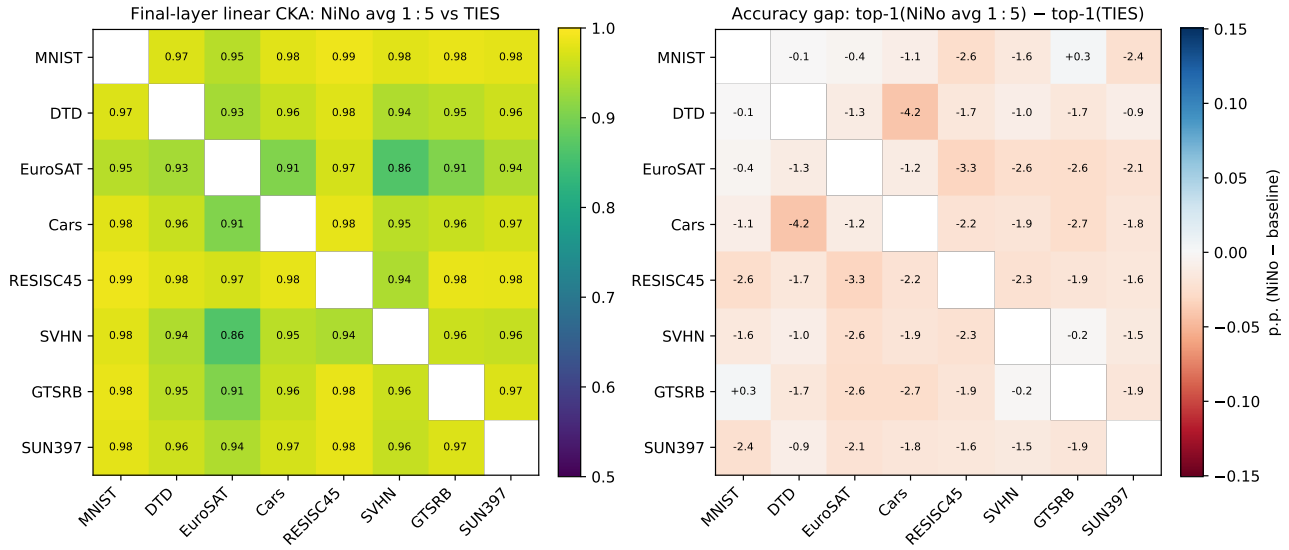

Figure 7. **NiNo avg** 1:5 **against TIES.** CKA remains high but the accuracy panel shifts more negative than the TA case, especially on DTD- and Cars-involving pairs.

Table 12. **8-task tree-merge accuracy under four input orderings.** Each row reports per-task top-1 plus the 8-task average and minimum. "First task" is the leftmost task in the input ordering (level-1 pair 1, position a). Orderings differ only in the input permutation; the recipe is the symmetric tree default `[zs,a,b,b,a]` at every node. Best per column in bold.

Method	First task	MNIST	DTD	EuroSAT	Cars	RESISC45	SVHN	GTSRB	SUN397	Avg	Min
tree, default	MNIST	**94.52**	39.52	53.09	50.25	55.78	**56.92**	39.37	54.15	55.45	39.37
tree, reverse	SUN397	84.17	40.21	52.47	49.99	53.44	54.03	**43.56**	56.08	54.24	40.21
tree, hard_first	Cars	85.37	40.43	55.84	**54.43**	55.83	46.56	39.18	55.12	54.09	39.18
tree, resisc_first	RESISC45	82.97	40.90	**57.19**	51.34	**61.33**	44.89	37.88	55.34	53.98	37.88

sion, we compare NiNo only to weight averaging and Task Arithmetic with $\lambda \in \{0.3, 0.4\}$.

For NiNo, pairwise merges use the paper's chosen `01221` recipe, and the 3-way merge uses the `01223` extension from `merge_qwen.py`. The pre-trained NiNo meta-network is unchanged from the vision setting; no Qwen-specific meta-training is performed.

Table 14 shows that TA is strongest on the macro-average for every merge configuration, so transfer to language is weaker than in the 28-pair vision benchmark. The macro-average gap from NiNo to TA (λ=0.3) widens to 5.6–9.2 p.p., much larger than the pairwise vision gap.

NiNo still retains selective strengths. It achieves the best 3-way HumanEval score at 46.3, exactly matching the standalone Code fine-tune, and it gives the strongest French score on the Fr+Code pair. We read this as partial cross-domain transfer: aggregate language merging is harder, but some task-specialized behavior still survives the meta-merge.

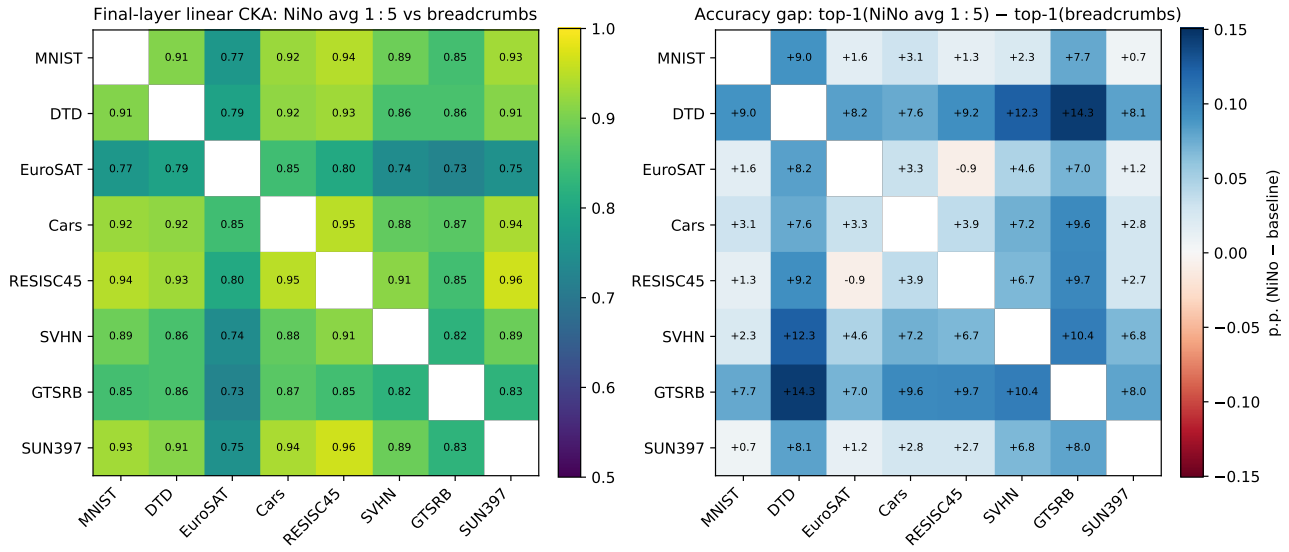

Figure 8. **NiNo avg** 1:5 **against breadcrumbs.** CKA is noticeably lower (cells on EuroSAT pairs drop below 0.65), and the accuracy panel is dominated by positive gaps because breadcrumbs underperforms on this grid.

Table 14. **Top-1 accuracy (%) of merged Qwen3-0.6B models, by merge configuration.** Per-config best result in bold. NiNo's per-config gap to TA $\lambda = 0.3$ widens to 5.6–9.2 p.p. on the macro-average, substantially above the 1.4 p.p. gap to TA on the 28-pair vision benchmark (Table 1).

Pair	Method	gsm8k	french	gsm8k-fr	humaneval	avg
zero-shot	Qwen3-0.6B base	19.5	25.1	21.1	38.4	26.0
3-way	Weight avg.	49.8	27.1	33.1	42.7	38.2
3-way	TA, $\lambda = 0.3$	**50.6**	**27.3**	**35.6**	42.7	**39.1**
3-way	TA, $\lambda = 0.4$	47.8	27.1	30.8	**43.9**	37.4
3-way	NiNo	39.1	24.4	23.5	**46.3**	33.3
Math+Fr	Weight avg.	49.3	**28.0**	29.3	31.7	34.6
Math+Fr	TA, $\lambda = 0.3$	**53.1**	27.5	**36.8**	**32.3**	**37.4**
Math+Fr	TA, $\lambda = 0.4$	53.0	27.7	35.3	32.9	37.2
Math+Fr	NiNo	44.0	26.1	28.8	26.8	31.4
Math+Code	Weight avg.	48.1	25.9	29.9	**42.7**	36.7
Math+Code	TA, $\lambda = 0.3$	**53.8**	25.3	**35.9**	**42.7**	**39.4**
Math+Code	TA, $\lambda = 0.4$	51.9	25.5	34.3	**43.3**	38.8
Math+Code	NiNo	42.2	25.4	27.2	26.2	30.2
Fr+Code	Weight avg.	43.3	26.9	29.9	**42.7**	35.7
Fr+Code	TA, $\lambda = 0.3$	**49.2**	26.8	**35.3**	42.1	**38.3**
Fr+Code	TA, $\lambda = 0.4$	48.8	27.1	33.3	40.9	37.5
Fr+Code	NiNo	37.8	**30.2**	27.6	35.4	32.7