DISSECTING SOFT-TARGET SUPERVISION IN KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation trains a student to match a teacher's predictive distribution. The probabilities assigned to non-target classes provide supervision beyond the ground-truth label. Permuting those probabilities changes their class assignments but still supplies an individual target for each class, so it does not directly test the effect of omitting those targets. This study compares unmodified and permuted targets with grouped and residual-uniform supervision. Grouped supervision keeps the retained-class targets and supervises the residual classes only through their total probability. Residual-uniform supervision keeps the same targets and assigns the residual mass equally across the remaining classes. All four conditions preserve the same retained probabilities and residual mass for a given training example and retained set. For each ground-truth class, we define two fixed sets of K non-target classes, denoted Top and Bottom, using the largest and smallest mean teacher probabilities, respectively. These means are computed over training images with that ground-truth label, rather than separately for each image. On CIFAR-100 and Tiny-ImageNet, the grouped-minus-permuted effect differs between Top and Bottom by +1.49 and +0.85 macro-F1 percentage points. Grouped supervision underperforms permutation for Bottom in both settings. For Top, the grouped-minus-permuted interval is above zero on CIFAR-100 but includes zero on Tiny-ImageNet. Residual-uniform targets improve over Bottom-grouped supervision by +1.06 and +1.12 points. Three additional CIFAR training seeds show the same interaction and negative Bottom $C - P$ effect. These comparisons evaluate changes in class assignment and omission of conditional residual supervision. The confidence intervals condition on the fitted models and mappings. The analysis reuses development sets from the preceding permutation analyses and does not match probability mass between Top and Bottom.

1 INTRODUCTION

Knowledge distillation trains a student with a teacher's predictive distribution (Hinton et al., 2015). Beyond the probability assigned to the ground-truth class, a soft target specifies a probability for every non-target class. These values are often interpreted as carrying information about relations among classes that is absent from the hard label. A basic empirical question is therefore not only whether soft targets help, but which parts of that class-level specification matter for student learning.

A common empirical strategy for addressing this question is to intervene on the teacher distribution. Prior work has permuted non-maximal predictions, separated target and non-target losses, retained selected probabilities, or otherwise transformed logits and targets (Furlanello et al., 2018; Zhao et al., 2022; Qin et al., 2024). Such interventions are informative only relative to what they preserve. Reassigning a probability value changes which class receives that target, but it still supplies an individual target to every class. Omitting that class-specific target is a different intervention. Conflating the two can make a reassignment result appear to answer a stronger question about whether the corresponding supervision was needed at all.

For a fixed group of non-target classes, this distinction can be made explicit. Their teacher probabilities can be permuted among the classes, supervised only through their sum, or replaced by equal values with the same sum. The grouped loss retains the residual mass without specifying the teacher's conditional distribution within the group. Permuted and residual-uniform targets both spec-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ify such a distribution, but with different class-level assignments. These controls separate sensitivity to reassignment from the effect of removing explicit supervision of the within-group conditional distribution.

This distinction matters beyond a single ablation because distillation methods routinely filter, mask, or restructure teacher outputs. A drop under permutation does not show that the original nonuniform residual ratios are necessary: the student still receives individual class targets, only reassigned. A direct omission control asks the stronger question of whether those ratios need to be specified at all. The comparison therefore concerns the information carried by supervision rather than a new optimizer or student architecture.

We study this comparison using fixed class-conditioned retained sets. For each ground-truth class, the Top and Bottom sets contain the K non-target classes with the largest and smallest mean teacher probabilities over training images with that label. These are not per-image ranks. Within each retained-set condition, all four supervision operators preserve the ground-truth probability, the retained probabilities, and the total residual mass. They differ only in how the residual classes are supervised.

The key contribution of this study is a controlled empirical separation of three interventions that are often conflated in soft-target ablations: reassigning residual probabilities, omitting their class-specific supervision, and replacing them with a uniform residual target. Holding the retained probabilities and residual mass fixed within each retained-set condition, we show that permutation sensitivity does not determine the effect of omission. The grouped-minus-permuted effect changes between Top and Bottom in both systems, while residual-uniform targets improve over Bottom-grouped supervision in both systems without restoring the teacher's original nonuniform residual values. Three additional CIFAR training seeds show the same interaction and negative Bottom $C - P$ effect. These results clarify what a reassignment ablation establishes and what requires a direct omission control.

2 Related work

Soft targets and reassignment. Soft-target distillation transfers a teacher's predictive distribution to a student (Hinton et al., 2015). Born Again Neural Networks include Dark Knowledge with Permuted Predictions (DKPP), which keeps the teacher's maximal output and permutes the others (Furlanello et al., 2018, Section 3.2). Here, the retained class set contains the ground-truth class and non-target classes selected from training-set means. Since the ground-truth class need not be the teacher's argmax, this selection rule differs from DKPP's.

Non-target distillation and target ablations. Decoupled Knowledge Distillation (DKD) separates target-versus-rest divergence from conditional non-target divergence and weights them independently (Zhao et al., 2022). The grouped objective uses the same KL decomposition for a partition with several individually retained classes and one residual group. Qin et al. (2024) study soft-label structure in dataset distillation. Their January 2025 arXiv revision describes per-example probability-rank swaps and top-k retention with the remaining probabilities zeroed without renormalization. In contrast, the present comparison uses fixed training-derived class sets and retains the residual group's total probability when omitting its individual targets. This permits reassignment and grouped supervision to be compared at the same retained probabilities and residual mass.

Uniform targets and training dynamics. Müller et al. (2019) examine label smoothing, learned representations, and subsequent distillation. Yuan et al. (2020) relate distillation to learned label smoothing and study teacher-free targets, including manually specified distributions. The residual-uniform control here still uses teacher-derived retained sets, retained probabilities, and residual total probability. It is therefore not a teacher-free target. Cho & Hariharan (2019) show that more accurate teachers do not necessarily produce better distilled students. Stanton et al. (2021) distinguish student generalization from fidelity to the teacher and identify optimization limitations. Beyer et al. (2022) study the effects of consistent teacher–student input views and longer training. The current experiments instead use fixed 60-epoch training with teacher targets cached before random augmentation and augmented student inputs.

Logit transformations and masking. Kim et al. (2021) analyze the effect of temperature on the KL objective and compare it with logit mean squared error. Logit Standardization applies a weighted Z-score to teacher and student logits before softmax (Sun et al., 2024). Refined Logit Distillation aligns the teacher's maximum-probability confidence with the student's true-class confidence and separately matches a distribution normalized over unmasked classes (Sun et al., 2025). Its mask depends on each example's teacher logits and ground-truth label. DeepKD uses separate gradient-momentum paths and dynamic masking of non-target logits before softmax (Huang et al., 2025). These methods modify logits or the optimization procedure. The experiments here instead modify already-softened teacher probabilities or aggregate them in the loss, while keeping the optimizer fixed within each system.

3 METHODS

3.1 RETAINED CLASS SETS AND RESIDUAL MASS

Let L be the number of classes, y the ground-truth label, and $q(x)$ the cached teacher distribution at temperature $\tau > 0$. For student logits z, write $p = \mathrm{softmax}(z/\tau)$. Teacher targets are computed from normalized images without random augmentation. The student receives randomly augmented images. For each true class, define

$$\mu_y(k) = \frac{1}{|\mathcal{T}_y|} \sum_{i \in \mathcal{T}_y} q_i(k), \qquad \mathcal{T}_y = \{i \in \mathrm{TRAIN} : y_i = y\}, \quad k \neq y. \tag{1}$$

The Top retained set S_y contains the K largest non-target means, and the Bottom retained set contains the K smallest, with class-index tie breaking. The retained set is fixed for each ground-truth class, rather than selected separately for each example. $K = 25$ for CIFAR-100 and $K = 50$ for Tiny-ImageNet.

Let $A_y = \{y\} \cup S_y$ be the retained class set including the ground-truth class and $R_y = \{1, \ldots, L\} \setminus A_y$ the residual group. Write $m = |R_y|$, $Q_R = \sum_{k \in R_y} q_k$, and $P_R = \sum_{k \in R_y} p_k$, suppressing y below. The residual group has 74 or 149 classes. Top and Bottom match retained-set size, not retained probability mass. For a fixed retained class set, the four supervision conditions share A, q_A, and Q_R.

Figure 1 summarizes the distribution comparison in each condition. Grouped supervision compares the residual sums, whereas the other conditions specify an individual target for every class. The student retains all L output classes. The label-dependent retained sets are used only to construct training supervision, not to select outputs at inference.

3.2 SUPERVISION CONDITIONS

Unmodified and permuted targets. Unmodified teacher targets (O) use q. For permuted supervision P, π_y fixes A and forms a Sattolo cycle on R (Sattolo, 1986). In source-to-destination notation,

$$q^P_j = q_j \quad (j \in A), \qquad q^P_{\pi_y(k)} = q_k \quad (k \in R). \tag{2}$$

Each mapping is shared within a true class and remains fixed over training. Permutation preserves the probability multiset, target-class probability, residual mass, and entropy. It changes which residual class receives each value, without deleting any value.

Grouped supervision. Grouped supervision (C) uses grouped vectors $\bar{q} = (q_A, Q_R)$ and $\bar{p} = (p_A, P_R)$. Its soft cross-entropy is

$$\ell_C(q, p) = -\sum_{k \in A} q_k \log p_k - Q_R \log P_R. \tag{3}$$

This loss does not specify the conditional distribution q_R/Q_R within the residual group. Residual logits still receive gradients through P_R, the hard-label loss, and shared parameters. Here, *omission* refers only to removing explicit supervision of that conditional distribution.

(a) Class-conditional selection

$$\mu_y(k) = \frac{1}{|\mathcal{T}_y|} \sum_{i \in \mathcal{T}_y} q_i(k), \quad k \neq y$$

$$\mu_y(k_1) \geq \cdots \geq \mu_y(k_{L-1})$$

$$S_y^{\mathrm{Top}} = \{k_1, \ldots, k_K\}$$

$$S_y^{\mathrm{Bottom}} = \{k_{L-K}, \ldots, k_{L-1}\}$$

$$A_y = \{y\} \cup S_y$$

$$R_y = \{1, \ldots, L\} \setminus A_y$$

(b) Distributions compared by the soft loss

Condition	Teacher	Student
Unmodified (O)	$(q_A,\ q_R)$	$(p_A,\ p_R)$
Permuted (P)	$(q_A,\ q_R^{\pi})$	$(p_A,\ p_R)$
Grouped (C)	$(q_A,\ Q_R)$	$(p_A,\ P_R)$
Uniform (U)	$(q_A,\ \frac{Q_R}{m}\mathbf{1}_m)$	$(p_A,\ p_R)$

$$Q_R = \textstyle\sum_{k \in R} q_k, \qquad P_R = \textstyle\sum_{k \in R} p_k, \qquad m = |R|$$

Figure 1: Class-set selection and the distributions compared during distillation. (a) Non-target classes are ranked by their training-set class-conditional mean probabilities, fixing either the Top or Bottom set for each label y. (b) q_R^{π} denotes the permuted residual vector and $\mathbf{1}_m$ the all-ones vector. Grouped supervision compares sums only in the loss, without changing the student's L outputs. All four comparisons preserve q_A and Q_R for a fixed retained set.

Uniform residual targets. Uniform supervision U uses

$$q_j^U = \begin{cases} q_j, & j \in A, \\ Q_R/m, & j \in R. \end{cases} \tag{4}$$

The probabilities on A and the total residual mass are unchanged, while the conditional target distribution on R is uniform. Unlike permutation, residual-uniform targets generally change the probability multiset and entropy. Neither transformation guarantees preservation of the teacher's argmax class when that class lies in the residual group. It is not teacher-free. The subscripts in P_{Top} and P_{Bottom} identify the retained set, with the same convention for C and U.

3.3 Loss decomposition and training

For normalized distributions and $Q_R > 0$, the KL chain rule gives

$$\mathrm{KL}(q\|p) = \mathrm{KL}(\bar{q}\|\bar{p}) + Q_R\,\mathrm{KL}(q_R/Q_R\|p_R/P_R). \tag{5}$$

At fixed A, permutation retains the first term and replaces the teacher's conditional residual distribution in the second. Grouped supervision retains the grouped term without prescribing that conditional distribution. Uniform supervision replaces it with $u_R = (1/m, \ldots, 1/m)$, so

$$\mathrm{KL}(q^U\|p) - \mathrm{KL}(\bar{q}\|\bar{p}) = Q_R\,\mathrm{KL}(u_R\|p_R/P_R). \tag{6}$$

These identities describe the objectives, not the final performance ordering. In soft cross-entropy, $\ell_U - \ell_C$ additionally contains the teacher-only constant $Q_R \log m$. Appendix A gives the full scalar and gradient expressions.

For a minibatch of B examples, the training objective is

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \left[(1-\lambda)\ \mathrm{CE}(z_i, y_i) + \lambda \tau^2 d_i \right]. \tag{7}$$

Here $\lambda \in [0, 1]$ weights the distillation term, z_i is the student logit vector, and CE is hard-label cross-entropy on the unscaled logits. The sample-level soft loss d_i excludes the factor τ^2. All main comparisons use $\lambda = 0.9$ and $\tau = 4$, corresponding to hard-label and soft-loss weights of 0.1 and 0.9. These weights are fixed across the four supervision conditions, not tuned separately for each condition. CIFAR uses soft cross-entropy for $O/P/U$ and Eq. 3 for C. The main Tiny-ImageNet comparisons use full KL for $O/P/U$ and grouped cross-entropy for C. Each soft loss sums over classes or groups before the single batch average in Eq. 7. Both terms are averaged once over the minibatch, and the soft term is multiplied by the squared temperature once. Teacher-only constants leave exact student gradients unchanged, but need not yield identical mixed-precision training trajectories. Stored FP32 targets are not renormalized after transformation. Temperature, loss weights, retained-set sizes, and training duration are fixed across conditions.

Table 1: Training settings. Both datasets use $\tau = 4$ and $\lambda = 0.9$ in Eq. 7, giving hard-label and soft-loss weights of 0.1 and 0.9. The main seeds are 19, 23, and 29. Tiny evaluation excludes ten exact-content overlaps identified after the permutation experiments.

Setting	CIFAR-100	Tiny-ImageNet
Classes	100	200
Student train / development	40,000 / 10,000	80,000 / 19,990
Teacher training images	40,000	72,000
Teacher / student	ResNet32(w2) / 20(w1)	ResNet34 / 18
Retained non-target classes K	25	50
Residual classes	74	149
Mappings per retained set	2	3
Student epochs / batch	60 / 128	60 / 256
Student precision	FP32	FP16 AMP

4 Experimental Setup

Data and models. The experiments use partitions of the CIFAR-100 (Krizhevsky, 2009) and Tiny-ImageNet (Le & Yang, 2015) training sets. Table 1 lists the settings shared across conditions within each system and the differences between systems. Teacher and student ResNets (He et al., 2016) start without pretrained weights. Tiny students train on 80,000 images, including the 8,000 inner-validation images excluded from the teacher's 72,000-image training set. The development set contains 19,990 images after the post-hoc exclusion described below. Appendix B gives architecture and normalization details.

Training and paired seeds. Students use 60 epochs of SGD with initial learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and epoch-stepped cosine decay. CIFAR uses reflect padding, random 32×32 crops, and horizontal flips. Tiny uses zero padding, random 64×64 crops, and flips. Teacher targets are cached from normalized images without random augmentation. All student conditions use the same augmentation procedure.

The primary comparisons use seeds 19, 23, and 29. Each training seed determines initialization, minibatch order, and augmentation randomness, which are paired across supervision conditions. A permutation mapping is a separate fixed reassignment of class coordinates. There are two mappings per CIFAR retained set and three per Tiny retained set, with a separately trained model for each seed–mapping combination. Grouped, uniform, and unmodified supervision use one model per seed and retained-set condition, with the unmodified model shared across retained sets.

The uniform controls add twelve students across the two systems. Three additional CIFAR seeds each add six students: Top/Bottom grouped and Top/Bottom permuted under both fixed mappings. All thirty final checkpoints were fixed before development evaluation, without intermediate development-based model selection. The five-seed Tiny experiment in Appendix F used a separate training setup.

Metric and comparisons. Macro-F1 is computed per model over all classes and reported as a percentage. A class with zero F1 denominator contributes zero. Differences are percentage points (pp). Let $\mathcal{F}(a)$ denote the mean macro-F1 for condition a, obtained by averaging permutation-model scores over mappings within each seed and then over three seeds. Other conditions enter once per seed. Scores, not predictions, are averaged. Define

$$\Delta_T = \mathcal{F}(C_{\mathrm{Top}}) - \mathcal{F}(P_{\mathrm{Top}}), \quad \Delta_B = \mathcal{F}(C_{\mathrm{Bottom}}) - \mathcal{F}(P_{\mathrm{Bottom}}), \quad I = \Delta_T - \Delta_B. \tag{8}$$

The interaction I directly compares the grouped-minus-permuted effects for Top and Bottom. Differences such as $C - P$ in plots and tables use this score shorthand.

Uncertainty. The analysis uses 20,000 paired class-stratified bootstrap resamples of development images. Each resample is shared across the relevant models, and model F1 is recomputed before aggregation. Teachers, trained models, and mappings remain fixed. The confidence intervals therefore describe evaluation-sample uncertainty, not variation over new training seeds or teachers. Separate

families contain six main CIFAR contrasts, three Tiny contrasts, and ten uniform/additional-seed contrasts, giving adjusted levels of approximately 99.17%, 98.33%, and 99.5%.

The CIFAR $O-C_{\mathrm{Top}}$ contrast uses a one-sided bound and a prespecified 0.5 pp non-inferiority margin. For $P-U$, equivalence under the prespecified criterion requires the entire adjusted interval to lie inside $[-0.5, 0.5]$ pp. The 0.5 pp margin is a prespecified comparison tolerance. Appendix B.2 gives the interval construction; random seeds are retained in the supplement. These separate adjustments do not provide a joint correction across the full sequence of studies.

Development-set reuse. The development sets had already been evaluated in the preceding permutation analyses before the grouped and uniform comparisons were specified. These analyses are summarized in Section 5.1 and Appendix F. Ten exact-content Tiny training/development overlaps were identified after the permutation experiments. Their development images were excluded from every reported condition, but the models were not retrained. The additional runs did not access the official test set. Appendix E documents separate CIFAR100-C feature and probe analyses and the limits of the available records.

5 Results

5.1 Probability reassignment

On Tiny-ImageNet, unmodified targets outperformed permuted targets by 3.01 pp in a five-seed experiment and by 2.85 pp in a separate 20-mapping experiment. The two experiments used different training settings and uncertainty estimates, as detailed in Appendix F. Both comparisons show lower student macro-F1 under reassignment. The following comparisons ask how this sensitivity differs from the effect of omitting individual residual targets.

Table 2 reports macro-F1 for all four supervision conditions in the main three-seed comparisons. Unmodified targets have the highest mean scores in these rows, so an improvement over grouped supervision does not by itself establish preservation of unmodified-target performance.

5.2 Reassignment and grouped supervision

On CIFAR-100, Top-grouped minus Top-permuted is +0.76 pp, with adjusted interval $[+0.21, +1.31]$, whereas Bottom-grouped minus Bottom-permuted is -0.73 pp, $[-1.39, -0.06]$. Their difference, the interaction I, is +1.49 pp, $[+0.63, +2.36]$.

On Tiny-ImageNet, the Top effect is +0.05 pp, $[-0.26, +0.36]$, the Bottom effect is -0.80 pp, $[-1.21, -0.39]$, and the interaction is +0.85 pp, $[+0.34, +1.36]$. Figure 2 shows a positive interaction and a negative Bottom effect in both systems, but the Tiny Top interval includes zero. Thus, the retained-set dependence repeats across systems, whereas a Top advantage for grouped supervision is supported only on CIFAR-100. The Top and Bottom sets have equal counts but different probability mass.

Table 2: Mean macro-F1 (%) for the main three-seed comparisons. Permutation scores are averaged over mappings within each seed, then over seeds. The unmodified-target reference is shared between the Top and Bottom rows. The Tiny reference uses the same training settings as its main comparisons. Means are descriptive.

Dataset	Set	Unmodified (O)	Permuted (P)	Grouped (C)	Uniform (U)
CIFAR-100	Top	66.33	65.10	65.86	65.23
	Bottom	66.33	62.50	61.77	62.83
Tiny-ImageNet	Top	64.21	63.96	64.01	64.11
	Bottom	64.21	61.99	61.19	62.31

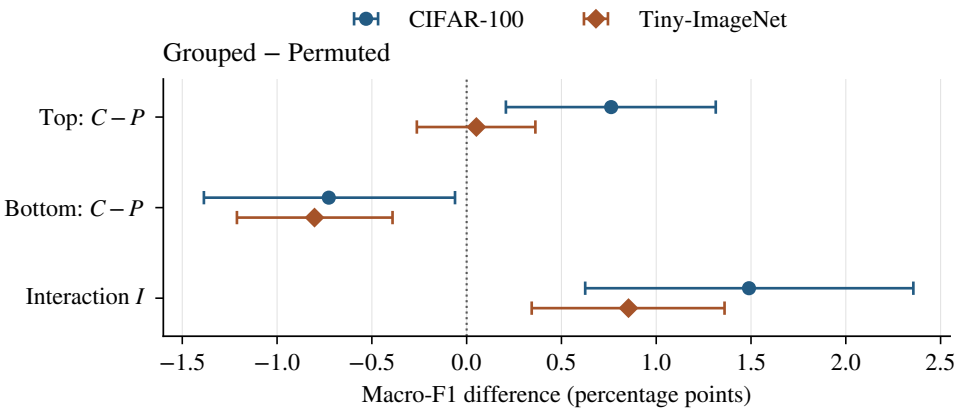


Figure 2: Grouped minus permuted supervision in the main three-seed comparisons. Points are mean macro-F1 differences after averaging permutation scores over mappings within each seed. Bars are adjusted evaluation-bootstrap intervals: 99.17% for CIFAR-100 ($N = 10{,}000$) and 98.33% for Tiny-ImageNet ($N = 19{,}990$). All trained models and mappings are held fixed. I is the Top effect minus the Bottom effect.

5.3 Uniform residual supervision

The uniform control tests whether equal residual targets improve over supervising only the residual sum. Bottom-uniform exceeds Bottom-grouped by +1.06 pp on CIFAR, $[+0.29, +1.83]$, and +1.12 pp on Tiny, $[+0.57, +1.67]$. For Top, the differences are -0.63 pp, $[-1.33, +0.04]$, and +0.09 pp, $[-0.34, +0.54]$, respectively. Both Top intervals include zero. Figure 3 places these effects beside the direct permutation–uniform comparisons.

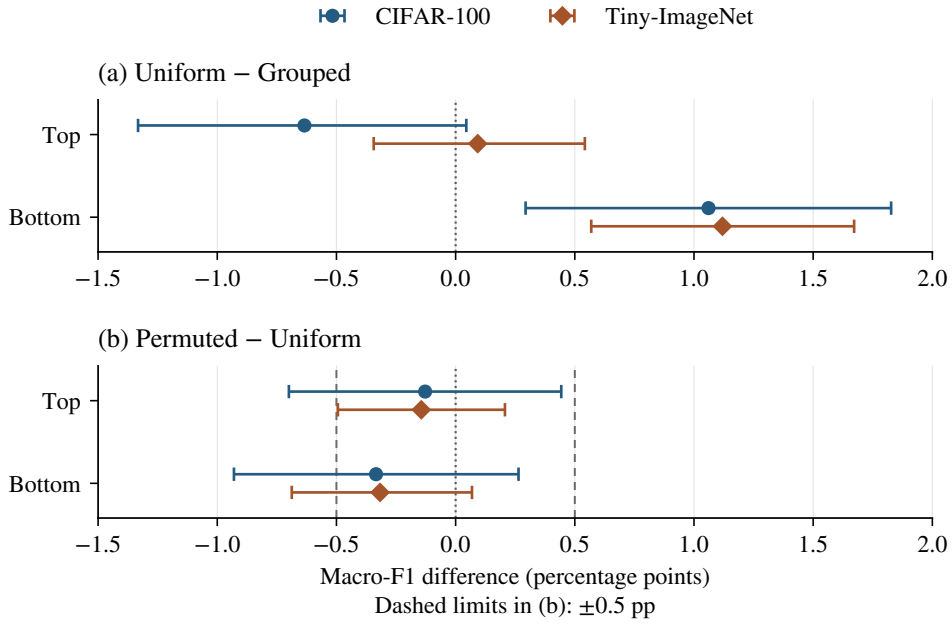


Figure 3: Uniform controls: (a) uniform minus grouped and (b) permuted minus uniform. Points are three-seed mean macro-F1 differences after within-seed mapping averages. Bars are 99.5% evaluation-bootstrap intervals with all models and mappings fixed. Dashed vertical lines in (b) mark the prespecified equivalence limits of ± 0.5 pp. Evaluation sets match Figure 2.

All four permutation-minus-uniform intervals include zero. Only Tiny Top lies wholly inside the prespecified ± 0.5 pp equivalence limits: its estimate is -0.144 pp and interval $[-0.494, +0.207]$. It narrowly meets this conditional criterion. Its lower endpoint lies close to -0.5 pp, making the finite bootstrap resolution relevant. Neither CIFAR retained set nor Tiny Bottom meets the criterion. Uniform targets therefore improve over Bottom-grouped supervision without reassigning the teacher's nonuniform residual values, but the results do not establish general interchangeability of uniform and permuted targets.

5.4 Training-seed repeatability

To test repeatability with respect to training randomness, we repeat the CIFAR comparison with three additional training seeds, changing initialization, minibatch order, and augmentation randomness while holding the teacher, retained sets, mappings, and development images fixed. The mean interaction is +1.46 pp, $[+0.56, +2.36]$, and Bottom-grouped minus Bottom-permuted is -0.98 pp, $[-1.68, -0.30]$. All three interactions are positive and all three Bottom contrasts are negative.

Figure 4 shows the seed-level contrasts for the primary and additional comparisons. Each permutation score averages two mapping-specific models. The bootstrap holds the 18 additional trained models fixed and quantifies evaluation-sample uncertainty, while the individual points show variation across training seeds. The evaluation set and mappings are unchanged. Across the six CIFAR seeds, the descriptive interaction mean is 1.475 pp with sample standard deviation 0.590 pp. The Bottom mean is -0.855 pp with standard deviation 0.485 pp. The supplement gives the individual values. These summaries do not replace the separate prespecified confidence intervals.

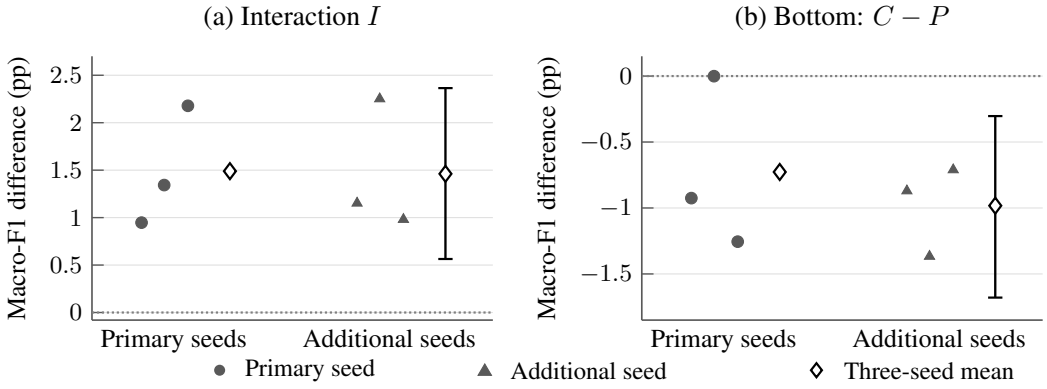


Figure 4: CIFAR seed-level contrasts for (a) interaction I and (b) Bottom $C - P$. Each point uses all models trained with one seed, averaging permutation scores over two mappings. Circles show the primary seeds (19, 23, 29), and triangles show the three additional seeds. Open diamonds mark group means. Error bars on the additional means are 99.5% evaluation-bootstrap intervals, not intervals over training seeds. All models use the same 10,000 development images. Dotted lines mark zero.

Non-inferiority to unmodified targets. The CIFAR unmodified-minus-Top-grouped estimate is +0.47 pp. Its adjusted one-sided upper bound is +1.02 pp, above the 0.5 pp margin. The criterion is therefore not met, even though the mean macro-F1 deficit is below the margin. The additional-seed experiment did not train new unmodified-target models. Appendix C reports all primary contrasts; the supplement provides seed-level differences.

Secondary metric. The saved predictions show the same direction for the Bottom comparisons under top-1 accuracy. Grouped, permuted, and residual-uniform supervision obtain 61.77%, 63.12%, and 63.57% on CIFAR-100, and 61.32%, 62.25%, and 62.66% on Tiny-ImageNet. This descriptive check is consistent with the Bottom macro-F1 results and is reported without a separate inferential analysis. The supplement provides scores for all conditions and additional seeds. Accuracy was examined post hoc, without confidence intervals or new hypothesis tests.

6 DISCUSSION AND LIMITATIONS

Reassignment versus omission. A permutation ablation alone cannot determine how performance would change if the corresponding individual targets were omitted.

What the uniform control adds. Uniform targets improve over Bottom-grouped supervision in both systems, so the improvement does not require reassigning the teacher's original nonuniform residual values. Teacher information remains in the retained-set selection, retained probabilities, and residual total mass. The comparison does not isolate regularization as the explanation. Only Tiny Top meets the specified permutation–uniform equivalence criterion, and only narrowly.

Permutation averaging. For a uniformly sampled residual cycle, $\mathbb{E}[q_j^P] = (Q_R - q_j)/(m-1)$, rather than Q_R/m. The comparisons use fixed cycles and separately trained students. Residual-uniform supervision is not assumed to equal the mean target of the executed mappings. Appendix A derives the cycle expectation and distinguishes it from averaging all permutations.

Mass and optimization. Bottom leaves more residual mass than Top at the same retained count (Table 3). Combining Eqs. 5 and 7, the conditional residual divergence has coefficient $\lambda\tau^2 Q_R$. A fixed λ does not equalize this coefficient across Top and Bottom. The interaction therefore does not isolate the effect of retained class identity from the amount of probability mass assigned to those classes. Gradient magnitude, direction, and other optimization effects are not separated.

Dataset	Retained set	$\overline{q_y}$	$\overline{M}$	$\overline{Q_R}$	$\overline{F}$
CIFAR-100	Top	0.2829	0.3589	0.3581	0.5097
	Bottom	0.2829	0.0791	0.6380	0.1069
Tiny-ImageNet	Top	0.1131	0.3324	0.5545	0.3763
	Bottom	0.1131	0.1497	0.7372	0.1680

Table 3: Mean teacher probability mass at $\tau = 4$ over 40,000 CIFAR-100 and 80,000 Tiny-ImageNet training images. $M = \sum_{k \in S_y} q_k$ is the retained non-target mass, Q_R is the residual mass, and $F = M/(1-q_y)$ is computed per example before averaging. O/P/C/U share these quantities within each retained set. Appendix D gives computation details.

Generalization across training settings. Each system uses one teacher, a fixed mapping set, retained-set size, loss weighting, and temperature. The results therefore characterize the tested teacher–student models and fixed training settings. Architectures, image resolutions, and precision also differ between systems, so cross-system differences combine these factors. Tiny Top includes zero, and CIFAR Top-grouped does not meet the 0.5 pp non-inferiority criterion against unmodified targets.

Evaluation limitations. The additional seeds test training-randomness repeatability on the same development images, not new evaluation data. Tiny's ten overlapping development images were excluded, but the models were not retrained. The adjusted intervals hold the models fixed and do not account for research decisions informed by previous results on these development sets. Appendix E details the data reuse.

When interpreting permutation ablations, compare them with both grouped and residual-uniform supervision to distinguish reassignment from the omission of individual targets.

7 CONCLUSION

Reassignment and omission of residual teacher probabilities produce different student outcomes. Across both systems, grouped-minus-permuted effects differ between Top and Bottom, while residual-uniform targets improve over Bottom-grouped supervision. Additional CIFAR seeds show the same interaction and negative Bottom $C - P$ effect. These results motivate a grouped control when permutation is used to assess omission of individual residual targets.

REPRODUCIBILITY STATEMENT

Appendices A–F give the derivations, experimental details, and supporting results. The supplement provides the supervision code, saved predictions, and a CPU script that recalculates all 19 primary comparisons with their reported confidence intervals or one-sided bound. It reproduces analyses from saved predictions, not teacher or student training.

REFERENCES

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10925–10934, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Beyer_Knowledge_Distillation_A_Good_Teacher_Is_Patient_and_Consistent_CVPR_2022_paper.html.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.html.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. DeepKD: A deeply decoupled and denoised knowledge distillation trainer. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0915. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2756ffa0487af8d276a0b1baf78732f4-Abstract-Conference.html.

Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2628–2635, 2021. doi: 10.24963/ijcai.2021/362. URL https://www.ijcai.org/proceedings/2021/362.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, CS231n course project report, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4194. URL https://papers.nips.cc/paper_files/paper/2024/hash/ee45939f4403e8abd63a15a29a9c055b-Abstract-Conference.html. Operational details follow arXiv:2406.10485v2 (2025), Sections 4.2–4.3 and Appendix C.

Sandra Sattolo. An algorithm to generate a random cyclic permutation. *Information Processing Letters*, 22(6):315–317, 1986. doi: 10.1016/0020-0190(86)90073-6. URL https://doi.org/10.1016/0020-0190(86)90073-6.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Logit_Standardization_in_Knowledge_Distillation_CVPR_2024_paper.html.

Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1110–1119, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Sun_Knowledge_Distillation_with_Refined_Logits_ICCV_2025_paper.html.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Revisiting_Knowledge_Distillation_via_Label_Smoothing_Regularization_CVPR_2020_paper.html.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_Decoupled_Knowledge_Distillation_CVPR_2022_paper.html.

A Loss identities and implementation

For normalized targets and $Q_R > 0$, splitting the KL sum over A and R gives Eq. 5. For soft cross-entropy, the differences additionally contain teacher-only entropy terms. Writing $r = q_R/Q_R$, $s_R = p_R/P_R$, and $H(r) = -\sum_j r_j \log r_j$,

$$\ell_O - \ell_C = Q_R\{H(r) + \mathrm{KL}(r\|s_R)\}, \tag{9}$$

$$\ell_U - \ell_C = Q_R\{\log m + \mathrm{KL}(u_R\|s_R)\}. \tag{10}$$

Removing the teacher entropy gives the full-KL versions in Eqs. 5–6. Constants independent of the student do not change exact gradients, but finite-precision training trajectories need not coincide.

The implementation does not renormalize stored FP32 targets. Let $v = z/\tau$ and $s = \sum_j q_j$. Differentiating the grouped loss gives

$$\frac{\partial \ell_C}{\partial v_j} = \begin{cases} sp_j - q_j, & j \in A, \\ sp_j - Q_R p_j/P_R, & j \in R. \end{cases} \tag{11}$$

The soft-loss contribution to Eq. 7 is $\lambda\tau/B$ times this derivative. Residual gradients are generally nonzero. Log-softmax and log-sum-exp are used for numerical stability. Permuted targets are gathered with π_y^{-1}; uniform targets use Q_R/m without subsequent renormalization.

For $Q_R = 0$, the weighted conditional term is zero by continuity. The experiments use nonempty residual groups ($m = 74$ or 149). Averaging all single m-cycles gives $\mathbb{E}[q_j^P] = (Q_R - q_j)/(m-1)$ because a coordinate cannot receive its own value. This differs from the all-permutation average Q_R/m and need not equal the average of the fixed mappings used here.

B Experimental and statistical details

B.1 Models and training

CIFAR ResNets use learned projection shortcuts: the depth-32 width-two teacher has five blocks per stage and widths 32/64/128, and the depth-20 width-one student has three blocks per stage and widths 16/32/64. Normalization uses means $(0.5071, 0.4867, 0.4408)$ and standard deviations $(0.2675, 0.2565, 0.2761)$. The teacher checkpoint is epoch 60.

Tiny uses ResNet34/18 with basic-block layouts $(3, 4, 6, 3)/(2, 2, 2, 2)$, widths 64/128/256/512, a 3×3 stride-one stem, no max-pooling, and 200 outputs. Normalization uses means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$; the teacher checkpoint is epoch 80. The five-seed permutation experiment used loader-generated augmentation randomness, whereas the main comparisons used sample/epoch-keyed augmentation and separate minibatch-order generators. Their unmodified references are not pooled.

Training used PyTorch (Paszke et al., 2019) 2.6.0, CUDA 12.4, and torchvision 0.21.0. CIFAR disabled TF32; Tiny disabled matrix-multiplication TF32 and retained its recorded cuDNN policy. Tiny's six grouped runs had 28 GradScaler skips inferred from scale reductions; its six uniform runs had 29, also counted by optimizer hooks. No compensating updates were added. Exact training seeds and settings are supplied in `data/settings.json`.

B.2 Statistical analysis

Macro-F1 is calculated separately for each model. Permutation scores are averaged over the two CIFAR or three Tiny mappings within each seed, then paired differences are averaged over three seeds. Other conditions contribute one model per seed. Predictions are not ensembled, and reused models are not counted as extra runs.

Each analysis uses 20,000 paired, class-stratified bootstrap resamples. The same sampled image weights are applied to all relevant models within a dataset, with teachers, models, and mappings fixed. For family size m_f and $\alpha = 0.05$, two-sided percentile limits use $\alpha/(2m_f)$ and $1 - \alpha/(2m_f)$; the CIFAR non-inferiority bound uses $1 - \alpha/m_f$. The family sizes are 6 for main CIFAR, 3 for main Tiny, and 10 jointly for uniform and additional-seed comparisons. The original random seeds and draw order are retained in the supplied settings and analysis code.

The displayed CIFAR interaction is the negative of stored H5, so its interval is transformed from $[l, u]$ to $[-u, -l]$. For $P - U$, the entire adjusted interval must lie within $[-0.5, 0.5]$ pp. Tiny Top meets this criterion narrowly; only about 50 of 20,000 draws fall in each 0.25% tail. The analyses do not correct for the full history of development-set use. The separate permutation intervals in Appendix F retain their original uncertainty scopes.

C Complete main comparisons

Table 4 includes all nineteen primary comparisons, including null results and the failed unmodified-target non-inferiority criterion. Seed-level differences and individual model scores are available from the supplement's `reproduce.py`; Figure 4 displays the main and additional CIFAR seed-level contrasts.

Table 4: All 19 primary differences (pp), with family-wise adjusted intervals. The $O - C_{\mathrm{Top}}$ row reports a one-sided upper bound. Interactions use the orientation in Eq. 8.

Contrast	Estimate	Lower	Upper
CIFAR-100, main comparisons, family size 6			
$P_{\mathrm{Top}} - P_{\mathrm{Bottom}}$	$+2.599$	$+2.090$	$+3.119$
$C_{\mathrm{Top}} - C_{\mathrm{Bottom}}$	$+4.088$	$+3.395$	$+4.787$
$C_{\mathrm{Top}} - P_{\mathrm{Top}}$	$+0.762$	$+0.207$	$+1.314$
$C_{\mathrm{Bottom}} - P_{\mathrm{Bottom}}$	-0.727	-1.386	-0.061
I	$+1.489$	$+0.625$	$+2.356$
$O - C_{\mathrm{Top}}$	$+0.470$	—	$+1.024$
Tiny-ImageNet, main comparisons, family size 3			
$C_{\mathrm{Top}} - P_{\mathrm{Top}}$	$+0.051$	-0.263	$+0.363$
$C_{\mathrm{Bottom}} - P_{\mathrm{Bottom}}$	-0.802	-1.212	-0.391
I	$+0.854$	$+0.343$	$+1.360$
CIFAR-100, uniform controls, family size 10			
$U_{\mathrm{Top}} - C_{\mathrm{Top}}$	-0.634	-1.332	$+0.045$
$U_{\mathrm{Bottom}} - C_{\mathrm{Bottom}}$	$+1.061$	$+0.294$	$+1.827$
$P_{\mathrm{Top}} - U_{\mathrm{Top}}$	-0.128	-0.700	$+0.443$
$P_{\mathrm{Bottom}} - U_{\mathrm{Bottom}}$	-0.334	-0.930	$+0.264$
Tiny-ImageNet, uniform controls, family size 10			
$U_{\mathrm{Top}} - C_{\mathrm{Top}}$	$+0.093$	-0.344	$+0.542$
$U_{\mathrm{Bottom}} - C_{\mathrm{Bottom}}$	$+1.119$	$+0.569$	$+1.672$
$P_{\mathrm{Top}} - U_{\mathrm{Top}}$	-0.144	-0.494	$+0.207$
$P_{\mathrm{Bottom}} - U_{\mathrm{Bottom}}$	-0.317	-0.687	$+0.069$
CIFAR-100, additional seeds, family size 10			
I	$+1.461$	$+0.564$	$+2.365$
$C_{\mathrm{Bottom}} - P_{\mathrm{Bottom}}$	-0.983	-1.679	-0.303

D Probability mass and secondary accuracy

The statistics in Table 3 use the main-comparison targets and retained sets. FP32 targets are accumulated in float64 without renormalization; residual mass is the stored row sum minus the true-class and retained masses. These quantities are descriptive summaries. The supplement also computes model-level macro-F1, top-1 accuracy, and seed-level differences. Accuracy is secondary and has no separate intervals or tests.

The reported comparisons do not include a matched hard-label-only student baseline or teacher development accuracy. The comparisons therefore assess differences between teacher-supervised conditions, not the gain over hard-label-only training.

E Data reuse and exclusions

CIFAR uses 40,000 training and 10,000 development images; no shared exact-content hashes were found in the current split audit. Tiny teachers use 72,000 training images and students use 80,000.

After the permutation and preservation-budget analyses, file-byte and decoded-RGB checks found ten training/development pairs with identical content but conflicting labels; their training images were also used by the teacher. All affected development rows were removed from every condition, seed, and mapping, leaving 19,990 images, without retraining. Duplicate screening used file-byte and decoded-RGB equality rather than near-duplicate detection.

Separate CIFAR100-C feature/probe analyses used official-test-derived identities, including label-dependent evaluation over 10,000 identities (Hendrycks & Dietterich, 2019). The inspected probe-fitting paths use official training labels. These historical feature/probe analyses are separate from the documented 40,000-training/10,000-development CIFAR split used for the comparisons in this paper.

Grouped and uniform comparisons followed the permutation analyses in Section 5.1 and Appendix F. Their conditions and analyses were fixed before the corresponding new outcomes. Uniform and additional-seed checkpoints were fixed before development evaluation, with no new teacher inference or official test-set evaluation. The overlap-excluded analysis reuses the same trained models and is not an independent replication.

F Supporting permutation comparisons

These permutation comparisons preceded the grouped and uniform comparisons and used different training settings and resampling units. Table 5 retains their original 95% intervals: evaluation-image resampling for the five-seed Tiny and CIFAR comparisons, whole-mapping resampling with three fixed seeds for Tiny's 20-mapping comparison, and mapping/seed resampling with fixed images for Tiny Top/Bottom. They are not pooled with the primary comparisons.

Table 5: Permutation comparisons (pp), with their original 95% intervals. Resampling units differ as specified above.

Comparison	Estimate	Lower	Upper
Tiny, five-seed Unmodified minus P	3.013	2.726	3.307
Tiny, Unmodified minus 20 maps	2.854	2.789	2.915
Tiny, Top minus Bottom	1.973	1.600	2.426
CIFAR, Top minus Bottom	2.599	2.220	2.981

G The Use of Large Language Models (LLMs)

Large language models assisted with manuscript drafting and revision.