# DOES REPRESENTATION SIMILARITY RANK KD GAINS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Limited training budgets require prioritizing students for knowledge distillation (KD). We evaluate early representation similarity through global gain ranking, selected-subset benefit, and gain-order repeatability using seed-paired improvements over task-only training. Five metrics at initialization and after task-only warm-up cover four 12-student Ped2 pools and a separate eight-student CIFAR-100 pool, with teacher and recipe fixed within each pool. Ped2 Spearman estimates range from $-0.336$ to $+0.434$ with wide intervals; CIFAR-100 estimates range from $-0.333$ to $+0.238$ despite positive gains in all 24 paired runs. A complete post-hoc evaluation on original Ped2 B0 assesses student selection across metrics, measurement phases, and subset sizes using three independent holdout seeds. Mean gains relative to uniform-random expectation range from $-0.037$ to $+0.306$ AUC percentage points, with 78/110 combinations above zero. In that pool, disjoint pilot and holdout blocks of three seeds each agree on 28/66 student-pair orders. Sample-size and sensitivity analyses characterize the resolution of the ranking measurements. Supporting Ped2 and ImageNet studies find identical selections for noise-aware and matched mean rules. The results support assessing early KD signals through selection benefit and gain-order repeatability alongside global ranking.

## 1 INTRODUCTION

Knowledge distillation (KD) transfers outputs or features from a teacher to a student (Hinton et al., 2015; Romero et al., 2015). For a practitioner choosing among students, a useful prediction is how much each student gains from enabling KD under a fixed training recipe. Representation similarity offers a natural candidate signal: it describes how closely teacher and student features align before or early in training. With the teacher and training recipe fixed, we ask whether that signal orders the seed-paired KD increment across a prespecified student candidate set.[1]

This question arises when several lightweight students satisfy a deployment constraint and only a subset can receive additional distillation evaluations. The desired gain is better abnormal-frame discrimination in anomaly detection or recognition accuracy in classification. Students have different task-only baselines and can benefit differently from the same teacher. An early similarity score could prioritize a subset for KD; its value depends on the gains obtained by that subset, the cost of measuring the score, and the repeatability of the estimated gains. Final deployment also depends on attained performance and inference cost.

Prior work evaluates representation metrics against functional differences (Ding et al., 2021), links KD outcomes to class-level agreement (Zhu et al., 2022), and uses compatibility scores to select teachers or students (Panigrahi et al., 2026; Anonymous, n.d.). Here, student-specific task-only baselines define the paired KD increment. We assess whether early representation alignment predicts this increment across students, how the selected subsets benefit, and how the gain ordering varies across training seeds.

Figure 1 illustrates the problem on original Ped2 B0 and CIFAR-100. Positive mean KD gains vary across students (a), while initialization linear CKA has weak observed rank associations with

[1] Anonymized code: `https://anonymous.4open.science/r/kd-distillability-audit-12C2/`.



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.

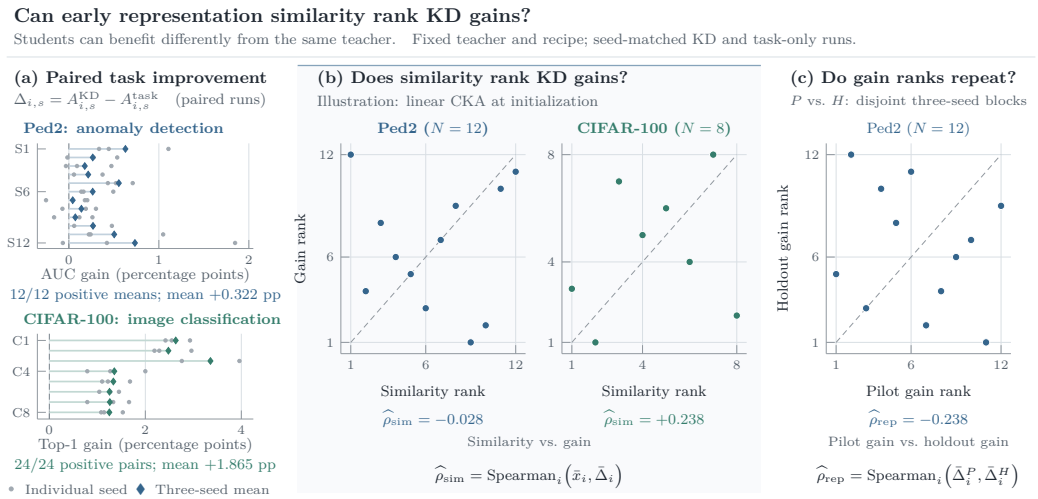


Figure 1: **KD gains, similarity ranking, and gain-order repeatability.** (a) Paired KD-minus-task-only gains for all 12 original B0 Ped2 and eight CIFAR-100 students. Dots show individual seeds; diamonds show three-seed means, in percentage points. (b) Mean initialization linear CKA ($\bar{x}_i$) versus mean paired gain ($\bar{\Delta}_i$), each averaged over three seeds. (c) Ped2 pilot ($P$) versus holdout ($H$) mean gains from disjoint three-seed blocks. Here $A$ denotes AUC or top-1 accuracy, $i$ indexes student configurations, $s$ indexes seeds, and $N$ counts configurations. In (b,c), each point represents a configuration; $\widehat{\rho}_{\mathrm{sim}}$ and $\widehat{\rho}_{\mathrm{rep}}$ are the respective Spearman estimates. Ped2 probes and outcomes use disjoint seeds; CIFAR uses shared seeds. Higher ranks indicate larger values; dashed diagonals indicate equal ranks.

these gains (b). In original Ped2 B0, disjoint three-seed pilot and holdout blocks give different gain orderings (c). These examples motivate examining both the predictive signal and the repeatability of its target ordering.

**Contributions.** Our primary contribution is a joint empirical evaluation of early similarity through global gain ranking, selected-subset gains, and gain-order repeatability. Five metrics at two phases cover four 12-student Ped2 pools under three fixed teachers and a separate eight-student CIFAR-100 pool. Global association estimates have wide uncertainty. On three independent holdout seeds for original Ped2 B0, complete post-hoc selection yields above-random gains in 78/110 combinations of metric, phase, and nontrivial subset size; differences from uniform-random expectation across all combinations range from $-0.037$ to $+0.306$ AUC percentage points. Comparisons include parameter-count baselines on Ped2 and paired-pilot ranking on original B0.

Student-pair differences and complete seed-subset analyses characterize gain-order variability. On original Ped2 B0, disjoint pilot and holdout blocks of three seeds each agree on 28/66 student-pair orders. Sample-size calibration and probe sensitivity characterize measurement resolution. Supporting Ped2 and ImageNet-1k decision evaluations find identical selections under fixed noise-aware and matched mean rules.

## 2 Related Work

**Distillation outcomes and training conditions.** KD outcomes depend on student capacity, supervision, and training. Stanton et al. (2021) distinguish teacher–student agreement from student generalization; Mirzadeh et al. (2020) address capacity gaps through teacher assistants; and Beyer et al. (2022) establish the importance of training duration and consistent teacher–student views. Recent methods modify the training objective through logit standardization (Sun et al., 2024) and distribution matching (Lv et al., 2024; Wang et al., 2025), or reduce teacher computation through token masking (Son et al., 2024). This dependence on both the student and training conditions motivates measuring KD gains within a fixed teacher–recipe setting.

**Representation similarity and task behavior.** Centered kernel alignment (CKA) (Kornblith et al., 2019) and singular vector canonical correlation analysis (SVCCA) (Raghu et al., 2017) quantify relations between learned representations. Ding et al. (2021) assess representation distances by their sensitivity to functional changes and specificity to behavior-preserving changes, using rank correlations with probing accuracy and robustness. Davari et al. (2023) show that CKA can change substantially under transformations that largely preserve functional behavior. In KD, Zhu et al. (2022) use converged CKA to examine class-specific representation mismatch, relate changes in prediction agreement to class-level accuracy changes, and develop selective class distillation. These studies evaluate representation measurements against functional differences between models and class-specific outcomes after distillation.

**Predicting transfer and selecting candidates.** LEEP (Nguyen et al., 2020) and LogME (You et al., 2021) estimate transfer performance from pretrained predictions or features and target labels, avoiding full fine-tuning of every candidate. Compatibility scores also guide choices within KD. For a fixed object-detection student, Cao et al. (2023) fit linear mappings between frozen representations and use validation adaptation costs to construct teacher sequences. In language-model distillation, GRACE scores teachers from student-gradient statistics on teacher-generated data and evaluates both rank correlation and selection regret (Panigrahi et al., 2026). For a fixed teacher, the C-MoNA public manuscript proposes cross-model neural tangent kernel alignment to select student architectures by predicting their post-distillation performance (Anonymous, n.d.). These approaches connect inexpensive candidate scores to outcomes after training. Our evaluation targets each student's seed-paired KD increment over its own task-only baseline, with teacher and recipe fixed within each pool. We assess early representation similarity through global gain ordering, the gains obtained by selected subsets, and the repeatability of the estimated ordering across training seeds.

## 3 PAIRED-GAIN RANKING DESIGN

### 3.1 PROBLEM FORMULATION

Let $T$ be a fixed teacher and $S$ a student. Under matched initial states and training budgets, task-only training yields performance $A_{\text{task}}(S)$ and KD yields $A_{\text{KD}}(T, S)$. Suppressing the seed index, the *KD gain* is

$$\Delta A(T, S) \;=\; A_{\text{KD}}(T, S) - A_{\text{task}}(S). \tag{1}$$

Here $A$ is frame-level AUC for video anomaly detection (VAD) or top-1 accuracy for classification, expressed as a fraction in $[0, 1]$; $100\Delta A$ is the gain in percentage points. The conditioning unit is a teacher–student–recipe configuration. Paired training matches seed, initialization, and data order; study-specific execution deviations are reported with the experimental results. Repeated seeds expose uncertainty in the paired increment.

For student configuration $i$ and seed $s$, write the paired increment as $\Delta_{i,s} = A^{\text{KD}}_{i,s} - A^{\text{task}}_{i,s}$. Its mean over a nonempty seed block $\mathcal{B}$ is

$$\bar{\Delta}^{\mathcal{B}}_i = \frac{1}{|\mathcal{B}|} \sum_{s \in \mathcal{B}} \Delta_{i,s}, \qquad \mu_i = \mathbb{E}_s[\Delta_{i,s}]. \tag{2}$$

The finite-seed mean $\bar{\Delta}^{\mathcal{B}}_i$ estimates the expected gain $\mu_i$ under the fixed teacher, student, and recipe. Figure 1(a) shows the individual increments and their three-seed means.

Students have different task-only baselines. The identity $A^{\text{KD}}_{i,s} - A^{\text{KD}}_{j,s} = (A^{\text{task}}_{i,s} - A^{\text{task}}_{j,s}) + (\Delta_{i,s} - \Delta_{j,s})$ separates their baseline gap from their difference in KD gain; these terms can have opposite signs. Gain ranking therefore concerns the benefit of adding KD, rather than final accuracy or the accuracy–cost trade-off.

For $N$ student configurations, metric $m$, and phase $t$, let $x_{ismt}$ denote the similarity of configuration $i$ at seed $s$. Let $P$ and $H$ be the nonempty probe and outcome seed sets. The similarity–gain association is

$$\widehat{\rho}_{\text{sim},mt} = \text{Spearman}^{N}_{i=1}\left( \frac{1}{|P|} \sum_{s \in P} x_{ismt}, \bar{\Delta}^{H}_i \right). \tag{3}$$

Disjoint seed sets separate probe randomness from outcome estimation; shared sets describe the association within the same training repetitions. Spearman uses average ranks for ties and requires nonconstant inputs. Figure 1(b) abbreviates this estimate as $\widehat{\rho}_{\mathrm{sim}}$ for linear CKA at initialization. All five metrics and both phases are evaluated without choosing their direction from the outcomes: positive correlation means that greater similarity accompanies greater KD gain. Statistical resampling uses student configurations as units. Seed repetitions estimate each configuration's mean and do not increase $N$.

To assess gain-order repeatability, the original Ped2 B0 pool uses the pilot seed block $P$ and a disjoint holdout block $H$. We compare its gain estimates across these blocks:

$$\widehat{\rho}_{\mathrm{rep}} = \mathrm{Spearman}_{i=1}^{N}\left(\bar{\Delta}_i^P, \bar{\Delta}_i^H\right), \qquad P \cap H = \varnothing. \tag{4}$$

Figure 1(c) uses three seeds in each block. This measures agreement between two finite-seed estimates of the same expected-gain ordering. It complements the similarity–gain association by assessing the repeatability of its measured target.

The additive model $\Delta_i = \mu_i + \varepsilon_i$ relates single-seed gain variation to reliability. Assume finite second moments, positive variances for the scalar metric signal $x$, expected gain $\mu$, and observed gain $\Delta$, and zero-mean noise $\varepsilon$ with variance $\sigma^2_{\mathrm{seed}}$, independent of both $x$ and $\mu$. Classical Pearson attenuation (Spearman, 1904) gives

$$\rho^2(x, \Delta) = \mathrm{ICC}\,\rho^2(x, \mu), \qquad \mathrm{ICC} = \frac{\mathrm{Var}(\mu)}{\mathrm{Var}(\mu) + \sigma^2_{\mathrm{seed}}}. \tag{5}$$

Here $\rho$ denotes Pearson correlation and ICC is the intraclass correlation under the additive model, with $\mathrm{Var}(\mu)$ measuring variation across configurations. Averaging $k$ independent, equal-variance seeds replaces $\sigma^2_{\mathrm{seed}}$ by $\sigma^2_{\mathrm{seed}}/k$. Adding a common positive gain leaves this reliability unchanged: average benefit and separation between students are different quantities. The ranking study measures Spearman directly; shared probe and outcome seeds can violate the metric–noise independence assumed above.

### 3.2 Representation measurements

The ranking question concerns five representation measurements: linear and radial basis function (RBF) CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), orthogonal Procrustes (Schönemann, 1966), and linear-regression $R^2$, evaluated at initialization and after task-only warm-up. These phases are analyzed separately. The ranking unit is a student configuration within a fixed teacher and recipe, rather than an individual seed replicate.

Initialization and warm-up probes measure alignment before and after task-only adaptation. Teacher features, probe examples, and measurement settings are fixed within each pool.

Linear and RBF CKA compare linear and kernel similarities across examples. SVCCA compares dominant feature subspaces; Procrustes measures orthogonal alignment; $R^2$ measures linear predictability between representations.

Let $d_T, d_S$ be the original feature widths and $N_{\mathrm{probe}}$ the number of probe examples. Centering and PCA reduction give $F_T, F_S \in \mathbb{R}^{N_{\mathrm{probe}} \times q}$, where $q = \min(d_T, d_S, N_{\mathrm{probe}} - 1)$. The normalized Procrustes score is

$$\Gamma = \frac{2\|F_T^\top F_S\|_*}{\|F_T\|_F^2 + \|F_S\|_F^2}. \tag{6}$$

Here $\|\cdot\|_*$ and $\|\cdot\|_F$ are the nuclear and Frobenius norms. For nonzero denominator, $0 \le \Gamma \le 1$, with equality at one exactly when $F_T R = F_S$ for an orthogonal $q \times q$ matrix $R$. All five scores are oriented so that larger values indicate greater similarity.

For the Ped2 probes, each stage-3 feature tensor is flattened per example. Linear CKA uses centered feature Gram matrices. RBF CKA centers its kernel matrices and uses the median off-diagonal squared distance as the squared bandwidth (one if the median is zero). SVCCA averages canonical correlations between the leading $\min(20, N_{\mathrm{probe}} - 1, d_T, d_S)$ left-singular vectors. Linear-regression $R^2$ averages both prediction directions after PCA to $\min(20, \lfloor N_{\mathrm{probe}}/4 \rfloor, d_T, d_S)$ components, with ridge coefficient $10^{-3}$, evaluated on the fitted probe sample. CKA, Procrustes, and $R^2$ denominators include $10^{-8}$ stabilization (the half-sum for Procrustes); scores are clipped to $[0, 1]$. Examples and settings are fixed across configurations.

### 3.3 SELECTION CONSEQUENCES AND EMPIRICAL ORDER STABILITY

A post-hoc analysis uses the complete existing outcomes to evaluate selection consequences. For each metric and phase, descending similarity defines the first $K$ students for every $K = 1, \ldots, N$. At cutoff ties, inclusion weights $w_i(K) \in [0, 1]$ average uniformly over tied subsets and sum to $K$. Selected mean gain and its advantage over uniform-random selection are

$$U_K = \frac{1}{K} \sum_{i=1}^{N} w_i(K) \bar{\Delta}_i^H, \qquad B_K = U_K - \frac{1}{N} \sum_{i=1}^{N} \bar{\Delta}_i^H. \tag{7}$$

The pool mean is the exact expected gain of a uniformly sampled $K$-subset. All curves include $K = N$, where $B_K = 0$. Baselines rank larger parameter counts in Ped2 and original B0 pilot mean gains. Every metric, phase, selection size, and outcome seed is retained, together with all seven nonempty probe-seed subsets.

Original B0 evaluates probe-based selections on independent holdout seeds. Other pools additionally use all three cross-seed folds: two probe seeds determine the order and the omitted outcome seed evaluates it. We average the three fold utilities. Overlapping folds and selection sizes yield dependent descriptive comparisons. Equal $K$ fixes candidate count; architecture-dependent compute and probe costs remain separate.

For every student pair $(i, j)$, $d_{ij,s} = \Delta_{i,s} - \Delta_{j,s}$ records its seed-wise gain difference. We report mean differences, their sample spread, and whether all three seeds give the same strict sign. Original B0 also compares pilot and holdout mean-order signs across all $\binom{N}{2}$ pairs. These finite-seed agreement summaries describe observed order stability.

### 3.4 BOUNDED DECISION EVALUATION

The Distillability Audit Protocol (DAP) uses a finite-budget paired pilot to make a launch/no-launch decision. Scope checks and structural deferral define the eligible candidate pool before the pilot. Within that pool, the pilot criterion is

$$\bar{\Delta}_i^P > \delta, \qquad k = |P| \geq 2, \qquad \frac{\widehat{\sigma}_i^P}{\bar{\Delta}_i^P - \delta} \leq \nu. \tag{8}$$

Here $\delta \geq 0$ is the gain threshold, $\nu > 0$ is fixed, and $\widehat{\sigma}_i^P$ is the sample standard deviation of $\{\Delta_{i,s} : s \in P\}$, with denominator $k - 1$. A pilot consumes both KD and task-only training.

Corrected Ped2 uses three pilot seeds, $\delta = 0$, and $\nu = 1$; the structural check (S2) defers teacher/student parameter ratios in $[0.5, 2]$. DAP and mean Top-$K$ share the S2-eligible pool and launch count $K = 6$. ImageNet uses $\delta = 0.005$ (0.5 top-1 points), $\nu = 1$, and three pilot seeds; its mean-only comparator requires $\bar{\Delta}_i^P > \delta$. Holdout outcomes evaluate these fixed selections and observed missed gains; unavailable outcomes remain missing.

## 4 EXPERIMENTS

RQ1 asks whether early similarity ranks paired KD gains and identifies subsets with gains above random expectation (§4.1–4.3). RQ2 examines gain-order repeatability across training seeds (§4.4). RQ3 compares noise-aware and mean-based selections and the holdout gains of selected and excluded students (§4.5). Each pool is analyzed separately, with student configurations as ranking units and gains in AUC or top-1 percentage points.

### 4.1 SIMILARITY RANKING ACROSS TEACHERS AND STUDENTS

On UCSD Ped2 (Mahadevan et al., 2010), three fixed PVTv2 teachers (Wang et al., 2022) (B0, B2, B4) share 12 PVTv2 student architectures; a separate B0 pool adds 12 students. Three paired outcome seeds estimate each configuration mean. Original B0 uses retrospective pilot-seed probes and existing independent holdout gains. The other pools use 108 new KD/task-only pairs, with shared probe and outcome seeds. The experiment matrix was fixed before these new runs.

Table 1: Paired gains and similarity–gain ranking (RQ1). Each pool has $N$ configurations with three outcome seeds each. Spearman ranges span five metrics and two phases. Positive means count configurations; gains use AUC points on Ped2 and top-1 points on CIFAR-100.

| Pool | $N$ | Mean gain (pp) | Positive means | Spearman range |
|---|---|---|---|---|
| Ped2 B0: original | 12 | $+0.3223$ | $12/12$ | $[-0.084, +0.434]$ |
| Ped2 B0: additional | 12 | $+0.1503$ | $8/12$ | $[-0.308, +0.112]$ |
| Ped2 B2 | 12 | $+0.0231$ | $10/12$ | $[-0.322, -0.056]$ |
| Ped2 B4 | 12 | $-0.0581$ | $9/12$ | $[-0.336, +0.336]$ |
| CIFAR-100 | 8 | $+1.865$ | $8/8$ | $[-0.333, +0.238]$ |

All Ped2 branches use ten epochs, $256 \times 256$ inputs, batch size 32, and AdamW with learning rate $10^{-4}$ and weight decay $5 \times 10^{-5}$. The task loss is target-frame mean squared error; KD adds $0.5$ times the squared error to frozen-teacher frame predictions. A cosine learning-rate schedule spans all ten epochs. Paired branches match initial weights and epoch-wise data order. Evaluation uses 1,962 target-frame mask labels, including 1,620 positive frames. Offline reconstruction from archived per-frame PSNR reproduces every paired AUC, with identical label fingerprints across branches. Each probe uses flattened stage-3 features from the same 320 evenly spaced training examples, at initialization and after two task-only epochs. The replay preserves the ten-epoch schedule and matches original B0 pilot initialization and data-order hashes. Warm-up supplies only the probe: both outcome branches restart from matched initial weights, with KD active from epoch one.

We correlate three-seed metric and gain means across configurations, reporting all ten metric–phase combinations per pool. Two-sided tests use 99,999 random permutations plus the observed ordering, with Holm correction over ten tests. Exploratory percentile intervals use 10,000 configuration-bootstrap resamples to measure candidate-composition sensitivity. Seed subsets and disjoint blocks assess seed variation; repeated students across teachers are not pooled as independent observations.

Post-hoc sample-size calibration under an i.i.d. Gaussian copula reaches 80% power at population Spearman correlations of $0.75$–$0.76$ for $N = 12$ and $0.88$–$0.89$ for $N = 8$, using an unadjusted two-sided $\alpha = 0.05$ permutation test. These model-based thresholds describe the resolution at each configuration count.

Figure 2 reports all 40 Ped2 estimates, ranging from $-0.336$ to $+0.434$. All within-pool Holm-adjusted $p$ values equal one. Original B0 initialization SVCCA has the largest positive estimate ($+0.434$, unadjusted $p = 0.1632$, interval $[-0.232, +0.807]$). The interval spans negative through strong positive associations, leaving the strength of global ranking unresolved at this configuration count.

Tables 1 and 2 report pool summaries and individual Ped2 gains, comparing original students across teachers and listing additional B0 students separately. In B4, three negative configuration means outweigh nine positive means, producing a negative pool average.

As a capacity baseline, parameter-count correlations with mean gain are $+0.172$, $+0.007$, $-0.032$, and $-0.091$ for original B0, additional B0, B2, and B4, respectively. The frozen secondary analysis enumerates all seven nonempty subsets of three seeds. Holding the full outcome mean fixed, single-probe-seed correlations across metrics and phases range from $-0.203$ to $+0.406$ (original B0), $-0.566$ to $+0.559$ (additional B0), $-0.587$ to $+0.434$ (B2), and $-0.524$ to $+0.413$ (B4).

With full-probe scores fixed, varying outcome subsets produces correlation sign changes within every Ped2 pool. Matched-subset analysis varies both on the same subset in pools with shared seeds. Overlapping subsets describe sensitivity rather than independent replication.

To assess sensitivity to the probe examples, a post-hoc replay partitions all 320 cached stage-3 examples into two complementary interleaved sets of 160. Recomputing all five metrics at both phases for all 144 jobs gives median absolute changes in similarity–gain correlation of $0.017$–$0.031$ across the four pools, with a maximum of $0.364$. Original-B0 initialization SVCCA changes from $0.434$ to $0.385$ and $0.070$ on the two subsets.

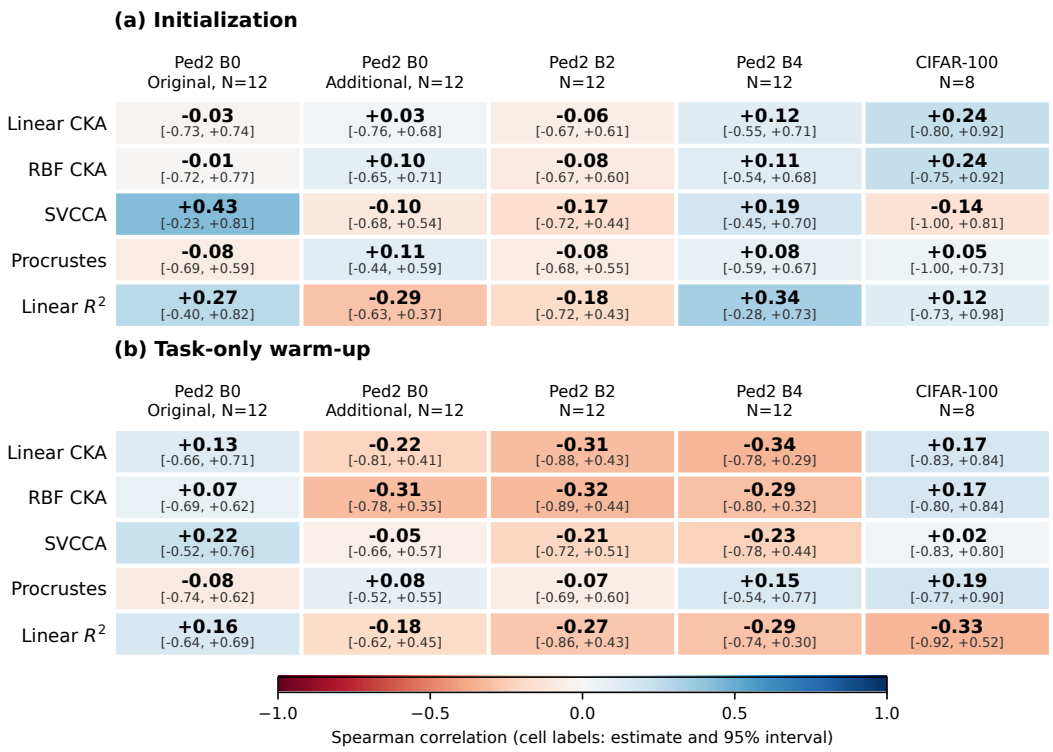


Figure 2: Similarity–gain ranking across all primary pools, metrics, and phases. Colors and bold numbers show Spearman estimates; brackets give exploratory 95% configuration-bootstrap intervals. Columns are separate pools and panels are measurement phases. Original Ped2 B0 uses independent holdout outcomes; other pools share probe and outcome seeds. All intervals include zero.

Table 2: Ped2 paired gains: mean ± sample standard deviation over three outcome seeds (AUC points). The left block compares original students across teachers; the right contains additional students. Rows follow fixed configuration order.

| | Original student pool | | | Additional student pool | |
|---|---|---|---|---|---|
| Student | B0 | B2 | B4 | Student | B0 |
| Tiny | $+0.629 \pm 0.418$ | $+0.198 \pm 0.150$ | $+0.175 \pm 0.110$ | 12/24/64 | $+0.174 \pm 0.022$ |
| Front-narrow | $+0.266 \pm 0.277$ | $+0.573 \pm 0.344$ | $+0.498 \pm 0.389$ | 24/48/128 | $-0.084 \pm 0.781$ |
| Narrow | $+0.177 \pm 0.263$ | $+0.488 \pm 0.069$ | $+0.456 \pm 0.163$ | 24/64/160 | $+0.469 \pm 0.343$ |
| Compact | $+0.215 \pm 0.159$ | $-1.192 \pm 3.226$ | $-0.820 \pm 2.290$ | 40/80/200 | $-0.183 \pm 0.581$ |
| Deep | $+0.555 \pm 0.141$ | $+0.146 \pm 0.140$ | $+0.054 \pm 0.208$ | 56/112/280 | $+0.136 \pm 0.202$ |
| Shallow | $+0.265 \pm 0.199$ | $+0.435 \pm 0.143$ | $+0.664 \pm 0.805$ | Back-heavy | $+0.172 \pm 0.267$ |
| Few heads | $+0.043 \pm 0.257$ | $+0.031 \pm 0.853$ | $+0.645 \pm 0.566$ | Front-heavy | $-0.486 \pm 1.003$ |
| SR 2 | $+0.139 \pm 0.191$ | $+0.243 \pm 0.161$ | $-0.026 \pm 0.095$ | Depth 1/2/2 | $+0.551 \pm 0.333$ |
| Standard | $+0.072 \pm 0.217$ | $+0.123 \pm 0.204$ | $+0.055 \pm 0.149$ | Depth 2/3/3 | $+0.996 \pm 1.343$ |
| L2 width | $+0.268 \pm 0.210$ | $+0.323 \pm 0.222$ | $+0.165 \pm 0.246$ | Heads 1/4/8 | $-0.308 \pm 0.462$ |
| Teacher-match | $+0.504 \pm 0.471$ | $+1.027 \pm 0.308$ | $+0.366 \pm 0.254$ | SR 1 | $+0.105 \pm 0.165$ |
| L2 mid-depth | $+0.734 \pm 0.994$ | $-2.119 \pm 3.953$ | $-2.931 \pm 5.114$ | SR 8 | $+0.263 \pm 0.265$ |

Student labels abbreviate configuration names. Numeric triplets denote stage widths unless marked as depth or heads; SR denotes spatial reduction.

### 4.2 Similarity ranking on CIFAR-100

The separate CIFAR-100 (Krizhevsky, 2009) study uses a DeiT-Base teacher (Touvron et al., 2021) initialized from ImageNet-1k (Russakovsky et al., 2015) and fine-tuned for 20 epochs (test top-1 89.46%), eight ViT students, and paired seeds 17, 29, and 71. Two task-only warm-up epochs

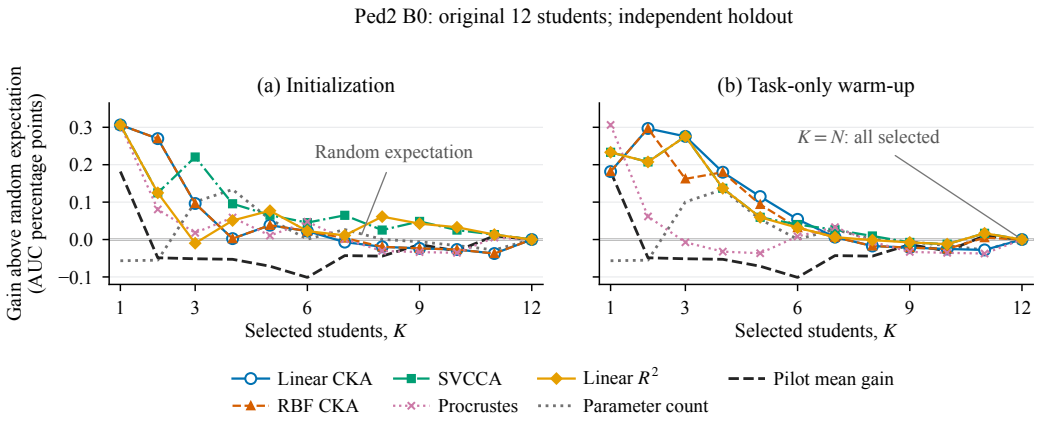


Figure 3: Post-hoc selection on original Ped2 B0 using three probe and three independent holdout seeds. Zero denotes uniform-random expectation; $K = N$ selects all students. All five metrics, both phases, and every $K$ are shown. Parameter count and pilot mean provide baselines; equal $K$ does not equalize costs.

Table 3: Post-hoc selection above random expectation (RQ1), over all metrics, phases, and $K < N$. Full uses three probe/outcome seeds (independent for original B0, shared otherwise); cross-seed averages three two-to-one folds. Counts are dependent; units are AUC or top-1 points.

| | Full-score evaluation | | Cross-seed evaluation | |
|---|---|---|---|---|
| Pool | Above random | Min–max (pp) | Above random | Min–max (pp) |
| Ped2 B0: original | 78/110 | $[-0.037, +0.306]$ | – | – |
| Ped2 B0: additional | 55/110 | $[-0.333, +0.845]$ | 73/110 | $[-0.339, +0.845]$ |
| Ped2 B2 | 47/110 | $[-2.142, +1.004]$ | 58/110 | $[-2.142, +1.004]$ |
| Ped2 B4 | 38/110 | $[-2.873, +0.556]$ | 50/110 | $[-2.873, +0.556]$ |
| CIFAR-100 | 40/70 | $[-0.615, +0.565]$ | 41/70 | $[-0.615, +0.992]$ |

precede 46 epochs per branch from the shared warmed checkpoint; KD uses temperature 4 and mixing weight 0.9. Similarity uses pre-classifier embeddings from the first 1,024 test images with deterministic transforms, fixed order, and no labels. Probe and outcome seeds are shared. Tests enumerate all $8!$ permutations, with the same two-sided criterion, ten-test Holm correction, and 10,000-resample intervals as Ped2.

All $24/24$ paired gains are positive, with mean $+1.865$ top-1 points. Spearman estimates range from $-0.333$ to $+0.238$, and every Holm-adjusted $p$ equals 1. Positive gains thus coexist with uncertain similarity–gain ranking (Figure 2).

### 4.3 Selection consequences across candidate pools

The complete post-hoc replay (§3.3) evaluates every selection size, metric, and phase. On independent original B0 holdout, 78/110 nontrivial metric–phase–size combinations exceed random expectation; initialization SVCCA does so for all $K = 1, \ldots, 11$, by 0.013–0.306 AUC points (Figure 3). Parameter-count and pilot-mean selection exceed random expectation for $6/11$ and $2/11$ sizes, with advantage ranges $[-0.057, +0.134]$ and $[-0.101, +0.182]$ AUC points, respectively. These comparisons share students and selection sizes.

Global correlation depends on the full ordering; selection benefit depends on membership in the chosen subset. Their empirical assessments can therefore differ. Table 3 shows positive and negative advantages in every pool, including cross-seed evaluation in the other four pools.

Original B0 selection uses $12 \times 3 \times 2 = 72$ student epochs for warm-up or 720 for paired pilots, plus $60K$ epochs for three-seed outcome evaluation. Initialization requires feature inference and metric computation. Epoch counts do not equate GPU time across architectures.

Table 4: Student-pair gain differences (RQ2). Medians cover absolute mean gaps and between-seed sample standard deviations over all unordered pairs, in AUC or top-1 points. Consistent pairs retain one strict order across three outcome seeds; pairs share students.

| Pool | Pairs | Median gap (pp) | Median variation (pp) | Consistent pairs |
|---|---|---|---|---|
| Ped2 B0: original | 66 | 0.228 | 0.406 | 17/66 |
| Ped2 B0: additional | 66 | 0.391 | 0.667 | 20/66 |
| Ped2 B2 | 66 | 0.439 | 0.512 | 24/66 |
| Ped2 B4 | 66 | 0.457 | 0.560 | 24/66 |
| CIFAR-100 | 28 | 1.137 | 0.563 | 16/28 |

Table 5: Fixed pilot decisions (RQ3). Gains use three evaluation seeds, in AUC points for Ped2 and top-1 points for ImageNet. The first row averages the selected set; other entries are individual student means.

| Study | Decision or subset | Mean gain (pp) |
|---|---|---|
| Ped2 | Same six students selected | $+0.2815$ |
| | Noise exclusion / one structural deferral | $+0.1393,\ +0.5554$ |
| ImageNet | Same three ViTs selected | $+2.497,\ +2.042,\ +3.849$ |
| | Two scope-deferred CNNs | $-2.20,\ -3.07$ |

### 4.4 Gain reliability and repeated ordering

In each Ped2 pool, the median sample standard deviation of student-pair differences exceeds their median absolute mean gap (Table 4); original B0 has a gap of $0.228$ versus variation of $0.406$ AUC points. CIFAR-100 has the reverse pattern ($1.137$ versus $0.563$ top-1 points), with $16/28$ pairs retaining one strict order across three seeds. Larger gaps relative to seed variation thus coexist with weak similarity–gain associations in CIFAR-100.

Original B0 pilot and holdout means have Spearman $-0.238$ and Pearson $-0.274$ correlation (Figure 1(c)), with 28/66 pair orders agreeing. Pilot seeds agree unanimously on 29/66 pairs; only three pairs retain one strict direction across all six seeds. These descriptive counts share students. All 12 holdout means are positive, yet their order changes between blocks, separating positive benefit from repeatable gain ordering.

### 4.5 Outcomes of bounded pilot decisions

Table 5 compares the fixed rules in §3.4. DAP and matched mean rules select identical sets and obtain identical accepted-set gains. All 12 Ped2 holdout means are positive ($30/36$ positive seed gains); additional noise exclusion and one structural deferral omit positive gains.

Selected ImageNet students are d320-L16, d384-L8, and d384-L16, in table order; all nine seed gains exceed $0.5$ points. Stopped d320-L8 has no holdout. Scope-deferred ResNet-18 (He et al., 2016) and MobileNetV3-Small (Howard et al., 2019) have six negative gains. ResNet-18 includes a data-loader amendment and imperfect pairing.

## 5 Conclusion

Within the tested pools, uncertain global similarity–gain estimates coexist with observed subset-selection value. Complete post-hoc evaluation on independent original Ped2 B0 holdout seeds includes above-random gains, while repeated seed blocks give different student orderings. Noise-aware and matched mean rules select identical sets in the supporting studies. These findings support evaluating early KD signals through global ranking, subset benefit, and gain-order repeatability under the specified teachers, measurement protocols, and training budgets.

### AI USE STATEMENT

Generative AI tools were used to assist in searching for candidate references. The Related Work section was written by the authors, and every cited reference was checked by the authors before inclusion. We take responsibility for the final content of this work.

### REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the paired estimand, statistical assumptions, configuration and seed counts, and evaluation procedures. The experiment descriptions distinguish pilot and holdout observations and report outcome availability and execution deviations.

## REFERENCES

Anonymous. Could student selection be the missing piece for efficient distillation? OpenReview public manuscript, n.d. URL https://openreview.net/forum?id=x4kreAhV1c. Accessed September 25, 2026.

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.

Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, and Liang-Yan Gui. Learning lightweight object detectors via multi-teacher progressive distillation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3577–3598, 2023.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019.

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.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals Kullback-Leibler divergence for knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305, 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good GRACEs: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, 2026.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Seungwoo Son, Jegwang Ryu, Namhoon Lee, and Jaeho Lee. The role of masking for efficient supervised knowledge distillation of vision transformers. In *Computer Vision – ECCV 2024*, volume 15125 of *Lecture Notes in Computer Science*, pp. 379–396, 2024.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

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, pp. 6906–6919, 2021.

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.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via $\alpha$-$\beta$-divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65167–65212, 2025.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32011–32024, 2022.

## A DATA PROVENANCE AND STUDENT CONFIGURATIONS

The Ped2 analyses use target-frame mask labels and seed-paired records. Earlier runs with shifted range labels are excluded. CIFAR-100 and ImageNet are analyzed separately. The 144 Ped2 jobs comprise 36 original-B0 probe jobs and 108 new paired jobs. They yield four 12-configuration ranking pools, each with three outcome seeds per configuration.

Tables 6 and 7 identify the students in Table 2. Unspecified depths are $[2, 2, 2]$, heads $[2, 4, 8]$, and spatial-reduction ratio 4. Original-pool probe seeds are 926101, 926203, and 926307; additional-pool seeds are 928101, 928203, and 928307. Original B0 uses an independent outcome block; B2, B4, and additional B0 pair probes and outcomes within seeds.

The accompanying numerical supplement retains all seed gains, metric measurements, seed-subset results, selection outcomes, pair differences, and simulation points. Its file index and checksums identify the complete records. The two B0 pools remain separate in all primary analyses. Their descriptive 24-configuration combination has 20 positive means, mean $+0.2363$ AUC points, and ten Spearman estimates in $[-0.092, +0.147]$; its full statistics are included in the numerical supplement.

Table 6: Original students (B0, B2, B4). Architecture specifications; defaults are stated in the text.

| Index | Name | Specification |
|---|---|---|
| 1 | T0_tiny | dims: [8, 16, 40] |
| 2 | T11_width_front_narrow | dims: [16, 128, 320] |
| 3 | T1_width_narrow | dims: [16, 32, 80] |
| 4 | T2_width_compact | dims: [16, 48, 120] |
| 5 | T4_depth_deep | depths: [2, 2, 4], dims: [32, 64, 160] |
| 6 | T4_depth_shallow | depths: [1, 1, 1], dims: [32, 64, 160] |
| 7 | T4_heads_few | dims: [32, 64, 160], num heads: [1, 2, 4] |
| 8 | T4_sr_small | dims: [32, 64, 160], sr ratio: 2 |
| 9 | T4_width_standard | dims: [32, 64, 160] |
| 10 | T5_width_l2 | dims: [48, 96, 240] |
| 11 | T6_teacher_match | dims: [32, 64, 128] |
| 12 | T8_l2_depth_mid | depths: [1, 1, 2], dims: [48, 96, 240] |

Table 7: Additional students (B0). Architecture specifications; defaults are stated in the text.

| Index | Name | Specification |
|---|---|---|
| 1 | E01_width_12_24_64 | dims: [12, 24, 64] |
| 2 | E02_width_24_48_128 | dims: [24, 48, 128] |
| 3 | E03_width_24_64_160 | dims: [24, 64, 160] |
| 4 | E04_width_40_80_200 | dims: [40, 80, 200] |
| 5 | E05_width_56_112_280 | dims: [56, 112, 280] |
| 6 | E06_width_back_heavy | dims: [24, 80, 240] |
| 7 | E07_width_front_heavy | dims: [48, 80, 160] |
| 8 | E08_depth_1_2_2 | depths: [1, 2, 2], dims: [32, 64, 160] |
| 9 | E09_depth_2_3_3 | depths: [2, 3, 3], dims: [32, 64, 160] |
| 10 | E10_heads_1_4_8 | dims: [32, 64, 160], num heads: [1, 4, 8] |
| 11 | E11_sr_1 | dims: [32, 64, 160], sr ratio: 1 |
| 12 | E12_sr_8 | dims: [32, 64, 160], sr ratio: 8 |

## B COMPLETE RANKING STATISTICS

Each Ped2 pool contains $N = 12$ configuration means; CIFAR-100 contains $N = 8$. Each mean uses three outcome seeds. Original Ped2 B0 uses disjoint probe and outcome blocks; the other pools share seeds. Intervals are exploratory configuration-bootstrap 95% intervals. Two-sided permutation tests use configurations as units, with Holm adjustment over the ten metric–phase comparisons within each pool. CIFAR-100 enumerates all $8!$ outcome permutations and uses 10,000 bootstrap resamples with analysis seed 928001. All 50 primary metric–phase results are reported below.

Table 8: B0 original: complete metric–phase results.

| Phase | Metric | Spearman | Pearson | 95% interval | $p$ | Holm $p$ |
|---|---|---|---|---|---|---|
| Init | Linear CKA | $-0.028$ | $+0.247$ | $[-0.730, +0.742]$ | 0.9393 | 1.000 |
| Init | RBF CKA | $-0.014$ | $+0.207$ | $[-0.720, +0.766]$ | 0.9740 | 1.000 |
| Init | SVCCA | $+0.434$ | $+0.495$ | $[-0.232, +0.807]$ | 0.1632 | 1.000 |
| Init | Procrustes | $-0.077$ | $+0.220$ | $[-0.685, +0.594]$ | 0.8168 | 1.000 |
| Init | Linear $R^2$ | $+0.266$ | $+0.416$ | $[-0.397, +0.819]$ | 0.4058 | 1.000 |
| Warm-up | Linear CKA | $+0.126$ | $+0.293$ | $[-0.658, +0.714]$ | 0.6994 | 1.000 |
| Warm-up | RBF CKA | $+0.070$ | $+0.309$ | $[-0.693, +0.620]$ | 0.8347 | 1.000 |
| Warm-up | SVCCA | $+0.224$ | $+0.287$ | $[-0.523, +0.764]$ | 0.4868 | 1.000 |
| Warm-up | Procrustes | $-0.084$ | $+0.015$ | $[-0.745, +0.624]$ | 0.7990 | 1.000 |
| Warm-up | Linear $R^2$ | $+0.161$ | $+0.280$ | $[-0.638, +0.691]$ | 0.6196 | 1.000 |

Table 9: B0 additional: complete metric–phase results.

| Phase | Metric | Spearman | Pearson | 95% interval | $p$ | Holm $p$ |
|---|---|---|---|---|---|---|
| Init | Linear CKA | $+0.028$ | $+0.388$ | $[-0.763, +0.683]$ | 0.9391 | 1.000 |
| Init | RBF CKA | $+0.098$ | $+0.415$ | $[-0.646, +0.707]$ | 0.7674 | 1.000 |
| Init | SVCCA | $-0.105$ | $+0.067$ | $[-0.682, +0.540]$ | 0.7500 | 1.000 |
| Init | Procrustes | $+0.112$ | $+0.065$ | $[-0.443, +0.591]$ | 0.7323 | 1.000 |
| Init | Linear $R^2$ | $-0.287$ | $+0.084$ | $[-0.629, +0.367]$ | 0.3688 | 1.000 |
| Warm-up | Linear CKA | $-0.217$ | $-0.217$ | $[-0.809, +0.409]$ | 0.4970 | 1.000 |
| Warm-up | RBF CKA | $-0.308$ | $-0.258$ | $[-0.781, +0.355]$ | 0.3281 | 1.000 |
| Warm-up | SVCCA | $-0.049$ | $-0.056$ | $[-0.664, +0.566]$ | 0.8849 | 1.000 |
| Warm-up | Procrustes | $+0.084$ | $-0.072$ | $[-0.520, +0.554]$ | 0.8021 | 1.000 |
| Warm-up | Linear $R^2$ | $-0.182$ | $-0.134$ | $[-0.620, +0.447]$ | 0.5704 | 1.000 |

Table 10: B2 original: complete metric–phase results.

| Phase | Metric | Spearman | Pearson | 95% interval | $p$ | Holm $p$ |
|---|---|---|---|---|---|---|
| Init | Linear CKA | $-0.056$ | $+0.297$ | $[-0.674, +0.610]$ | 0.8698 | 1.000 |
| Init | RBF CKA | $-0.084$ | $+0.246$ | $[-0.674, +0.598]$ | 0.7996 | 1.000 |
| Init | SVCCA | $-0.168$ | $-0.114$ | $[-0.721, +0.438]$ | 0.6032 | 1.000 |
| Init | Procrustes | $-0.077$ | $-0.007$ | $[-0.682, +0.553]$ | 0.8188 | 1.000 |
| Init | Linear $R^2$ | $-0.182$ | $-0.015$ | $[-0.718, +0.432]$ | 0.5697 | 1.000 |
| Warm-up | Linear CKA | $-0.308$ | $-0.218$ | $[-0.875, +0.429]$ | 0.3292 | 1.000 |
| Warm-up | RBF CKA | $-0.322$ | $-0.166$ | $[-0.891, +0.441]$ | 0.3071 | 1.000 |
| Warm-up | SVCCA | $-0.210$ | $-0.214$ | $[-0.715, +0.511]$ | 0.5130 | 1.000 |
| Warm-up | Procrustes | $-0.070$ | $-0.215$ | $[-0.688, +0.604]$ | 0.8334 | 1.000 |
| Warm-up | Linear $R^2$ | $-0.266$ | $-0.138$ | $[-0.863, +0.433]$ | 0.4040 | 1.000 |

Table 11: B4 original: complete metric–phase results.

| Phase | Metric | Spearman | Pearson | 95% interval | $p$ | Holm $p$ |
|---|---|---|---|---|---|---|
| Init | Linear CKA | $+0.119$ | $+0.324$ | $[-0.548, +0.708]$ | 0.7142 | 1.000 |
| Init | RBF CKA | $+0.112$ | $+0.308$ | $[-0.541, +0.679]$ | 0.7311 | 1.000 |
| Init | SVCCA | $+0.189$ | $-0.073$ | $[-0.445, +0.702]$ | 0.5588 | 1.000 |
| Init | Procrustes | $+0.077$ | $-0.165$ | $[-0.588, +0.671]$ | 0.8154 | 1.000 |
| Init | Linear $R^2$ | $+0.336$ | $+0.107$ | $[-0.285, +0.730]$ | 0.2851 | 1.000 |
| Warm-up | Linear CKA | $-0.336$ | $-0.448$ | $[-0.785, +0.290]$ | 0.2894 | 1.000 |
| Warm-up | RBF CKA | $-0.294$ | $-0.483$ | $[-0.804, +0.317]$ | 0.3572 | 1.000 |
| Warm-up | SVCCA | $-0.231$ | $-0.266$ | $[-0.775, +0.441]$ | 0.4750 | 1.000 |
| Warm-up | Procrustes | $+0.154$ | $-0.280$ | $[-0.545, +0.773]$ | 0.6321 | 1.000 |
| Warm-up | Linear $R^2$ | $-0.287$ | $-0.282$ | $[-0.745, +0.304]$ | 0.3706 | 1.000 |

Table 12: CIFAR-100: complete metric–phase results ($N = 8$).

| Phase | Metric | Spearman | Bootstrap 95% | Exact $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Init | Linear CKA | $+0.238$ | $[-0.800, +0.923]$ | 0.5821 | 1.000 |
| Init | RBF CKA | $+0.238$ | $[-0.747, +0.923]$ | 0.5821 | 1.000 |
| Init | SVCCA | $-0.143$ | $[-1.000, +0.811]$ | 0.7520 | 1.000 |
| Init | Procrustes | $+0.048$ | $[-1.000, +0.728]$ | 0.9349 | 1.000 |
| Init | Linear $R^2$ | $+0.119$ | $[-0.730, +0.975]$ | 0.7930 | 1.000 |
| Warm-up | Linear CKA | $+0.167$ | $[-0.829, +0.842]$ | 0.7033 | 1.000 |
| Warm-up | RBF CKA | $+0.167$ | $[-0.797, +0.842]$ | 0.7033 | 1.000 |
| Warm-up | SVCCA | $+0.024$ | $[-0.829, +0.800]$ | 0.9768 | 1.000 |
| Warm-up | Procrustes | $+0.190$ | $[-0.772, +0.905]$ | 0.6646 | 1.000 |
| Warm-up | Linear $R^2$ | $-0.333$ | $[-0.924, +0.519]$ | 0.4279 | 1.000 |

# C Gain reliability and independent-seed decisions

## C.1 Classical attenuation and paired variation

Under the assumptions of Eq. 5, independence gives $\mathrm{Cov}(x, \Delta) = \mathrm{Cov}(x, \mu)$ and $\mathrm{Var}(\Delta) = \mathrm{Var}(\mu) + \sigma^2_{\mathrm{seed}}$. Thus

$$\rho^2(x, \Delta) = \frac{\mathrm{Cov}(x, \mu)^2}{\mathrm{Var}(x)[\mathrm{Var}(\mu) + \sigma^2_{\mathrm{seed}}]} = \rho^2(x, \mu)\,\mathrm{ICC}. \tag{9}$$

If $k \geq 1$ seed errors have equal variance, are independent across seeds, and are jointly independent of $x$ and $\mu$, their average has variance $\sigma^2_{\mathrm{seed}}/k$. Writing $\mathrm{ICC}_1 = \mathrm{ICC}$ yields

$$\rho^2(x, \bar{\Delta}_k) = \mathrm{ICC}_k\, \rho^2(x, \mu), \qquad \mathrm{ICC}_k = \frac{k\,\mathrm{ICC}_1}{1 + (k-1)\mathrm{ICC}_1}. \tag{10}$$

Here $\rho$ is Pearson correlation; the empirical Spearman statistics are evaluated directly. For $k \geq 2$ paired observations, the sample variance identity is

$$\widehat{\sigma}^2_{\Delta} = \widehat{\sigma}^2_{A^{\mathrm{KD}}} + \widehat{\sigma}^2_{A^{\mathrm{task}}} - 2\widehat{\mathrm{Cov}}(A^{\mathrm{KD}}, A^{\mathrm{task}}). \tag{11}$$

All terms use the same paired observations and denominator $k - 1$. Paired-gain variation therefore depends on both branches and their covariance.

## C.2 Repeated gains and fixed decisions

Table 13 reports the paired gains underlying the repeatability analysis in §4.4. Both disjoint three-seed blocks use the same teacher, data, 12 students, and ten-epoch budget.

Table 13: Gain-order repeatability on corrected Ped2 B0 (RQ2). All 12 students are listed in the original-pool order of Table 2. Entries are mean ± sample standard deviation of paired AUC gains in percentage points, with three seeds in each disjoint block. Positive counts are seed-level holdout gains. The two columns of means have Spearman $-0.238$ and Pearson $-0.274$ correlation.

| Student index | Pilot gain (pp) | Holdout gain (pp) | Positive holdout seeds |
|---|---|---|---|
| 1 | $+0.2321 \pm 0.1592$ | $+0.6285 \pm 0.4175$ | 3/3 |
| 2 | $+0.6610 \pm 0.3228$ | $+0.2656 \pm 0.2770$ | 2/3 |
| 3 | $+0.4709 \pm 0.1345$ | $+0.1769 \pm 0.2634$ | 2/3 |
| 4 | $-1.2747 \pm 3.1678$ | $+0.2149 \pm 0.1588$ | 3/3 |
| 5 | $+0.1151 \pm 0.1340$ | $+0.5554 \pm 0.1407$ | 3/3 |
| 6 | $+0.4890 \pm 0.2565$ | $+0.2650 \pm 0.1985$ | 3/3 |
| 7 | $+0.7614 \pm 0.7019$ | $+0.0430 \pm 0.2574$ | 2/3 |
| 8 | $+0.0463 \pm 0.0875$ | $+0.1393 \pm 0.1911$ | 2/3 |
| 9 | $+0.2737 \pm 0.1430$ | $+0.0717 \pm 0.2166$ | 2/3 |
| 10 | $+0.1988 \pm 0.1850$ | $+0.2685 \pm 0.2096$ | 3/3 |
| 11 | $+1.1209 \pm 0.6551$ | $+0.5039 \pm 0.4712$ | 3/3 |
| 12 | $-1.2025 \pm 2.2531$ | $+0.7343 \pm 0.9943$ | 2/3 |

Table 5 gives the selected-set and excluded-student means. The S2-eligible mean-only rule launches seven students with positive holdout means, averaging $+0.2612$ AUC points. In Table 13, student 8 is the additional noise exclusion and student 5 the reported structural deferral.

Table 14: Frozen same-family ImageNet-1k decisions. Entries are three-seed mean gains in top-1 percentage points, using disjoint pilot and holdout seeds. A dash denotes an unavailable holdout.

| Student | Pilot gain | Holdout gain |
|---|---|---|
| d320-L8 | $+0.145$ | — |
| d320-L16 | $+2.855$ | $+2.497$ |
| d384-L8 | $+1.963$ | $+2.042$ |
| d384-L16 | $+3.989$ | $+3.849$ |

ImageNet selection outcomes are reported in §4.5 and Table 5; Table 14 provides the ViT pilot and holdout means. Final student weights were not archived for this runner. In the CNN comparison, ResNet-18 seed 123 changed data-loader workers, and seed 42 lacks a matched augmentation trace. These execution deviations limit the pairing of that comparison.

# D Seed, probe, and sample-size sensitivity

## D.1 Seed-budget sensitivity

The frozen secondary analysis enumerates every nonempty subset of three seeds. Table 15 summarizes all 280 metric–phase–subset records. The probe-subset column varies probe seeds against the full outcome mean; the outcome-subset column varies outcome seeds against the full probe mean. The matched-subset column varies both within matching seeds and is unavailable for original B0, whose blocks are disjoint. Each range spans all metrics and phases at that budget. The numerical supplement retains the individual records and per-metric ranges. Parameter-count correlations with full mean gains are $+0.172$, $+0.007$, $-0.032$, and $-0.091$ for original B0, additional B0, B2, and B4.

Table 15: Seed-budget sensitivity: full ranges across all five metrics, both phases and every subset of the stated size. These are descriptive ranges, not confidence intervals.

| Pool | Seeds | Probe subset | Outcome subset | Matched subset |
|---|---|---|---|---|
| B0 original | 1 | $[-0.203, +0.406]$ | $[-0.280, +0.692]$ | — |
| | 2 | $[-0.224, +0.490]$ | $[-0.140, +0.545]$ | — |
| | 3 | $[-0.084, +0.434]$ | $[-0.084, +0.434]$ | — |
| B0 additional | 1 | $[-0.566, +0.559]$ | $[-0.580, +0.573]$ | $[-0.664, +0.476]$ |
| | 2 | $[-0.399, +0.273]$ | $[-0.476, +0.350]$ | $[-0.399, +0.350]$ |
| | 3 | $[-0.308, +0.112]$ | $[-0.308, +0.112]$ | $[-0.308, +0.112]$ |
| B2 original | 1 | $[-0.587, +0.434]$ | $[-0.503, +0.476]$ | $[-0.629, +0.469]$ |
| | 2 | $[-0.538, +0.329]$ | $[-0.336, +0.273]$ | $[-0.587, +0.294]$ |
| | 3 | $[-0.322, -0.056]$ | $[-0.322, -0.056]$ | $[-0.322, -0.056]$ |
| B4 original | 1 | $[-0.524, +0.413]$ | $[-0.566, +0.462]$ | $[-0.566, +0.357]$ |
| | 2 | $[-0.385, +0.448]$ | $[-0.329, +0.490]$ | $[-0.385, +0.238]$ |
| | 3 | $[-0.336, +0.336]$ | $[-0.336, +0.336]$ | $[-0.336, +0.336]$ |

## D.2 Probe-partition sensitivity

A post-hoc replay divides the 320 cached Ped2 training examples into complementary interleaved sets at zero-based positions $0, 2, \ldots, 318$ and $1, 3, \ldots, 319$, without labels or outcome-based selection. All five frozen metric implementations are recomputed at initialization and after two task-only epochs for all 144 jobs, giving 2,880 values. Feature-file digests match the original records; training and checkpoints are unchanged.

Tables below retain every metric and phase in all four pools. Gain columns correlate three-seed similarity means with the original three-seed gain means; score-order columns compare each reduced

probe with the full-probe similarity ordering. Original B0 retains independent holdout outcomes; other pools retain shared probe/outcome seeds. Centering, PCA and RBF bandwidth are recomputed within each subset. The maximum Procrustes PCA rank changes from 319 to 159; SVCCA and regression retain their 20-component cap. Thus probe size and example composition vary together, with the stage-3 feature definition fixed.

Table 16: Ped2 B0 original: complete probe-partition sensitivity (Spearman correlations).

| | | Similarity–gain $\widehat{\rho}_{\mathrm{sim}}$ | | | Score-order agreement | |
|---|---|---|---|---|---|---|
| Phase | Metric | Full 320 | Even 160 | Odd 160 | Even/full | Odd/full |
| Init | Linear CKA | $-0.028$ | $+0.028$ | $-0.028$ | $+0.993$ | $+1.000$ |
| Init | RBF CKA | $-0.014$ | $-0.028$ | $-0.028$ | $+0.993$ | $+0.993$ |
| Init | SVCCA | $+0.434$ | $+0.385$ | $+0.070$ | $+0.657$ | $+0.727$ |
| Init | Procrustes | $-0.077$ | $-0.077$ | $-0.077$ | $+1.000$ | $+1.000$ |
| Init | Linear $R^2$ | $+0.266$ | $+0.420$ | $+0.238$ | $+0.895$ | $+0.937$ |
| Warm-up | Linear CKA | $+0.126$ | $+0.161$ | $+0.175$ | $+0.972$ | $+0.979$ |
| Warm-up | RBF CKA | $+0.070$ | $+0.105$ | $+0.098$ | $+0.979$ | $+0.972$ |
| Warm-up | SVCCA | $+0.224$ | $+0.217$ | $+0.189$ | $+0.993$ | $+0.965$ |
| Warm-up | Procrustes | $-0.084$ | $-0.084$ | $-0.084$ | $+1.000$ | $+1.000$ |
| Warm-up | Linear $R^2$ | $+0.161$ | $+0.154$ | $+0.168$ | $+0.993$ | $+0.993$ |

Table 17: Ped2 B0 additional: complete probe-partition sensitivity (Spearman correlations).

| | | Similarity–gain $\widehat{\rho}_{\mathrm{sim}}$ | | | Score-order agreement | |
|---|---|---|---|---|---|---|
| Phase | Metric | Full 320 | Even 160 | Odd 160 | Even/full | Odd/full |
| Init | Linear CKA | $+0.028$ | $+0.014$ | $-0.049$ | $+0.993$ | $+0.986$ |
| Init | RBF CKA | $+0.098$ | $+0.063$ | $+0.021$ | $+0.993$ | $+0.986$ |
| Init | SVCCA | $-0.105$ | $-0.217$ | $-0.336$ | $+0.909$ | $+0.902$ |
| Init | Procrustes | $+0.112$ | $+0.112$ | $+0.112$ | $+1.000$ | $+1.000$ |
| Init | Linear $R^2$ | $-0.287$ | $-0.273$ | $-0.294$ | $+0.951$ | $+0.972$ |
| Warm-up | Linear CKA | $-0.217$ | $-0.238$ | $-0.252$ | $+0.993$ | $+0.986$ |
| Warm-up | RBF CKA | $-0.308$ | $-0.266$ | $-0.308$ | $+0.972$ | $+0.979$ |
| Warm-up | SVCCA | $-0.049$ | $-0.091$ | $-0.168$ | $+0.993$ | $+0.972$ |
| Warm-up | Procrustes | $+0.084$ | $+0.084$ | $+0.084$ | $+1.000$ | $+1.000$ |
| Warm-up | Linear $R^2$ | $-0.182$ | $-0.182$ | $-0.182$ | $+0.965$ | $+1.000$ |

Table 18: Ped2 B2: complete probe-partition sensitivity (Spearman correlations).

| | | Similarity–gain $\widehat{\rho}_{\mathrm{sim}}$ | | | Score-order agreement | |
|---|---|---|---|---|---|---|
| Phase | Metric | Full 320 | Even 160 | Odd 160 | Even/full | Odd/full |
| Init | Linear CKA | $-0.056$ | $+0.014$ | $-0.056$ | $+0.986$ | $+1.000$ |
| Init | RBF CKA | $-0.084$ | $-0.084$ | $-0.084$ | $+1.000$ | $+1.000$ |
| Init | SVCCA | $-0.168$ | $+0.007$ | $-0.126$ | $+0.427$ | $+0.594$ |
| Init | Procrustes | $-0.077$ | $-0.105$ | $-0.133$ | $+0.951$ | $+0.958$ |
| Init | Linear $R^2$ | $-0.182$ | $+0.077$ | $-0.014$ | $+0.888$ | $+0.860$ |
| Warm-up | Linear CKA | $-0.308$ | $-0.280$ | $-0.259$ | $+0.972$ | $+0.972$ |
| Warm-up | RBF CKA | $-0.322$ | $-0.308$ | $-0.308$ | $+0.993$ | $+0.993$ |
| Warm-up | SVCCA | $-0.210$ | $-0.231$ | $-0.231$ | $+0.979$ | $+0.965$ |
| Warm-up | Procrustes | $-0.070$ | $-0.196$ | $-0.196$ | $+0.972$ | $+0.972$ |
| Warm-up | Linear $R^2$ | $-0.266$ | $-0.259$ | $-0.231$ | $+0.993$ | $+0.965$ |

Table 19: Ped2 B4: complete probe-partition sensitivity (Spearman correlations).

| | | Similarity–gain $\widehat{\rho}_{\mathrm{sim}}$ | | | Score-order agreement | |
|---|---|---|---|---|---|---|
| Phase | Metric | Full 320 | Even 160 | Odd 160 | Even/full | Odd/full |
| Init | Linear CKA | $+0.119$ | $+0.098$ | $+0.077$ | $+0.993$ | $+0.993$ |
| Init | RBF CKA | $+0.112$ | $+0.112$ | $+0.112$ | $+1.000$ | $+1.000$ |
| Init | SVCCA | $+0.189$ | $+0.259$ | $+0.455$ | $+0.601$ | $+0.727$ |
| Init | Procrustes | $+0.077$ | $+0.077$ | $+0.077$ | $+1.000$ | $+1.000$ |
| Init | Linear $R^2$ | $+0.336$ | $+0.357$ | $+0.371$ | $+0.860$ | $+0.958$ |
| Warm-up | Linear CKA | $-0.336$ | $-0.357$ | $-0.245$ | $+0.993$ | $+0.972$ |
| Warm-up | RBF CKA | $-0.294$ | $-0.294$ | $-0.294$ | $+1.000$ | $+1.000$ |
| Warm-up | SVCCA | $-0.231$ | $-0.182$ | $-0.140$ | $+0.986$ | $+0.972$ |
| Warm-up | Procrustes | $+0.154$ | $+0.154$ | $+0.154$ | $+1.000$ | $+1.000$ |
| Warm-up | Linear $R^2$ | $-0.287$ | $-0.287$ | $-0.252$ | $+1.000$ | $+0.979$ |

### D.3 Model-based detection calibration

A separate post-hoc simulation uses independent continuous pairs from a Gaussian copula with latent Pearson correlation $r = 2\sin(\pi\rho_S/6)$, giving population Spearman correlation $\rho_S$. At each $N \in \{8, 12\}$ and $\rho_S \in \{0, 0.01, \ldots, 0.99\}$, 200,000 simulated samples are tested against a two-sided rank-permutation threshold at each of two levels. The null is exact over $8! = 40{,}320$ permutations for $N = 8$ and uses 1,000,000 permutations for $N = 12$. The simulation seed is 202609251; common random draws are used across the correlation grid. All 400 power estimates are retained in the numerical supplement.

Table 20 gives the adjacent grid points bracketing 80% power, with Monte Carlo standard error approximately $0.0009$ at power $0.8$. The calibration describes test resolution under the stated i.i.d. model. The $\alpha = 0.005$ level is a conservative single-test reference for ten comparisons; joint Holm power is not simulated.

Table 20: Model-based population Spearman correlation required for 80% power, bracketed by adjacent simulated grid points.

| Configurations $N$ | Test level $\alpha$ | 80%-power crossing |
|---|---|---|
| 8 | 0.05 | $[0.88, 0.89]$ |
| 8 | 0.005 | $[0.96, 0.97]$ |
| 12 | 0.05 | $[0.75, 0.76]$ |
| 12 | 0.005 | $[0.87, 0.88]$ |

## E Complete post-hoc selection results

The post-hoc replay reports every metric, phase and selection size $K = 1, \ldots, N$ for all five primary pools. Entries are $100B_K$ from Eq. 7: selected mean gain minus the exact uniform-random expectation, in percentage points. Larger scores rank first; cutoff ties average uniformly over tied subsets. CKA, RBF, Proc. and $R^2$ denote linear CKA, RBF CKA, orthogonal Procrustes and linear regression. Param. ranks larger parameter counts; Pilot ranks original-B0 pilot gains over all 12 students. Dashes indicate unavailable baselines.

Full-score tables use all three probe and outcome seeds. Original B0 has disjoint blocks; other pools share seeds and additionally report all three cross-seed folds, each selecting with two probe seeds and evaluating the omitted outcome seed. Entries average the fold utilities. Overlapping folds and selection sizes yield dependent descriptive comparisons; $K = N$ has zero advantage for every rule.

The numerical supplement retains all 3,980 seed-subset selection rows, 1,320 cross-seed fold rows, and 292 student-pair differences, including per-seed gains, inclusion weights, and gaps to the best observed $K$-subset. On original B0, warm-up selection uses 72 student training epochs, paired-pilot selection uses 720, and paired evaluation adds $60K$. Initialization and parameter-count selection

require no student training. These epoch counts exclude feature inference, metric computation, and architecture-dependent runtime.

Table 21: Ped2 B0 original: three-seed probe means. All selection sizes; advantage over random selection in percentage points (pool mean $+0.3223$).

| | Initialization | | | | | Warm-up | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ | Param. | Pilot |
| 1 | $+0.306$ | $+0.306$ | $+0.306$ | $+0.306$ | $+0.306$ | $+0.182$ | $+0.182$ | $+0.233$ | $+0.306$ | $+0.233$ | $-0.057$ | $+0.182$ |
| 2 | $+0.270$ | $+0.270$ | $+0.125$ | $+0.080$ | $+0.125$ | $+0.297$ | $+0.297$ | $+0.207$ | $+0.062$ | $+0.207$ | $-0.055$ | $-0.049$ |
| 3 | $+0.096$ | $+0.096$ | $+0.221$ | $+0.018$ | $-0.010$ | $+0.276$ | $+0.162$ | $+0.276$ | $-0.007$ | $+0.276$ | $+0.101$ | $-0.051$ |
| 4 | $+0.002$ | $+0.002$ | $+0.096$ | $+0.059$ | $+0.051$ | $+0.180$ | $+0.180$ | $+0.137$ | $-0.032$ | $+0.137$ | $+0.134$ | $-0.053$ |
| 5 | $+0.038$ | $+0.038$ | $+0.065$ | $+0.010$ | $+0.077$ | $+0.115$ | $+0.094$ | $+0.059$ | $-0.037$ | $+0.059$ | $+0.054$ | $-0.071$ |
| 6 | $+0.022$ | $+0.022$ | $+0.045$ | $+0.048$ | $+0.022$ | $+0.054$ | $+0.032$ | $+0.040$ | $+0.008$ | $+0.032$ | $+0.001$ | $-0.101$ |
| 7 | $-0.007$ | $+0.004$ | $+0.065$ | $+0.001$ | $+0.011$ | $+0.006$ | $+0.006$ | $+0.026$ | $+0.033$ | $+0.006$ | $+0.027$ | $-0.043$ |
| 8 | $-0.019$ | $-0.019$ | $+0.025$ | $-0.031$ | $+0.061$ | $-0.017$ | $-0.017$ | $+0.009$ | $-0.006$ | $-0.002$ | $+0.000$ | $-0.044$ |
| 9 | $-0.024$ | $-0.024$ | $+0.048$ | $-0.034$ | $+0.042$ | $-0.022$ | $-0.022$ | $-0.008$ | $-0.033$ | $-0.008$ | $-0.006$ | $-0.014$ |
| 10 | $-0.027$ | $-0.027$ | $+0.025$ | $-0.036$ | $+0.033$ | $-0.025$ | $-0.025$ | $-0.012$ | $-0.036$ | $-0.012$ | $-0.016$ | $-0.030$ |
| 11 | $-0.037$ | $-0.037$ | $+0.013$ | $+0.005$ | $+0.013$ | $-0.028$ | $+0.005$ | $+0.017$ | $-0.037$ | $+0.017$ | $-0.028$ | $+0.010$ |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ |

Table 22: Ped2 B0 additional: three-seed probe means. All selection sizes; advantage over random selection in percentage points (pool mean $+0.1503$).

| | Initialization | | | | | Warm-up | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ | Param. | Pilot |
| 1 | $+0.845$ | $+0.845$ | $-0.015$ | $+0.023$ | $-0.015$ | $+0.401$ | $-0.333$ | $+0.401$ | $+0.023$ | $+0.021$ | $-0.015$ | – |
| 2 | $+0.194$ | $+0.194$ | $-0.174$ | $-0.105$ | $-0.124$ | $+0.034$ | $+0.034$ | $+0.083$ | $+0.068$ | $+0.211$ | $+0.049$ | – |
| 3 | $+0.018$ | $+0.124$ | $-0.131$ | $+0.211$ | $-0.098$ | $-0.190$ | $-0.190$ | $+0.063$ | $+0.151$ | $+0.063$ | $+0.040$ | – |
| 4 | $+0.010$ | $+0.010$ | $-0.157$ | $+0.147$ | $-0.157$ | $-0.137$ | $-0.137$ | $+0.043$ | $-0.001$ | $+0.043$ | $-0.054$ | – |
| 5 | $+0.071$ | $+0.071$ | $-0.062$ | $+0.182$ | $-0.217$ | $-0.156$ | $-0.156$ | $+0.026$ | $-0.128$ | $-0.032$ | $+0.126$ | – |
| 6 | $+0.052$ | $+0.052$ | $+0.015$ | $+0.075$ | $-0.040$ | $+0.011$ | $-0.207$ | $-0.085$ | $+0.034$ | $-0.133$ | $-0.001$ | – |
| 7 | $+0.011$ | $+0.102$ | $-0.052$ | $-0.027$ | $+0.023$ | $+0.003$ | $-0.056$ | $+0.048$ | $-0.018$ | $-0.179$ | $-0.066$ | – |
| 8 | $+0.060$ | $+0.060$ | $+0.060$ | $-0.009$ | $+0.060$ | $-0.055$ | $-0.035$ | $-0.015$ | $-0.018$ | $-0.051$ | $-0.008$ | – |
| 9 | $+0.066$ | $+0.066$ | $+0.056$ | $+0.036$ | $+0.056$ | $-0.036$ | $-0.036$ | $-0.001$ | $-0.042$ | $-0.033$ | $+0.028$ | – |
| 10 | $+0.061$ | $+0.061$ | $+0.052$ | $-0.001$ | $-0.014$ | $-0.034$ | $-0.001$ | $-0.034$ | $+0.002$ | $-0.034$ | $+0.002$ | – |
| 11 | $-0.002$ | $-0.002$ | $+0.058$ | $+0.001$ | $-0.002$ | $-0.002$ | $-0.002$ | $-0.029$ | $-0.002$ | $-0.002$ | $-0.002$ | – |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | – |

Table 23: Ped2 B0 additional: mean of all three cross-seed folds. All selection sizes; advantage over random selection in percentage points (pool mean $+0.1503$).

| | Initialization | | | | | Warm-up | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ |
| 1 | $+0.845$ | $+0.845$ | $+0.119$ | $+0.023$ | $+0.119$ | $-0.085$ | $-0.339$ | $+0.367$ | $+0.114$ | $-0.004$ |
| 2 | $+0.327$ | $+0.194$ | $+0.080$ | $-0.105$ | $+0.080$ | $+0.141$ | $-0.089$ | $-0.032$ | $-0.145$ | $-0.098$ |
| 3 | $+0.137$ | $+0.137$ | $+0.013$ | $+0.176$ | $+0.176$ | $-0.036$ | $-0.077$ | $+0.044$ | $+0.179$ | $+0.004$ |
| 4 | $+0.140$ | $+0.083$ | $+0.067$ | $+0.152$ | $+0.149$ | $-0.180$ | $-0.237$ | $+0.021$ | $+0.116$ | $+0.110$ |
| 5 | $+0.211$ | $+0.152$ | $+0.039$ | $+0.182$ | $+0.159$ | $-0.151$ | $-0.007$ | $+0.030$ | $+0.016$ | $-0.025$ |
| 6 | $+0.161$ | $+0.163$ | $+0.067$ | $+0.109$ | $+0.235$ | $+0.026$ | $+0.004$ | $+0.042$ | $-0.072$ | $-0.011$ |
| 7 | $+0.107$ | $+0.109$ | $+0.004$ | $-0.028$ | $+0.178$ | $-0.041$ | $+0.021$ | $+0.154$ | $-0.032$ | $-0.063$ |
| 8 | $+0.105$ | $+0.105$ | $+0.107$ | $+0.018$ | $+0.107$ | $-0.024$ | $-0.013$ | $+0.042$ | $-0.027$ | $-0.071$ |
| 9 | $+0.094$ | $+0.094$ | $+0.066$ | $+0.036$ | $+0.046$ | $+0.011$ | $+0.020$ | $-0.017$ | $-0.043$ | $-0.076$ |
| 10 | $+0.098$ | $+0.052$ | $+0.068$ | $-0.001$ | $+0.056$ | $-0.015$ | $-0.006$ | $+0.012$ | $+0.001$ | $+0.007$ |
| 11 | $-0.003$ | $-0.003$ | $+0.059$ | $+0.001$ | $+0.004$ | $-0.002$ | $+0.006$ | $-0.018$ | $-0.002$ | $-0.020$ |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ |

Table 24: Ped2 B2: three-seed probe means. All selection sizes; advantage over random selection in percentage points (pool mean $+0.0231$).

| | Initialization | | | | | Warm-up | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ | Param. | Pilot |
| 1 | $+0.175$ | $+0.175$ | $+0.175$ | $+0.123$ | $+0.175$ | $+1.004$ | $+1.004$ | $+1.004$ | $-2.142$ | $+1.004$ | $+0.550$ | – |
| 2 | $+0.149$ | $+0.149$ | $+0.363$ | $+0.564$ | $+0.363$ | $-0.569$ | $-0.569$ | $+0.564$ | $-0.796$ | $+0.564$ | $+0.425$ | – |
| 3 | $+0.133$ | $+0.133$ | $-0.472$ | $+0.409$ | $+0.244$ | $-0.338$ | $-0.338$ | $-0.338$ | $-0.498$ | $-0.338$ | $-0.431$ | – |
| 4 | $+0.101$ | $+0.101$ | $-0.279$ | $+0.309$ | $+0.214$ | $-0.557$ | $-0.557$ | $-0.229$ | $-0.270$ | $-0.557$ | $-0.292$ | – |
| 5 | $+0.282$ | $+0.282$ | $-0.222$ | $+0.291$ | $-0.072$ | $-0.444$ | $-0.444$ | $-0.181$ | $-0.015$ | $-0.444$ | $-0.223$ | – |
| 6 | $+0.327$ | $+0.327$ | $-0.116$ | $+0.040$ | $-0.043$ | $-0.354$ | $-0.354$ | $-0.354$ | $+0.008$ | $-0.354$ | $-0.177$ | – |
| 7 | $+0.311$ | $+0.107$ | $-0.273$ | $+0.093$ | $+0.107$ | $-0.237$ | $-0.237$ | $-0.237$ | $+0.008$ | $-0.237$ | $-0.008$ | – |
| 8 | $+0.121$ | $+0.121$ | $-0.224$ | $+0.140$ | $+0.145$ | $-0.180$ | $-0.180$ | $-0.138$ | $-0.145$ | $-0.156$ | $+0.020$ | – |
| 9 | $+0.141$ | $+0.153$ | $-0.087$ | $+0.157$ | $-0.109$ | $-0.099$ | $-0.114$ | $-0.077$ | $-0.077$ | $-0.105$ | $+0.064$ | – |
| 10 | $+0.168$ | $+0.168$ | $-0.069$ | $-0.073$ | $-0.069$ | $-0.048$ | $-0.048$ | $-0.040$ | $-0.040$ | $-0.040$ | $-0.064$ | – |
| 11 | $+0.195$ | $+0.195$ | $-0.042$ | $-0.016$ | $-0.042$ | $-0.016$ | $-0.016$ | $-0.020$ | $-0.016$ | $-0.016$ | $-0.016$ | – |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | – |

Table 25: Ped2 B2: mean of all three cross-seed folds. All selection sizes; advantage over random selection in percentage points (pool mean $+0.0231$).

| | Initialization | | | | | Warm-up | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ |
| 1 | $+0.152$ | $+0.152$ | $+0.204$ | $+0.466$ | $+0.336$ | $+1.004$ | $+1.004$ | $+0.846$ | $-2.142$ | $+0.661$ |
| 2 | $+0.149$ | $+0.309$ | $+0.295$ | $+0.564$ | $+0.252$ | $-0.569$ | $-0.569$ | $+0.484$ | $-0.773$ | $+0.566$ |
| 3 | $+0.289$ | $+0.289$ | $+0.486$ | $+0.473$ | $+0.363$ | $-0.246$ | $-0.246$ | $+0.450$ | $-0.336$ | $+0.543$ |
| 4 | $+0.329$ | $+0.180$ | $+0.419$ | $+0.309$ | $+0.364$ | $-0.106$ | $-0.106$ | $+0.477$ | $-0.199$ | $+0.468$ |
| 5 | $+0.226$ | $+0.256$ | $-0.018$ | $+0.320$ | $+0.441$ | $-0.115$ | $-0.381$ | $-0.119$ | $-0.095$ | $+0.108$ |
| 6 | $+0.329$ | $+0.278$ | $+0.056$ | $+0.343$ | $+0.409$ | $-0.341$ | $-0.331$ | $-0.057$ | $-0.042$ | $-0.273$ |
| 7 | $+0.332$ | $+0.332$ | $+0.112$ | $+0.093$ | $+0.038$ | $-0.251$ | $-0.251$ | $+0.018$ | $-0.015$ | $-0.242$ |
| 8 | $+0.108$ | $+0.121$ | $+0.072$ | $+0.125$ | $+0.068$ | $-0.189$ | $-0.192$ | $-0.151$ | $+0.065$ | $-0.191$ |
| 9 | $+0.158$ | $+0.158$ | $+0.085$ | $-0.107$ | $-0.103$ | $-0.140$ | $-0.145$ | $-0.100$ | $+0.107$ | $-0.133$ |
| 10 | $-0.050$ | $+0.179$ | $+0.109$ | $-0.073$ | $-0.057$ | $-0.077$ | $-0.092$ | $-0.056$ | $-0.040$ | $-0.072$ |
| 11 | $-0.022$ | $-0.035$ | $-0.020$ | $-0.016$ | $-0.029$ | $-0.025$ | $-0.025$ | $-0.020$ | $-0.016$ | $-0.017$ |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ |

Table 26: Ped2 B4: three-seed probe means. All selection sizes; advantage over random selection in percentage points (pool mean $-0.0581$).

| | Initialization | | | | | Warm-up | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ | Param. | Pilot |
| 1 | $+0.233$ | $+0.233$ | $+0.233$ | $+0.556$ | $+0.233$ | $-2.873$ | $-2.873$ | $+0.424$ | $-2.873$ | $+0.424$ | $+0.556$ | – |
| 2 | $+0.173$ | $+0.173$ | $+0.395$ | $-1.158$ | $+0.395$ | $-1.224$ | $-1.224$ | $+0.268$ | $-1.158$ | $+0.268$ | $+0.390$ | – |
| 3 | $+0.153$ | $+0.153$ | $-0.694$ | $-0.698$ | $+0.498$ | $-0.779$ | $-0.779$ | $-0.779$ | $-0.531$ | $-0.779$ | $-0.698$ | – |
| 4 | $+0.290$ | $+0.290$ | $-0.465$ | $-0.343$ | $+0.401$ | $-0.408$ | $-0.408$ | $-0.774$ | $-0.370$ | $-0.774$ | $-0.495$ | – |
| 5 | $+0.317$ | $+0.317$ | $-0.231$ | $-0.252$ | $+0.344$ | $-0.304$ | $-0.479$ | $-0.597$ | $-0.156$ | $-0.479$ | $-0.315$ | – |
| 6 | $+0.357$ | $+0.357$ | $-0.072$ | $-0.093$ | $+0.357$ | $-0.380$ | $-0.380$ | $-0.380$ | $-0.059$ | $-0.380$ | $-0.194$ | – |
| 7 | $+0.311$ | $+0.197$ | $-0.001$ | $-0.063$ | $+0.409$ | $-0.253$ | $-0.253$ | $-0.253$ | $-0.035$ | $-0.253$ | $-0.106$ | – |
| 8 | $+0.177$ | $+0.177$ | $-0.096$ | $-0.051$ | $+0.263$ | $-0.217$ | $-0.131$ | $-0.131$ | $-0.125$ | $-0.131$ | $-0.089$ | – |
| 9 | $+0.237$ | $+0.237$ | $-0.073$ | $+0.002$ | $-0.085$ | $-0.113$ | $-0.113$ | $-0.054$ | $-0.087$ | $-0.113$ | $+0.002$ | – |
| 10 | $+0.236$ | $+0.236$ | $-0.055$ | $-0.075$ | $-0.055$ | $-0.046$ | $-0.046$ | $-0.027$ | $-0.027$ | $-0.046$ | $-0.075$ | – |
| 11 | $+0.261$ | $+0.261$ | $-0.047$ | $-0.021$ | $-0.047$ | $-0.021$ | $-0.021$ | $-0.003$ | $-0.021$ | $-0.021$ | $-0.021$ | – |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | – |

Table 27: Ped2 B4: mean of all three cross-seed folds. All selection sizes; advantage over random selection in percentage points (pool mean $-0.0581$).

| | Initialization | | | | | Warm-up | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ |
| 1 | $+0.150$ | $+0.150$ | $+0.231$ | $+0.556$ | $+0.227$ | $+0.246$ | $+0.246$ | $+0.341$ | $-2.873$ | $+0.341$ |
| 2 | $+0.173$ | $+0.173$ | $+0.528$ | $-1.158$ | $+0.487$ | $-1.243$ | $+0.346$ | $+0.385$ | $-1.158$ | $+0.362$ |
| 3 | $+0.325$ | $+0.206$ | $+0.510$ | $-0.698$ | $+0.393$ | $-0.642$ | $+0.418$ | $+0.359$ | $-0.637$ | $+0.359$ |
| 4 | $+0.272$ | $+0.350$ | $-0.232$ | $-0.301$ | $+0.486$ | $-0.414$ | $-0.414$ | $+0.336$ | $-0.310$ | $+0.336$ |
| 5 | $+0.335$ | $+0.382$ | $-0.110$ | $-0.218$ | $-0.142$ | $-0.222$ | $-0.531$ | $-0.226$ | $-0.238$ | $+0.069$ |
| 6 | $+0.389$ | $+0.361$ | $-0.035$ | $-0.093$ | $-0.100$ | $-0.348$ | $-0.344$ | $-0.167$ | $-0.073$ | $-0.354$ |
| 7 | $+0.345$ | $+0.362$ | $+0.044$ | $-0.063$ | $+0.007$ | $-0.267$ | $-0.273$ | $-0.272$ | $-0.001$ | $-0.273$ |
| 8 | $+0.179$ | $+0.177$ | $+0.060$ | $-0.051$ | $+0.026$ | $-0.215$ | $-0.217$ | $-0.170$ | $+0.037$ | $-0.240$ |
| 9 | $+0.245$ | $+0.240$ | $+0.075$ | $+0.002$ | $+0.055$ | $-0.170$ | $-0.165$ | $-0.075$ | $+0.066$ | $-0.122$ |
| 10 | $-0.034$ | $+0.257$ | $+0.079$ | $-0.075$ | $-0.032$ | $-0.099$ | $-0.076$ | $-0.036$ | $-0.027$ | $-0.060$ |
| 11 | $-0.017$ | $-0.030$ | $-0.015$ | $-0.021$ | $-0.029$ | $-0.020$ | $-0.020$ | $-0.003$ | $-0.021$ | $-0.021$ |
| 12 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ |

Table 28: CIFAR-100: three-seed probe means. All selection sizes; advantage over random selection in percentage points (pool mean $+1.8650$).

| | Initialization | | | | | Warm-up | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ | Param. | Pilot |
| 1 | $-0.612$ | $-0.612$ | $-0.615$ | $-0.612$ | $-0.615$ | $-0.612$ | $-0.612$ | $-0.612$ | $-0.612$ | $-0.612$ | – | – |
| 2 | $+0.440$ | $+0.440$ | $-0.610$ | $-0.572$ | $+0.000$ | $-0.562$ | $-0.562$ | $-0.562$ | $-0.562$ | $-0.562$ | – | – |
| 3 | $+0.116$ | $+0.116$ | $-0.202$ | $-0.552$ | $+0.497$ | $+0.123$ | $+0.123$ | $+0.123$ | $+0.123$ | $-0.169$ | – | – |
| 4 | $+0.241$ | $+0.241$ | $+0.041$ | $-0.041$ | $+0.565$ | $-0.041$ | $-0.041$ | $-0.041$ | $-0.041$ | $-0.281$ | – | – |
| 5 | $+0.090$ | $+0.090$ | $+0.331$ | $+0.090$ | $+0.346$ | $+0.090$ | $+0.090$ | $+0.090$ | $+0.121$ | $+0.074$ | – | – |
| 6 | $+0.203$ | $+0.203$ | $+0.187$ | $+0.203$ | $+0.187$ | $+0.203$ | $+0.203$ | $-0.027$ | $+0.203$ | $-0.039$ | – | – |
| 7 | $+0.086$ | $+0.086$ | $+0.087$ | $+0.086$ | $+0.073$ | $+0.086$ | $+0.086$ | $+0.086$ | $+0.086$ | $-0.110$ | – | – |
| 8 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | – | – |

Table 29: CIFAR-100: mean of all three cross-seed folds. All selection sizes; advantage over random selection in percentage points (pool mean $+1.8650$).

| | Initialization | | | | | Warm-up | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | CKA | RBF | SVCCA | Proc. | $R^2$ | CKA | RBF | SVCCA | Proc. | $R^2$ |
| 1 | $+0.295$ | $+0.295$ | $-0.615$ | $-0.612$ | $+0.992$ | $-0.612$ | $-0.612$ | $-0.612$ | $-0.612$ | $-0.612$ |
| 2 | $+0.012$ | $+0.012$ | $-0.450$ | $-0.572$ | $+0.188$ | $-0.033$ | $-0.033$ | $-0.033$ | $-0.562$ | $-0.312$ |
| 3 | $+0.274$ | $+0.274$ | $-0.202$ | $-0.402$ | $+0.203$ | $-0.048$ | $-0.048$ | $+0.123$ | $-0.048$ | $-0.244$ |
| 4 | $+0.223$ | $+0.223$ | $+0.027$ | $-0.041$ | $+0.283$ | $-0.041$ | $-0.041$ | $-0.041$ | $-0.041$ | $-0.078$ |
| 5 | $+0.038$ | $+0.038$ | $+0.331$ | $+0.140$ | $+0.332$ | $+0.108$ | $+0.108$ | $-0.034$ | $+0.108$ | $+0.007$ |
| 6 | $+0.051$ | $+0.051$ | $+0.240$ | $+0.203$ | $+0.152$ | $+0.203$ | $+0.203$ | $-0.027$ | $+0.203$ | $+0.101$ |
| 7 | $+0.093$ | $+0.093$ | $+0.087$ | $+0.086$ | $+0.119$ | $+0.086$ | $+0.086$ | $+0.028$ | $+0.086$ | $-0.047$ |
| 8 | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ | $+0.000$ |