WFD: TEACHER FEATURE WHITENING AND ASYMMETRIC STUDENT DECORRELATION FOR CROSS-ARCHITECTURE KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-architecture knowledge distillation transfers knowledge from large teacher models to structurally different student models, but existing methods do not address the distributional anisotropy of teacher features across heterogeneous architectures. We show that teacher features from different architecture families exhibit substantially different distributional characteristics, and that this heterogeneity distorts the alignment signal received by the student, leading to inconsistent performance across teacher–student pairs. We propose WFD (Whitened Feature Distillation), which addresses this problem through two complementary components: (1) ZCA whitening applied to frozen teacher features, which flattens the teacher's eigenvalue spectrum toward an isotropic alignment target without modifying the teacher or introducing additional parameters; and (2) asymmetric student decorrelation, which suppresses representational redundancy in the student projector to improve the diversity of learned features. Evaluated on CIFAR-100 across a wide range of heterogeneous teacher–student pairs spanning Transformer, CNN, and MLP architectures, WFD yields consistent improvements over prior methods. On ImageNet-1K, WFD remains competitive across the evaluated teacher–student pairs, with the most consistent gains observed for MLP students. Code is available at https://anonymous.4open.science/r/WFD-7D80.

1 INTRODUCTION

Knowledge distillation (KD) is an effective model compression technique that transfers knowledge from a large teacher model to a small student model (Hinton et al., 2015). As diverse architectures—Vision Transformer (Dosovitskiy et al., 2021), Swin Transformer (Liu et al., 2021), ConvNeXt (Liu et al., 2022), and MLP-Mixer (Tolstikhin et al., 2021)—have achieved strong performance, leveraging them as teachers to train lightweight CNN-based student models such as ResNet (He et al., 2016) and MobileNet (Howard et al., 2017) has become increasingly important. This setting, known as *cross-architecture* knowledge distillation, has attracted growing research attention.

Cross-architecture KD is substantially more difficult than same-architecture KD because teacher and student models possess different inductive biases, making their internal representation spaces structurally heterogeneous. In particular, we observe that teacher features from different architectures exhibit markedly different distributional characteristics in terms of their *eigenvalue spectra*. Analyzing Swin-T, ConvNeXt-T, MLP-Mixer-B/16, and ViT-S, we find that the ratio of the largest eigenvalue to the mean eigenvalue ($\lambda_{\max}/\bar{\lambda}$) varies widely across architectures—from as low as $8\times$ (ViT-S) to as high as $31\times$ (ConvNeXt-T)—indicating that teacher feature energy is concentrated along a small number of directions in a highly non-uniform fashion. This *distributional heterogeneity* acts as a confounding factor that prevents students from uniformly absorbing the alignment signal provided by the teacher.

Existing cross-architecture KD methods address the architectural gap through various strategies. OFA (Hao et al., 2023) maps teacher and student features into a common, architecture-agnostic logit space via per-stage, architecture-specific projectors. ScaleKD (Fan et al., 2024) handles three types of heterogeneity with dedicated modules. FBT (Li et al., 2025) and PAT (Lin et al., 2025) introduce bridge models or prompt tuning to close the architectural gap. RSD (Zhang et al., 2025) performs

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.

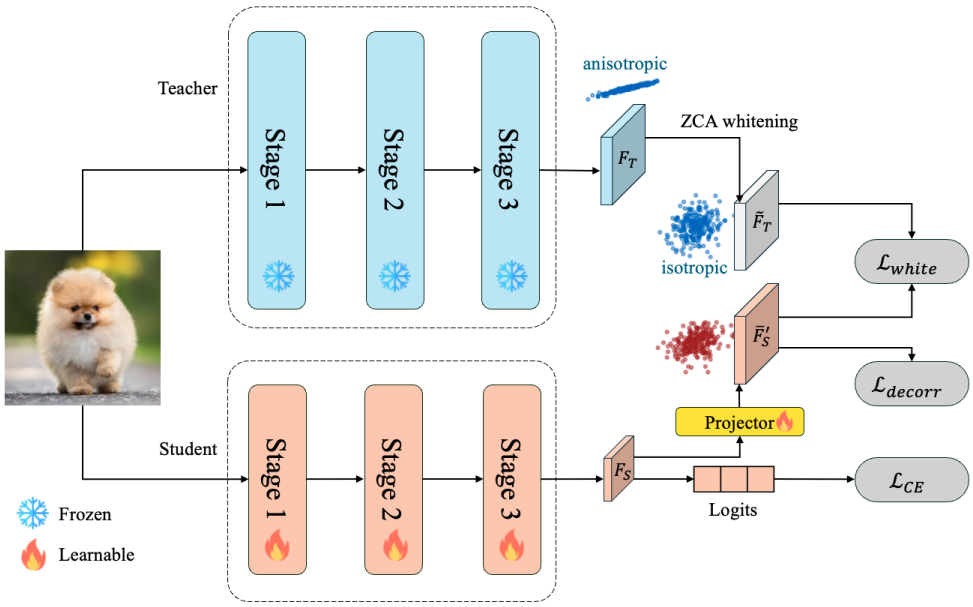


Figure 1: Overview of WFD. The frozen teacher extracts anisotropic features $\mathbf{F}_T$, which are transformed into isotropic features $\tilde{\mathbf{F}}_T$ via ZCA whitening. Student features $\mathbf{F}_S$ are mapped to the teacher space through a projector, yielding $\bar{\mathbf{F}}'_S$, and regularized by the self-decorrelation loss $\mathcal{L}_{\text{decorr}}$ applied to the student side only, making it *asymmetric*. Whitened teacher features and projected student features are then aligned via $\mathcal{L}_{\text{white}}$, and the full objective $\mathcal{L}_{\text{total}}$ combines cross-entropy with both distillation terms.

architecture-invariant alignment by encouraging the cross-correlation matrix between teacher and student features to approximate the identity matrix. While each of these methods achieves meaningful performance gains, none explicitly normalizes the distributional heterogeneity of teacher features before computing the alignment objective.

In this paper, we propose WFD, which is built on the key insight that normalizing the distribution of teacher features yields a more consistent and reliable alignment signal for the student. Specifically, WFD comprises two complementary components. First, we apply *ZCA whitening* (Bell & Sejnowski, 1997) to teacher features to remove distributional heterogeneity and flatten their eigenvalue spectra toward an isotropic representation. Second, we apply *asymmetric student decorrelation*. The two components act on opposite sides of the same cross-correlation matrix: whitening removes redundancy from the teacher by construction, the decorrelation term from the student by penalty. They contribute largely independently, and combining them yields a larger gain than either component alone.

Evaluated on CIFAR-100 across a broad range of heterogeneous teacher–student architecture combinations spanning Transformers, CNNs, and MLPs, WFD consistently outperforms existing methods including RSD (Zhang et al., 2025). On the Swin-T $\rightarrow$ ResNet-18 pair, WFD achieves 84.85%, a gain of $+0.93$ percentage points over RSD.

Contributions. We show that the penultimate representations of CNN, Transformer, and MLP teachers differ systematically in their eigenvalue spectra, with anisotropy ranging from $8\times$ to $31\times$. This heterogeneity distorts identity-target alignment. Whitening removes it, leaving a residual governed by the batch size and the feature dimension rather than by the teacher. That residual is not specific to our method: any batch estimate of a D-dimensional covariance with $B < D$ is rank-deficient in the same way. Building on this, we propose WFD, which adds no learnable parameters on the teacher path. Experiments on CIFAR-100 across 12 heterogeneous teacher–student pairs demonstrate consistent improvements over all baselines, with an average gain of $+11.82$ pp over the student baseline and $+1.14$ pp over the previous state-of-the-art.

2 Related Work

Knowledge Distillation. Knowledge distillation was introduced by Hinton et al. (2015), who proposed transferring knowledge via soft label matching between teacher and student output distributions. FitNet (Romero et al., 2015) introduced feature-based distillation by matching intermediate feature maps through a regression loss. AT (Zagoruyko & Komodakis, 2017) proposed attention map alignment, and CRD (Tian et al., 2020) maximizes mutual information between teacher and student representations via contrastive learning. Relational methods transfer structure across samples rather than individual activations: RKD (Park et al., 2019) matches pairwise distances and angles among samples, while CC (Peng et al., 2019) preserves the inter-sample correlation congruence of the teacher. DKD (Zhao et al., 2022) decouples the distillation loss into target and non-target class components for more precise supervision. DIST (Huang et al., 2022) relaxes exact logit matching by preserving inter-class and intra-class relations in the teacher's predictions. SimKD (Chen et al., 2022) directly reuses teacher classifier weights to improve knowledge transfer efficiency. While these methods perform well when teacher and student share the same architecture family, they degrade substantially in cross-architecture settings (Hao et al., 2023; Zhang et al., 2025) due to structural heterogeneity between representation spaces.

Cross-Architecture KD. Cross-architecture KD has attracted increasing attention as diverse architecture families—Transformers, CNNs, and MLPs—achieve complementary strengths. OFA (Hao et al., 2023) projects teacher and student features into a common, architecture-agnostic logit space through architecture-specific per-stage projectors, mitigating the dimension mismatch. RSD (Zhang et al., 2025) trains the cross-correlation matrix between batch-normalized teacher and student features to approximate the identity matrix, enabling architecture-invariant alignment without handcrafted correspondence. ScaleKD (Fan et al., 2024) addresses three types of heterogeneity—feature computation paradigm, model scale, and knowledge density—with dedicated modules. FBT (Li et al., 2025) fuses teacher and student module functions (convolution, attention, MLP) into an intermediate bridge representation before knowledge transfer. PAT (Lin et al., 2025) inserts prompt-tuning blocks into the teacher to adapt its features towards the student's architecture. HCD (Xu et al., 2026) combines a complementary feature mapper with a sub-logit orthogonality loss to reduce cross-architecture gaps. None of the above explicitly addresses the distributional anisotropy of teacher features as a source of misalignment.

Feature Whitening and Teacher Refinement. ZCA whitening (Bell & Sejnowski, 1997) produces an isotropic distribution and, among all whitening transforms, uniquely minimizes the expected squared distance to the original features (Kessy et al., 2018). DBN (Huang et al., 2018) integrates whitening into batch normalization; IterNorm (Huang et al., 2019) proposes Newton-Schulz iterations for efficient whitening. W-MSE (Ermolov et al., 2021) employs a whitened MSE objective in self-supervised learning. Soft-ZCA (Diera et al., 2025) controls whitening degree via eigenvalue regularization. VICReg (Bardes et al., 2022) decorrelates embedding pairs via a covariance term while separately preserving per-dimension variance to prevent collapse. SiNGER (Yu et al., 2026) suppresses high-norm artifacts in ViT teacher features via nullspace-guided perturbation. VkD (Miles et al., 2024) combines orthogonal projections with task-specific normalization. WhitenMTD (Ma et al., 2024) applies PCA whitening to unify multi-teacher distributions in retrieval. TCS (Zhou et al., 2025) defines a PCA coordinate system for student projection. Our work is the first to apply ZCA whitening to teacher feature preprocessing specifically for cross-architecture KD.

3 Method

3.1 Preliminaries

Given a mini-batch of B samples, let $\mathbf{F}_T \in \mathbb{R}^{B\times D_T}$ and $\mathbf{F}_S \in \mathbb{R}^{B\times D_S}$ denote the final embeddings of the teacher and student. Student features are mapped to the teacher space via a projector g and column-wise standardized, yielding $\bar{\mathbf{F}}'_S \in \mathbb{R}^{B\times D_T}$. The projector is defined as:

$$g(\mathbf{F}_S) = \mathrm{GELU}\big(\mathrm{BN}(\mathbf{F}_S\mathbf{W}_1^\top)\big)\mathbf{W}_2^\top, \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\gamma D_S \times D_S}$, $\mathbf{W}_2 \in \mathbb{R}^{D_T \times \gamma D_S}$, and γ is the projector expansion ratio. Throughout, $\bar{(\cdot)}$ denotes column-wise batch standardization:

$$\bar{\mathbf{f}}_d = \frac{\mathbf{f}_d - \mu_d}{\sigma_d + 10^{-5}}, \tag{2}$$

where $\mathbf{f}_d$ is the d-th column of a feature matrix and μ_d, σ_d are its batch mean and standard deviation. We write $\bar{\mathbf{F}}'_S = \overline{g(\mathbf{F}_S)}$ to denote the standardized projected student features.

RSD (Zhang et al., 2025) constructs a Pearson cross-correlation matrix $\mathbf{P} \in \mathbb{R}^{D_T \times D_T}$ between teacher and student feature dimensions across the batch:

$$\mathbf{P} = \frac{1}{B} \bar{\mathbf{F}}_T^\top \bar{\mathbf{F}}'_S, \tag{3}$$

where $P_{dd'}$ is the Pearson correlation between the d-th and d'-th projected feature dimensions across the batch; diagonal entries ($d = d'$) measure per-dimension alignment and off-diagonal entries ($d \neq d'$) measure spurious cross-dimensional correlations. All three losses in this paper share a common redundancy suppression objective rooted in the redundancy-reduction principle of Barlow Twins (Zbontar et al., 2021) and adapted for cross-architecture KD by RSD (Zhang et al., 2025):

$$\mathcal{L}_{\text{bt}}(\mathbf{M}) = \underbrace{\sum_d (M_{dd} - 1)^2}_{\text{invariance term}} + \kappa \underbrace{\sum_{d \neq d'} M_{dd'}^2}_{\text{redundancy-reduction term}} . \tag{4}$$

RSD minimizes $\mathcal{L}_{\text{bt}}(\mathbf{P})$ to maximize cross-architecture alignment.

The anisotropy problem. RSD implicitly assumes the teacher self-correlation $\mathbf{C}_T = (1/B)\bar{\mathbf{F}}_T^\top \bar{\mathbf{F}}_T \approx \mathbf{I}$. In practice, however, the anisotropy ratio $\lambda_{\max}/\bar{\lambda}$ of $\mathbf{C}_T$ varies widely across architecture families, as we analyze in Sec. 3.2. When $\mathbf{C}_T \neq \mathbf{I}$, the identity target of Eq. (4) is not well-matched to the teacher's actual feature geometry: dimensions along high-variance directions dominate the diagonal loss, while low-variance dimensions contribute little. As a result, $\mathbf{P}$ reflects the teacher's eigenvalue imbalance rather than the intended semantic alignment between teacher and student representations.

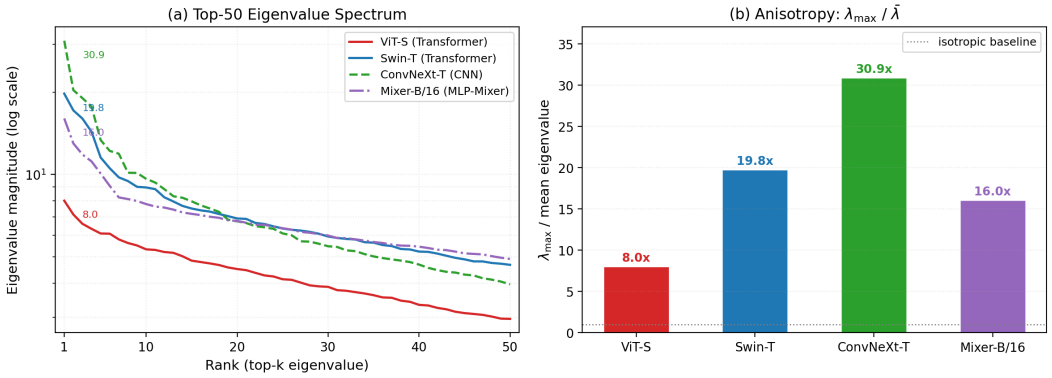


Figure 2: Teacher feature anisotropy analysis. **(a)** Top-50 eigenvalue spectra (log scale) of the self-correlation matrix $\mathbf{C}_T$ for four teacher architectures; isotropic features would show flat spectra. **(b)** Anisotropy index $\lambda_{\max}/\bar{\lambda}$: ConvNeXt-T $= 30.9\times$, Swin-T $= 19.8\times$, Mixer-B/16 $= 16.0\times$, ViT-S $= 8.0\times$.

3.2 Teacher Feature Distribution Analysis

For each teacher, we extract penultimate-layer features from the CIFAR-100 test set (10,000 images) using the same checkpoints as in our main experiments, and compute the self-correlation matrix $\mathbf{C}_T = \frac{1}{N}\bar{\mathbf{F}}_T^\top \bar{\mathbf{F}}_T$ over all $N = 10{,}000$ test samples (rather than a single mini-batch) to obtain a stable estimate of the teacher's feature distribution. Figure 2 visualizes the eigenvalue spectra of $\mathbf{C}_T$ for all four teacher architectures evaluated in our study. The anisotropy index $\lambda_{\max}/\bar{\lambda}$ reveals

Algorithm 1 PyTorch-style pseudocode for WFD Loss

```
# feat_s, feat_t: student/teacher feats [B,D]
feat_s_n = Normalize(feat_s)
# ZCA whitening of teacher (no_grad)
feat_n = Normalize(feat_t.float())
cov = feat_n.T @ feat_n / B
U, S, _ = torch.linalg.svd(cov, full_matrices=False)
W = U @ diag(1/(S+eps).sqrt()) @ U.T
feat_t_w = feat_n @ W
feat_t_w /= feat_t_w.std(dim=0) + 1e-5
idt = torch.eye(D)
rcc = feat_t_w.T @ feat_s_n / B
loss_align = (rcc - idt).pow(2)
loss_align[(1-idt).bool()] *= kappa
# Asymmetric self-decorrelation of student
rcc_s = feat_s_n.T @ feat_s_n / B
loss_decorr = (rcc_s - idt).pow(2)
loss_decorr[(1-idt).bool()] *= kappa
loss = loss_align.sum() + beta*loss_decorr.sum()
return loss
```

stark inter-architecture differences: ConvNeXt-T $= 30.9\times$, Swin-T $= 19.8\times$, Mixer-B/16 $= 16.0\times$, ViT-S $= 8.0\times$.

All four architectures are anisotropic, but the degree varies widely. Isotropic features would produce a flat spectrum; instead, all four teachers exhibit a rapid eigenvalue decay, indicating that most feature energy is concentrated in a small number of principal directions. This is especially pronounced for ConvNeXt-T, where the top eigenvalue is $31\times$ the mean—reflecting the inductive bias of local convolution towards low-dimensional, spatially structured representations. In contrast, ViT-S features, shaped by global self-attention, are comparatively more isotropic, yet still far from uniform. These indices are computed over the full test set; single-batch estimates are larger (Appendix B).

Anisotropy is not specific to any single architecture paradigm but is a universal property of deep feature representations, and its degree can differ by up to $4\times$ across teacher families. When a fixed identity-target objective such as RSD is applied without accounting for this variation, the alignment signal is distorted by the teacher's eigenvalue imbalance, producing inconsistent improvements across teacher–student pairs. This motivates our use of ZCA whitening to normalize each teacher's feature distribution before computing the alignment signal.

3.3 Teacher Feature Whitening

We apply ZCA whitening to teacher features to directly address the anisotropy problem. Among isotropic transformations, ZCA whitening is uniquely suited to our setting because it produces an isotropic distribution while minimizing the expected squared distance to the original features among all whitening transforms (Kessy et al., 2018). This means the whitened features $\tilde{\mathbf{F}}_T$ retain the semantic geometry of the original teacher representation as closely as possible, merely rescaling the principal directions to equalize their variance.

Whitening transform. Given the eigendecomposition $\mathbf{C}_T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ contains the eigenvectors and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{D_T})$ the eigenvalues, the whitening matrix is:

$$\mathbf{W} = \mathbf{U}\,(\mathbf{\Lambda} + \varepsilon\mathbf{I})^{-1/2}\,\mathbf{U}^\top, \tag{5}$$

where $\varepsilon = 10^{-2}$ is not a numerical-stability constant (Eqs. (2) and (6) use 10^{-5}) but bounds the amplification of low-variance directions at $1/\sqrt{\varepsilon}$, giving a partial rather than complete whitening in the sense of Soft-ZCA (Diera et al., 2025). Because $\mathbf{W}$ is estimated from a mini-batch of size B rather than the full dataset, the estimate of $\mathbf{C}_T$ is rank-deficient whenever $B < D_T$ and residual column-wise scaling may remain (Appendix B); we therefore define the final whitened features as:

$$\tilde{\mathbf{F}}_T = \frac{\bar{\mathbf{F}}_T\mathbf{W}}{\mathrm{std}_{\mathrm{col}}(\bar{\mathbf{F}}_T\mathbf{W}) + 10^{-5}}, \tag{6}$$

which ensures each dimension of $\tilde{\mathbf{F}}_T$ has unit variance. Since the teacher is frozen during training, all whitening computations are performed inside a `no_grad` context and introduce no additional learnable parameters and no gradient overhead; Appendix D quantifies the remaining compute cost.

Alignment loss. We define the whitened cross-correlation matrix between teacher and student as:

$$\tilde{\mathbf{C}} = \frac{1}{B}\tilde{\mathbf{F}}_T^\top \bar{\mathbf{F}}'_S, \tag{7}$$

and minimize:

$$\mathcal{L}_{\text{white}} = \mathcal{L}_{\text{bt}}(\tilde{\mathbf{C}}). \tag{8}$$

Eq. (6) fixes the diagonal of $(1/B)\tilde{\mathbf{F}}_T^\top \tilde{\mathbf{F}}_T$ to unity and whitening suppresses its off-diagonal entries, so the identity target is approached directly rather than assumed. Any identity-target objective computed on unwhitened teacher features instead inherits the teacher's spectral imbalance.

3.4 Asymmetric Student Decorrelation

While teacher whitening makes the alignment target substantially more isotropic, the projected student features $\bar{\mathbf{F}}'_S$ can still develop internal correlations during training. When the projector is optimized purely to minimize $\mathcal{L}_{\text{white}}$, different output dimensions may learn redundant directions, reducing the effective capacity of the student representation. This phenomenon is related to dimensional collapse in contrastive self-supervised learning (Jing et al., 2022), where representation dimensions become correlated and fail to span the full feature space; decorrelation objectives such as Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022) are designed to counteract it.

To address this, we introduce an asymmetric student decorrelation loss. Defining the student self-correlation matrix as:

$$\mathbf{C}_S = \frac{1}{B}\bar{\mathbf{F}}'^\top_S \bar{\mathbf{F}}'_S \in \mathbb{R}^{D_T \times D_T}, \tag{9}$$

we minimize:

$$\mathcal{L}_{\text{decorr}} = \mathcal{L}_{\text{bt}}(\mathbf{C}_S). \tag{10}$$

As in Eq. (4), this loss combines an invariance term, which drives each dimension of $\bar{\mathbf{F}}'_S$ toward unit variance, with a redundancy-reduction term that suppresses cross-dimensional correlations; together they push $\mathbf{C}_S$ toward the identity matrix. We term this *asymmetric* decorrelation because it is applied only to the student side: the teacher is frozen and whitened without gradients, so a teacher-side decorrelation term would have no parameters to act on. The student decorrelation loss thus acts as a complementary regularizer that diversifies the student projector output independently of the cross-architecture alignment objective.

3.5 Full Objective

The total training loss combines cross-entropy with both distillation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda\big(\mathcal{L}_{\text{white}} + \beta\,\mathcal{L}_{\text{decorr}}\big), \tag{11}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, λ controls the overall distillation strength, and β balances the student decorrelation term relative to the whitening alignment term. The loss hyperparameters ($\lambda = 0.1$, $\beta = 0.1$, $\kappa = 0.01$) are tuned on a single teacher–student pair (Swin-T $\rightarrow$ ResNet-18) and held fixed across all 12 CIFAR-100 pairs, demonstrating that WFD does not require per-pair tuning. Besides the student, only the projector g is trained; the teacher is frozen, and $\mathbf{W}$ is computed on the fly from each mini-batch's teacher features and receives no gradient.

Algorithm 1 summarizes the complete training procedure.

4 Experiments

We evaluate WFD on CIFAR-100 across 12 heterogeneous teacher–student pairs and compare against eleven baselines spanning logit-based and feature-based methods.

4.1 Experimental Setup

Dataset. All experiments use CIFAR-100 (Krizhevsky, 2009) (100 classes, 50K train / 10K test images). Images are resized to 224×224 to match the teacher architectures' expected input resolution.

Table 1: Top-1 accuracy (%) on CIFAR-100. **Bold** indicates the best result and underline the second-best.

		Scratch		Logit-based				Feature-based							Ours
Teacher	Student	T.	S.	KD	DKD	DIST	OFA	FitNet	CC	RKD	CRD	FBT	PAT	RSD	WFD
CNN students															
Swin-T	ResNet-18	89.26	74.01	78.74	80.26	77.75	80.54	78.87	74.19	74.11	77.63	81.61	81.22	83.92	**84.85**
ViT-S	ResNet-18	92.44	74.01	77.26	78.10	76.49	80.15	77.71	74.26	73.72	76.60	81.93	80.11	81.50	**83.38**
Mixer-B/16	ResNet-18	87.62	74.01	77.79	78.67	76.36	79.39	77.15	74.26	73.75	76.42	81.90	80.07	81.85	**82.49**
Swin-T	MobileNetV2	89.26	73.68	74.68	71.07	72.89	80.98	74.28	71.19	69.00	79.80	81.28	78.78	83.68	**84.16**
ViT-S	MobileNetV2	92.44	73.68	72.77	69.80	72.54	78.45	73.54	70.67	68.46	78.14	82.10	78.87	81.68	**83.38**
Mixer-B/16	MobileNetV2	87.62	73.68	73.33	70.20	73.26	78.78	73.78	70.73	68.95	78.15	80.83	78.62	81.74	**82.09**
Transformer students															
ConvNeXt-T	DeiT-T	88.42	68.00	72.99	74.60	73.55	75.76	60.78	68.01	69.79	65.94	79.57	79.59	82.46	**82.89**
Mixer-B/16	DeiT-T	87.62	68.00	71.36	73.44	71.67	73.90	71.05	68.13	69.89	65.35	74.40	74.66	78.50	**79.86**
ConvNeXt-T	Swin-Pico	88.42	72.63	76.44	76.80	76.41	78.32	24.06	72.63	71.73	67.09	80.73	80.74	82.21	**83.34**
Mixer-B/16	Swin-Pico	87.62	72.63	75.93	76.39	75.85	78.93	75.20	73.32	70.82	67.03	78.44	78.44	81.28	**82.46**
MLP students															
ConvNeXt-T	ResMLP-12	88.42	66.56	72.25	73.22	71.93	81.22	45.47	67.70	65.82	63.35	78.03	83.50	84.21	**85.58**
Swin-T	ResMLP-12	89.26	66.56	71.89	72.82	11.05	80.63	63.12	68.37	64.66	61.72	77.20	80.94	82.67	**84.85**
Avg. Gain over Scratch Student		89.03	71.45	+3.17	+3.16	−2.31	+7.47	−5.20	−0.33	−1.40	−0.02	+8.38	+8.17	+10.69	**+11.82**

Models. Teachers: Swin-T (Liu et al., 2021), ConvNeXt-T (Liu et al., 2022), ViT-S (Dosovitskiy et al., 2021), Mixer-B/16 (Tolstikhin et al., 2021). Students: ResNet-18 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), ResMLP-12 (Touvron et al., 2022), DeiT-T (Touvron et al., 2021), Swin-Pico (Liu et al., 2021). All 12 pairs in Table 1 cover Transformer, CNN, and MLP teacher–student combinations. Teachers are pretrained on CIFAR-100 and frozen.

Training. CNN students (ResNet-18, MobileNetV2): SGD (momentum $= 0.9$, wd $= 2 \times 10^{-3}$), cosine annealing, 300 epochs. Transformer/MLP students (ResMLP-12, DeiT-T, Swin-Pico): AdamW, 300 epochs. Hyperparameters: $\lambda = 0.1$, $\beta = 0.1$, $\kappa = 0.01$, $\gamma = 2$, tuned on Swin-T $\rightarrow$ ResNet-18 and fixed across all 12 pairs. Full training configurations are given in Appendix A.

4.2 Main Results

Table 1 compares WFD against eleven baselines across all 12 heterogeneous teacher–student pairs on CIFAR-100. WFD achieves the best accuracy in every single row, surpassing both logit-based methods (KD, DKD, DIST, OFA) and feature-based methods (FitNet, CC, RKD, CRD, FBT, PAT, RSD). WFD achieves an average accuracy of 83.28%, a gain of $+11.82$ pp over the student trained from scratch (71.45%), and outperforms the previous best method (RSD, 82.14%) by $+1.14$ pp.

CNN students. On the six CNN student pairs, WFD consistently outperforms RSD by $+0.35$ to $+1.88$ pp. The largest gains appear on ViT-S as teacher ($+1.88$ pp on ResNet-18, $+1.70$ pp on MobileNetV2); ViT-S also has the lowest anisotropy ($8\times$) and, uniquely, a narrower feature dimension ($D = 384$ vs. $D = 768$) among our teachers, though this trend does not hold monotonically across all teachers (see Limitations).

Transformer and MLP students. On Transformer and MLP students, gains reach up to $+2.18$ pp (Swin-T $\rightarrow$ ResMLP-12). MLP students (ResMLP-12) benefit most, likely because the representational gap to CNN/Transformer teachers is larger. WFD also outperforms PAT, which adapts the teacher by prompt tuning, by $+2.08$ pp on ConvNeXt-T $\rightarrow$ ResMLP-12, suggesting that teacher-side normalization is more effective than teacher-side adaptation here.

4.3 ImageNet-1K Results

To validate WFD at a larger scale, we evaluate on ImageNet-1K following RSD's benchmark, which at this scale uses a different pair set (DeiT-T in place of ViT-S, Swin-N in place of Swin-Pico); we report its 12 pairs excluding the three with a ResNet-50 teacher. Table 2 compares WFD against RSD (Zhang et al., 2025).

WFD shows the most consistent gains on MLP students, improving over RSD by $+0.20$ pp on both ConvNeXt-T and Swin-T teachers. On CNN students, results are more mixed: WFD wins on two of

Table 2: Top-1 accuracy (%) on ImageNet-1K. **Bold** indicates the best result and underline the second-best. PAT reports only 4 of the 12 pairs; '—' denotes results not reported and its average gain is omitted.

		Scratch		Logit-based				Feature-based							Ours
Teacher	Student	T.	S.	KD	DKD	DIST	OFA	FitNet	CC	RKD	CRD	FBT	PAT	RSD	WFD
CNN students															
DeiT-T	ResNet-18	72.19	69.75	70.22	69.39	70.64	71.34	70.44	69.77	69.47	69.25	71.22	—	71.70	**71.87**
Swin-T	ResNet-18	81.35	69.75	71.14	71.10	70.91	71.85	71.18	70.07	68.89	69.09	**72.21**	71.54	72.13	72.17
Mixer-B/16	ResNet-18	76.58	69.75	70.89	69.89	70.66	71.38	70.78	70.05	69.46	68.40	**71.44**	—	71.41	71.39
DeiT-T	MobileNetV2	72.19	68.87	70.87	70.14	71.08	71.39	70.95	70.69	69.72	69.60	71.78	—	**72.18**	72.16
Swin-T	MobileNetV2	81.35	68.87	72.05	71.71	71.76	72.32	71.75	70.69	67.52	69.58	**72.54**	—	72.36	72.12
Mixer-B/16	MobileNetV2	76.55	68.87	71.92	70.93	71.74	72.12	71.59	70.79	69.86	68.89	**72.31**	72.22	71.90	71.74
Transformer students															
ConvNeXt-T	DeiT-T	82.05	72.17	74.00	73.95	74.07	74.41	70.45	73.12	71.47	69.18	**75.26**	74.44	74.46	74.77
Mixer-B/16	DeiT-T	76.46	72.17	74.16	72.82	74.22	74.46	74.38	72.82	72.24	68.23	**75.00**	—	74.26	73.57
ConvNeXt-T	Swin-N	82.05	75.53	77.15	77.00	77.25	77.50	74.81	75.79	75.48	74.15	**77.73**	—	77.70	77.70
Mixer-B/16	Swin-N	76.55	75.53	76.26	75.03	76.54	76.63	76.17	75.81	75.52	73.38	**76.87**	—	76.55	76.02
MLP students															
ConvNeXt-T	ResMLP-12	82.05	76.65	76.84	77.23	77.24	77.53	74.69	75.79	75.28	73.57	77.33	—	78.41	**78.61**
Swin-T	ResMLP-12	81.35	76.65	76.67	76.99	77.25	77.31	76.48	76.15	75.10	73.40	77.42	77.59	77.61	**77.81**
Avg. Gain over Scratch Student															
		78.39	72.05	+1.47	+0.97	+1.57	+1.97	+0.76	+0.58	−0.38	−1.49	**+2.21**	—	+2.18	+2.11

six pairs, is within 0.02 pp on two, and loses on two (up to -0.24 pp on Swin-T $\rightarrow$ MobileNetV2), for an average change of -0.04 pp. Among Transformer students WFD wins on ConvNeXt-T $\rightarrow$ DeiT-T, is essentially tied on ConvNeXt-T $\rightarrow$ Swin-N (77.70% vs. RSD's 77.70%), and trails on the two Mixer-B/16 pairs; averaged over all 12 pairs it is slightly behind RSD ($+2.11$ vs. $+2.18$ pp gain over scratch). PAT (Lin et al., 2025) reports ImageNet-1K results on only 4 of these 12 pairs, citing computational cost; WFD outperforms PAT on 3 of them, trailing only on Mixer-B/16 $\rightarrow$ MobileNetV2. FBT (Li et al., 2025) reports full results on all 12 pairs and is stronger overall at this scale, outperforming both RSD and WFD on 8 of 12 pairs, most notably on Transformer-student pairs; FBT achieves this with a heavier fusion-based bridge model, whereas WFD uses only a lightweight two-layer projector (see Limitations). Overall, WFD remains competitive with RSD at ImageNet-1K scale, though the gains are smaller and less consistent than on CIFAR-100.

4.4 Ablation Study

Table 3 isolates the contribution of each component on Swin-T $\rightarrow$ ResNet-18, using the RSD baseline (83.92%) as the starting point.

Teacher whitening alone ($+0.67$ pp) provides a larger individual gain than student decorrelation alone ($+0.29$ pp), suggesting that normalizing the alignment target is the more critical of the two interventions. Combining both components yields $+0.93$ pp, which is near-additive ($+0.67 + 0.29 = +0.96$ pp), indicating that the two components address largely orthogonal sources of suboptimality. Teacher whitening corrects the biased alignment target, while student decorrelation reduces representational redundancy in the projector.

Between the first two rows only the teacher's feature distribution changes, so the $+0.67$ pp gain is attributable to whitening itself. Appendix C adds a batch-size sweep: *for a fixed teacher*, the margin over RSD widens as B grows and the whitened target approaches $\mathbf{I}$, while RSD improves far less over the same range.

Table 3: Component ablation on Swin-T $\rightarrow$ ResNet-18 (CIFAR-100 Top-1, %).

White	Decorr	Accuracy	Δ
		83.92	—
✓		84.59	+0.67
	✓	84.21	+0.29
✓	✓	**84.85**	**+0.93**

4.5 Hyperparameter Analysis

Student decorrelation weight β. Table 4 reports sensitivity to β on Swin-T $\rightarrow$ ResNet-18. The optimal value is $\beta = 0.10$; all values in $[0.01, 0.20]$ exceed 84.5%, demonstrating robustness. $\beta \geq 0.50$ causes decorrelation to dominate, degrading performance.

Distillation loss weight λ. Table 5 shows sensitivity to λ. The optimal is $\lambda = 0.10$; deviating in either direction leads to noticeable performance drops.

Table 4: β sensitivity on Swin-T $\rightarrow$ ResNet-18 ($\lambda = 0.1$, %).

β	0.01	0.05	0.10	0.20	0.50	1.00
Acc.	84.58	84.66	**84.85**	84.57	83.93	83.83

Table 5: λ sensitivity on the same pair ($\beta = 0.1$, %).

λ	0.05	0.10	0.20
Acc.	83.75	**84.85**	83.86

5 Conclusion

We presented WFD, a cross-architecture knowledge distillation method that addresses the distributional heterogeneity of teacher features via two complementary components: ZCA whitening of frozen teacher features and asymmetric decorrelation of student projector outputs.

Our core insight is that existing cross-architecture KD methods do not address the anisotropy of teacher features, despite all major architecture families exhibiting anisotropy ratios ranging from $8\times$ to $31\times$. Methods that implicitly rely on an isotropic alignment target, such as RSD's identity-matching cross-correlation loss, are directly affected by this violated assumption. By applying ZCA whitening to teacher features, WFD transforms this untested assumption into an explicitly enforced alignment target, at no additional parameter or gradient cost. Asymmetric student decorrelation further improves results by suppressing redundancy in the student representation, and the two components interact near-additively.

Experiments on CIFAR-100 across 12 heterogeneous teacher–student pairs spanning Transformer, CNN, and MLP architectures demonstrate that WFD achieves the best accuracy in all 12 settings, with an average gain of $+11.82$ pp over the student trained from scratch and $+1.14$ pp over the previous state-of-the-art. On ImageNet-1K, WFD remains competitive with RSD across the evaluated teacher–student pairs. These results confirm that explicitly normalizing teacher feature distributions is an effective and generalizable strategy for cross-architecture knowledge distillation.

Limitations and future work. WFD is not tied to the penultimate layer—the same transform applies at any stage—but we restrict it there to isolate the effect of whitening, leaving per-stage extension to future work. Anisotropy predicts gain magnitude only *within* a teacher, not across teachers: the attainable floor is set by B and D_T (Appendix B), so ViT-S's narrower D_T confounds cross-teacher comparison. WFD leads the unwhitened baseline at every batch size we test, including the smallest ($B = 128$); larger batches lower this floor, with the largest margins at the larger batch sizes (Appendix C). Our ImageNet-1K study is also limited in scope: hyperparameters at this scale were selected per student family from a small set of candidates guided by CIFAR-100 proxy runs, rather than tuned per pair. The remaining gap to FBT (Li et al., 2025) at this scale is therefore partly tuning and partly scope: FBT distills all four stages through a fused bridge model, whereas WFD aligns a single embedding. Combining whitening with such fusion strategies is a natural next step.

Reproducibility Statement

Full training configurations for CIFAR-100 and ImageNet-1K—optimizer, batch size, schedule, and the method hyperparameters λ, β, κ, γ and ε—are given in Appendix A. Because $\mathbf{W}$ is estimated from the batch covariance, the batch size affects how closely the alignment target approaches the identity; this dependence is quantified in Appendix B. Source code, configuration files, and the teacher checkpoints used are released at https://anonymous.4open.science/r/WFD-7D80.

AI USE STATEMENT

We used generative AI tools to assist with method ideation, implementation and experiments, literature retrieval, and writing and editing the manuscript. We did not use them to generate synthetic data or to produce or verify mathematical proofs, as neither is part of this work. The authors verified all outputs and take full responsibility for the content of this work.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. arXiv:2105.04906.

Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, 2022.

Andor Diera, Lukas Galke, and Ansgar Scherp. Isotropy matters: Soft-ZCA whitening of embeddings for semantic code search. In *ESANN*, 2025. arXiv:2411.17538.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021. arXiv:2007.06346.

Jiawei Fan, Chao Li, Xiaolong Liu, and Anbang Yao. ScaleKD: Strong vision transformers could be excellent teachers. In *NeurIPS*, 2024.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *NeurIPS*, 2023. arXiv:2310.19444.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *CVPR*, 2018.

Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening of neural networks. In *CVPR*, 2019.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022. arXiv:2110.09348.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Guopeng Li, Qiang Wang, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Fuse before transfer: Knowledge fusion for heterogeneous distillation. In *ICCV*, 2025. arXiv:2410.12342.

Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Perspective-aware teaching: Adapting knowledge for heterogeneous distillation. In *ICCV*, 2025. arXiv:2501.08885.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022.

Zhe Ma, Jianfeng Dong, Shouling Ji, Zhenguang Liu, Xuhong Zhang, Zonghui Wang, Sifeng He, Feng Qian, Xiaobo Zhang, and Lei Yang. Let all be whitened: Multi-teacher distillation for efficient visual retrieval. In *AAAI*, 2024. arXiv:2312.09716.

Roy Miles, Ismail Elezi, and Jiankang Deng. VkD: Improving knowledge distillation using orthogonal projections. In *CVPR*, 2024. arXiv:2403.06213.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, 2019.

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

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Liuchi Xu, Hao Zheng, Lu Wang, Lisheng Xu, and Jun Cheng. Heterogeneous complementary distillation. In *AAAI*, 2026. arXiv:2511.10942.

Geunhyeok Yu, Sunjae Jeong, Yoonyoung Choi, Jaeseung Kim, and Hyoseok Hwang. SiNGER: A clearer voice distills vision transformers further. In *ICLR*, 2026. arXiv:2509.20986.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Weijia Zhang, Yuehao Liu, Wu Ran, and Chao Ma. Cross-architecture distillation made simple with redundancy suppression. In *ICCV*, 2025. arXiv:2507.21844.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022.

Junjie Zhou, Ke Zhu, and Jianxin Wu. All you need in knowledge distillation is a tailored coordinate system. In *AAAI*, 2025. arXiv:2412.09388.

A Implementation Details

We follow RSD (Zhang et al., 2025) and OFA (Hao et al., 2023) for dataset splits, teacher checkpoints, and the baseline numbers reported in Tables 1 and 2. Training configurations are our own and are listed in Table 6.

Unlike most training hyperparameters, the batch size B enters WFD directly: $\mathbf{W}$ is estimated from the batch covariance $\mathbf{C}_T$, so B relative to the teacher embedding dimension D_T governs how closely the whitened target approaches $\mathbf{I}$ (Appendix B). Batch size follows the student family rather than the teacher–student pair and was not tuned per pair.

Table 6: Training configurations on CIFAR-100.

	CNN students (ResNet-18, MobileNetV2)	Transformer / MLP students (DeiT-T, Swin-Pico, ResMLP-12)
optimizer	SGD ($m = 0.9$)	AdamW
learning rate	0.05	5×10^{-4}
weight decay	2×10^{-3}	0.05
batch size	256	128
GPUs	1	1
schedule / epochs	cosine / 300	
input resolution	224×224	
AMP / model EMA	yes / yes	
λ / β / κ / γ / ε	0.1 / 0.1 / 0.01 / 2 / 10^{-2}	

ImageNet-1K. Table 7 lists the ImageNet-1K configurations. Epoch budgets follow the benchmark protocol used by RSD (Zhang et al., 2025) and FBT (Li et al., 2025): CNN students are trained for 100 epochs and Transformer and MLP students for 300. The projector width γ and the off-diagonal weight κ were selected per student family from a small set of candidates; λ, β and ε are shared across all twelve pairs. Learning rates follow linear scaling from the CIFAR-100 setting. Because $\mathbf{W}$ is estimated from the batch, we state separately the effective batch entering the whitening statistics: the CNN runs use $B \geq 1024$, which exceeds $D_T = 768$, so the covariance estimate is full rank there, whereas the Transformer and MLP runs operate in the rank-deficient regime quantified in Appendix B. Most CNN runs compute these statistics per process, so the effective batch is the per-GPU value rather than the product of the per-GPU batch size and GPU count in Table 7. Whitening statistics are exchanged with a single all-gather of a $B \times D_T$ matrix (under 0.4 MB at $B = 128$, $D_T = 768$), which is negligible next to the activations of the forward pass. The MLP-student runs estimate $\mathbf{W}$ from per-process batches rather than the gathered batch and still improve over RSD by $+0.20$ pp on both teachers, indicating that the local approximation is not a limiting factor at these batch sizes. Settings not listed here follow the configuration files released with the code.

Table 7: Training configurations on ImageNet-1K. Two values in the CNN column refer to ResNet-18 and MobileNetV2 respectively.

	CNN students (ResNet-18, MobileNetV2)	Transformer students (DeiT-T, Swin-N)	MLP students (ResMLP-12)
optimizer	SGD ($m = 0.9$)	AdamW	AdamW
learning rate	1.0 / 0.8	5×10^{-4}	1×10^{-3}
weight decay	1×10^{-4}	0.05	0.05
batch size per GPU	1280 / 1024	128	342
GPUs	2	2	3
B for $\mathbf{W}$	1024–2048	256	342
schedule / epochs	step / 100	cosine / 300	cosine / 300
γ	4	6	4
κ	0.05	0.01	0.01
λ / β / ε	0.01 / 0.1 / 10^{-2}		
input resolution	224×224		

Code. Source code, configuration files, and the teacher checkpoints used are available at https://anonymous.4open.science/r/WFD-7D80.

B Whitening Quality as a Function of Batch Size

$\mathbf{W}$ is estimated from the batch covariance, so the quality of the resulting alignment target depends on B relative to D_T. We quantify this with the anisotropy index $\lambda_{\max}/\bar{\lambda}$ of $(1/B)\tilde{\mathbf{F}}_T^\top\tilde{\mathbf{F}}_T$: for a matrix with unit diagonal this index equals 1 if and only if the matrix is $\mathbf{I}$, so it measures directly how far the whitened target is from the identity. Features are the penultimate-layer activations of the CIFAR-100 test set and values are averaged over 20 random batches.

Table 8: Anisotropy index of the whitened teacher correlation $(1/B)\tilde{\mathbf{F}}_T^\top\tilde{\mathbf{F}}_T$ at several batch sizes. Bold entries are the batch sizes used in our CIFAR-100 experiments. The raw row is the unwhitened correlation at B=128.

	Swin-T	ConvNeXt-T	Mixer-B/16	ViT-S
D_T	768	768	768	384
raw (B=128)	32.33	37.85	25.41	13.88
$B = 128$	**7.20**	**6.88**	**7.32**	**3.41**
$B = 256$	**4.22**	**3.64**	**4.28**	**1.77**
$B = 512$	2.64	1.96	2.53	1.20
$B = 1024$	1.79	1.26	1.52	1.10

Whitening removes inter-architecture variation. Across the three 768-dimensional teachers the spread of the anisotropy index falls from 1.49× to 1.06× at B=128. The residual off-diagonal RMS is identical (0.081) for all three and matches the rank-induced floor $\sqrt{(D_T/r - 1)/(D_T - 1)}$ with $r = B - 1$. What remains after whitening is therefore governed by B and D_T rather than by the teacher architecture, which is precisely the heterogeneity removal WFD is designed to achieve. ViT-S sits lower because $D_T = 384$ gives it a lower floor.

A lower bound on the residual. The identity target cannot be reached exactly at these batch sizes. When $B < D_T$ the batch estimate of $\mathbf{C}_T$ has rank at most $B - 1$, and any positive semi-definite matrix with unit diagonal and rank r satisfies

$$\|\mathbf{C} - \mathbf{I}\|_F^2 \;\geq\; \frac{D_T^2}{r} - D_T, \tag{12}$$

by Cauchy–Schwarz applied to its eigenvalues, whose sum is fixed at D_T by the unit diagonal. For $D_T = 768$ and $B = 128$ this gives $\|\mathbf{C} - \mathbf{I}\|_F \geq 62.26$, and the measured value is 62.34: the whitening operates at the limit of what the batch rank permits. Together with the ε regularizer of Eq. (5), this keeps WFD in the partial-whitening regime discussed in Sec. 3.3 — enough spectral flattening to remove the inter-architecture variation of Sec. 3.2, without promoting low-variance directions to equal footing with informative ones.

C Batch Size and the Benefit of Whitening

Appendix B shows that the whitened target approaches $\mathbf{I}$ as B grows. If that is the mechanism by which whitening helps, then increasing B should benefit WFD more than it benefits a method that does not whiten. Table 9 tests this directly. All runs share identical hyperparameters ($\gamma = 2$, $\kappa = 0.01$, $\lambda = 0.1$, $\beta = 0.1$), a single GPU, 300 epochs, and a learning rate scaled linearly with B. Learning rate and batch size therefore vary together, and the two effects are not separated here; within each column, however, both methods see the same setting.

WFD is ahead at all twelve settings, and the margin is largest at the larger batch sizes: it rises from +0.13 to +0.98 for Swin-T and from +1.21 to +2.49 for ViT-S. The two methods also respond to B differently. Going from $B = 128$ to $B = 512$ improves RSD by +0.38, +0.47 and +0.67 pp on the three teachers, but improves WFD by +1.08, +1.75 and +1.16 pp. The ViT-S column is the clearest

Table 9: CIFAR-100 Top-1 (%) versus batch size with a ResNet-18 student.

	B=128	B=256	B=384	B=512
learning rate	0.05	0.10	0.15	0.20
Swin-T teacher ($D_T = 768$)				
RSD	83.06	83.62	83.40	83.44
WFD	**83.19**	**84.13**	**84.38**	**84.27**
Δ	+0.13	+0.51	+0.98	+0.83
ViT-S teacher ($D_T = 384$)				
RSD	80.35	80.74	80.90	80.82
WFD	**81.56**	**82.36**	**83.20**	**83.31**
Δ	+1.21	+1.62	+2.30	+2.49
Mixer-B/16 teacher ($D_T = 768$)				
RSD	80.52	81.20	80.98	81.19
WFD	**80.99**	**81.65**	**82.27**	**82.15**
Δ	+0.47	+0.45	+1.29	+0.96

case: RSD is essentially flat across a fourfold increase in batch size, whereas WFD gains $+1.75$ pp over the same range. This is what the whitening account predicts. A larger batch raises the rank of the covariance estimate and lowers the residual of Table 8, so the alignment target becomes more isotropic and the student receives a cleaner signal; a method that never whitens has no such target to sharpen. The batch size is therefore not a ceiling imposed on WFD but a lever that whitening makes usable.

The CIFAR-100 results of Table 1 use $B = 256$ for CNN students and $B = 128$ for Transformer and MLP students, so the advantage reported there does not depend on unusually large batches. The CNN runs there use the benchmark default of 0.05 at every batch size, so their absolute numbers are not directly comparable with this table.

D Computational Overhead

Whitening adds no learnable parameters, only one eigendecomposition per step. Table 10 isolates that cost, since all entries use the same pair, batch size, precision and projector.

Table 10: Training cost on Swin-T $\rightarrow$ ResNet-18, CIFAR-100, $B = 256$, one V100, AMP. Best of three 80-step runs after warm-up, each in a separate process; noise only adds time.

	s / step	peak memory	trainable params
KD	0.264	3389 MiB	11.23 M
RSD	0.270	4370 MiB	12.54 M
WFD (SVD)	0.397	4368 MiB	12.54 M
WFD (symmetric eigendecomposition)	0.282	4368 MiB	12.54 M

Peak memory is indistinguishable from RSD (within 0.1% for both decompositions). The time cost depends entirely on how the decomposition is computed. Our released implementation calls `torch.linalg.svd`, which costs $+47\%$ per step. Since $\mathbf{C}_T$ is symmetric positive semi-definite, `torch.linalg.eigh` yields the same transform at a fraction of the cost: across the batch sizes we use, the two whitening matrices agree to a relative Frobenius error below 3×10^{-4} and the anisotropy index of the resulting target matches to four decimal places, while the per-step overhead drops to $+4.4\%$ over RSD. We report the SVD figures because they correspond to the code that produced our results, but the intrinsic cost of whitening is the smaller number. The two also agree end to end, with full 300-epoch training at $B = 384$ reaching 84.38% with SVD and 84.23% with the symmetric decomposition, within run-to-run spread.