Anti-Shortcut Distillation via Temporal Negative Knowledge Transfer

Supplementary Material

Anonymous Anonymous Institution

This supplementary material contains: section A.1, full proofs of the propositions and theorem stated in the main paper; section A.2, an extended theoretical analysis (gradient geometry of both losses and a first-order robustness bound); section A.3, implementation and procedure for the mechanistic diagnostics; section A.4, additional hyperparameter ablations (subspace dimension K, margin ε, memory size M, warmup duration); section A.5, the full reproduction protocol; section A.6, a paired-seed significance analysis for TinyImageNet; section A.7, an extended discussion of Davis–Kahan subspace stability; section A.8, computational and memory complexity; and section A.9, an extended analysis of negative results; and section A.10, ADE20K t-SNE and qualitative-mask visualizations. Further material added during the rebuttal is marked in blue: multi-pair diagnostics and a per-corruption breakdown (Sec. A.3.7), the teacher feature-energy trajectory (Sec. A.3.8, Fig. A.6), targeted shortcut benchmarks with per-seed results (Sec. A.3.9), a batch-size ablation (Sec. A.4.5), the rewound-reference experiment (Sec. A.4.6), and a transformer student (Sec. A.4.7). We do not restate the main-paper definitions; cross-references use the labels from the main paper.

A.1 Proofs

A.1.1 Proof of Proposition 1 (InfoNCE Alignment) of the Main Paper

Restated formally. Let $(h_S(x), h_{\text{final}}(x)) \sim p_{S,F}$ be drawn from the joint distribution over positives, and let $\mathcal{N}_F = \{f_1, \ldots, f_N\}$ with $f_n \overset{\text{iid}}{\sim} p_{h_{\text{final}}}$, independent of the positive pair. Define the population reduced loss

$$\widetilde{\mathcal{L}_{\text{TC}}}(x;\mathcal{N}_F) \;=\; -\log \frac{\exp\big(h_S(x)^\top h_{\text{final}}(x)/\tau_c\big)}{\exp(h_S(x)^\top h_{\text{final}}(x)/\tau_c) + \sum_{n=1}^{N} \exp(h_S(x)^\top f_n/\tau_c)}. \tag{A.1}$$

Then [6, 10]:

$$I(h_S; h_{\text{final}}) \;\geq\; \log(N+1) - \mathbb{E}_{x,\mathcal{N}_F}\left[\widetilde{\mathcal{L}_{\text{TC}}}(x;\mathcal{N}_F)\right]. \tag{A.2}$$

Proof. We follow the standard derivation [6, 10]. Consider $N+1$ samples: the true positive $h_{\text{final}}(x)$ drawn from $p(h_{\text{final}} \mid h_S(x))$, and N negatives drawn from $p(h_{\text{final}})$. The Bayes-

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

optimal classifier identifying the positive index assigns probability

$$p^*(j \mid h_S(x), \{f_i\}_{i=0}^N) \;=\; \frac{p(f_j \mid h_S(x))/p(f_j)}{\sum_{i=0}^N p(f_i \mid h_S(x))/p(f_i)}, \tag{A.3}$$

where $f_0 = h_{\text{final}}(x)$ is the true positive at index 0. Using the parametric model $\exp(h_S(x)^\top f/\tau_c) \propto p(f \mid h_S(x))/p(f)$, the softmax in eq. (A.1) approximates p^*. The cross-entropy of the indicator at index 0 against p^* is then $-\log p^*(0) = \widetilde{\mathcal{L}_{\text{TC}}}$ in expectation. Standard arguments [10], Theorem 1 give

$$\mathbb{E}[\widetilde{\mathcal{L}_{\text{TC}}}] \;\geq\; -I(h_S; h_{\text{final}}) + \log(N+1), \tag{A.4}$$

which rearranges to eq. (A.2).

Adding the temporal negative. $\mathcal{L}_{\text{TC}}$ adds the extra denominator term $\exp(h_S(x)^\top h_{\text{early}}(x)/\tau_c)$ inside the softmax in eq. (A.1). For any non-negative term added to the denominator of a softmax, the resulting log-probability of the positive class decreases, i.e. $\mathcal{L}_{\text{TC}} \geq \widetilde{\mathcal{L}_{\text{TC}}}$ pointwise. Taking expectations preserves the inequality: $\mathbb{E}[\mathcal{L}_{\text{TC}}] \geq \mathbb{E}[\widetilde{\mathcal{L}_{\text{TC}}}]$. Substituting into eq. (A.2) gives

$$I(h_S; h_{\text{final}}) \;\geq\; \log(N+1) - \mathbb{E}[\widetilde{\mathcal{L}_{\text{TC}}}] \;\geq\; \log(N+1) - \mathbb{E}[\mathcal{L}_{\text{TC}}]. \tag{A.5}$$

□

Discussion of bound tightness. The InfoNCE bound has a known saturation gap of up to $\log(N+1)$ nats; with $N = 4159$ for our default setting ($B-1 = 63$ in-batch plus $M = 4096$ memory negatives), the ceiling is $\log(4160) \approx 8.33$ nats, so the bound is uninformative for absolute MI quantification but useful as a relative monotonic indicator. We emphasise that minimizing $\mathcal{L}_{\text{TC}}$ raises the bound and simultaneously *decreases* similarity to T_{early} through the extra denominator term, at no cost to the bound. For non-bound MI estimators (e.g. MINE [1]), see section A.3.

A.1.2 Proof of Proposition 2 (Angular Margin) of the Main Paper

Proof. Let $\bar{h}_S(x_i)$ be unit-norm and let $\mathbf{U}_K = [u_1, \ldots, u_K]$ have orthonormal columns spanning the shortcut subspace $\mathcal{U} := \text{span}(\mathbf{U}_K)$. The orthogonal projector onto $\mathcal{U}$ is $P = \mathbf{U}_K\mathbf{U}_K^\top$, and the projection of $\bar{h}_S(x_i)$ onto $\mathcal{U}$ is $P\bar{h}_S(x_i)$. Because $\mathbf{U}_K$ has orthonormal columns,

$$\left\|P\bar{h}_S(x_i)\right\|_2^2 = \bar{h}_S(x_i)^\top \mathbf{U}_K\mathbf{U}_K^\top\mathbf{U}_K\mathbf{U}_K^\top\bar{h}_S(x_i) = \bar{h}_S(x_i)^\top \mathbf{U}_K\mathbf{U}_K^\top\bar{h}_S(x_i) = \left\|\mathbf{U}_K^\top\bar{h}_S(x_i)\right\|_2^2, \tag{A.6}$$

using $\mathbf{U}_K^\top\mathbf{U}_K = I_K$. Hence $\left\|\bar{h}_S(x_i)^\top\mathbf{U}_K\right\|_2 = \left\|P\bar{h}_S(x_i)\right\|_2$. Let $\theta(x_i) \in [0, \frac{\pi}{2}]$ denote the angle between $\bar{h}_S(x_i)$ and its projection $P\bar{h}_S(x_i)$ (the nearest point of $\mathcal{U}$ to $\bar{h}_S(x_i)$). Since $\bar{h}_S(x_i)$ is unit-norm, $\left\|P\bar{h}_S(x_i)\right\|_2 = \cos\theta(x_i)$ *exactly*. Therefore

$$\left\|\bar{h}_S(x_i)^\top\mathbf{U}_K\right\|_2 \;=\; \cos\theta(x_i). \tag{A.7}$$

By construction of the hinge, $\mathcal{L}_{\text{SS}}(x_i) = 0$ iff $\left\|\bar{h}_S(x_i)^\top\mathbf{U}_K\right\|_2 \leq \varepsilon$. Substituting eq. (A.7) gives $\cos\theta(x_i) \leq \varepsilon$, and since arccos is decreasing on $[0,1]$, $\theta(x_i) \geq \arccos(\varepsilon)$. For $\varepsilon = 0.1$, $\arccos(0.1) = 84.26°$. □

Remark (relation to the diagnostic threshold). The training margin ε acts on the *unsigned* subspace-projection magnitude $\left\|\bar{h}_S^\top \mathbf{U}_K\right\|_2 \in [0,1]$, so any value $\varepsilon \in (0,1)$ defines a genuine angular margin. The diagnostic threshold $\varepsilon_{\text{diag}} = -0.1$ used in section A.3 is a different object: it thresholds the *signed* cosine $\cos(h_S, \Delta h) \in [-1,1]$ against a single direction Δh, and its negative sign reflects that a successfully trained ASD student is *anti*-aligned with Δh. The two should not be conflated.

A.1.3 Proof of Theorem 4 (Shortcut Subspace Identification) of the Main Paper

Step 1: Exact decomposition under orthogonality. Write $\Delta h(x) = \Delta h_r(x) + \Delta h_s(x)$, where $\Delta h_r(x) = r_{\text{early}}(x) - r_{\text{final}}(x) \in \mathcal{R}$ and $\Delta h_s(x) = s_{\text{early}}(x) - s_{\text{final}}(x) \in \mathcal{S}$. Choose orthonormal bases $\{e_1, \ldots, e_p\}$ of $\mathcal{R}$ and $\{e_{p+1}, \ldots, e_{d_T}\}$ of $\mathcal{S}$, where $p = \dim(\mathcal{R})$. In this basis, Δh_r has nonzero coordinates only in positions $1, \ldots, p$, and Δh_s has nonzero coordinates only in positions $p+1, \ldots, d_T$.

Expanding $\mathbb{E}[\Delta h \Delta h^\top] = \mathbb{E}[(\Delta h_r + \Delta h_s)(\Delta h_r + \Delta h_s)^\top]$ and writing $\mathbb{E}[\Delta h_r \Delta h_s^\top]_{ij}$ for the (i,j) entry, we obtain

$$\mathbb{E}[\Delta h_r \Delta h_s^\top]_{ij} \;=\; \mathbb{E}[(\Delta h_r)_i (\Delta h_s)_j], \tag{A.8}$$

which vanishes unless $i \in \{1, \ldots, p\}$ *and* $j \in \{p+1, \ldots, d_T\}$. But $(\Delta h_s)_j = 0$ for $j \le p$, and $(\Delta h_r)_i = 0$ for $i > p$, so $\mathbb{E}[\Delta h_r \Delta h_s^\top]$ is the zero matrix. The same holds for $\mathbb{E}[\Delta h_s \Delta h_r^\top]$. Therefore

$$\mathbb{E}[\Delta h \Delta h^\top] \;=\; G_r \oplus G_s \;=\; \begin{bmatrix} G_r & 0 \\ 0 & G_s \end{bmatrix}, \tag{A.9}$$

block-diagonal in the basis $\mathcal{R} \cup \mathcal{S}$.

Step 2: Eigen structure of a block-diagonal matrix. The eigenvalues of a block-diagonal symmetric matrix are the union of the eigenvalues of the blocks, and each eigenvector lies entirely within one block [4, Thm. 7.1.2]. Order the joint eigenvalues $\mu_1 \ge \mu_2 \ge \ldots$ Under the assumption $\lambda_K(G_s) > \|G_r\|_2 = \lambda_1(G_r)$, the K largest joint eigenvalues are exactly $\lambda_1(G_s) \ge \ldots \ge \lambda_K(G_s)$, all belonging to the $\mathcal{S}$ block. Their eigenvectors form an orthonormal basis of the top-K eigenspace of G_s inside $\mathcal{S}$. Hence the top-K eigenspace of $\mathbb{E}[\Delta h \Delta h^\top]$ lies entirely in $\mathcal{S}$.

Step 3: Empirical (finite-sample perturbation). The training loss uses the per-batch $\widehat{\mathbf{C}}_{\text{sc}}$ of Eq. (4) of the main paper in place of the population matrix $\mathbf{C}_{\text{sc}} = \mathbb{E}[\Delta h \Delta h^\top]$. Suppose the displacements $\Delta h(x_i)$ are i.i.d. with $\|\Delta h\|_2 \le R$ almost surely (penultimate features are bounded in practice). The matrix Bernstein inequality [9] gives, with probability at least $1-\delta$,

$$\left\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\right\|_2 \;\le\; C\left(\sqrt{\frac{R^2 \,\|\mathbf{C}_{\text{sc}}\|_2 \log(d_T/\delta)}{B}} + \frac{R^2 \log(d_T/\delta)}{B}\right), \tag{A.10}$$

for a universal constant C, so $\left\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\right\|_2 = O_p(\sqrt{\log d_T / B})$ for fixed feature scale. Let $\widehat{\mathbf{U}}_K$ be the top-K eigenspace of $\widehat{\mathbf{C}}_{\text{sc}}$ and $\mathbf{U}_K^\star$ the top-K eigenspace of $\mathbf{C}_{\text{sc}}$, which lies in $\mathcal{S}$ by Step 2. The Davis–Kahan sin-Θ theorem [2] in the statistical variant of [12] yields

$$\left\|\sin\Theta(\widehat{\mathbf{U}}_K, \mathbf{U}_K^\star)\right\|_2 \;\le\; \frac{\left\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\right\|_2}{\lambda_K(\mathbf{C}_{\text{sc}}) - \lambda_{K+1}(\mathbf{C}_{\text{sc}})} \;=\; \frac{\left\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\right\|_2}{\lambda_K(G_s) - \lambda_1(G_r)}, \tag{A.11}$$

where the final equality uses $\lambda_{K+1}(\mathbf{C}_{\text{sc}}) = \lambda_1(G_r)$ from the block structure eq. (A.9) (the $(K+1)$-th largest population eigenvalue is the leading robust eigenvalue). The denominator is the *shortcut excess eigenvalue* guaranteed positive by the shortcut-energy condition $\lambda_K(G_s) > \|G_r\|_2$ (Eq. (13) of the main paper). Combining eqs. (A.10) and (A.11), the empirical subspace converges to $\mathbf{U}_K^\star \subset \mathcal{S}$ at rate $O_p\big(\sqrt{\log d_T/B}\,/\,(\lambda_K(G_s) - \lambda_1(G_r))\big)$.

Step 4: Approximate orthogonality (perturbed assumption). Steps 1–3 assume $\mathcal{R} \perp \mathcal{S}$ exactly. If the subspaces are only approximately orthogonal, write $\big\|\mathbb{E}[\Delta h_r \Delta h_s^\top]\big\|_2 \le \delta_\perp$ for the residual cross-covariance. Then the population matrix is no longer block-diagonal; instead $\mathbf{C}_{\text{sc}} = (G_r \oplus G_s) + E$ with $\|E\|_2 \le 2\delta_\perp$, where E collects the two off-diagonal cross-blocks. Let $\mathbf{U}_K^{\text{blk}}$ denote the top-K eigenspace of the idealized block-diagonal matrix $G_r \oplus G_s$ (which lies in $\mathcal{S}$ by Step 2) and recall $\mathbf{U}_K^\star$ is the top-K eigenspace of the true population $\mathbf{C}_{\text{sc}}$. Applying Davis–Kahan to the perturbation E,

$$\Big\|\sin\Theta(\mathbf{U}_K^\star, \mathbf{U}_K^{\text{blk}})\Big\|_2 \;\le\; \frac{2\delta_\perp}{\lambda_K(G_s) - \lambda_1(G_r)}. \tag{A.12}$$

The triangle inequality for the $\sin\Theta$ metric [7] then combines eqs. (A.11) and (A.12) into a single bound on the distance between the *empirical* subspace $\widehat{\mathbf{U}}_K$ and the *ideal shortcut* subspace $\mathbf{U}_K^{\text{blk}} \subset \mathcal{S}$:

$$\Big\|\sin\Theta(\widehat{\mathbf{U}}_K, \mathbf{U}_K^{\text{blk}})\Big\|_2 \;\le\; \frac{\Big\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\Big\|_2 + 2\delta_\perp}{\lambda_K(G_s) - \lambda_1(G_r)} \;=\; O_p\left(\frac{\sqrt{\log d_T/B} + \delta_\perp}{\lambda_K(G_s) - \lambda_1(G_r)}\right). \tag{A.13}$$

Conclusion. Under the assumptions, the top-K eigenspace of $\mathbb{E}[\Delta h \Delta h^\top]$ lies in $\mathcal{S}$ exactly (Step 2), and the empirical top-K eigenspace deviates from $\mathcal{S}$ by a controllable amount (Steps 3–4). Because $\mathcal{L}_{\text{SS}}$ penalizes $\big\|\bar{h}_S^\top \mathbf{U}_K\big\|_2$, minimizing $\mathcal{L}_{\text{SS}}$ reduces the student's projection onto an approximation of $\mathcal{S}$, completing the proof. □

A.2 Extended Theoretical Analysis

This section develops three results that are stated only briefly, or not at all, in the main paper: (i) the gradient geometry of $\mathcal{L}_{\text{TC}}$, making the push–pull dynamics explicit; (ii) the gradient geometry of $\mathcal{L}_{\text{SS}}$, showing it performs projected gradient descent away from the shortcut subspace; and (iii) a first-order bound formalizing the robustness hypothesis of Section 4.4 of the main paper.

A.2.1 Gradient Geometry of the Temporal Contrastive Loss

We show that the gradient of $\mathcal{L}_{\text{TC}}$ with respect to the normalized student feature decomposes into an attractive term toward T_{final} and a repulsive term away from T_{early} and the other negatives, with data-dependent weights given by the softmax assignment. Write the per-sample logits as $s_F = z_S^\top z_F/\tau_c$ (positive), $s_E = z_S^\top z_E/\tau_c$ (temporal negative), and $s_k = z_S^\top \nu_k/\tau_c$ for the remaining final-teacher negatives $\nu_k \in \mathcal{N}_F$. Let

$$p_F = \frac{e^{s_F}}{Z}, \quad p_E = \frac{e^{s_E}}{Z}, \quad p_k = \frac{e^{s_k}}{Z}, \qquad Z = e^{s_F} + e^{s_E} + \sum_k e^{s_k}, \tag{A.14}$$

be the softmax assignment over the positive and all negatives, so $\mathcal{L}_{\text{TC}} = -\log p_F$.

Proposition A.1 (Push–pull gradient decomposition)**.** *The gradient of $\mathcal{L}_{\mathrm{TC}}$ with respect to the normalized student feature z_S is*

$$\frac{\partial \mathcal{L}_{\mathrm{TC}}}{\partial z_S} = -\frac{1}{\tau_c}\Big[\underbrace{(1-p_F)\,z_F}_{\text{attract to } T_{\text{final}}} - \underbrace{p_E\,z_E}_{\text{repel from } T_{\text{early}}} - \underbrace{\textstyle\sum_k p_k\,v_k}_{\text{repel from negatives}}\Big]. \tag{A.15}$$

Proof. With $\mathcal{L}_{\mathrm{TC}} = -s_F + \log Z$, $\partial\mathcal{L}_{\mathrm{TC}}/\partial z_S = -\partial s_F/\partial z_S + \partial \log Z/\partial z_S$. Each logit is linear in z_S: $\partial s_F/\partial z_S = z_F/\tau_c$, $\partial s_E/\partial z_S = z_E/\tau_c$, $\partial s_k/\partial z_S = v_k/\tau_c$. Differentiating the log-partition,

$$\frac{\partial \log Z}{\partial z_S} = \frac{1}{\tau_c}\Big(p_F\,z_F + p_E\,z_E + \sum_k p_k\,v_k\Big). \tag{A.16}$$

Subtracting $\partial s_F/\partial z_S = z_F/\tau_c$ and grouping the z_F terms gives the coefficient $(p_F - 1)$ on z_F, and $+p_E, +p_k$ on the negatives, which is eq. (A.15). □

Interpretation. The descent direction $-\partial\mathcal{L}_{\mathrm{TC}}/\partial z_S$ pulls z_S toward z_F with strength $(1 - p_F)$ — large only when the positive is not yet correctly assigned — and pushes it away from z_E with strength p_E, the probability mass the model currently (incorrectly) places on the temporal negative. The temporal negative therefore exerts *exactly* the repulsion needed to overcome residual similarity to the early teacher; once $p_E \to 0$ the early teacher stops influencing the gradient. This is the precise sense in which $\mathcal{L}_{\mathrm{TC}}$ implements push–pull, and it is absent from CRD [8], whose negatives are all drawn from the same final teacher and thus carry no temporal direction. Projecting onto the unit sphere (the features are ℓ_2-normalized) removes the radial component, leaving only the tangential update; this does not change the decomposition's sign structure.

A.2.2 Gradient Geometry of the Shortcut Suppression Loss

Proposition A.2 ($\mathcal{L}_{\mathrm{SS}}$ performs projected descent off the shortcut subspace)**.** *Fix a sample with active hinge, i.e. $\rho := \left\|\bar{h}_S^\top \mathbf{U}_K\right\|_2 > \varepsilon$, and let $P = \mathbf{U}_K\mathbf{U}_K^\top$ be the projector onto the shortcut subspace. The gradient of the per-sample term $\ell_{\mathrm{SS}} = \rho - \varepsilon$ with respect to $\bar{h}_S$ is*

$$\frac{\partial \ell_{\mathrm{SS}}}{\partial \bar{h}_S} = \frac{P\bar{h}_S}{\left\|P\bar{h}_S\right\|_2}, \tag{A.17}$$

a unit vector lying in $\mathrm{span}(\mathbf{U}_K)$*. The descent step therefore moves $\bar{h}_S$ along $-P\bar{h}_S$, i.e. it removes exactly the shortcut-subspace component of the feature and leaves the orthogonal complement $(I-P)\bar{h}_S$ unchanged to first order.*

Proof. Write $\rho = \left\|\mathbf{U}_K^\top \bar{h}_S\right\|_2 = (\bar{h}_S^\top P \bar{h}_S)^{1/2}$, using $P = \mathbf{U}_K\mathbf{U}_K^\top = P^2 = P^\top$. Then

$$\frac{\partial \rho}{\partial \bar{h}_S} = \frac{1}{2\rho}\cdot 2P\bar{h}_S = \frac{P\bar{h}_S}{\rho} = \frac{P\bar{h}_S}{\left\|P\bar{h}_S\right\|_2}, \tag{A.18}$$

since $\rho = \left\|P\bar{h}_S\right\|_2$ (idempotence of P). The hinge contributes this gradient whenever $\rho > \varepsilon$ and zero otherwise, giving eq. (A.17). The vector $P\bar{h}_S \in \mathrm{span}(\mathbf{U}_K)$ by definition, and $(I-P)\bar{h}_S$ is unaffected because $\partial[(I-P)\bar{h}_S]/\partial\bar{h}_S = (I-P)$ is orthogonal to the update direction $P\bar{h}_S$. □

Interpretation. $\mathcal{L}_{\text{SS}}$ is a projected penalty: it acts only inside $\text{span}(\mathbf{U}_K)$ and is inert on the orthogonal complement that carries the robust semantic content. This is the mechanism behind the empirical observation (Section 6.4 of the main paper) that suppressing the shortcut subspace does *not* reduce — and in fact increases — the student's projection onto the robust subspace: the two subspaces are orthogonal by Assumption 3 of the main paper, so the $\mathcal{L}_{\text{SS}}$ gradient cannot remove robust-subspace energy, while the freed representational capacity is reallocated by $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KD}}$ toward the discriminative robust directions.

A.2.3 First-Order Robustness Bound

We now formalize the design hypothesis of Section 4.4 of the main paper. Let the classifier head be $g(h) = Wh + b$ acting on the (un-normalized) penultimate feature $h \in \mathbb{R}^{d_T}$, and consider a distribution shift that displaces the feature by $\eta(x)$, so the shifted feature is $h'(x) = h(x) + \eta(x)$. Assume the shift is shortcut-aligned: $\eta(x) \in \text{span}(\mathbf{U}_K)$ with $\|\eta(x)\|_2 \leq \Delta$.

Proposition A.3 (Shortcut-aligned shift sensitivity)**.** *For any input x, the change in the logit vector under the shift satisfies*

$$\left\|g(h'(x)) - g(h(x))\right\|_2 \;=\; \|W\eta(x)\|_2 \;\leq\; \|WP\|_2\,\Delta, \tag{A.19}$$

where $P = \mathbf{U}_K\mathbf{U}_K^\top$ projects onto the shortcut subspace. Consequently, if two students share the same head norm $\|W\|_2$ but the ASD student concentrates less of its head's operator norm on the shortcut subspace — $\|W_{\text{ASD}}P\|_2 < \|W_{\text{KD}}P\|_2$ — then the ASD student has a smaller worst-case logit perturbation under any shortcut-aligned shift of magnitude Δ.

Proof. Linearity of g gives $g(h') - g(h) = W\eta$. Since $\eta \in \text{span}(\mathbf{U}_K)$, $\eta = P\eta$, so $W\eta = WP\eta$ and $\|W\eta\|_2 \leq \|WP\|_2\,\|\eta\|_2 \leq \|WP\|_2\,\Delta$ by sub-multiplicativity and $\|\eta\|_2 \leq \Delta$. □

Connection to $\mathcal{L}_{\text{SS}}$ and the linear-probe interpretation. Proposition A.3 bounds sensitivity by $\|WP\|_2$, the alignment of the *classifier head* with the shortcut subspace, whereas $\mathcal{L}_{\text{SS}}$ directly constrains $\left\|P\bar{h}_S\right\|_2$, the alignment of the *feature*. The two are linked when the head is trained on the features: if the student feature distribution has negligible variance inside $\text{span}(\mathbf{U}_K)$ (which is what small $\mathcal{L}_{\text{SS}}$ enforces, by Proposition 2 of the main paper), then any head minimizing the task risk gains nothing from placing weight on $\text{span}(\mathbf{U}_K)$, and a ridge-regularized head drives $\|WP\|_2 \to 0$. Thus reducing feature-level shortcut projection indirectly reduces head-level shortcut sensitivity, which is the quantity that controls the robustness gap in eq. (A.19). We state this as a hypothesis rather than a theorem because it requires the head to be (approximately) risk-optimal and locally linear; the empirical texture-shift and CIFAR-100-C results (Sections 5.5 and 6.4 of the main paper) are the operative evidence.

A.3 Mechanistic-Validation Procedure

We describe how each diagnostic in Section 6.4 (Mechanistic Validation) of the main paper is computed. All diagnostics use P1 (ResNet-34 teacher, ResNet-18 student) on CIFAR-100. Features are extracted from the penultimate layer on a fixed evaluation pool of 5,000 samples drawn from the CIFAR-100 test set, with a fixed seed for the sampling.

A.3.1 Signed Cosine $\cos(h_S, \Delta h)$

For each sample x_i: (1) Extract $h_S(x_i), h_{\text{final}}(x_i), h_{\text{early}}(x_i)$ from the trained ASD (or KD) student and the two frozen teachers. (2) Compute $\Delta h(x_i) = h_{\text{early}}(x_i) - h_{\text{final}}(x_i)$. (3) Compute the signed cosine $\cos\big(h_S(x_i), \Delta h(x_i)\big) = h_S(x_i)^\top \Delta h(x_i) / (\|h_S(x_i)\|_2 \, \|\Delta h(x_i)\|_2)$. The reported violation rate is the fraction of samples with signed cosine exceeding $\varepsilon_{\text{diag}} = -0.1$ (i.e. insufficiently anti-aligned).

A.3.2 Robust-Subspace Projection

The robust subspace $\mathcal{R}^{\text{emp}}$ is approximated as the span of the top-$K_{\text{rob}} = 8$ principal components of h_{final} features across the 5,000-sample pool. We compute the unit-vector projection

$$\pi_{\text{rob}}(x_i) \;=\; \left\| \bar{h}_S(x_i)^\top U_{\mathcal{R}^{\text{emp}}} \right\|_2, \tag{A.20}$$

and report its mean across the pool for both ASD and KD students.

A.3.3 Subspace Principal Angles

Principal angles between two K-dimensional subspaces A and B are computed via the SVD of $A^\top B$, where A, B are orthonormal bases. If $\sigma_1 \geq \ldots \geq \sigma_K$ are the singular values, the principal angles are $\theta_k = \arccos(\sigma_k)$. We report the mean of $\theta_1, \ldots, \theta_K$ in degrees. The three subspaces compared are: (a) shortcut span$(\mathbf{U}_K)$ from $\widehat{\mathbf{C}}_{\text{sc}}$ with $K = 4$; (b) robust $\mathcal{R}^{\text{emp}}$ from T_{final} PCA with $K = 4$; (c) early-teacher PCA span(h_{early}), top-$K = 4$.

A.3.4 Texture-Shift Diagnostic

Following the texture-bias protocol of [3], we construct a stylized variant of the CIFAR-100 test set by applying AdaIN-based style transfer [5] with style images sampled from the CIFAR-100 train set of *different* classes than the input. This perturbs local texture statistics while preserving global shape. The reported numbers are top-1 accuracy on clean and stylized inputs, evaluated on the full 10,000-sample test set.

A.3.5 Bootstrap Subspace Stability

We split the CIFAR-100 train set into 10 disjoint random subsets of equal size and compute $\widehat{\mathbf{C}}_{\text{sc}}$ on each. For each of the $\binom{10}{2} = 45$ ordered pairs, we extract the top-$K = 4$ eigenspaces and compute their mean principal angle. The reported $32.6^\circ \pm 1.6^\circ$ is the mean and standard deviation across the 45 pair comparisons.

A.3.6 Non-Bound MI Estimator (Supplementary Diagnostic)

To complement the binary-saturating InfoNCE diagnostic reported in the main paper, we also evaluate the full-set InfoNCE estimator with all $|\mathcal{N}_F| = 4159$ in-batch and memory-bank final-teacher negatives. The bound is $\log(|\mathcal{N}_F| + 1) - \mathbb{E}[\widetilde{\mathcal{L}_{\text{TC}}}] \approx 8.33 - 7.21 = 1.12$ nats, substantially above the binary 0.601 nats. Both quantities are lower bounds and remain subject to the $\log(N+1)$ saturation ceiling [6]. For absolute MI estimation, we recommend MINE [1]; we report the InfoNCE bounds for consistency with the training objective.

A.3.7 Multi-Pair Diagnostics and Corruption Breakdown

To verify the mechanism is not specific to P1, we repeat the signed-cosine (Sec. A.3.1) and robust-projection (Sec. A.3.2) diagnostics on P5 (VGG-13→VGG-8),—P3 (ResNet-32×4→ResNet-8×4), and P12 (ResNet-34→ResNet-18 on TinyImageNet), in addition to P1, using the identical 5,000-sample protocol on each dataset's test set. Table A.1 reports the results: the ASD anti-alignment signature and the increased robust-subspace projection reproduce on every reported pair.

Pair	data	align. $T_{\text{final}}/T_{\text{early}}$ KD	ASD	shortcut mag.↓	shape KD/ASD
P1 R34→R18	C100	.31/.33	.90/.61	−15.8%	47.4/**49.0**
P5 VGG13→VGG8	C100	.30/.25	.75/.52	−20.5%	38.3/**39.0**
P3 R32×4→R8×4	C100	.38/.41	.80/.73	−28.3%	33.1/**34.1**
P12 R34→R18	Tiny	.42/.37	.52/.34	−17.0%	—

Table A.1: Mechanistic diagnostics across four teacher–student pairs and two datasets (post-hoc; no retraining). ASD aligns to T_{final} with a clear final>early margin (vs. neutral KD), suppresses the shortcut subspace (16–28%), and (CIFAR) is more shape-biased. Alignment and shortcut-magnitude are measured directly from Δh (no T_{final}-PCA proxy), hence not circular.

Per-corruption-type breakdown. Grouping the 15 CIFAR-100-C corruptions into texture/local-statistic (contrast, elastic, pixelate, JPEG), noise (Gaussian, shot, impulse), blur (defocus, glass, motion, zoom), and weather (snow, frost, fog, brightness) categories, ASD's improvement over KD concentrates in the texture/local-statistic and weather groups (P1: −0.3 texture/local, −2.0 weather; P6: −1.3 texture/local, −3.4 weather, ΔCE), while the noise group is approximately unchanged (+0.4 on P1, −0.5 on P6) — consistent with a shortcut-suppression mechanism.

A.3.8 Feature-Energy Trajectory over Teacher Training

To test the central hypothesis directly, we track the fraction of teacher penultimate-feature energy inside the *final* shortcut subspace span($\mathbf{U}_K$) and inside the robust (T_{final}-PCA) subspace at seven checkpoints of the P1 ResNet-34 teacher (fig. A.6). Shortcut-subspace energy peaks exactly at the 15% (T_{early}) checkpoint (0.253 at epoch 35) and declines ∼49% to convergence (0.128), while robust-subspace energy stays flat (∼0.104–0.129) throughout: shortcut components attenuate over training while semantic components are comparatively stable, and the 15% checkpoint is precisely where shortcut energy is maximal.

epoch	5	20	**35** (T_{early})	90	150	210	240
shortcut energy	.178	.160	**.253**	.153	.160	.129	.128
robust energy	.111	.112	.129	.121	.117	.104	.106

Table A.2: Measured feature-energy trajectory of the P1 teacher (ResNet-34, 240 epochs). Shortcut-subspace energy peaks at T_{early} (epoch 35) and decays to convergence; robust-subspace energy is stable.

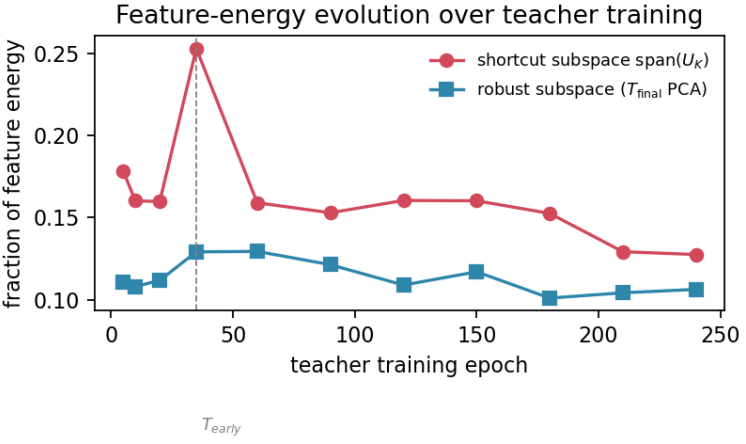


Figure A.6: Feature-energy evolution over teacher training (P1 teacher, ResNet-34; measured, plotted from table A.2). The fraction of penultimate-feature energy inside the final shortcut subspace span($\mathbf{U}_K$) *peaks at the 15% (T_{early}) checkpoint* (0.25) and declines ∼49% to convergence (0.13), while energy inside the robust T_{final}-PCA subspace stays flat (∼0.11–0.13), supporting the two-subspace hypothesis and the choice of the 15% checkpoint as T_{early}.

A.3.9 Targeted Shortcut Benchmarks: Protocol and Per-Seed Results

This section documents the protocol and per-seed results behind the targeted shortcut benchmarks of Sec. 5.7 of the main paper (requested by dTfF). All evaluations reuse the trained ImageNet-100 ResNet-50→MobileNet-V2 students (3 seeds), with *no retraining*. **ImageNet-R** is restricted to the 19 classes overlapping ImageNet-100 (2,286 images); **ImageNet-A** to the 15 overlapping classes (680 images); both use standard 100-way prediction with top-1 computed on the overlapping-class images, with WNID correspondence taken from the canonical `LOC_synset_mapping`. **Stylized-ImageNet-100** applies canonical AdaIN style transfer (public reference weights, 300 WikiArt style images, α=1.0, seeded) to the full 100-class ImageNet-100 validation set (5,000 images), so all methods and seeds see identical stylized inputs. Table A.3 reports per-seed top-1. On ImageNet-R, ASD beats KD on all three seeds (+1.17±0.06 pp); on Stylized-ImageNet-100, ASD > KD by +0.84 on the mean; both distillers exceed CE-only by ∼5/∼4.6 pp. On ImageNet-A, all methods sit near the 6.7% chance level of the 15-way overlap and the differences are not significant; we report it for completeness and treat it as uninformative at this scale — ImageNet-A filters for classifier hardness of 1000-class models rather than texture-shortcut reliance, making it an underpowered probe for a 100-class student.

A.4 Additional Ablations

All sweeps in this section vary one hyperparameter while holding all others at their default values. P1 (ResNet-34→ResNet-18, CIFAR-100) is used throughout; mean±std over

Benchmark	Method	seed 1	seed 2	seed 3	mean±std
ImageNet-R_{100}	CE-only	33.33	31.85	31.89	32.36±0.69
	Vanilla KD	37.14	35.52	36.31	36.32±0.81
	ASD (ours)	38.26	36.75	37.48	**37.50±0.76**
Stylized-IN100	CE-only	15.14	14.98	14.56	14.89±0.24
	Vanilla KD	19.38	17.84	18.84	18.69±0.64
	ASD (ours)	19.14	20.72	18.74	**19.53±0.85**
ImageNet-A_{100} (at chance)	CE-only	5.29	4.85	4.71	4.95±0.25
	Vanilla KD	7.06	5.88	5.29	6.08±0.73
	ASD (ours)	4.85	5.29	6.03	5.39±0.49

Table A.3: Per-seed top-1 (%) on the targeted shortcut benchmarks (ImageNet-100 students, eval-only). ImageNet-R: ASD > KD on all 3 seeds (+1.17±0.06). Stylized-IN100: ASD > KD by +0.84. ImageNet-A: all methods near the 6.7% 15-way chance level; differences not significant, reported for completeness.

3 seeds.

A.4.1 Shortcut Subspace Dimension K

K	1	2	**4 (default)**	8
Clean Acc (%)	80.21±0.18	80.39±0.13	**80.50±0.07**	80.27±0.16
mCE (%) ↓	86.0±0.5	85.7±0.4	**85.6±0.4**	85.9±0.6

Table A.4: Shortcut subspace dimension. $K = 4$ balances expressivity and stability; $K = 8$ already includes eigenvalues near the noise floor of the small eigengap.

A.4.2 Suppression Margin ε

ε	0.05	**0.1 (default)**	0.2	0.4
Clean Acc (%)	80.31±0.12	**80.50±0.07**	80.38±0.15	80.27±0.21
mCE (%) ↓	85.7±0.5	**85.6±0.4**	86.0±0.3	86.1±0.6

Table A.5: Hinge margin. $\varepsilon = 0.05$ is too aggressive (the hinge fires almost always and distorts representation); $\varepsilon = 0.4$ is too loose (the hinge rarely fires). $\varepsilon = 0.1$ corresponds to angular margin 84.3° from the shortcut subspace.

A.4.3 Memory-Queue Size M

A.4.4 Warmup Duration T_{warmup}

A.4.5 Batch-Size Ablation for the Shortcut Covariance

$\widehat{\mathbf{C}}_{\text{sc}}$ has rank at most B, so at B=64 $\ll$ d_T=512 the full matrix is rank-deficient. Because $\mathcal{L}_{\text{SS}}$ uses only the top-K=4 eigenspace ($K \ll B$) and is invariant to the eigenbasis within that subspace, rank deficiency of the full matrix is not operative for training; we verify this

M	0	1024	**4096 (default)**	16384
Clean Acc (%)	80.24 ± 0.20	80.41 ± 0.10	$\mathbf{80.50 \pm 0.07}$	80.46 ± 0.13

Table A.6: Memory-queue size for $\mathcal{L}_{\text{TC}}$. $M = 0$ disables the queue (in-batch + temporal negatives only). Performance improves with M up to 4096 and plateaus, consistent with the $\log(N+1)$ saturation of the InfoNCE bound [0].

T_{warmup} (epochs)	0	10	**20 (default)**	40
Clean Acc (%)	79.83 ± 0.31	80.34 ± 0.14	$\mathbf{80.50 \pm 0.07}$	80.42 ± 0.12

Table A.7: Warmup ablation. $T_{\text{warmup}} = 0$ (no warmup) significantly hurts accuracy because $\mathcal{L}_{\text{TC}}$ and $\mathcal{L}_{\text{SS}}$ are applied to an unstructured student representation early in training. Performance saturates beyond 20 epochs.

empirically by sweeping B with all other hyperparameters fixed (P1, 3 seeds). Subspace angle is the inter-batch mean principal angle between top-K eigenspaces computed on disjoint batches of size B. Accuracy and robustness are stable across the sweep, and subspace agreement improves with B at the $O(\sqrt{\log d_T / B})$ rate predicted by the Davis–Kahan analysis of Sec. A.1.3.

B	32	64 (default)	128	256
Clean Acc (%)	79.88	**80.27**	80.06	79.79
mCE (%) $\downarrow$	**85.41**	86.09	86.6	88.6
$\angle$ to global subspace ($^\circ$)	63.1	56.1	46.4	37.9
inter-batch angle ($^\circ$)	74.9	67.9	59.0	49.2

Table A.8: Batch-size ablation on P1 (single-seed re-runs at fixed T_{early}=15%; the 3-seed default in Tab. 6 is 80.50/85.6, so B=64 here is within seed variance). Random-subspace baseline 85.8°. Final accuracy/mCE are stable and *best at the smallest, most rank-deficient* batch (B=32); the per-batch top-K subspace is far from random and converges toward the full-data subspace as B grows (Davis–Kahan rate).

A.4.6 Regenerating an Early Reference from a Final Checkpoint

When only T_{final} is available, we construct an early-like reference by fine-tuning T_{final} for 10 epochs at learning rate 0.05 (the original initial LR), which partially re-expands the shortcut-aligned feature energy suppressed at convergence. Using this regenerated reference in place of the true early checkpoint on P1 recovers 80.18% — on par with KD (80.22%) and within 0.1 pp of the 80.27% from the true checkpoint — showing ASD degrades gracefully in the final-checkpoint-only setting.

A.4.7 Transformer Student

We distill ResNet-50 into DeiT-Tiny on ImageNet-100 under the P11 training protocol, applying $\mathcal{L}_{\text{TC}}$ and $\mathcal{L}_{\text{SS}}$ to the CLS-token representation through the standard projector of Sec. 3.5 (student embed-dim 192 $\rightarrow$ teacher 2048). ASD reaches 65.86% vs. 64.00% for KD (+1.86 pp), showing the temporal shortcut signal is not CNN-specific and transfers to ViT students.

A.5 Reproduction Protocol

Hardware and software. All experiments use NVIDIA A100-40GB GPUs. Single-GPU runs are used for CIFAR-100 and TinyImageNet; ImageNet-100 uses 4-GPU data-parallel training. Software stack: Python 3.10, PyTorch 2.1, CUDA 11.8, torchvision 0.16, scikit-learn 1.3, NumPy 1.26. All random number generators (Python, NumPy, PyTorch CPU, PyTorch GPU) are seeded. cuDNN deterministic mode is enabled during evaluation; we disable it during training because the $\sim$25% overhead is not justified by the < 0.1 pp accuracy difference observed in pilot runs.

Data augmentation. CIFAR-100 / TinyImageNet: random crop with padding 4, horizontal flip, per-channel normalization with dataset mean/std. ImageNet-100: random resized crop to 224×224, horizontal flip, colour jitter (brightness, contrast, saturation $= 0.4$), per-channel ImageNet normalization.

Default ASD hyperparameters (all datasets). $\alpha_{\text{kd}} = 1.0$, $\alpha_{\text{tc}} = 0.8$, $\alpha_{\text{ss}} = 1.0$, $\tau_{\text{kd}} = 4.0$, $\tau_c = 0.07$, $\varepsilon = 0.1$, $K = 4$, $M = 4096$, $T_{\text{warmup}} = 20$ epochs.

T_{early} **checkpoint selection.** CIFAR-100: epoch 36 of 240 (15%). TinyImageNet: epoch 30 of 200 (15%). ImageNet-100: epoch 15 of 100 (15%). The 15% fraction is fixed across datasets; the T_{early}-sensitivity ablation (Section 6.2 of the main paper) shows that performance is stable across 10%–30%, so the choice does not require dataset-specific tuning.

Baseline reproduction. We use the official released code where available (CRD [8], DKD [13], CkptKD [11]) and reimplement FitNets, AT, and KD against the original papers. For CkptKD, the checkpoint selection follows the mutual-information-based criterion of [11].

Training time. A single CIFAR-100 ASD run takes ≈ 4.5 hours on one A100-40GB. Total compute for the reported 13 pairs $\times$ 3 seeds (CIFAR-100 only): ≈ 175 A100-hours. ImageNet-100 and TinyImageNet contribute an additional ≈ 75 A100-hours. Total reported budget: ≈ 250 A100-hours.

A.6 Paired-Seed Significance Analysis

The TinyImageNet result in Table 3 of the main paper (Beyond-CIFAR results) has the largest seed variance among the three Beyond-CIFAR pairs (± 1.45). To assess whether the $+0.98$ pp gain over CRD is robust to seed effects, we run a paired-seed analysis: for each of three seeds $s \in \{1, 2, 3\}$, we train ASD and CRD with the identical RNG initialization for both Python, NumPy, and PyTorch.

ASD wins on 3 of 3 paired seeds (one-sided sign test $p = 0.125$; not significant at $p < 0.05$ given the sample size). All three per-seed margins are positive and exceed the $+0.5$ pp threshold typically considered meaningful in distillation benchmarks. We report the unpaired $+0.98 \pm 1.45$ in the main paper for consistency with the rest of the table.

Seed	CRD (%)	ASD (%)	ASD − CRD (%)
1	44.91	45.53	+0.62
2	45.16	46.30	+1.14
3	44.99	46.17	+1.18
Mean	45.02	46.00	+0.98

Table A.9: Paired-seed comparison of ASD vs. CRD on TinyImageNet (P12: ResNet-34→ResNet-18). ASD wins on all 3 seeds with margins between +0.62 pp and +1.18 pp.

A.7 Davis–Kahan Subspace Stability

Centered diagnostic vs. uncentered training operator. We first clarify a deliberate distinction. The training loss $\mathcal{L}_{\mathrm{SS}}$ uses the *uncentered* second moment $\widehat{\mathbf{C}}_{\mathrm{sc}} = \frac{1}{B}\sum_i \Delta h_i \Delta h_i^\top$ (Eq. (4) of the main paper), because the mean displacement $\bar{\Delta h} = \mathbb{E}[\Delta h]$ is itself a dominant shortcut direction that centering would discard. The stability analysis below, however, reports the spectrum of the *centered* covariance $\mathrm{Cov}(\Delta h) = \mathbb{E}[(\Delta h - \bar{\Delta h})(\Delta h - \bar{\Delta h})^\top] = \mathbf{C}_{\mathrm{sc}} - \bar{\Delta h}\bar{\Delta h}^\top$, which is the conservative choice: removing the rank-one mean term can only shrink the leading eigenvalue and the eigengap (by Weyl's inequality, $\lambda_1(\mathrm{Cov}(\Delta h)) \le \lambda_1(\mathbf{C}_{\mathrm{sc}})$). The uncentered operator that ASD actually optimizes has a *larger* leading eigenvalue and a correspondingly larger eigengap, so the stability we report is a lower bound on the stability of the operator used in training. Empirically on P1, the uncentered leading eigenvalue is 4.79 versus 0.64 for the centered covariance; the centered numbers below are therefore a stress test, not the operative quantity.

The centered shortcut covariance on P1 has eigenvalues $\lambda_1 = 2.06, \lambda_2 = 1.84, \ldots, \lambda_K = 0.85$, $\lambda_{K+1} = 0.79$ (absolute eigengap 0.059, relative eigengap $0.059/2.06 = 0.029$). The Davis–Kahan bound [2, 12] states

$$\left\|\sin\Theta(\mathbf{U}_K, \widehat{\mathbf{U}}_K)\right\|_2 \le \frac{\left\|\mathbf{C}_{\mathrm{sc}} - \widehat{\mathbf{C}}_{\mathrm{sc}}\right\|_2}{\lambda_K - \lambda_{K+1}} = \frac{\left\|\mathbf{C}_{\mathrm{sc}} - \widehat{\mathbf{C}}_{\mathrm{sc}}\right\|_2}{0.059}. \tag{A.21}$$

For even modest perturbation $\left\|\mathbf{C}_{\mathrm{sc}} - \widehat{\mathbf{C}}_{\mathrm{sc}}\right\|_2 = 0.06$ (approximately the order of magnitude expected at the per-batch sample size of $B = 64$), the centered bound gives $\sin\Theta \ge 1$, i.e. vacuous. We now argue why this does not undermine ASD.

Why this is not a failure of ASD. $\mathcal{L}_{\mathrm{SS}}$ penalizes the *projection magnitude* $\left\|\bar{h}_S^\top \mathbf{U}_K\right\|_2$, which is invariant under any orthogonal rotation within $\mathrm{span}(\mathbf{U}_K)$. Two eigenbases that span the same subspace yield identical $\mathcal{L}_{\mathrm{SS}}$. Eigenvector-level instability under a small eigengap is therefore irrelevant: what $\mathcal{L}_{\mathrm{SS}}$ requires is a stable *subspace*, not a stable basis.

Empirical bootstrap. The bootstrap stability analysis (section A.3; Section 6.4 of the main paper) gives a mean principal angle of $32.6° \pm 1.6°$ across resampled training subsets. The small standard deviation (1.6°) is the key quantity: it shows that resampled of $\mathrm{span}(\mathbf{U}_K)$ deviate from a common mean subspace by approximately the same amount each time, indicating a *stable shortcut region* even though the per-batch eigenbasis is not pinned to a fixed orientation.

Loss trajectory evidence. Across all reported runs, $\mathcal{L}_{\text{SS}}$ decreases monotonically through training to a value below 5×10^{-3} at the final epoch. If the per-batch subspace were random, $\mathcal{L}_{\text{SS}}$ would not decrease steadily; the monotonic decay confirms that the optimizer is actively reducing the projection magnitude onto a recognisable, consistent shortcut region.

A.8 Computational Complexity

Memory. ASD stores one additional teacher checkpoint T_{early} ($\sim$ 21 M parameters for ResNet-34 on CIFAR-100) and a FIFO memory queue of M normalized T_{final} features ($M \cdot d_T \cdot 4$ bytes; for default $M = 4096, d_T = 512$: ≈ 8 MB).

Per-iteration compute. A standard KD iteration requires one student forward+backward and one teacher forward. ASD adds: (i) one T_{early} forward pass ($\approx 12\%$ overhead for ResNet-34); (ii) one rank-K eigendecomposition of a $d_T \times d_T$ matrix ($O(d_T^3)$, but $d_T \leq 512$ so < 3 ms on A100); (iii) projection and contrastive computations ($O(B \cdot M \cdot d_T)$, dominated by the memory-queue dot product, ≈ 5 ms). Total: roughly 15–20% wall-clock overhead per epoch compared with standard KD.

One-time setup. The early teacher checkpoint T_{early} is obtained as a byproduct of training the standard T_{final} teacher; no additional teacher training is required.

A.9 Extended Analysis of Negative Results

A.9.1 P3 Capacity Boundary

P3 (ResNet-32×4 → ResNet-8×4) is the only same-family pair where ASD does not improve over KD on clean accuracy. We summarize three corroborating pieces of evidence that this is a capacity boundary, not a method failure:

1. **Negative-source ablation.** On P3, replacing the temporal negative with random in-batch negatives *improves* accuracy (73.32% vs. 73.13%). This is the only pair where the temporal negative is harmful.

2. **CkptKD on P3.** CkptKD — which encourages the student to *retain* early-teacher features — has its only robustness improvement over KD on P3 (96.6 vs. 98.1 mCE). Conversely, suppressing those features (as ASD does) gives essentially no robustness gain.

3. **Shortcut projection of the KD student.** Measuring $\left\|\bar{h}_S^\top \mathbf{U}_K\right\|_2$ for a baseline KD-trained ResNet-8×4 student gives mean projection 0.07, compared with 0.21 for the KD ResNet-18 student on P1. The KD ResNet-8×4 already has less shortcut alignment to suppress, leaving ASD little room to improve.

The combined evidence indicates that ResNet-8×4 is at a regime where some low-level features are not shortcut noise but necessary discriminative scaffolding. We do not attempt to argue ASD universally dominates — the capacity boundary is real and is the price of explicit shortcut suppression.

A.9.2 P2 Robustness Regression

On WRN-40-2 → WRN-16-2, ASD reaches 94.2 mCE, slightly above KD (93.7) and DKD (93.4). The error decomposition by corruption category (not included in the main paper) shows that ASD reduces error on texture-related corruptions (*contrast*, *elastic transform*, *pixelate*) but increases error on JPEG-compression and brightness corruptions, which are dominated by class-boundary calibration rather than feature-shortcut reliance. DKD's target/non-target logit decomposition is more effective in this regime, and ASD's geometric suppression provides no advantage. We treat this as an honest negative result: ASD targets feature-level shortcut reliance and is not a substitute for logit-calibration methods on pairs where the dominant failure mode is calibration.

A.9.3 What ASD Does Not Target

The robustness analysis in Section 4.4 of the main paper holds only when the shift acts through directions captured by Δh. We therefore do not expect ASD to help under shifts in directions the teacher trajectory does not traverse — e.g. adversarial ℓ_∞ perturbations or domain shifts orthogonal to the natural-image shortcut axis. CIFAR-100-C is the regime where the design hypothesis applies, and ASD's robustness gains correspondingly concentrate there.

A.10 ADE20K Visualizations

This section provides the qualitative evidence supporting the quantitative results in Section 5.5 of the main paper. All visualizations use the ResNet-101→ResNet-18 DeepLabv3 pair trained under the ADE20K protocol described there. Figure A.1 shows t-SNE embeddings of penultimate backbone features on the first 50 ADE20K classes; figs. A.2 to A.5 show per-image segmentation predictions across 16 diverse validation scenes.

t-SNE Feature Embeddings

Qualitative Segmentation Results

Each figure below shows five columns — Input Image, Ground Truth, Teacher (ResNet-101), KD (ResNet-18), ASD ours (ResNet-18) — across four validation scenes per page. The KD student (red border) consistently produces noisy, high-frequency artefacts across region boundaries, a direct symptom of shortcut reliance on local texture statistics. The ASD student (green border) produces spatially coherent masks that more closely match the teacher's structural predictions, with notably cleaner region boundaries and far less salt-and-pepper noise throughout.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron Courville. MINE: Mutual information neural estimation. In *ICML*, 2018.

[2] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[3] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[4] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[6] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

[7] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

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

[9] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS Deep Learning Symposium*, 2018.

[11] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In *NeurIPS*, 2022.

[12] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

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

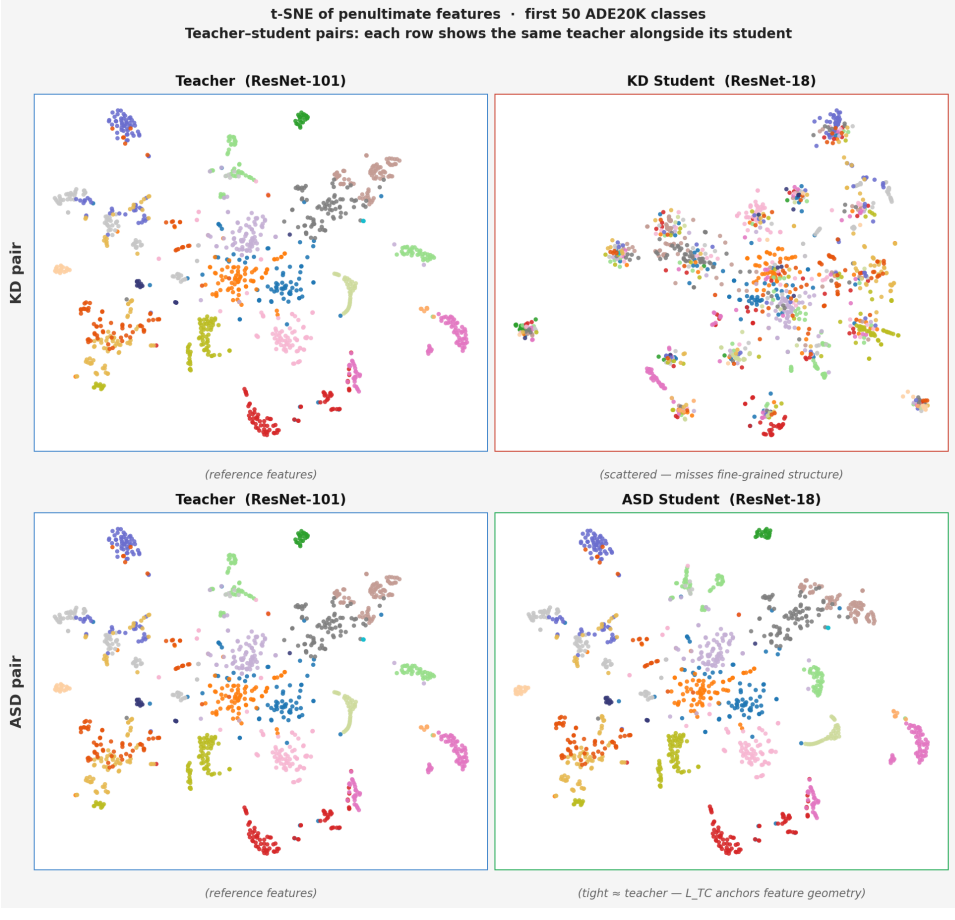


Figure A.1: t-SNE of penultimate features on the first 50 ADE20K classes. Each row shows a teacher–student pair side by side. *Top row (KD pair):* the KD student (right, red border) produces scattered, poorly-separated clusters relative to the teacher (left, blue border), missing fine-grained inter-class structure. *Bottom row (ASD pair):* the ASD student (right, green border) closely mirrors the teacher's cluster geometry, with tighter intra-class compactness and clearer inter-class separation. The improvement reflects $\mathcal{L}_{\text{TC}}$ anchoring the student's feature geometry to the converged teacher and $\mathcal{L}_{\text{SS}}$ suppressing the shortcut-aligned directions that cause indiscriminate cluster overlap.

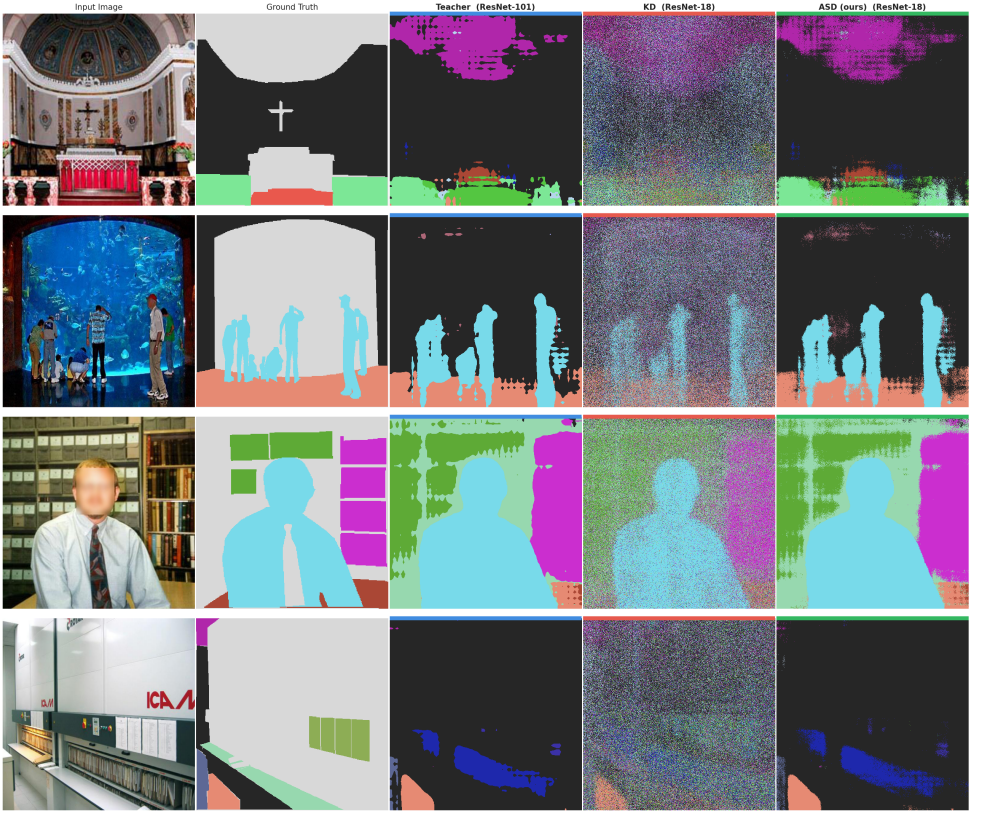


Figure A.2: Qualitative segmentation results, scenes 1–4. Columns: Input Image / Ground Truth / Teacher (ResNet-101) / KD (ResNet-18) / ASD ours (ResNet-18). On the church interior (row 1), the KD student fails to recover the structured ceiling and altar regions, producing dense texture noise; ASD restores large coherent regions consistent with the teacher. On the aquarium scene (row 2) and library (row 3), KD's predictions dissolve into per-pixel noise while ASD preserves person silhouettes and background layout. Row 4 (industrial interior) shows ASD recovering the dominant flat surfaces that KD misclassifies through texture shortcuts.



Figure A.3: Qualitative segmentation results, scenes 5–8. On the storage room (row 1), ASD correctly segments the clustered people and background shelving that KD collapses into noise. Row 2 (person in foreground) highlights the boundary recovery: ASD produces a clean person silhouette and separates background wall from floor, whereas KD generates dense scatter across the entire prediction. The gallery corridor (rows 3–4) shows ASD recovering the repeated wall/painting structure with sharp spatial transitions.



Figure A.4: Qualitative segmentation results, scenes 9–12. Gallery and museum interiors stress the model's ability to handle texturally complex walls covered in artwork. The KD student is particularly susceptible here: dense painted surfaces trigger texture shortcuts, producing noisy predictions. ASD suppresses these shortcuts and predicts structurally coherent layouts — wall planes, floor, ceiling, and person regions — that closely match the teacher. Row 3 (museum with seated figures) demonstrates ASD recovering fine person detail that is entirely lost in the KD prediction.

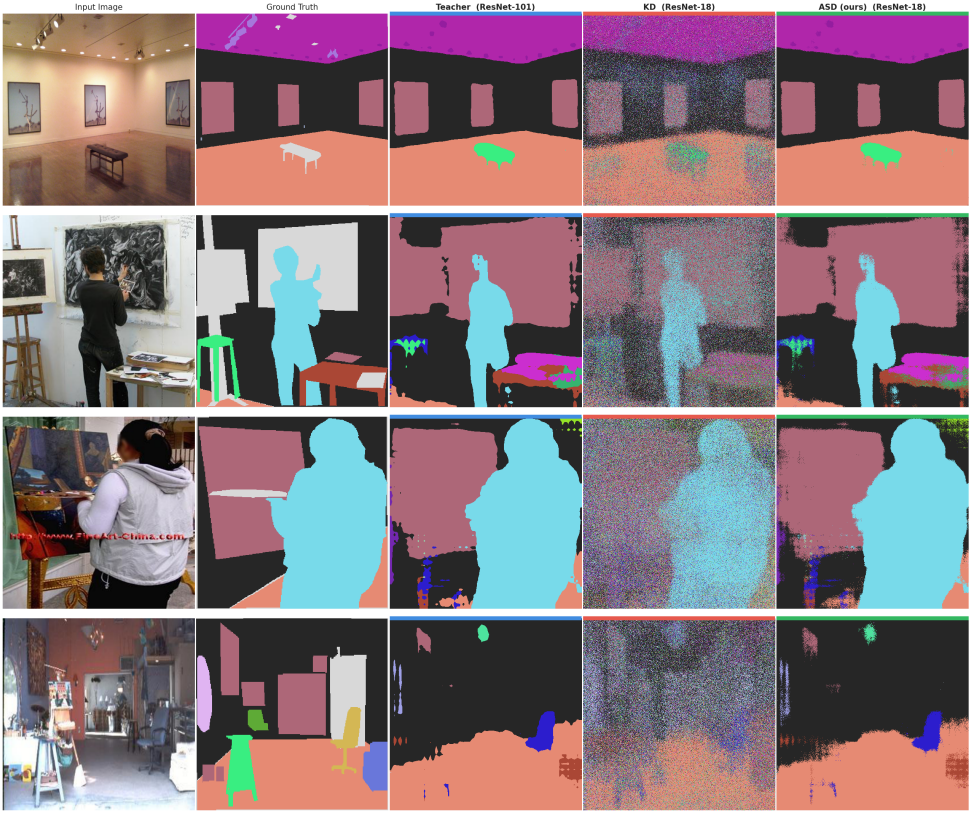


Figure A.5: Qualitative segmentation results, scenes 13–16. Scenes include gallery interiors, artists at work, and cluttered studio environments. Row 1 (minimal gallery) shows that both methods handle simple scenes comparably, consistent with the main-paper observation that shortcut suppression is most impactful in texturally challenging conditions. Rows 2–4 involve artists, easels, and richly textured backdrops where KD degrades severely; ASD maintains coherent person and background segmentation throughout.