RESIDUAL ROUTING OVER SHAPLEY PRIORS: STUDENT-ADAPTIVE MULTI-TEACHER DISTILLATION UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-teacher knowledge distillation (MKD) promises to transfer the complementary expertise of specialised models to a compact student and, with it, better out-of-distribution (OOD) detection. We report an overlooked failure mode of this promise: measured against the label-only supervised student that MKD studies usually omit, static per-sample teacher weighting, whether derived from exact Shapley attribution or from confidence and entropy priors, does not improve the student's OOD detection and in most cases degrades it. The cause is not imperfect expert identification but the rigidity of a static allocation that cannot adapt to the student's learning state. We propose Shapley-Anchored Residual Routing (SRR), which keeps a fixed teacher prior and trains a lightweight router, jointly with the student, to predict a per-sample residual gate and a mixing coefficient, using no outlier data for training or checkpoint selection. Across OOD benchmarks SRR raises mean energy-score AUROC by 3.5% ($+3.0$ points) over uniform distillation and is the only distillation variant that matches a budget-matched supervised student instead of falling below it, with routing gains of $+1.0$ to $+1.7$ points on label-conditioned and label-free priors alike and the feature-space readout left intact. The routing coefficient is also a deployable dividend: a lightweight probe that regresses it from frozen features, even of an untouched supervised student, matches feature-space k-NN detection (2.7–4.0% above the energy score) with about $350\times$ less storage and no per-query feature-bank retrieval.

1 INTRODUCTION

Deploying deep networks in safety-critical settings demands both in-distribution (ID) accuracy and dependable out-of-distribution (OOD) detection, read from the deployed model through logit-space scores such as the energy (Hendrycks & Gimpel, 2017; Liu et al., 2020) or feature-space scores such as k-nearest-neighbour distance (Sun et al., 2022). Deep ensembles are among the strongest detectors (Lakshminarayanan et al., 2017), but their cost grows with the number of members, and multi-teacher knowledge distillation (MKD) has become the appealing alternative (Hinton et al., 2015; You et al., 2017; Malinin et al., 2020): a pool of teachers with complementary expertise, combined well, should transfer richer supervision, and with it better uncertainty, to a single student at single-model cost. Class-specialist pools make the promise concrete. Each teacher excels on its home classes and is confidently wrong elsewhere, so a sample-dependent gate that routes each input to the teachers that know it is widely expected to transmit sharper supervision than uniform averaging, whether the gate comes from prediction entropy (Cheng et al., 2024), true-class confidence (Zhang et al., 2022), or cooperative-game credit assignment with Shapley values (Shapley, 1953; Rozemberczki et al., 2022). Relative to uniform distillation the expectation is met: exact per-sample Shapley values recover the coverage structure of a specialist pool without observing teacher identity, and their non-uniformity statistic q_{shap} orders the gain over uniform averaging across pools of dialled specialisation (Section 4).

Whether the promise holds hinges on the baseline. MKD studies usually benchmark a gating design against uniform distillation and leave the label-only supervised student unexamined. Against that reference, under both an energy-score and a feature-space readout, a different picture emerges: *no static teacher gate, whether uniform, log-probability, entropy, exact Shapley, or one of three recent*

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.

expert-mixture methods, produces a student whose energy-score OOD detection reaches that of the same architecture trained from labels alone (0.855–0.877 vs. 0.880 mean AUROC). The four expert-mixture baselines lower the energy readout by 2.2 points on average, and most static variants also lose on the feature-space readout.

The collapse is not a failure of expert identification. For our value function, exact per-sample Shapley attribution and elementary log-probability weighting rank the teachers identically on every input (Proposition 1), and both recover the coverage design. The breakdown stems from the rigidity of a static allocation: a fixed, label-derived prior settles which teachers understand a sample in the abstract, but cannot say *how far the student, at its current learning state, should follow that supervision*.

To resolve this mismatch we propose **Shapley-Anchored Residual Routing (SRR)**. Rather than learning a gate from scratch or following a static allocation, SRR anchors on a fixed teacher prior and trains, jointly with the student, a lightweight router on the student's penultimate features; the router predicts a candidate gate and a per-sample mixing coefficient $r(x) \in [0, r_{\max}]$ that interpolates between the prior and the student-adaptive gate. Under our primary protocol neither training nor checkpoint selection uses OOD or pseudo-unknown data. Three findings organise the paper.

- **A family-level collapse and its decomposition (Figure 1, Table 1).** Against the supervised reference the four expert-mixture baselines lose 2.2 points on average while the gating and routing variants average $+0.04$. Writing the energy change as a feature term (Δk-NN) plus a logit-readability term ($-\Delta\text{gap}$) shows that the static Shapley gate keeps features ($+0.20$) but loses readability (-0.55). Residual routing is the only variant that keeps features while turning readability positive ($+0.62$): it raises mean energy AUROC to 0.888, a 3.5% **relative gain** ($+3.0$ **points) over uniform distillation**, and is the **only distillation variant that matches the supervised student at equal training budget** (0.888 vs. 0.889; $+0.7$ over the 100-epoch supervised student, intervals in Section 5), with k-NN AUROC and ID accuracy within 0.2 points and the lowest FPR@95 of all distillation variants.
- **The prior is an input; the correction is the method (Table 2).** Anchoring the same router on three priors, two label-conditioned (Shapley, LogP) and one label-free (entropy), yields routing gains of $+1.1$**,** $+1.0$ **and** $+1.7$ **points** over the respective static gates; the ordering of the priors is preserved, so the correction repairs the student–prior mismatch rather than a particular gating rule.
- **A deployable dividend: a bank-free detector (Tables 3, 4).** The training-time coefficient $r(x)$ is normally discarded. A lightweight probe fitted to frozen features on ID data alone reproduces it and scores 0.912 **AUROC on the routed student and** 0.915 **on an untouched supervised student**, matching feature-space k-NN (0.916 / 0.915) with a 2.7–4.0% gain over the energy score, at about $350\times$ less storage than the feature bank and the latency of the energy score.

Supporting analyses show that the router's intervention is selective, that the gain is not explained by training budget or checkpoint selection, and that no distillation variant beats the supervised student on a compact MobileNetV2, which we report as a limitation (Section 7).

2 Related Work

OOD detection and its readouts. Post-hoc scores read a trained classifier without retraining: maximum softmax probability (Hendrycks & Gimpel, 2017), energy (Liu et al., 2020), max-logit (Hendrycks et al., 2022), GEN (Liu et al., 2023), and feature-space methods such as Mahalanobis (Lee et al., 2018), k-NN (Sun et al., 2022), ViM (Wang et al., 2022), ReAct (Sun et al., 2021) and ASH (Djurisic et al., 2023). Feature-space scores typically exceed logit-space scores by a few points on the same model (Sun et al., 2022; Lafon et al., 2023); we report both and decompose their difference rather than treat them as independent readouts. Training-time methods add auxiliary outliers (Hendrycks et al., 2019) or reshape logits (Wei et al., 2022). Learned-confidence branches (DeVries & Taylor, 2018; Corbière et al., 2019) regress a model's own success signal from its features; our routing head shares the mechanism but regresses a quantity produced by a training-only module, and regressing the student's own confidence does not reproduce its behaviour.

Ensembles remain a strong reference (Lakshminarayanan et al., 2017); ensemble distillation transfers their uncertainty to one student (Malinin et al., 2020; Fathullah & Gales, 2023).

Multi-teacher distillation. Early multi-teacher work averages teachers (Fukuda et al., 2017; You et al., 2017). Per-sample weighting uses the teacher's confidence on the true label (CA-MKD; Zhang et al., 2022), prediction entropy with decoupled KD (Zhao et al., 2022) (DE-MKD; Cheng et al., 2024), or a reinforcement-learning agent rewarded by the student's loss (MTKD-RL; Yang et al., 2025); BOMD adds bi-level optimisation with feature-level alignment (Gong & Wen, 2024). All of these produce a gate that is a function of the teachers and the label. Ours is a function of the *student*: a residual over a fixed prior, in the spirit of mixture-of-experts gating (Jacobs et al., 1991; Shazeer et al., 2017), and we show the residual matters more than the choice of prior. A rarely reported control in this literature is the label-only student; we make it the reference point of every table.

Shapley values in machine learning. The Shapley value is the unique credit assignment satisfying efficiency, symmetry, linearity and the null-player axiom (Shapley, 1953), used for feature attribution (Lundberg & Lee, 2017) and data valuation (Ghorbani & Zou, 2019); see Rozemberczki et al. (2022). We apply it per sample over teachers, as one of three priors (two label-conditioned, one label-free), and we show that under our value function its teacher ordering coincides with a much cheaper log-probability gate.

3 Setup, priors, and protocol

Distillation template. We are given M teachers $\{t_i\}_{i=1}^{M}$ over C in-distribution (ID) classes and a training set $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^{N}$. Teacher i produces logits $z_i(x)$ and a softened distribution $p_i(x) = \mathrm{softmax}(z_i(x)/\tau)$. Every distillation method below produces a per-sample gate $g(x) \in \Delta^{M-1}$ and trains the student s on

$$p_T(x) = \textstyle\sum_i g_i(x)\, p_i(x), \qquad \mathcal{L}_{\mathrm{KD}}(g) = \mathrm{CE}\big(y, p_s(x)\big) + \lambda_{\mathrm{kd}}\, \tau^2\, \mathrm{KL}\big(p_T(x) \,\|\, p_s(x)\big). \quad (1)$$

The *supervised* student uses the same recipe with $\lambda_{\mathrm{kd}} = 0$ and is the reference of every comparison. At test time only the student runs; the default OOD score is the energy $E(x) = \mathrm{logsumexp}\, z_s(x)$ (Liu et al., 2020), and we always also report k-NN on the student's penultimate features (Sun et al., 2022) as a second, feature-space readout.

Specialist pools. Pools contain $M = 5$ ResNet-18 teachers that share architecture and budget and differ only in which classes each sees. In the **cov**k pool each teacher trains on a cyclic block of $2k$ of the 10 CIFAR-10 classes so that every class is covered by exactly k teachers (cov2, cov3, cov4); the **generalist** pool trains all five on all classes. Our main pool is cov3. Each cov3 teacher reaches 93–97% on its home classes and 0% elsewhere while remaining confident (0.85–0.91 maximum probability on non-home inputs); the uniform ensemble of the five reaches 94.5% accuracy and an energy mean AUROC of 0.897.

Two label-conditioned priors. The *Shapley gate* treats teachers as players with coalition value $v(S; x) = -\mathrm{CE}\big(y, \frac{1}{|S|}\sum_{i\in S} p_i(x)\big)$, $v(\emptyset; x) = -\log C$, computes each teacher's exact per-sample Shapley value $\phi_i(x)$ by enumerating the 2^M coalitions, and sets $g^{\mathrm{S}}(x) = \mathrm{softmax}(\phi(x)/\tau_g)$. Its non-uniformity $q_{\mathrm{shap}}(x) = 1 - H(g^{\mathrm{S}}(x))/\log M$, averaged over $\mathcal{D}$ at fixed τ_g, is a statistic of the pool available before any student is trained. The *LogP gate* is the one-line alternative $g^{\mathrm{P}}(x) = \mathrm{softmax}\big(\log p_i(y \mid x)/\tau_g\big)$, a softmax over each teacher's log-probability of the true label.[1]

Proposition 1 (Ordering). *For the value function above, the order of $\phi_i(x)$ over teachers equals the order of $p_i(y \mid x)$ for every x (proof in Appendix C).*

The two priors therefore agree on which teachers to trust on every sample and differ only in how sharply they weight them; on cov3 both concentrate on the three home teachers and leak $\leq 2\%$ of

[1] This is the gate that an earlier version of this work labelled CA-MKD. The original CA-MKD weight is $(1 - \mathrm{softmax}_i(\mathrm{CE}_i))/(M-1)$ without a temperature and caps every weight at $1/(M-1)$; we now report the faithful implementation under the name CA-MKD and the softmax-of-log-probability gate as LogP (Appendix B).

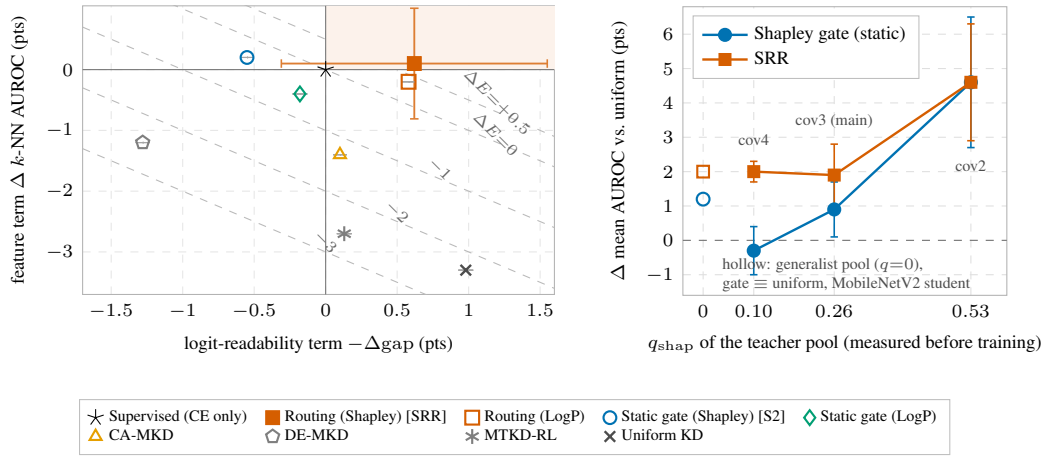


Figure 1: **Left: decomposition of the energy gain.** With $\text{gap} = \text{AUROC}_{k\text{-NN}} - \text{AUROC}_E$ per row, every row satisfies $\Delta E = \Delta k\text{-NN} + (-\Delta\text{gap})$: a *feature term* ($y$: what k-NN reads from the penultimate features) plus a *logit-readability term* (x: how much of it the classifier head passes to the energy score). Dashed anti-diagonals are iso-ΔE lines; the 100-epoch supervised student is the origin. The expert-mixture baselines lose 1.2–3.3 points on the feature term; the static Shapley gate keeps features ($+0.20$) but loses readability (-0.55); the two routed rows (filled) are the only ones with a positive readability term ($+0.62$, $+0.58$). ResNet-34, cov3, means over four OOD sets and three seeds. Error bars on the routed Shapley point are paired 95% intervals over seeds: vertical, Δk-NN ± 0.91; horizontal, ΔE ± 0.93 placed on the readability axis at fixed Δk-NN. **Right: static gating looks appealing against the wrong reference.** Over the specialisation dial the static Shapley gate's gain over *uniform* distillation grows with q_{shap} (-0.3, $+0.9$, $+4.6$ points at $q_{\text{shap}} = 0.10, 0.26, 0.53$), yet on the main pool it still ends 0.35 points below the supervised student.

the mass elsewhere (Figure 2 in Appendix D). Exact enumeration (< 1 min at $M = 5$) adds no ordering information the LogP gate lacks; Section 6.1 says why we still anchor on it. As a label-free control, Section 6.1 also uses the *entropy gate* $g^{\text{E}}(x) = \text{softmax}\big(-H(p_i(x))/\tau_g\big)$, the weighting of DE-MKD (Cheng et al., 2024) without its decoupled loss, which trusts whichever teacher is most certain.

Protocol and statistics. ID data is CIFAR-10 (45k/5k train/validation split); OOD test sets are SVHN, CIFAR-100, TinyImageNet and DTD. Every checkpoint is selected by ID validation accuracy; no OOD or pseudo-OOD data enters training or selection (an OOD-aware selection variant is analysed in Appendix A). All methods use the same three seeds $\{42, 123, 3407\}$, and every comparison against the supervised student is *paired by seed*: for method m and seed s we form $d_s = \frac{1}{4}\sum_j \big(\text{AUROC}_{m,s,j} - \text{AUROC}_{\text{CE},s,j}\big)$ over the four OOD sets and report the mean of d_s with a 95% t-interval over seeds. We call a difference an *improvement* only if the interval excludes zero, and two methods *equivalent within* δ on a readout if two one-sided tests at the 5% level reject $|d| \geq \delta$, i.e. the 90% interval lies inside $(-\delta, \delta)$; we pre-set $\delta = 0.5$ points and report the smallest $\delta \in \{0.5, 1\}$ attained. Four OOD sets are not four seeds and are never used to inflate significance: the sampling error of a single OOD test set is 0.4–0.5 points, but a paired comparison on the same test set reduces it to 0.05–0.11, so seeds, not test sets, are the bottleneck; a 95% half-width of 0.3 points needs 5, 7, 10 or 13 seeds for a seed standard deviation of 0.2, 0.3, 0.4 or 0.5 points, and differences of 0.1–0.2 points are only ever reported as equivalences. The k-NN readout follows Sun et al. (2022): the bank is the penultimate, global-average-pooled 512-d features of the 45,000 unshuffled ID training images, ℓ_2-normalised; the score is the cosine similarity to the $k = 50$-th nearest bank feature. Baseline details are in Appendix B.

Table 1: **Main comparison** (ResNet-34, cov3, val-acc selection, 3 seeds). Energy AUROC averaged over four OOD sets (SVHN, CIFAR-100, TinyImageNet, DTD) and over the three semantic sets (CIFAR-100, TinyImageNet, DTD); FPR@95 and k-NN AUROC over four sets; $-\Delta\text{gap}$ is the logit-readability term of Figure 1; "Probe" is the routing-head score of Section 6.2. Parenthesised values are differences to the 100-epoch supervised student in points. Static rows train for 100 epochs; the routed rows add SRR's 30+50-epoch routing stage, so the supervised student continued for 50 epochs is their budget-matched reference. Routing (Shapley) vs. supervised, paired over three seeds (paired sd 0.37): 95% CI $[-0.21, +1.65]$ on energy and $[-0.81, +1.01]$ on k-NN; other differences are point estimates; "–": not computed. †Faithful CA-MKD weights (Zhang et al., 2022). Per-set numbers in Appendix F.

Method	ID acc	Energy, 4 sets	Energy, 3 sets	FPR@95	k-NN, 4 sets	$-\Delta$gap	Probe
Supervised (CE only)	94.9	0.880	0.874	**0.429**	0.915	–	0.915
Supervised, 150 ep.	95.0	**0.889** (+0.9)	–	–	–	–	–
Routing (Shapley) [SRR]	94.7	0.888 (+0.72)	**0.883** (+0.93)	0.511	0.916 (+0.10)	+0.62	**0.912**
Routing (LogP)	94.6	0.884 (+0.38)	0.879 (+0.50)	0.515	0.913 (−0.20)	+0.58	–
Static gate (Shapley) [S2]	94.6	0.877 (−0.35)	0.873 (−0.07)	0.525	**0.917** (+0.20)	−0.55	–
Static gate (LogP)	94.5	0.875 (−0.58)	0.875 (+0.14)	0.535	0.911 (−0.40)	−0.18	–
CA-MKD†	94.3	0.867 (−1.30)	–	0.542	0.901 (−1.40)	+0.10	–
DE-MKD	93.6	0.856 (−2.48)	0.845 (−2.83)	0.539	0.903 (−1.20)	−1.28	–
MTKD-RL	94.1	0.855 (−2.58)	0.837 (−3.69)	0.525	0.888 (−2.70)	+0.13	–
Uniform KD	94.1	0.857 (−2.33)	0.843 (−3.11)	0.524	0.882 (−3.30)	+0.98	–

4 STATIC GATING DOES NOT BEAT THE SUPERVISED STUDENT

Lead finding: a family-level effect. Grouping the eight distillation rows of Table 1 by what they do with the teacher pool separates them cleanly. The four expert-mixture baselines (uniform KD, CA-MKD, DE-MKD, MTKD-RL) lower the energy readout by 2.17 points on average relative to the supervised student (range -2.58 to -1.30); the four gating and routing variants (static gates on the Shapley and LogP priors, and residual routing on either) average $+0.04$ (range -0.58 to $+0.72$). The 2.2-point separation between the two families is four times the sampling error of a single OOD test set and is the largest effect in the table. The label-free entropy gate of Section 6.1 (-2.52) falls in the first family: the split is about label-conditioning, and mixing specialist teachers without it costs OOD detection.

The appeal of static gating. Relative to uniform averaging, per-sample gating does what it promises. On cov3 every training sample has $q_{\text{shap}} \geq 0.17$ (mean 0.259), the Shapley gate assigns home teachers 0.326 on average (ideal $1/3$) versus 0.011 elsewhere, and the student distilled through it gains $+2.0$ energy points over uniform KD (0.877 vs. 0.857). Over the specialisation dial the gain over uniform is monotone in q_{shap} (Figure 1, right; Table 5 in Appendix D): $+4.6$ points at $q_{\text{shap}} = 0.53$, $+0.9$ at 0.26, -0.3 at 0.10, and zero by construction when the gate is uniform. A gate-geometry intervention (Appendix E) shows the gain needs both the graded home-teacher weights and the small non-home leak.

The ceiling, on two averages. Measured against the supervised student, no static gate is an improvement: the label-only ResNet-34 reaches 0.880 energy AUROC, and the static Shapley gate ends 0.35 points below it, the LogP gate 0.58 below, and the four baselines 1.3–2.6 below. The four-set mean deserves one caveat. The SVHN column is nearly orthogonal to the other three (its correlation with CIFAR-100, TinyImageNet and DTD across the nine rows is -0.16, -0.20 and -0.32, whereas those three correlate at 0.85–0.98), and the method that maximises it, MTKD-RL (0.907), is the worst on the mean. Table 1 therefore also reports the three-set mean over the semantic OOD sets. The top tier (supervised, the two routed rows, the two static gates) is the same on both averages, and the routed Shapley row's margin over the supervised student grows from $+0.72$ to $+0.93$. Three explanations are excluded: expert identification (by Proposition 1 the Shapley and LogP gates rank teachers identically and both recover the coverage design, while the faithful CA-MKD gate, forced to put about 25% of its mass on non-home teachers, does worse); budget (static rows and the supervised student train for 100 epochs; the supervised student continued for our rout-

ing stage's 50 extra epochs reaches 0.889, so budget alone is worth at most +0.9, the reference used for the routed rows in Section 5); and the pool (its uniform ensemble scores 0.897).

Two readouts, one decomposition. k-NN on the penultimate features exceeds the energy score on every row, by 3.43 points on average (sd 0.64), and the two columns are strongly correlated across rows (Pearson 0.86, Spearman 0.80); the size of the gap is a known property of feature-space detectors (Sun et al., 2022; Lafon et al., 2023), and the choice of readout (3.5 points) moves AUROC more than any single method does (≤ 2.6). We therefore decompose rather than compare: with $\text{gap} = \text{AUROC}_{k\text{-NN}} - \text{AUROC}_E$,

$$\Delta E \;=\; \underbrace{\Delta k\text{-NN}}_{\text{feature term}} \;+\; \underbrace{(-\Delta\text{gap})}_{\text{logit-readability term}} \quad , \tag{2}$$

where Δ is the difference to the supervised student (Figure 1, left; last columns of Table 1). The identity holds row by row, and it sorts the failures. The expert-mixture baselines lose on the feature term (-1.2 to -3.3): distillation from the mixture degrades the representation itself, and no readout recovers it. The static Shapley gate does the opposite: it improves the feature term (+0.20, the best k-NN in the table) but loses 0.55 on readability, so the energy score does not see the better features; the LogP gate loses a little on both. Only the two routed rows have a positive readability term (+0.62 and +0.58) while holding the feature term within 0.2 of the supervised student.

5 Residual routing: correcting the prior with the student's state

A fixed, label-derived prior answers "which teachers know this sample" but not "how far should *this* student, now, follow them". SRR learns the second question as a residual over the first.

Router and blend. A two-head MLP reads the student's penultimate feature $f(x)$ (detached, so router losses do not shape the representation) and outputs a candidate gate $g^{\text{c}}(x) = \text{softmax}(Wf(x)/\tau_r)$ and a mixing coefficient $r(x) = r_{\min} + (r_{\max} - r_{\min})\,\sigma(w_r^\top f(x))$. With a prior $g^0 \in \{g^{\text{S}}, g^{\text{P}}\}$, the distillation gate is

$$g(x) = \big(1 - r(x)\big)\, g^0(x) + r(x)\, g^{\text{c}}(x), \tag{3}$$

so $r(x) \to 0$ follows the prior and $r(x) \to r_{\max}$ follows the student-adaptive gate ($r_{\max} = 0.4$ throughout).

Joint objective. Student and router are trained together on

$$\begin{aligned}\mathcal{L}_{\text{SRR}} = \mathcal{L}_{\text{KD}}(g) + \lambda_{\text{s}}(t)\, \underbrace{q(x)\, \text{CE}\big(g^0(x), g^{\text{c}}(x)\big)}_{\mathcal{L}_{\text{sup}}} \\ + \lambda_{\text{a}}\, \underbrace{c_a(x)\, \big(1 - r(x)\big)\, \text{KL}\big(p_a(x) \,\|\, p_s(x)\big)}_{\mathcal{L}_{\text{anchor}}} + \lambda_r\, \underbrace{\text{BCE}\big(r(x), r_0\big)}_{\mathcal{L}_r},\end{aligned} \tag{4}$$

where $q(x)$ is the prior's per-sample non-uniformity, so the candidate gate is pulled towards the prior in proportion to how decisive the prior is and the router deviates first where the prior is least sure; λ_{s} anneals from 1.0 to 0.2 ("trust the prior, then the router"). $\mathcal{L}_{\text{anchor}}$ ties the student to a frozen reference p_a pre-trained on the static prior, weighted by the reference's confidence $c_a(x)$ and by $1 - r$, and $\mathcal{L}_r$ is a weak prior $r_0 = 0.2$ on the mixing coefficient. The student starts from the converged static-prior student; the router is pre-trained for 30 epochs and then trained jointly for 50 (Appendix A). At test time router, teachers and prior are discarded.

Result. Routing on the Shapley prior reaches 0.888 energy AUROC and 0.916 k-NN AUROC. Against the 100-epoch supervised student the paired energy difference is +0.72 points (95% CI $[-0.21, +1.65]$, $p \approx 0.08$, three seeds); against the supervised student continued for SRR's extra 50 epochs (0.889) it is -0.15, a tie within seed noise. We therefore claim not that routing beats supervised training but that it is the only distillation variant that does not fall below it at equal budget: every static variant loses 0.35–2.6 points, and continuing the static-prior student for the

Table 2: **Prior × routing** (ResNet-34, cov3, val-acc, 3 seeds; energy mean AUROC over four OOD sets). The identical router, losses and schedule are anchored on three priors: two label-conditioned (Shapley, LogP) and one label-free (entropy). The routing gain is positive on all three and largest on the weakest prior; the level after routing follows the prior, and only the label-conditioned priors end above the supervised reference (0.880).

Prior g^0	static gate	+ residual routing	routing gain	vs. CE (static / routed)
Shapley g^{S} (label-conditioned)	0.877	**0.888**	+1.1	−0.35 / +0.72
LogP g^{P} (label-conditioned)	0.875	0.884	+1.0	−0.58 / +0.38
Entropy g^{E} (label-free)	0.855	0.872	+1.7	−2.52 / −0.78
gap Shapley − LogP	0.2	0.35		
gap Shapley − Entropy	2.2	1.5		

same 50 epochs without a router recovers nothing (0.867 vs. 0.873, two-set mean; Appendix A), so the +1.1 points that the routing stage adds over the static Shapley gate are not budget. SRR is also the only row of Table 1 at or above the 100-epoch supervised student on each of the four OOD sets, +0.93 on the three semantic sets, and lowest in FPR@95 among distillation variants (0.511). On k-NN the difference to the supervised student is +0.10 (95% CI $[-0.81, +1.01]$), equivalent within ±1 point by TOST (Section 3). In the decomposition of Eq. (2) the gain is almost entirely readability: feature term +0.10, readability term +0.62. ID accuracy is 94.7% against 94.9% and 95.0% for the two supervised references: routing gives no ID advantage, and we claim none.

What the router does. On the ResNet-34 student the mixing coefficient is selective: on the val-acc checkpoint its mean is 0.19 on easy ID test inputs, 0.27 on hard ones, 0.31 on CIFAR-100 and 0.35 on SVHN (three seeds; the groups are stochastically ordered, Appendix J). The correction is small in gate space (mean $r \cdot \|g^{\mathrm{c}} - g^{\mathrm{S}}\|_1 \approx 0.02$) yet decisive on the energy readout: the mismatch is about *how much* to follow the prior on hard and unfamiliar inputs, not about *whom*. On a compact MobileNetV2 the router saturates at $r_{\max}$ and learns a constant mixture; there no distillation variant beats the supervised student (Section 7).

6 Empirical validation and deployment analyses

6.1 The correction is not specific to the prior

If the value of SRR came from the Shapley prior being unusually well informed, replacing the prior should remove the gain. It does not (Table 2). Anchoring the identical router, losses and schedule on the LogP prior (initialisation, candidate-gate target and anchor all taken from the LogP student) raises it from 0.875 to 0.884, a routing gain of +1.0 against +1.1 for Shapley; the two routed students differ by 0.35 points (paired per-seed differences −1.15, +0.10, +0.07: a tie). As Proposition 1 predicts, the two anchors differ only in sharpness; we keep Shapley as the default anchor only because it also keeps the feature-space readout (+0.10 vs. −0.20 on k-NN). Anchoring on the label-free entropy gate, whose static student (0.855) lands in the expert-mixture family of Section 4, raises it by +1.7 to 0.872: the largest routing gain of the three, on the weakest prior. In the decomposition of Eq. (2) the two label-conditioned routed rows are the only rows with a positive readability term (+0.62, +0.58).

Two conclusions follow. First, the correction is not specific to the prior: it is positive on a Shapley, a LogP and an entropy anchor, and it repairs what all three share, a fixed allocation that ignores the student's state. Second, routing does not substitute for the prior: the ordering of the three priors is unchanged by routing, the gap between the strongest and the weakest narrows from 2.2 to 1.5 points but does not close, and the routed entropy student stays 0.78 below the supervised reference. The prior sets the floor, routing adds one to two points on top, and only priors that concentrate on the correct experts end above the label-only student. One qualification: the routed LogP student does not keep the k-NN readout (−0.20), so readout consistency in Figure 1 is so far a property of the Shapley anchor.

Table 3: **Routing-head probe** (ResNet-34, cov3, val-acc, 3 seeds; mean AUROC over four OOD sets). Probes are two-layer MLPs (128 hidden units) fitted on frozen penultimate features of the ID training set, regressing either the SRR router's $r(x)$ or the minimum teacher cross-entropy of the sample; no OOD data is used. The probe on the *frozen supervised* student, which never saw a router, reaches the same parity with k-NN at a hold-out fit of only $R^2 = 0.41$.

Student (frozen)	Probe target	Probe AUROC	k-NN	Energy	Retrained?
Routing (Shapley)	router $r(x)$	**0.912**	0.916	0.888	no
Routing (Shapley)	min. teacher CE	0.894	0.916	0.888	no
Supervised	**router $r(x)$ of the routed run**	**0.915**	0.915	0.880	no
Supervised	min. teacher CE	0.906	0.915	0.880	no

Table 4: **Deployment cost** of the detectors in Table 3 (ResNet-34, $d = 512$, $N = 45{,}000$ ID training features). Storage is analytic (fp32). Latency is end-to-end (backbone forward plus detector) for 10^4 CIFAR-10 test images on one RTX 4090 at batch 256, mean$\pm$sd over 5 repeats after warm-up; throughput is 10^4 divided by that time.

Detector	Extra storage	ms per 10^4	img/s	Build cost
Energy	0	668.4 ± 2.5	14,960	none
Routing-head probe	0.26 MB (66K params)	667.9 ± 2.6	14,970	probe fit ($\sim$2 min)
k-NN	92.16 MB ($N \times d$ bank)	686.1 ± 2.6	14,575	bank + neighbour search
Mahalanobis	1.07 MB (C means, d^2 cov.)	670.0 ± 1.9	14,925	covariance inverse

Protocol robustness. Selecting checkpoints by an OOD-aware score instead of validation accuracy lifts the static gates by 1–2 points at an ID cost of 3–6 points, while SRR's energy result is unchanged (0.884 vs. 0.884 on the two-set mean) at an ID cost of 1.4; uniform KD and MTKD-RL also move little (Appendix A). SRR's gain is thus not a selection artefact, and it needs no pseudo-OOD data to be realised.

6.2 The routing coefficient as a feature-bank-free detector

The mixing coefficient $r(x)$ is produced by a module that exists only during training. It is, however, a deterministic function of the student's feature $f(x)$, so it can be kept: we fit a probe h_ψ (two layers, 128 hidden units, 66K parameters) on the frozen student's penultimate features to regress the normalised $r(x)$ over the ID training set, and use $\hat{r}(x) = \sigma(h_\psi(f(x)))$ as an OOD score (higher = more OOD). Fitting takes about two minutes and uses no OOD data; a head co-trained during SRR on detached features gives the same numbers (Appendix H).

Accuracy. On every student in Table 1, feature-space k-NN exceeds the energy score, by 3.4 points on average and by 2.85 on the routed student; a gap of this size is a known property of k-NN on CIFAR-scale students (Sun et al., 2022; Lafon et al., 2023) and is not a contribution of ours. The probe is. On the routed student it scores 0.912 mean AUROC over the four OOD sets (SVHN 0.955, CIFAR-100 0.889, TinyImageNet 0.883, DTD 0.922), recovering 2.4 of the 2.85-point gap to k-NN (0.916) with 66K parameters and no feature bank (Table 3); we read this as parity with feature-space retrieval: the difference is 0.4 points, inside the ± 1-point equivalence margin adopted in Section 3. A probe regressing the student's own maximum softmax probability, or Gaussian noise, does not reproduce this (Appendix H), so the effect is not "any auxiliary head helps".

The routed backbone is not required. The decisive control is the third row of Table 3. The identical probe fitted on the *frozen supervised* student, whose backbone never saw a router, regressing the $r(x)$ of the routed run of the same seed, reaches 0.915 mean AUROC: equal to k-NN on that student (0.915) and 3.5 points above its energy score (0.880), at a hold-out R^2 of only 0.41 against 0.997–0.999 on the routed backbone. The probe transfers the coarse structure of $r(x)$, not its per-sample detail, and the coarse structure carries the OOD signal. Hence the routing head needs no change to the backbone: an off-the-shelf supervised ResNet-34 acquires k-NN-grade detection from

a 66K-parameter head, with no retraining and no neighbour search. The signal still originates in a routed run, which must exist once; a router-free teacher signal, the minimum teacher cross-entropy, gives 0.906 on the same student (fourth row), 2.6 points above energy. The result also closes the loop with Section 4: the supervised student's features already carry separability its head does not read; routing passes it to the energy score during training (Eq. (2)) and the probe reads it out afterwards. We have shown this for one architecture.

6.3 Deployment latency and storage

Table 4 states what parity buys. Storage is the decisive difference: k-NN keeps the $N \times d$ ID feature bank (92 MB at $N = 45$k, $d = 512$), which grows with the training set and ships with the model, whereas the probe adds 66K parameters (0.26 MB, a $350\times$ reduction) and nothing that depends on N. Latency is not decisive on a modern GPU: the ResNet-34 forward pass dominates, all four detectors run at about 670 ms per 10^4 images ($\approx 15{,}000$ images/s), the probe is indistinguishable from energy (667.9 vs. 668.4 ms), and k-NN's blocked GPU search adds 18 ms (2.6%) at this bank size. The routing head thus matches feature-bank retrieval at the inference cost of the energy score, with no bank, no per-query search, and memory and compute that stay $O(1)$ in N where k-NN's are $O(N)$.

7 Discussion and limitations

What is and is not transferred. The specialist pool carries OOD-relevant knowledge (its ensemble scores 0.897) and every static gate identifies who knows what, yet no static student beats the label-only student. What routing adds is a per-sample, per-training-state decision on how far to follow the prior; by Eq. (2) it acts on the readability of the features by the head, not on the features, and it is the only intervention in the table that does so; at equal budget it restores parity with supervised training rather than exceeding it. The probe makes the same point from the other side: a supervised student's features already carry k-NN-grade separability its head does not read, and a 66K-parameter head reads it out without touching the backbone. We do not claim that Shapley attribution is irreplaceable; we claim that a fixed prior must pass through a student-adaptive correction to be useful, and that the correction's coefficient is itself a deployable signal.

Where the gain vanishes. On MobileNetV2 the supervised student (0.817 four-set energy) is not beaten by any distillation variant (SRR 0.813, uniform 0.803, Shapley gate 0.803), and the router saturates. A natural hypothesis is that residual correction compensates a capacity or budget gap between student and teachers; testing it needs a dial over student capacity or CE budget, which we have not run. Specialisation also breaks disagreement-based detection with the teachers themselves (mutual information AUROC 0.44 vs. 0.92 on a generalist pool; Appendix L).

Scope. Five teachers of one architecture, CIFAR-scale data, exact 2^M Shapley enumeration, three seeds. The family-level effect (2.2 points) is supported at any seed count; the 0.7-point gap to the 100-epoch supervised student has a paired spread of 0.37 points, so three seeds leave its interval touching zero, and the budget-matched comparison is a tie; effects of 0.1–0.2 points can only be stated as equivalences. Hyper-parameters ($r_{\max} = 0.4$, $\lambda_a = 0.5$, $r_0 = 0.2$, the annealing schedule) are empirical constants fixed once on the controlled setup and reused unchanged across priors, students and pools (Appendix J).

8 Conclusion

Against the label-only student that most multi-teacher studies omit, static per-sample gating of class-specialist teachers costs OOD detection, however well it identifies the experts. A residual router that decides per sample how far the student should follow a fixed prior removes that cost, matching a budget-matched supervised student without sacrificing the feature-space readout, for three priors alike; and its mixing coefficient can be refit from frozen features, even of an untouched supervised student, as a detector on par with k-NN without a feature bank. The prior is an input; the correction is the method; the coefficient is worth keeping.

ETHICS STATEMENT

This work uses only public benchmark datasets (CIFAR-10/100, SVHN, TinyImageNet, DTD) and involves no human subjects or personal data. Better OOD detection reduces silent failures of deployed classifiers; we see no direct negative societal impact beyond those generic to image classification.

REPRODUCIBILITY STATEMENT

All numbers are read from logged metric files of the released training and evaluation code (anonymised repository provided as supplementary material). Appendix A lists every hyperparameter, the class-coverage designs, the selection protocol and seeds; Appendix B describes the baseline implementations, including the distinction between the LogP gate and faithful CA-MKD; Appendix F gives per-set results behind every mean in the main text; the figure data files and the script that regenerates them from raw outputs are included.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for research ideation and execution: proposing and refining hypotheses and experiment designs (the supervised control, the selection-protocol, gate-geometry and specialisation-dial controls, the third prior and the frozen-student probe), implementing methods and experiment orchestration code (training, evaluation, result-aggregation and benchmarking scripts, all run and checked by the authors), and interpreting results in discussion with the authors. We also used generative AI tools to formulate Proposition 1 and to draft its proof (Appendix C), which the authors verified line by line. We have not used generative AI tools to generate synthetic datasets or to clean or reformat datasets; translation, survey or interview design, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to identify and summarise related literature, to draft and edit sections of the paper and improve its readability, to create and modify figure scripts and software code, and to format references. We have reviewed all AI-assisted work: every reported number was traced by the authors to a logged metric file, all AI-edited code was executed end-to-end by the authors, the proof was checked by the authors, and all citations were checked against the original publications. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Xin Cheng, Zhiqiang Zhang, Wei Weng, Wenxin Yu, and Jinjia Zhou. DE-MKD: Decoupled multi-teacher knowledge distillation based on entropy. *Mathematics*, 12(11):1672, 2024. doi: 10.3390/math12111672.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Yassir Fathullah and Mark J. F. Gales. Logit-based ensemble distribution distillation for robust autoregressive sequence uncertainties. In *Uncertainty in Artificial Intelligence (UAI), PMLR 216*, 2023.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, 2017.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019.

Shuyue Gong and Weigang Wen. Bi-level orthogonal multi-teacher distillation. *Electronics*, 13(16): 3345, 2024. doi: 10.3390/electronics13163345.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2015. arXiv:1503.02531.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Marc Lafon, Elias Ramzi, Clément Rambour, and Nicolas Thome. Hybrid energy based model in the feature space for out-of-distribution detection. In *International Conference on Machine Learning (ICML)*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The Shapley value in machine learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*, 2022.

Chuanguang Yang, Xinqiang Yu, Han Yang, Zhulin An, Chengqing Yu, Libo Huang, and Yongjun Xu. Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In *AAAI Conference on Artificial Intelligence*, pp. 9148–9156, 2025. doi: 10.1609/AAAI.V39I9.32990.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

A Training details and hyper-parameters

Split and data. CIFAR-10 train is split $90/10$ into 45k training and 5k validation images with a fixed index file shared by every run and every pool. OOD test sets: SVHN test (26,032), CIFAR-100 test (10,000), TinyImageNet validation (10,000, resized to 32×32) and DTD (5,640). Standard crop/flip augmentation.

Teachers. ResNet-18 (CIFAR variant), 100 epochs, SGD momentum 0.9, lr 0.1 cosine, weight decay 5×10^{-4}, batch 128. Coverage: cov$k$ assigns teacher $i \in \{0..4\}$ the classes $\{(2i+j) \bmod 10 : j < 2k\}$; for cov3 the home sets are $\{0..5\}, \{2..7\}, \{4..9\}, \{6,7,8,9,0,1\}, \{8,9,0,1,2,3\}$. Every teacher keeps a 10-way head. The generalist pool trains all five on all classes with seeds 0–4.

Priors. $\tau_{\text{prob}} = 1$ for the coalition value, $\tau_g = 1$ for both gate softmaxes; exact enumeration of 32 coalitions per sample from cached logits ($\sim$1.6M softmax evaluations, under a minute).

Students. ResNet-34 and MobileNetV2 (CIFAR variants), 100 epochs, SGD momentum 0.9, lr 0.1 cosine, weight decay 5×10^{-4}, batch 128, $\tau = 4$, $\lambda_{\text{kd}} = 1$ for all distillation methods and $\lambda_{\text{kd}} = 0$ for the supervised student (otherwise identical). The supervised-continuation control resumes the supervised student for 50 epochs at lr 0.01 cosine, matching SRR's joint stage.

SRR. Student initialised from the converged static-prior student of the same seed. Router: MLP with 256 hidden units and two heads (gate logits, r-logit), $\tau_r = 1$, reading the detached average-pooled penultimate feature. Router pre-training: 30 epochs, AdamW lr 10^{-3}, weight decay 10^{-4}, soft cross-entropy to the prior. Joint stage: 50 epochs; student SGD lr 0.01 cosine; router AdamW lr 2.5×10^{-5}. λ_{s} annealed linearly from 1.0 to 0.2 between 10% and 70% of the joint stage; anchor weight $\lambda_{\text{a}} = 0.5$ with $c_a(x)$ the reference's maximum softmax probability; $r_{\min} = 0$, $r_{\max} = 0.4$, $r_0 = 0.2$, $\lambda_r = 0.2$. The LogP-anchored variant changes only the prior, the initial/reference student and the router target.

Probe. Linear($d \to 128$)–ReLU–Linear($128 \to 1$), sigmoid output, MSE to $\tilde{r} = (r - r_{\min})/(r_{\max} - r_{\min})$ (or to the standardised minimum teacher cross-entropy), AdamW lr 10^{-3}, fitted on 80% of the ID training features with the remaining 20% used only to report fit quality. The co-trained variant attaches the same head to detached features during SRR.

Selection. Primary protocol: checkpoint with the best ID validation accuracy. Secondary (OSR) protocol: best $\text{acc}_{\text{ID}} + \text{AUROC} + \text{OSCR} - \text{FPR95}$ computed on the ID validation split against 10,000 SVHN-*training* images used as pseudo-unknowns, after epoch 10 and subject to $\text{acc}_{\text{ID}} \geq 0.85$. Under the OSR protocol (two-set mean over SVHN and CIFAR-100, 3 seeds, each run saving both checkpoints): uniform $0.865 \to 0.868$ (ID $94.2 \to 89.7$), LogP gate $0.853 \to 0.867$ ($94.6 \to 88.3$), DE-MKD $0.856 \to 0.873$ ($94.1 \to 89.7$), MTKD-RL $0.866 \to 0.868$ ($94.0 \to 92.4$), Shapley gate $0.871 \to 0.885$ ($94.6 \to 91.3$), Shapley-gate continuation ($+50$ epochs, no router) $0.867 \to 0.870$, SRR $0.884 \to 0.884$ ($94.7 \to 93.3$). Because the pseudo-unknowns share a source with one test set, the OSR protocol is not used for any headline number.

Readouts. Energy $=$ logsumexp of the student logits at temperature 1. k-NN (Sun et al., 2022): bank $=$ penultimate global-average-pooled 512-d features of the 45,000 unshuffled ID training images (no augmentation), ℓ_2-normalised; query features normalised likewise; score $=$ cosine similarity to the $k = 50$-th nearest bank feature (blocked matrix products, no approximate index). Mahalanobis: class means and a shared covariance on the same features. All readouts use the val-acc checkpoint; FPR@95 is the false-positive rate at 95% ID true-positive rate.

Compute. Single consumer GPU (RTX 2080 Ti / 4090 class). A ResNet-34 student takes about 1.5 h for 100 epochs; the SRR stages add about 1.2 h; the probe fit takes about two minutes. The latency benchmark of Table 4 runs on one RTX 4090 in fp32 with batch 256, the 10^4 query images preloaded to GPU memory, 10 warm-up batches, and 5 timed repeats bracketed by `cuda.synchronize`; k-NN uses the blocked cosine-similarity search of the evaluation code (no approximate index) and Mahalanobis the precomputed inverse covariance.

B Baselines

All baselines share Eq. (1) and differ only in the gate. **Entropy gate**: $\mathrm{softmax}_i(-H(p_i(x))/\tau_g)$, $\tau_g = 1$, label-free (`ebkd` in the code); its routed variant is trained exactly like the LogP-anchored one with the entropy gate as prior, initialisation and anchor. **LogP gate**: $\mathrm{softmax}_i(\log p_i(y|x)/\tau_g)$, $\tau_g = 1$; on cov3 its mean non-home mass is 0.0002 and its mean non-uniformity is 0.317 (analytically $1 - \log 3/\log 5 = 0.317$ for a uniform three-teacher gate). **CA-MKD** (Zhang et al., 2022): per-sample weights $w_i = (1 - \mathrm{softmax}_i(\mathrm{CE}_i))/(M-1)$, no temperature, maximum weight $1/(M-1)$, logit-level branch only; on cov3 this forces about 25% of the mass onto non-home teachers. **DE-MKD** (Cheng et al., 2024): entropy-based teacher weights combined with decoupled KD (Zhao et al., 2022) and a KD warm-up, following the original method. **MTKD-RL** (Yang et al., 2025): an agent reads teacher cross-entropies, teacher–student disagreement and teacher logits, outputs per-sample weights, and is updated by reward-weighted binary cross-entropy with reward $-(\mathrm{CE}+\mathrm{KD})$, following the original method. **Sel**: uniform average of the top-3 teachers by $\phi_i(x)$ (the home teachers on cov3), used in the geometry intervention.

C Proof of Proposition 1

Fix x and write $a_i = p_i(y|x)$. For any coalition $S \not\ni i$, $v(S \cup \{i\}; x) = \log\big(\frac{1}{|S|+1}(\sum_{j\in S} a_j + a_i)\big)$ is strictly increasing in a_i while $v(S; x)$ does not depend on a_i; with $v(\emptyset; x) = -\log C$ the singleton term $v(\{i\}; x) = \log a_i$ is increasing in a_i as well. Hence every marginal contribution $v(S \cup \{i\}; x) - v(S; x)$, and therefore the weighted sum $\phi_i(x)$, is a strictly increasing function of a_i when the other teachers' probabilities are held fixed. For two teachers i, k with $a_i > a_k$, exchanging their roles maps the coalitions excluding i onto those excluding k term by term, and each term for i dominates the corresponding term for k; thus $\phi_i(x) > \phi_k(x)$. □

D Per-sample signal and the specialisation dial

Table 5: Specialisation dial (ResNet-34, val-acc, 3 seeds; two-set mean energy AUROC over SVHN and CIFAR-100). †MobileNetV2 student. ‡Gate numerically uniform: "Shapley" re-runs uniform averaging.

Pool	classes/teacher	q_{shap} mean / p50	Uniform	Shapley gate	Δ	SRR	Δ
cov2	4 of 10 (2×)	0.529 / 0.532	0.839±.003	0.885±.019	+4.6	0.885±.017	+4.6
cov3 (main)	6 of 10 (3×)	0.259 / 0.261	0.865±.007	0.873±.004	+0.9	0.884±.006	+1.9
cov4	8 of 10 (4×)	0.100 / 0.101	0.878±.002	0.875±.007	−0.3	0.897±.002	+2.0
generalist†	10 of 10 (5×)	0.000 / 0.000	0.772±.008	0.784±.010‡	(+1.2)	0.792±.022	+2.0

The dial is designed to test whether the static gate's gain over *uniform* distillation is monotone in q_{shap}, so uniform KD is its baseline; the supervised reference is run for the main pool only (Table 1).

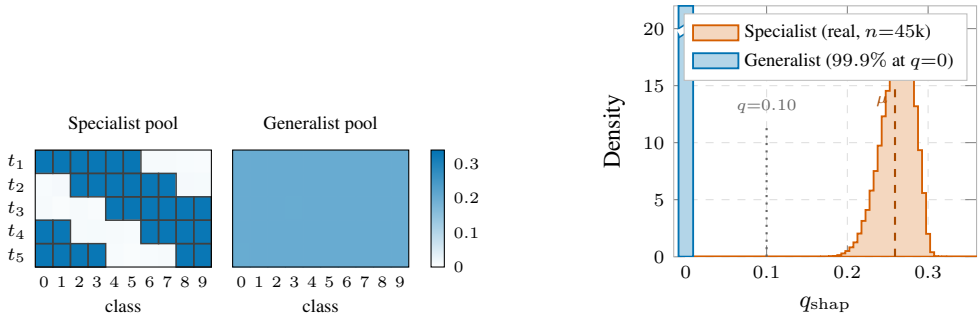


Figure 2: Left: mean Shapley-gate weight per (teacher, class) over the ID training set; on cov3 the gate recovers each teacher's six home classes (outlined) with a $29\times$ home/non-home contrast; on the generalist pool it is uniform. Right: per-sample q_{shap}; cov3 concentrates around $\mu = 0.259$ with every sample above 0.10, while 99.9% of generalist samples sit at 0.

E GATE-GEOMETRY INTERVENTION

Let $\mathcal{H}(x)$ be the three home teachers of x's class on cov3. *Oracle top-3*: $1/3$ on $\mathcal{H}(x)$ (identical to Sel). *Oracle + leak*: $(1-\ell)/3$ on $\mathcal{H}(x)$ and $\ell/2$ on the two non-home teachers with $\ell = 0.0222$, the Shapley gate's mean non-home mass. *Shapley, leak removed*: g^{S} restricted to $\mathcal{H}(x)$ and renormalised. Distilling against each (ResNet-34, val-acc, 3 seeds; two-set mean): uniform $0.865 \pm .007$, oracle top-3 $0.860 \pm .007$, oracle + leak $0.852 \pm .015$, Shapley without leak $0.855 \pm .016$, Shapley gate $0.873 \pm .004$. The gain over uniform requires both the graded weights and the leak; the differences among the three ablated gates are within one standard deviation.

F PER-SET RESULTS

Table 6: Energy AUROC per OOD set (ResNet-34, cov3, val-acc, 3 seeds). Per-set values are listed where the per-set metrics were exported; "–" otherwise (the four-set means of those rows come from the same runs).

Method	ID acc	SVHN	CIFAR-100	TIN	DTD	Mean (4)
Supervised (CE only)	94.9	–	–	–	–	0.880
Supervised +50 ep.	95.0	–	–	–	–	0.889
SRR (Shapley prior)	94.7	0.901±.010	0.868±.001	0.873±.001	0.908±.003	0.887±.003
SRR (LogP prior)	94.6	–	–	–	–	0.884
Shapley gate (static)	94.6	0.888±.007	0.859±.002	0.867±.003	0.893±.010	0.877±.004
LogP gate (static)	94.5	0.872±.033	0.857±.005	0.865±.005	0.904±.009	0.874±.013
Routing (entropy)	–	–	–	–	–	0.872
Entropy gate (static)	94.1	0.877±.016	0.834±.003	0.842±.002	0.867±.007	0.855
CA-MKD (faithful)	94.3	–	–	–	–	0.867
DE-MKD	93.6	0.886±.002	0.834±.003	0.838±.003	0.864±.001	0.856±.002
MTKD-RL	94.1	0.907±.010	0.825±.001	0.834±.002	0.851±.003	0.854±.003
Uniform KD	94.1	0.900±.013	0.829±.002	0.841±.001	0.858±.004	0.857±.004

Five-seed two-set means (seeds $\{42, 123, 3407, 2025, 31415\}$): SRR $0.887 \pm .006$, MTKD-RL $0.867 \pm .004$, Shapley gate $0.867 \pm .010$, uniform $0.863 \pm .010$, DE-MKD $0.861 \pm .005$, LogP gate $0.856 \pm .020$ (the supervised student was run at three seeds only).

Table 7: AUROC of seven post-hoc scores on the uniform, Shapley-gate and SRR ResNet-34 students (cov3, val-acc, 3 seeds; two-set mean over SVHN and CIFAR-100, with the two sets in parentheses).

Detector	Uniform student	Shapley-gate student	SRR student
MSP (Hendrycks & Gimpel, 2017)	0.798 (0.799 / 0.798)	0.888 (0.908 / 0.869)	0.892 (0.912 / 0.872)
Max-logit (Hendrycks et al., 2022)	0.866 (0.899 / 0.834)	0.873 (0.888 / 0.858)	0.884 (0.900 / 0.867)
Energy (Liu et al., 2020)	0.865 (0.900 / 0.829)	0.873 (0.888 / 0.859)	0.884 (0.901 / 0.868)
GEN (Liu et al., 2023)	0.871 (0.898 / 0.844)	0.879 (0.895 / 0.863)	0.887 (0.905 / 0.870)
k-NN (Sun et al., 2022)	0.896 (0.944 / 0.848)	0.922 (0.953 / 0.892)	0.920 (0.947 / 0.892)
Mahalanobis (Lee et al., 2018)	0.888 (0.945 / 0.831)	0.912 (0.948 / 0.876)	0.911 (0.945 / 0.877)
ViM (Wang et al., 2022)	0.829 (0.878 / 0.780)	0.905 (0.947 / 0.862)	0.906 (0.944 / 0.867)

G Post-hoc detectors on the same students

H Additional probe results and controls

Co-trained head. With a detached head trained during SRR the student's energy metrics are unchanged (0.885 vs. 0.884 two-set mean), confirming neutrality; the head gives AUROC 0.957 ± 0.005 (SVHN) and 0.888 ± 0.001 (CIFAR-100) alone. **Post-hoc probe** on the frozen val-acc student: hold-out R^2 of 0.999, 0.997, 0.998 over the three seeds (the target is a function of the same features, so a high R^2 is expected and is reported only as a sanity check); per-set AUROC 0.955 / 0.889 / 0.883 / 0.922. **Combination with energy.** A standardised combination $z(E) - \lambda z(\hat{r})$ with λ chosen on pseudo-unknowns selects the grid maximum on every seed and adds at most 0.2 points over the probe alone; we therefore do not report it. **Controls** (same probe, same student, 3 seeds): regressing the student's own training-set maximum softmax probability changes AUROC by $+1.3 \pm 1.0$ (SVHN) and -1.2 ± 0.7 (CIFAR-100) relative to energy; regressing $\mathcal{N}(0, 1)$ noise gives -13.0 and -28.6 (seed 42). **Minimum-teacher-CE probe** on the SRR student: per-set change over energy -0.2 / $+1.4$ / $+0.9$ / $+0.6$, mean AUROC 0.894. On the uniform and Shapley-gate students the same probe adds $+4.2$ / $+2.3$ CIFAR-100 points respectively. **Frozen supervised student.** Regressing the routed run's $r(x)$ (same seed) from the supervised student's features gives hold-out $R^2 = 0.41$ and mean AUROC 0.915 over the four OOD sets (3 seeds); regressing the minimum teacher cross-entropy gives 0.906; the student's own energy is 0.880 and its k-NN 0.915. **Compact student.** On MobileNetV2 the router saturates and $\hat{r}$ is uninformative (AUROC 0.34–0.39).

I Compact student

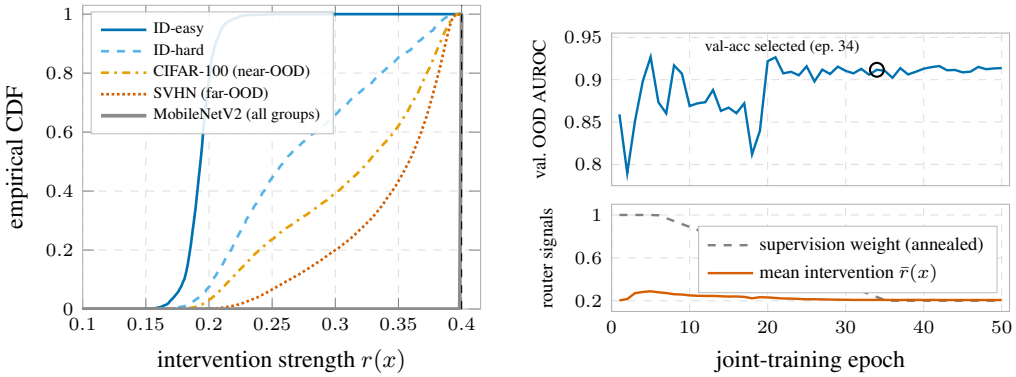
MobileNetV2, cov3, val-acc, 3 seeds, two-set mean energy AUROC (four-set mean in brackets): supervised 84.4% ID, [0.817]; SRR 84.2%, $0.823 \pm .000$ [0.813]; uniform 85.0%, $0.809 \pm .009$ [0.803]; Shapley gate 82.5%, $0.809 \pm .018$ [0.803]; Sel 81.4%, $0.771 \pm .027$. Five-seed two-set means: SRR $0.819 \pm .007$, uniform $0.811 \pm .008$, Shapley gate $0.800 \pm .017$, Sel $0.783 \pm .026$. The router's $r(x)$ equals $r_{\max}$ on every sample (std $< 10^{-5}$). Retrained MobileNetV2 runs on a rebuilt machine sit about two points below these runs for all methods, so only within-environment comparisons are made.

J Router diagnostics

Ablations (MobileNetV2, single seed, OSR-selected): removing the anchor term 0.844, removing the supervision term 0.840, $r_{\max} = 0.2$ or 1.0: 0.846 / 0.845, against 0.840 for full SRR; all within run-to-run noise. On ResNet-34 the same constants were used without re-tuning.

K Attribution analyses

On cov3, 100% of non-home teacher–sample pairs receive negative ϕ; in a synthetic game with $K \in \{3, 5\}$ players, 1–3 competent players and the rest drawing class-agnostic Dirichlet predictions (con-



(a) Empirical CDF of $r(x)$ by sample group.

(b) Joint-training dynamics (ResNet-34, seed 42).

Figure 3: (a) Per-sample mixing coefficient of the ResNet-34 router by group, val-acc checkpoint, seed 42. The four groups are stochastically ordered by OOD-ness (means 0.19 / 0.27 / 0.31 / 0.34; over three seeds 0.19 / 0.27 / 0.31 / 0.35), and the fraction driven to the cap ($r \geq 0.39$) is 0.0% / 0.5% / 3.2% / 5.1% (0 / 1.9 / 6.7 / 14.1% over three seeds): under the primary protocol the router is selective but far from saturated. The MobileNetV2 line (grey) shows that router saturating at $r_{\max}$ on every input (drawn from its OSR-selected checkpoint). (b) As the supervision weight anneals, the mean coefficient $\bar{r}$ rises during exploration and settles near 0.21; the marked epoch (34) is the val-acc-selected checkpoint deployed throughout the paper.

centration $\in \{0.1, 0.5, 2\}$), the expected ϕ of an uninformative player is negative in all 15 settings (range -0.14 to -2.66). A second, global attribution ϕ_i^{pu} with value $v(S) = -\text{KL}(\text{Unif} \,\|\, p_S(u))$ averaged over pseudo-unknowns u has Pearson correlation -0.40 with the classification attribution across teachers; blending it into the prior at $\beta = 0.5$ gives 0.879 ± 0.011 two-set mean (vs. 0.873 for the Shapley gate). Negative results: a signed-Shapley repulsion loss hurts (0.855 ± 0.016); a coalition-softmax gate does not help (0.862 ± 0.021).

L TEACHER DISAGREEMENT AS A DETECTOR

Table 8: Teacher-*ensemble* OOD scores (AUROC; all teachers at test time). Mutual information works on generalist teachers but collapses on specialists, which disagree on ID inputs by construction.

	Specialist (cov3)		Generalist	
Ensemble score	SVHN	CIFAR-100	SVHN	CIFAR-100
Ensemble energy	0.926	0.880	0.954	0.899
Total uncertainty $H(\bar{p})$	0.864	0.727	0.940	0.907
Mutual information	0.439	0.406	0.920	0.896

M SRR TRAINING LOOP

Algorithm 1 SRR training (with optional routing-head probe)

Require: Teachers $\{t_i\}_{i=1}^{M}$, training set $\mathcal{D}$, prior g^0
Cache $z_i(x)$ for all $x \in \mathcal{D}$; compute $g^0(x)$ and its non-uniformity $q(x)$
Train student against $g^0 \to$ frozen reference p_a and initialisation
Pre-train router on detached features to match g^0
for joint epoch $= 1$ to T **do**
  for mini-batch (x, y) **do**
    $f(x), p_s(x) \leftarrow$ student; $g^{\mathrm{c}}(x), r(x) \leftarrow$ router$(f(x)^{\mathrm{detached}})$
    $g \leftarrow (1-r)g^0 + rg^{\mathrm{c}}$; compute $\mathcal{L}_{\mathrm{SRR}}$ (Eqs. 3, 4)
    optional: $\mathcal{L}_{\hat{r}} \leftarrow \mathrm{MSE}\big(\sigma(h_\psi(f(x)^{\mathrm{detached}})), \tilde{r}(x)\big)$
    update student (SGD), router (AdamW), probe (AdamW)
  end for
  anneal λ_{s}
end for
Ensure: Student (and probe h_ψ); router, teachers, g^0 discarded