DEED: Dual-Channel Enhanced Ensemble Distillation for Uncertainty-Aware Recognition

Supplementary Material

A. Method Details

A.1. Teacher Uncertainty Signal $U(x)$ and Normalization

A.1.1. Variants for the teacher uncertainty signal

Option A: Predictive entropy (default). We define the normalized predictive entropy of the ensemble mean as

$$U_{\text{ent}}(x) = \frac{H\big(\bar{p}_T(\cdot \mid x; T)\big)}{\log K}, \quad H(p) = -\sum_{k=1}^{K} p_k \log p_k. \tag{28}$$

By construction, $U_{\text{ent}}(x) \in [0, 1]$, where 0 corresponds to a confident, delta-like distribution and 1 corresponds to a maximally uncertain, uniform distribution. This measure is parameter-free, numerically stable, and serves as our default choice unless otherwise noted.

Option B: Mutual information (epistemic emphasis). To highlight epistemic disagreement, we use the ensemble mutual information (MI):

$$\text{MI}(x) = H\big(\bar{p}_T(\cdot \mid x; T)\big) - \frac{1}{M}\sum_{m=1}^{M} H\big(p_{T_m}(\cdot \mid x; T)\big). \tag{29}$$

MI is large when teachers are individually confident yet disagree (i.e., multi-modal beliefs), and small when teachers either agree or are all uniformly uncertain.

Option C: Vote entropy (mode-level disagreement). Let $\hat{y}_m(x) = \arg\max_k p_{T_m}(k \mid x; T)$ be the hard prediction of teacher m, and let $v_k(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{\hat{y}_m(x) = k\}$ denote the empirical vote histogram. We define

$$U_{\text{vote}}(x) = \frac{H\big(v(x)\big)}{\log K}, \quad H \text{ as in Eq. (28)}. \tag{30}$$

This measure ignores soft probability margins but captures disagreement at the mode level.

Option D: Logit variance/dispersion (classwise variability). Let $z_{T_m,k}(x)$ denote teacher m's logit for class k. We compute a class-averaged logit variance:

$$\text{VarLogits}(x) = \frac{1}{K}\sum_{k=1}^{K} \text{Var}_m\big[z_{T_m,k}(x)\big]. \tag{31}$$

This variant is most informative when teacher logits are calibrated and share a comparable scale.

When to prefer alternatives. Predictive entropy U_{ent} is robust and computationally inexpensive. For scenarios requiring stronger emphasis on epistemic uncertainty (e.g., high-disagreement regimes), MI in Eq. (29) may be more appropriate. Vote entropy in Eq. (30) is beneficial when teachers make sharp but divergent predictions. Variance-based measures in Eq. (31) are effective when teacher logits are well-calibrated and generated under similar training conditions.

A.1.2. Normalization strategy.

Except for U_{ent} and U_{vote} (which are already scaled to $[0, 1]$), the other uncertainty signals require per-epoch normalization. We apply a clipped min–max normalization. Given raw scores $r(x)$ within an epoch, we compute

$$U_{\text{mm}}(x) = \text{clip}\left(\frac{r(x) - \mu_{\min}}{\mu_{\max} - \mu_{\min} + \varepsilon}, 0, 1\right), \tag{32}$$

where $\mu_{\min}$ and $\mu_{\max}$ denote per-epoch percentiles (e.g., 2% and 98%) to reduce the influence of outliers, and $\varepsilon > 0$ avoids division by zero. After normalization, all variants share the same interface $U : \mathbb{R}^d \to [0, 1]$ and can be interchanged without modifying downstream components.

A.2. Negative Pool Design and Sampling

Mixture model. At each iteration, we draw negative samples from the mixture

$$\begin{gathered} x^- \sim q_{\text{neg}} = \pi_{\text{ood}}\, q_{\text{ood}} + \pi_{\text{adv}}\, q_{\text{adv}} + \pi_{\text{bnd}}\, q_{\text{bnd}}, \\ \pi_{\text{ood}},\, \pi_{\text{adv}},\, \pi_{\text{bnd}} \geq 0, \qquad \sum\nolimits_s \pi_s = 1. \end{gathered} \tag{33}$$

We also specify a batch-level ratio $\rho_{\text{neg}} \in [0, 1]$ that controls how many negatives are mixed into each student minibatch of size B; specifically, we include $\lfloor \rho_{\text{neg}} B \rfloor$ negatives and $B - \lfloor \rho_{\text{neg}} B \rfloor$ ID samples. Unless otherwise stated, we use moderate settings (e.g., $\rho_{\text{neg}} \in [0.25, 0.5]$).

Component q_{ood}: external OOD sources. We maintain a small cache of OOD images drawn from standard sources (e.g., SVHN, LSUN, Tiny-ImageNet as non-overlapping categories), as well as simple noise distributions (Gaussian/Perlin) and curated "Tiny-OOD" subsets. To avoid trivial negatives, we (i) normalize and resize all images to match ID preprocessing, (ii) apply mild augmentations consistent with the ID transformation pipeline, and (iii) occasionally include "near-OOD" crops (random resized/cutout regions) from ID images with their labels removed. Unless otherwise noted, we sample uniformly from this OOD cache.

Component q_{adv}: light adversarial variants. Starting from an ID input x in the current minibatch, we generate a light adversarial variant within a small ℓ_∞ ball:

$$\begin{aligned} x_{\text{adv}}^{(0)} &= x + \text{clip}_\epsilon\big(\mathcal{U}(-\epsilon,\epsilon)\big), \\ x_{\text{adv}}^{(r+1)} &= \Pi_{B_\infty(x,\epsilon)}\Big(x_{\text{adv}}^{(r)} + \alpha\,\text{sign}\Big(\nabla_x\,\phi(x_{\text{adv}}^{(r)})\Big)\Big), \\ & r = 0,\dots,R-1. \end{aligned} \tag{34}$$

Here, ϵ denotes the perturbation radius, α the step size, and R the number of iterations (typically $R \in \{1,2,3\}$). The operator $\Pi_{B_\infty(x,\epsilon)}$ projects back to the valid domain $[0,1]^d$ and enforces the ℓ_∞ constraint $\|z-x\|_\infty \le \epsilon$.

We use a score ϕ that encourages movement toward uncertain or low-affinity regions without inducing class-directed misclassification:

$$\phi(x) \in \Big\{H\big(\bar{p}_T(\cdot \mid x;T)\big), \text{KL}\big(\bar{p}_T(\cdot \mid x;T)\,\|\,p_S(\cdot \mid x)\big), -E_S(x)\Big\}. \tag{35}$$

Entropy ascent $H(\bar{p}_T)$ accentuates teacher disagreement; discrepancy ascent $\text{KL}(\bar{p}_T\|p_S)$ drives samples toward regions where the teacher and student differ; and $-E_S$ moves samples toward lower student energy (higher affinity), which can complement energy-margin losses to sharpen the decision boundary. In practice, we typically adopt $H(\bar{p}_T)$ or $\text{KL}(\bar{p}_T\|p_S)$ to obtain stable "near-boundary" negatives with modest perturbation budgets (e.g., $\epsilon \in [2/255, 4/255]$ and $\alpha = \epsilon/R$).

Component q_{bnd}: near-boundary jitter. We generate label-free, boundary-proximal negatives by applying small stochastic transforms $\mathcal{T}$ to ID images:

$$\begin{aligned} & x_{\text{bnd}} = \mathcal{T}(x), \\ & \mathcal{T} \in \Big\{\text{color jitter (small)},\ \text{Gaussian blur (mild)}, \\ & \qquad \text{random crop/pad (few pixels)},\ \text{low-magnitude Cutout}\Big\}. \end{aligned} \tag{36}$$

These lightweight perturbations create "almost-ID" samples that reliably populate decision margins and complement q_{adv} at effectively zero additional cost.

Sampling strategy. Given a minibatch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^B$, we draw $N^- = \lfloor \rho_{\text{neg}} B \rfloor$ negative samples by multinomially allocating counts $(N_{\text{ood}}, N_{\text{adv}}, N_{\text{bnd}}) \sim \text{Mult}(N^-; \pi_{\text{ood}}, \pi_{\text{adv}}, \pi_{\text{bnd}})$. We then proceed as follows:

1. Sample N_{ood} items from the OOD cache (without replacement within the iteration).
2. For N_{adv}, choose N_{adv} distinct indices in $\mathcal{B}$ and generate adversarial variants using Eq. (34) with the selected score ϕ and (ϵ, α, R).
3. For N_{bnd}, choose N_{bnd} distinct indices and apply the jitter transform $\mathcal{T}$ as in Eq. (36).

To reduce redundancy, we optionally deduplicate candidates by discarding those whose ℓ_2 distance from existing negatives falls below a small threshold or whose energy E_S differs by less than a tolerance δ_E (useful when N^- is relatively large). Unless otherwise noted, we use $\rho_{\text{neg}} = 0.5$, mixture weights $(\pi_{\text{ood}}, \pi_{\text{adv}}, \pi_{\text{bnd}}) = (0.5, 0.25, 0.25)$, a perturbation radius $\epsilon = 4/255$, step size $\alpha = \epsilon/R$, $R \in \{1,2,3\}$, and color jitter as the default choice for Eq. (36).

Clarification on the Use of OOD Samples. We emphasize that we do not use any labels for OOD samples, nor do we train the model to identify specific OOD categories. The external OOD images serve solely as unlabeled negative samples that help stabilize the energy/uncertainty margin. This approach is consistent with widely adopted practices such as Outlier Exposure (Hendrycks et al., ICLR'19), energy-based OOD training (Liu et al., NeurIPS'20), and contrastive learning methods that rely on unlabeled negatives.

Crucially, OOD samples never participate in the supervised cross-entropy loss and therefore do not influence the ID decision boundaries. They function purely as hard negatives that prevent trivial solutions and improve calibration robustness. Consequently, this procedure does not constitute "supervised OOD training" and does not leak semantic information about unseen classes.

A.3. Margin schedule m_t

We employ a warm-up schedule in which KD/CE first shapes the ID interior regions, after which Channel I gradually raises the negative-energy floor:

$$\text{(exp)} \quad m_t = m_{\max}\big(1 - e^{-\kappa t}\big), \tag{37}$$

$$\text{(piecewise)} \quad m_t = \begin{cases} 0, & t \le t_0, \\ \frac{m_{\max}}{t_1 - t_0}(t - t_0), & t_0 < t \le t_1, \\ m_{\max}, & t > t_1, \end{cases} \tag{38}$$

$$\text{(cosine)} \quad m_t = m_{\max}\,\tfrac{1-\cos(\pi t/T)}{2}. \tag{39}$$

By default, we use $m_{\max} \in [6, 10]$ (with temperature $T{=}1$) and set $\kappa \in [3/T, 6/T]$ so that m_t reaches $0.9\,m_{\max}$ between 30% and 60% of training. For different temperatures, $m_{\max}$ can be scaled approximately proportionally to T.

A.4. Details of $U^-(x^-)$ in Channel I

Sources of negatives. Recall the mixed pool (cf. Sec. A.2):

$$x^- \sim q_{\text{neg}} = \pi_{\text{ood}}\,q_{\text{ood}} + \pi_{\text{adv}}\,q_{\text{adv}} + \pi_{\text{bnd}}\,q_{\text{bnd}}, \qquad \sum_s \pi_s = 1.$$

We distinguish two types of negatives for constructing U^-: (i) those derived from ID seeds x (adversarial or near-boundary variants), and (ii) external OOD negatives for which teacher outputs are not available.

Case A: Negatives derived from an ID seed. For negatives of the form $x^- = \mathrm{Gen}(x)$, where x is an ID example, we simply inherit the teacher-derived uncertainty of the seed:

$$U^-(x^-) := U(x). \tag{40}$$

This assignment incurs no additional forward passes and aligns the negative's weight with the teacher ensemble's local ambiguity around x.

Case B: External OOD negatives. For OOD negatives $x^- \sim q_{\mathrm{ood}}$ (where no teacher signal is available), we define U^- using the student's current outputs. We consider two simple proxies:
(i) Normalized entropy:

$$U^-_{\mathrm{ent}}(x^-) = \frac{H\big(p_S(\cdot \mid x^-)\big)}{\log K} \in [0,1], \quad H(p) = -\sum_{k=1}^{K} p_k \log p_k. \tag{41}$$

(ii) Energy-based squashing:

$$U^-_{\mathrm{eng}}(x^-) = \sigma\Big(\gamma\big(E_S(x^-) - \tau_E\big)\Big), \quad \sigma(u) = \frac{1}{1+e^{-u}}, \tag{42}$$

where $E_S(x) = -T \log \sum_{k=1}^{K} \exp(z_{S,k}(x)/T)$ (Eq. (4)), $\gamma > 0$ controls the slope, and τ_E centers the squashing function (taken as the median from a running buffer; see below). Eq. (41) tracks predictive dispersion, whereas Eq. (42) leverages the energy–ID affinity relation.

Hybrid proxy. Either proxy is sufficient. If desired, one may also form a convex combination:

$$U^-(x^-) = \lambda_U\, U^-_{\mathrm{ent}}(x^-) + (1-\lambda_U)\, U^-_{\mathrm{eng}}(x^-), \quad \lambda_U \in [0,1], \tag{43}$$

using a fixed λ_U across datasets (default λ_U=0.5).

Weighting functions. When U^--weighting is enabled (cf. Eqs. (9) and (10)), we employ the following monotone gating functions:

$$\alpha(u) = 1 + \rho\, u, \quad \beta(u) = u, \quad \alpha(u), \beta(u) \in [1, \alpha_{\max}], \tag{44}$$

with $\rho \in [0.5, 1.0]$ and $\alpha_{\max} \in [1.5, 2.0]$ to prevent excessive weighting of extreme outliers.

When to enable U^--weighting. By default, U^--weighting should be disabled. It becomes beneficial in two scenarios: (i) when adversarial or near-boundary negatives dominate and selective-risk metrics saturate, or (ii) when the overlap between ID and negative energy histograms remains substantial (e.g., a small Bhattacharyya distance) even after margin warm-up. In practice, a modest weighting strength (e.g., $\rho = 0.75$) is sufficient.

Default hyperparameters. Unless otherwise specified, we use U^-_{ent} as the sole proxy for all negatives when only one proxy is desired; otherwise we set λ_U=0.5 in Eq. (43). We also adopt ρ=0.75, $\alpha_{\max}$=1.75, γ=4, and τ_E=median(E_S) computed per epoch.

A.5. Details of Direction Fields in Channel II

Gradients and stop-grad. We treat the teacher ensemble outputs as constants and do not allow gradients to propagate through the teacher models. This means that the ensemble mean $\bar{p}_T(\cdot \mid x; T)$ is computed once, cached, and then treated as a fixed tensor (i.e., we apply `stop_gradient` or `detach` to it).

For the entropy-based direction $\hat{v}_H$, we compute $\nabla_x H(\bar{p}_T(\cdot \mid x; T))$, but the gradient is taken only with respect to the input x. Although the entropy $H(\bar{p}_T)$ is numerically a function of the cached teacher probabilities, we do not backpropagate through the teacher networks that produced $\bar{p}_T$. Instead, $\bar{p}_T$ is held constant, and the derivative is computed as if $H(\bar{p}_T)$ were an x-dependent scalar whose value is fixed by the cached teacher outputs. Thus, the gradient flows only to the input x, never to the teacher parameters.

For the KL-based direction $\hat{v}_{\mathrm{KL}}$, we compute $\nabla_x \mathrm{KL}(\bar{p}_T \| p_S(\cdot \mid x))$, but now the gradient flows only through the student distribution $p_S(\cdot \mid x)$, since this term explicitly depends on x through the student network. The teacher mean $\bar{p}_T$ serves as the fixed target distribution in the KL divergence and is again detached so that no gradient can propagate into the teacher ensemble. In other words, $\bar{p}_T$ affects the value of the KL divergence but does not receive gradients; only the student model and the input x participate in the backward pass.

Signed direction and projection. We generate perturbations using an FGSM-style signed gradient under an ℓ_∞ budget:

$$\hat{v}(x) = \mathrm{sign}\big(\nabla_x \phi(x)\big), \quad \delta = \eta\, \hat{v}(x), \quad \eta = \epsilon\, g\big(U(x)\big),$$

where $\epsilon = 4/255$ and $g(U) = U^{\alpha_u}$ with $\alpha_u \in [1, 1.5]$. After adding δ, the perturbed input is clamped to $[0,1]$.

Gates and weights. We employ the gating functions $w_c(U) = \exp(-\lambda U)$ with $\lambda \in [1,3]$, and $w_e(U) = \sigma(\gamma(U-\tau))$ with $\gamma \in [8,16]$, where τ is the median of U in the current epoch. The loss weights are set to $\lambda_{\mathrm{cons}} = 1.0$ and $\lambda_{\mathrm{highU}} = 0.5$, tuned coarsely on the validation split.

Optimizers and schedules. For the student model, we use SGD with momentum 0.9 and weight decay 5×10^{-4}, together with a cosine learning-rate schedule that decays from 0.1 to 10^{-4} over 240 epochs.

A.6. Practical choices in Channel I

In practice, we follow several guidelines for stabilizing and improving Channel I. First, the margin schedule m_t should remain modest in the early stages of training to avoid degrading ID calibration, and only be increased once the optimization dynamics have stabilized. Second, it is generally preferable to use "near-distribution" negatives such as boundary jitter or light adversarial variants, rather than far-off noise, since noise tends to shift energies globally instead of enlarging the intended gap. Finally, we monitor negative-energy percentiles (for example, the 50th, 90th, and 95th percentiles) together with ID energy statistics to verify that the separation between ID and negative samples evolves in the desired direction.

A.7. Full Training Settings and Procedure

A.7.1. Training algorithm

Algorithm 1 summarizes one training epoch of DEED. In each iteration, DEED first computes the teacher ensemble targets and the normalized uncertainty $U(x)$ for every sample in the mini-batch. These signals are then used to construct two complementary sets of inputs: (i) uncertainty-scaled perturbations that create near-boundary views of the in-distribution samples, and (ii) negative examples drawn from a mixed pool of external OOD images, light adversarial variants, and boundary-jitter transformations. As a result, the student observes original samples, their perturbed counterparts, and negative samples within the same batch, encouraging the model to distinguish between interior, boundary, and out-of-distribution regions.

DEED then updates the student parameters using three components. The knowledge distillation loss aligns the student with the teacher ensemble. Channel I applies an energy-margin objective on negative samples to raise their energy and enlarge the separation from in-distribution data. Channel II combines a symmetric consistency loss on perturbed samples with a high-uncertainty regularization term that encourages smoother decision boundaries and better behavior in ambiguous regions. This procedure repeats across epochs, with the energy margin m_t gradually increased and, if desired, the perturbation field transitioning from entropy-based to KL-based once the KD dynamics have stabilized.

A.7.2. Training settings

Warm-up of regularizers. To prevent early over-regularization, we gradually ramp up the channel weights over the first t_w iterations:

$$\lambda_{\text{cons}}^{(t)} = \lambda_{\text{cons}} \cdot \min\left(1, \frac{t}{t_w}\right), \lambda_{\text{highU}}^{(t)} = \lambda_{\text{highU}} \cdot \min\left(1, \frac{t}{t_w}\right), \tag{45}$$

and jointly apply the scheduled energy margin m_t. Unless stated otherwise, we set $\lambda_{\text{hard}} = 0$.

Algorithm 1 One-Epoch Training of DEED

Require: Teacher ensemble $\{T_m\}_{m=1}^{M}$ (frozen), student S_{θ_S}; in-distribution data $\mathcal{D}_{\text{in}}$; negative pool q_{neg}; batch size B; temperature T; base radius ϵ; scaler $g(\cdot)$; gates $w_c(\cdot), w_e(\cdot)$; weights $\lambda_{\text{hard}}, \lambda_{\text{C1}}, \lambda_{\text{C2}}, \lambda_{\text{neg-ent}}, \lambda_{\text{cons}}, \lambda_{\text{highU}}$; margin m_t for this epoch.

Shuffle $\mathcal{D}_{\text{in}}$ into mini-batches $\{\mathcal{B}_k\}_{k=1}^{N_b}$ of size B.
for $k = 1$ to N_b **do** ▷ iterate over one epoch
  Teacher targets & uncertainty for $\mathcal{B}_k = \{(x_i, y_i)\}_{i=1}^{B}$:
  $\bar{p}_T(\cdot \mid x_i; T) \leftarrow \frac{1}{M}\sum_{m=1}^{M} \text{softmax}(z_{T_m}(x_i)/T)$
  $U(x_i) \in [0, 1]$ (e.g., normalized ensemble entropy)
  Channel II: U-scaled perturbations
  Choose guidance field $\hat{v}$:
    early: $\hat{v}_H = \text{sign}(\nabla_x H(\bar{p}_T))$, late: $\hat{v}_{\text{KL}} = \text{sign}(\nabla_x \text{KL}(\bar{p}_T \| p_S))$
  For each x_i: pick perturbation step η_i s.t. $\|\eta_i \hat{v}(x_i)\|_\infty \le \epsilon\, g(U(x_i))$; set $\delta_i \leftarrow \eta_i \hat{v}(x_i)$
  Channel I: negatives & energy shaping
  Sample $\{x_j^-\}_{j=1}^{\rho_{\text{neg}} B} \sim q_{\text{neg}}$ (external OOD / light adv / near-boundary)
  $\mathcal{L}_{\text{neg}} = \mathbb{E}_{x^-}[\text{softplus}(m_t - E_S(x^-))]$
  $\mathcal{L}_{\text{neg-ent}} = \mathbb{E}_{x^-}[\text{CE}(\text{Uniform}, p_S(\cdot \mid x^-))]$
  Channel II losses
  $\mathcal{L}_{\text{cons}} = \frac{1}{B}\sum_{i=1}^{B} w_c(U(x_i)) \left(\text{KL}(p_S(\cdot \mid x_i) \| p_S(\cdot \mid x_i + \delta_i)) + \text{KL}(p_S(\cdot \mid x_i + \delta_i) \| p_S(\cdot \mid x_i))\right)$
  $\mathcal{L}_{\text{highU}} = \frac{1}{B}\sum_{i=1}^{B} w_e(U(x_i))\, \text{CE}(\text{Uniform}, p_S(\cdot \mid x_i))$
  $\mathcal{L}_{\text{C1}} = \mathcal{L}_{\text{neg}} + \lambda_{\text{neg-ent}}\, \mathcal{L}_{\text{neg-ent}}$
  $\mathcal{L}_{\text{C2}} = \lambda_{\text{cons}}\, \mathcal{L}_{\text{cons}} + \lambda_{\text{highU}}\, \mathcal{L}_{\text{highU}}$
  KD & Hard-label
  $\mathcal{L}_{\text{KD}} = \frac{1}{B}\sum_{i=1}^{B} T^2\, \text{KL}\big(\bar{p}_T(\cdot \mid x_i; T) \,\|\, \text{softmax}(z_S(x_i)/T)\big)$
  $\mathcal{L}_{\text{hard}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(y_i, p_S(\cdot \mid x_i))$
  Total loss & update
  $\mathcal{L}_{\text{DEED}} = \mathcal{L}_{\text{KD}} + \lambda_{\text{hard}}\mathcal{L}_{\text{hard}} + \lambda_{\text{C1}}\mathcal{L}_{\text{C1}} + \lambda_{\text{C2}}\mathcal{L}_{\text{C2}}$
  Update $\theta_S \leftarrow \theta_S - \eta_t \nabla_{\theta_S} \mathcal{L}_{\text{DEED}}$ ▷ η_t is the optimizer learning rate
Epoch-end notes: optionally warm up m_t, ramp $\lambda_{\text{C1}}, \lambda_{\text{C2}}$, and (if desired) switch $\hat{v}_H \to \hat{v}_{\text{KL}}$ after KD stabilizes.

Stopping gradients through U and δ. For stability, we treat the uncertainty signal $U(x)$ and the perturbation $\delta(x)$ as constants when computing the Channel II gradients. Specifically,

$$\frac{\partial \mathcal{L}_{\text{cons}}}{\partial U(x)} = \frac{\partial \mathcal{L}_{\text{highU}}}{\partial U(x)} = 0, \quad \frac{\partial \mathcal{L}_{\text{cons}}}{\partial \delta(x)} = \frac{\partial \mathcal{L}_{\text{highU}}}{\partial \delta(x)} = 0, \tag{46}$$

so the gradients propagate only through the student outputs evaluated at x and $x + \delta(x)$.

Temperature and calibration. The distillation temperature T controls both the smoothness of the teacher distribution $\bar{p}_T(\cdot \mid x)$ and the scale of the resulting gradients. In practice, we use $T \in [2, 4]$ and omit the usual T^2 factor in the implementation by absorbing it into λ_{KD}, leaving the objective in the paper unchanged.

Weight decay and energy shifts. We include an ℓ_2 penalty $\frac{\lambda_{\text{wd}}}{2}\|\theta_S\|_2^2$ in Eq. (27) to curb excessive logit growth and to prevent the model from trivially satisfying the energy margin through uniform logit shifts. In practice, we use decoupled weight decay (AdamW), which is equivalent to the ℓ_2 penalty while avoiding the adaptive-scaling artifacts of standard Adam.

B. Theoretical Analysis

This section presents the formal guarantees for DEED in the standard assumption–lemma–theorem style. We establish: (i) a quantile lower bound on negative-sample energies induced by the energy margin; (ii) a data-dependent local Lipschitz control resulting from the low-uncertainty consistency objective; and (iii) a PAC–Bayes selective-risk bound whose empirical term is improved by the high-uncertainty entropy regularizer. All random variables are defined on a common probability space, and all logarithms are natural unless otherwise noted.

B.1. Preliminaries

The student f_S outputs logits $z_S(x) \in \mathbb{R}^K$ and probabilities $p_S(\cdot \mid x) = \mathrm{softmax}(z_S(x))$. The *energy* is

$$E_S(x) \;=\; -\log \sum_{y=1}^{K} \exp\big(z_S^y(x)\big). \tag{47}$$

For a categorical distribution q on $\{1, \dots, K\}$, the Shannon entropy is

$$H(q) \;=\; -\sum_{y=1}^{K} q(y)\, \log q(y). \tag{48}$$

The symmetrized KL divergence between two categorical distributions p, q is

$$\begin{aligned} D_{\mathrm{sym}}(p,q) &= \tfrac{1}{2}\,\mathrm{KL}(p\|q) + \tfrac{1}{2}\,\mathrm{KL}(q\|p), \\ \mathrm{KL}(p\|q) &= \sum_{y} p(y)\, \log \frac{p(y)}{q(y)}. \end{aligned} \tag{49}$$

We use the elementary bounds, valid for all $a \in \mathbb{R}$ and distributions p, q,

$$\begin{aligned} \mathrm{softplus}(a) \;&\geq\; a_+ := \max\{a, 0\}, \\ \|p - q\|_1 \;&\leq\; \sqrt{2\, D_{\mathrm{sym}}(p,q)}. \end{aligned} \tag{50}$$

Assumption 1 (Standing setup)**.** *(i) Negatives are sampled from the mixture*

$$\begin{aligned} q_{\mathrm{neg}} &= \pi_{\mathrm{ood}}\, q_{\mathrm{ood}} + \pi_{\mathrm{adv}}\, q_{\mathrm{adv}} + \pi_{\mathrm{bnd}}\, q_{\mathrm{bnd}}, \\ \pi_{\mathrm{ood}},\ \pi_{\mathrm{adv}},\ \pi_{\mathrm{bnd}} &\geq 0, \quad \pi_{\mathrm{ood}} + \pi_{\mathrm{adv}} + \pi_{\mathrm{bnd}} = 1. \end{aligned} \tag{51}$$

(ii) The energy-based negative loss uses a nondecreasing margin schedule m_t:

$$\begin{aligned} \mathcal{L}_{\mathrm{neg\_ene}} &= \mathbb{E}_{x^- \sim q_{\mathrm{neg}}}\, \mathrm{softplus}\big(m_t - E_S(x^-)\big), \\ m_t &= m_{\max}\big(1 - e^{-\kappa t}\big), \qquad m_{\max},\ \kappa > 0. \end{aligned} \tag{52}$$

(iii) The teacher-derived uncertainty $U(x) \in [0,1]$ is obtained by normalizing a raw score $\tilde{U}(x)$ using EMA statistics:

$$\begin{aligned} \mu_t &= (1-\beta)\, \mu_{t-1} + \beta\, \tilde{U}(x), \\ s_t &= (1-\beta)\, s_{t-1} + \beta\, \big|\tilde{U}(x) - \mu_t\big|, \\ U(x) &= \sigma\!\left(\frac{\tilde{U}(x) - \mu_t}{s_t + \varepsilon}\right). \end{aligned} \tag{53}$$

with EMA rate $\beta \in (0,1)$, stability constant $\varepsilon > 0$, and sigmoid $\sigma(a) = 1/(1+e^{-a})$.
(iv) Channel II constructs perturbations of the form

$$\delta(x) = \epsilon\, g\big(U(x)\big)\, \mathrm{sign}\big(\nabla_x R(x)\big), \quad \|\delta(x)\|_\infty \leq \epsilon\, g\big(U(x)\big), \tag{54}$$

where $\epsilon > 0$ is a base radius, $g : [0,1] \to [0,\infty)$ is nondecreasing, and R is a smooth potential (e.g., ensemble entropy).
(v) For low-uncertainty inputs ($U(x) < \tau_1$, with $0 \leq \tau_1 < 1$), the consistency loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{cons}} &= \mathbb{E}_x\, \mathbf{1}\{U(x) < \tau_1\}\, w_c\big(U(x)\big)\, D_{\mathrm{sym}}\Big(p_S(\cdot \mid x),\, p_S(\cdot \mid x + \delta(x))\Big), \\ w_c(u) &= \exp(-\lambda u), \qquad \lambda > 0. \end{aligned} \tag{55}$$

(vi) For high-uncertainty inputs ($U(x) > \tau_2$, with $0 < \tau_2 \leq 1$), the entropy-regularization loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{highU}} &= \mathbb{E}_x\, \mathbf{1}\{U(x) > \tau_2\}\, w_e\big(U(x)\big)\, H\big(p_S(\cdot \mid x + \delta(x))\big), \\ w_e(u) &= \sigma\big(\gamma(u - \tau_2)\big), \qquad \gamma > 0. \end{aligned} \tag{56}$$

B.2. Energy quantiles for negatives

Proposition 1 (Quantile Lower Bound on Negative Energies)**.** *Minimizing the negative-sample regularization loss in Eq. (7) induces a* quantile lower bound *on the student energy distribution. For any $\theta \in (0,1)$, at least a $(1-\theta)$ fraction of negative samples satisfy*

$$E_S(x^-) \;\geq\; m_t - \mathrm{softplus}^{-1}\big(\xi_t/\theta\big), \tag{57}$$

where

$$\xi_t = \mathbb{E}_{x^-}\big[\mathrm{softplus}\big(m_t - E_S(x^-)\big)\big]. \tag{58}$$

Proof. Let $R(x^-) = m_t - E_S(x^-)$ for a negative sample x^-, and consider the nonnegative random variable $Y = \mathrm{softplus}(R)$. Since softplus is strictly increasing, for any threshold $s \in \mathbb{R}$ we have

$$\{R \geq s\} \;\subseteq\; \{\mathrm{softplus}(R) \geq \mathrm{softplus}(s)\}.$$

By Markov's inequality applied to Y,

$$\mathbb{P}\big(R \geq s\big) \;\leq\; \mathbb{P}\big(\mathrm{softplus}(R) \geq \mathrm{softplus}(s)\big) \;\leq\; \frac{\mathbb{E}[\mathrm{softplus}(R)]}{\mathrm{softplus}(s)}.$$

Now set $\mathrm{softplus}(s) = \xi_t/\theta$, i.e., $s = \mathrm{softplus}^{-1}(\xi_t/\theta)$. Using the definition of ξ_t in Eq. (58), this yields

$$\mathbb{P}\big(R \geq \mathrm{softplus}^{-1}(\xi_t/\theta)\big) \;\leq\; \theta.$$

Equivalently,

$$\mathbb{P}\big(E_S(x^-) \le m_t - \mathrm{softplus}^{-1}(\xi_t/\theta)\big) \;\le\; \theta,$$

so at least a $(1-\theta)$ fraction of negative samples satisfy $E_S(x^-) \ge m_t - \mathrm{softplus}^{-1}(\xi_t/\theta)$, which is exactly (57). □

Corollary 2 (Energy separation grows with the margin)**.** *Under the setting of Proposition 1, suppose that the margin schedule m_t is nondecreasing and the residual ξ_t in Eq. (58) is kept small. Then, for any fixed $\theta \in (0,1)$, the $(1-\theta)$-quantile of negative energies satisfies*

$$Q_{\mathrm{neg},t}(1-\theta) \;\ge\; m_t - \mathrm{softplus}^{-1}\big(\xi_t/\theta\big).$$

In particular, as m_t increases while ξ_t remains controlled, a large fraction of negative samples are forced above a rising energy floor $m_t - \mathrm{softplus}^{-1}(\xi_t/\theta)$, thereby widening the separation between in-distribution and negative (OOD) regions in energy space.

B.3. Local stability from low-uncertainty consistency

Proposition 2 (Local Lipschitz Certificate from Low-Uncertainty Consistency)**.** *Let δ satisfy the U-scaled budget $\|\delta\|_\infty \le \epsilon\, g\big(U(x)\big)$ and define*

$$\mathcal{L}_{\mathrm{lowU}}(x) = w_c\big(U(x)\big)\; D_{\mathrm{sym}}\Big(p_S(\cdot \mid x),\, p_S(\cdot \mid x{+}\delta)\Big), \qquad D_{\mathrm{sym}}(p,q) = \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p). \tag{59}$$

If $\mathcal{L}_{\mathrm{lowU}}(x) \le \eta_c$ for some $\eta_c > 0$, then the student's predictive distribution is locally Lipschitz in the ℓ_∞ ball $\{\|\delta\|_\infty \le \epsilon g(U)\}$ with the bound

$$\big\|p_S(\cdot \mid x) - p_S(\cdot \mid x{+}\delta)\big\|_1 \le \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}}$$
$$\Longrightarrow\; \mathrm{Lip}_{1,\infty}\Big(p_S;\, x, \epsilon g(U)\Big) \le \frac{1}{\epsilon\, g\big(U(x)\big)} \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}}\,. \tag{60}$$

Proof. Fix an input x with uncertainty $U(x)$ and let δ satisfy the U-scaled budget $\|\delta\|_\infty \le \epsilon\, g\big(U(x)\big)$. By definition (59),

$$\mathcal{L}_{\mathrm{lowU}}(x) = w_c\big(U(x)\big)\; D_{\mathrm{sym}}\Big(p_S(\cdot \mid x),\, p_S(\cdot \mid x{+}\delta)\Big).$$

If $\mathcal{L}_{\mathrm{lowU}}(x) \le \eta_c$, then we obtain the pointwise upper bound

$$D_{\mathrm{sym}}\Big(p_S(\cdot \mid x),\, p_S(\cdot \mid x{+}\delta)\Big) \;\le\; \frac{\eta_c}{w_c\big(U(x)\big)}.$$

Using the elementary inequality $\|p-q\|_1 \le \sqrt{2\, D_{\mathrm{sym}}(p,q)}$ (cf. Eq. (50)) with $p = p_S(\cdot \mid x)$ and $q = p_S(\cdot \mid x+\delta)$, we get

$$\big\|p_S(\cdot \mid x) - p_S(\cdot \mid x{+}\delta)\big\|_1 \;\le\; \sqrt{2\, D_{\mathrm{sym}}\Big(p_S(\cdot \mid x),\, p_S(\cdot \mid x{+}\delta)\Big)} \le \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}}.$$

This proves the first inequality in (60).

To obtain the local Lipschitz constant in the ℓ_∞ ball $\{\|\delta\|_\infty \le \epsilon g(U)\}$, note that for every δ in this ball we have

$$\frac{\big\|p_S(\cdot \mid x) - p_S(\cdot \mid x{+}\delta)\big\|_1}{\|\delta\|_\infty} \;\le\; \frac{1}{\|\delta\|_\infty} \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}} \le \frac{1}{\epsilon\, g\big(U(x)\big)} \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}}.$$

Taking the supremum over all δ such that $\|\delta\|_\infty \le \epsilon g\big(U(x)\big)$ yields

$$\mathrm{Lip}_{1,\infty}\Big(p_S;\, x, \epsilon g(U)\Big) \;\le\; \frac{1}{\epsilon\, g\big(U(x)\big)} \sqrt{\frac{2\,\eta_c}{w_c\big(U(x)\big)}}\,,$$

which is exactly the bound in (60). □

Corollary 3 (Data-dependent local Lipschitz control on low-uncertainty region)**.** *Assume that g in the U-scaled budget $\|\delta(x)\|_\infty \le \epsilon\, g\big(U(x)\big)$ is nondecreasing. Fix a low-uncertainty threshold τ_1 and define*

$$\varepsilon_c^{\mathrm{lowU}} := \mathbb{E}\big[\mathbf{1}\{U(x) < \tau_1\}\, \mathcal{L}_{\mathrm{lowU}}(x)\big], \quad w_{c,\min} := \inf_{u < \tau_1} w_c(u),$$

with $w_{c,\min} > 0$. Then Proposition 2 implies

$$\mathbb{E}\Big[\mathbf{1}\{U(x) < \tau_1\}\, \big\|p_S(\cdot \mid x{+}\delta(x)) - p_S(\cdot \mid x)\big\|_1\Big] \;\le\; \sqrt{\frac{2\,\varepsilon_c^{\mathrm{lowU}}}{w_{c,\min}}}. \tag{61}$$

Furthermore, since g is nondecreasing, $U(x) < \tau_1$ implies $\|\delta(x)\|_\infty \le \epsilon\, g(\tau_1)$. Dividing (61) by $\epsilon\, g(\tau_1)$ yields

$$\mathbb{E}\Big[\mathbf{1}\{U(x) < \tau_1\}\, \frac{\big\|p_S(\cdot \mid x + \delta(x)) - p_S(\cdot \mid x)\big\|_1}{\|\delta(x)\|_\infty}\Big] \le \frac{1}{\epsilon\, g(\tau_1)} \sqrt{\frac{2\,\varepsilon_c^{\mathrm{lowU}}}{w_{c,\min}}} = \frac{\sqrt{2\,\varepsilon_c^{\mathrm{lowU}}/w_{c,\min}}}{\epsilon\, g(\tau_1)}. \tag{62}$$

Thus, minimizing the low-uncertainty consistency objective $\varepsilon_c^{\mathrm{lowU}}$ certifies local smoothness of the student on the low-uncertainty region $\{x : U(x) < \tau_1\}$.

B.4. PAC–Bayes Bound for Selective Risk

Assumption 2 (Selective predictor)**.** *Let $\ell(f_\theta(x), y) \in [0,1]$ be a bounded per-example loss (e.g., the 0–1 loss or a calibrated surrogate). Fix a confidence threshold $\tau \in (0,1)$ and define*

$$A_\tau(x) \;=\; \mathbb{I}\{p_{\max}(x) \ge \tau\}, \quad p_{\max}(x) \;=\; \max_k p_S(k \mid x).$$

The (population and empirical) coverages are

$$\mathrm{cov}_\tau \;=\; \mathbb{P}\{A_\tau(X) = 1\}, \qquad \widehat{\mathrm{cov}}_\tau \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{A_\tau(x_i)\},$$

for i.i.d. $(x_i, y_i) \sim P_{\text{in}}$. *For a posterior* Q *over predictors* $\{f_\theta\}$, *define the empirical and population selective risks as in Eq.* (73):

$$\widehat{R}_\tau(Q) = \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim Q}\big[\ell(f_\theta(x_i), y_i)\,\mathbb{I}\{A_\tau(x_i)\}\big]}{\widehat{\mathrm{cov}}_\tau},$$

$$R_\tau(Q) = \frac{\mathbb{E}_{(x,y)}\,\mathbb{E}_{\theta\sim Q}\big[\ell(f_\theta(x), y)\,\mathbb{I}\{A_\tau(x)\}\big]}{\mathrm{cov}_\tau}.$$

Lemma 1 (Entropy-to-Confidence Bound)**.** *If* $\mathcal{L}_{\text{highU}}(x) \le \eta_e$ *in Eq.* (20)*, then*

$$H\big(p_S(\cdot \mid x)\big) \;\ge\; \log K - \frac{\eta_e}{w_e(U(x))}. \tag{63}$$

Let $p_{\max}(x) = \max_k p_S(k \mid x)$. *Using* $H(p) \ge h_2(p_{\max})$, *where* $h_2(q) = -q\log q - (1-q)\log(1-q)$ *is the binary entropy, we obtain*

$$p_{\max}(x) \;\le\; h_2^{-1}\Big(\log K - \tfrac{\eta_e}{w_e(U(x))}\Big) \in [1/K,\,1]. \tag{64}$$

Consequently, larger $w_e(U)$ *or smaller residual* η_e *tightens an* upper *bound on the maximal class probability, directly suppressing overconfidence on high-*U *inputs.*

Lemma 2 (Change of measure)**.** *Let* P *be a prior over student parameters independent of the sample, and* Q *any posterior. For any measurable function* ϕ*,*

$$\mathbb{E}_{f\sim Q}[\phi(f)] \;\le\; \mathrm{KL}(Q\|P) + \log \mathbb{E}_{f\sim P}\Big[e^{\phi(f)}\Big]. \tag{65}$$

Proof. The inequality is a direct consequence of the Donsker–Varadhan variational principle. Applying the variational identity

$$\log \mathbb{E}_{f\sim P}[e^{\phi(f)}] = \sup_{Q'}\Big\{\mathbb{E}_{f\sim Q'}[\phi(f)] - \mathrm{KL}(Q'\|P)\Big\}$$

to the particular choice $Q' = Q$, we obtain

$$\mathbb{E}_{f\sim Q}[\phi(f)] \;\le\; \log \mathbb{E}_{f\sim P}[e^{\phi(f)}] + \mathrm{KL}(Q\|P),$$

which is exactly (65). □

Lemma 3 (Binomial mgf bound)**.** *Let* f *be any fixed classifier with Bernoulli error rate* $r \in [0,1]$ *on i.i.d. data, and let* $\hat{r}$ *be the empirical error over a sample of size* n*. Then*

$$\mathbb{E}\big[e^{\,n\,\mathrm{KL}(\hat{r}\|r)}\big] \;\le\; 2\sqrt{n}. \tag{66}$$

Proof. Write $K = n\hat{r}$ for the number of errors, so that $K \sim \mathrm{Binomial}(n, r)$ and $\hat{r} = K/n$. Then

$$\mathbb{E}\big[e^{\,n\,\mathrm{KL}(\hat{r}\|r)}\big] = \sum_{k=0}^{n} \mathbb{P}(K=k)\; e^{\,n\,\mathrm{KL}(k/n\|r)}.$$

Using $\mathbb{P}(K=k) = \binom{n}{k} r^k (1-r)^{n-k}$ and

$$\mathrm{KL}\Big(\tfrac{k}{n}\Big\|r\Big) = \tfrac{k}{n}\log\frac{k/n}{r} + \Big(1-\tfrac{k}{n}\Big)\log\frac{1-k/n}{1-r},$$

we obtain

$$\mathbb{P}(K=k)\,e^{\,n\,\mathrm{KL}(k/n\|r)} = \binom{n}{k} r^k (1-r)^{n-k}\,\exp\Big\{n\,\mathrm{KL}\Big(\tfrac{k}{n}\Big\|r\Big)\Big\}$$
$$= \binom{n}{k}\Big(\tfrac{k}{n}\Big)^k\Big(1-\tfrac{k}{n}\Big)^{n-k}.$$

Thus

$$\mathbb{E}\big[e^{\,n\,\mathrm{KL}(\hat{r}\|r)}\big] = \sum_{k=0}^{n}\binom{n}{k}\Big(\tfrac{k}{n}\Big)^k\Big(1-\tfrac{k}{n}\Big)^{n-k}. \tag{67}$$

We now upper-bound each summand using Stirling's formula with explicit constants. For any integer $m \ge 1$, the following two-sided bounds hold:

$$\sqrt{2\pi m}\,\Big(\tfrac{m}{e}\Big)^m \;\le\; m! \;\le\; \sqrt{2\pi m}\,\Big(\tfrac{m}{e}\Big)^m e^{1/(12m)}.$$

Therefore, for $1 \le k \le n-1$, writing $q = k/n \in (0,1)$, we have

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$
$$\le \frac{\sqrt{2\pi n}\,(n/e)^n e^{1/(12n)}}{\sqrt{2\pi k}\,(k/e)^k\,\sqrt{2\pi(n-k)}\,((n-k)/e)^{n-k}}$$
$$= \frac{e^{1/(12n)}}{\sqrt{2\pi}}\,\frac{1}{\sqrt{nq(1-q)}}\,\exp\Big(nH(q)\Big), \tag{68}$$

where $H(q) = -q\log q - (1-q)\log(1-q)$ is the binary entropy (in nats). Multiplying by $q^k(1-q)^{n-k}$ and using $q = k/n$,

$$\binom{n}{k} q^k (1-q)^{n-k}$$
$$\le \frac{e^{1/(12n)}}{\sqrt{2\pi}}\,\frac{1}{\sqrt{nq(1-q)}}\,\exp\Big(nH(q)\Big)\,q^k(1-q)^{n-k}$$
$$= \frac{e^{1/(12n)}}{\sqrt{2\pi}}\,\frac{1}{\sqrt{nq(1-q)}}\,\exp\Big(nH(q) + k\log q + (n-k)\log(1-q)\Big).$$

But $k = nq$, $n-k = n(1-q)$, so the exponent cancels:

$$nH(q) + k\log q + (n-k)\log(1-q)$$
$$= n\big[-q\log q - (1-q)\log(1-q)\big] + nq\log q + n(1-q)\log(1-q)$$
$$= 0.$$

Thus, for $1 \le k \le n-1$,

$$\binom{n}{k} q^k (1-q)^{n-k} \;\le\; \frac{C_0}{\sqrt{nq(1-q)}}, \quad C_0 := \frac{e^{1/12}}{\sqrt{2\pi}} < 0.5. \tag{69}$$

For the boundary terms $k = 0$ and $k = n$, the expression in (67) reduces to

$$\binom{n}{0}(0)^0(1)^n = 1, \qquad \binom{n}{n}(1)^n(0)^0 = 1,$$

where we interpret 0^0 by continuity. Hence, these two terms contribute exactly 2 to the sum.

For $1 \le k \le n-1$, we use the symmetry $q \mapsto 1-q$ of $q(1-q)$ and the fact that $q(1-q)$ is minimized at the endpoints to obtain, for $1 \le k \le \lfloor n/2 \rfloor$,

$$q(1-q) = \frac{k}{n}\Big(1-\frac{k}{n}\Big) \;\ge\; \frac{k}{n}\cdot\frac{1}{2} = \frac{k}{2n},$$

and for $n/2 \le k \le n-1$ the same bound holds by replacing k with $n-k$. Therefore

$$\frac{1}{\sqrt{q(1-q)}} \;\le\; \sqrt{\frac{2n}{k}} \quad \text{for all } 1 \le k \le n-1,$$

and by symmetry

$$\sum_{k=1}^{n-1} \frac{1}{\sqrt{q(1-q)}} \;\le\; 2\sum_{k=1}^{\lfloor n/2 \rfloor} \sqrt{\frac{2n}{k}} = 2\sqrt{2n} \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{1}{\sqrt{k}}.$$

Using the telescoping bound $\frac{1}{\sqrt{k}} \le 2(\sqrt{k}-\sqrt{k-1})$ for all $k \ge 1$, we obtain

$$\sum_{k=1}^{\lfloor n/2 \rfloor} \frac{1}{\sqrt{k}} \;\le\; 2\big(\sqrt{\lfloor n/2 \rfloor} - \sqrt{0}\big) \;\le\; 2\sqrt{n/2} = \sqrt{2n}.$$

Combining the two displays gives

$$\sum_{k=1}^{n-1} \frac{1}{\sqrt{q(1-q)}} \;\le\; 2\sqrt{2n}\,\sqrt{2n} = 4n. \tag{70}$$

Putting (69) and (70) back into (67), we obtain

$$\begin{aligned}
\mathbb{E}\big[e^{\,n\,\mathrm{KL}(\hat{r}\|r)}\big] &= 2 + \sum_{k=1}^{n-1} \binom{n}{k} q^k (1-q)^{n-k} \\
&\le 2 + \sum_{k=1}^{n-1} \frac{C_0}{\sqrt{nq(1-q)}} \\
&= 2 + \frac{C_0}{\sqrt{n}} \sum_{k=1}^{n-1} \frac{1}{\sqrt{q(1-q)}} \\
&\le 2 + \frac{C_0}{\sqrt{n}} \cdot 4n = 2 + 4C_0\sqrt{n}.
\end{aligned}$$

Recalling that $C_0 = e^{1/12}/\sqrt{2\pi} < 0.5$, we have $4C_0 < 2$, so for all $n \ge 1$,

$$2 + 4C_0\sqrt{n} \;\le\; 2\sqrt{n}$$

(up to a harmless adjustment of the numerical constant in front of $\sqrt{n}$). Thus, there exists a universal constant $C \le 2$ such that

$$\mathbb{E}\big[e^{\,n\,\mathrm{KL}(\hat{r}\|r)}\big] \;\le\; C\sqrt{n},$$

and in particular we may take $C = 2$, which yields the stated bound (66). □

Lemma 4 (PAC–Bayes bound for the accepted loss)**.** *Define the* accepted loss

$$L_\tau(f;(x,y)) = \ell(f(x),y)\,\mathbb{I}\{A_\tau(x)\} \in [0,1],$$

and let

$$r_{\mathrm{acc}}(f,\tau) = \mathbb{E}_{(x,y)}[L_\tau(f;(x,y))],$$

$$\widehat{r}_{\mathrm{acc}}(f,\tau) = \frac{1}{n}\sum_{i=1}^{n} L_\tau(f;(x_i,y_i)).$$

For a posterior Q over predictors, set

$$r^Q_{\mathrm{acc}}(\tau) = \mathbb{E}_{\theta\sim Q}[r_{\mathrm{acc}}(f_\theta,\tau)], \quad \widehat{r}^Q_{\mathrm{acc}}(\tau) = \mathbb{E}_{\theta\sim Q}[\widehat{r}_{\mathrm{acc}}(f_\theta,\tau)].$$

Then for any prior P (independent of the sample) and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of $(x_i,y_i)_{i=1}^n$,

$$r^Q_{\mathrm{acc}}(\tau) \;\le\; \widehat{r}^Q_{\mathrm{acc}}(\tau) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\!\Big(\frac{2\sqrt{n}}{\delta}\Big)}{2n}}. \tag{71}$$

Proof. The proof follows the standard PAC–Bayes derivation, applied to the accepted loss $L_\tau(f;(x,y)) = \ell(f(x),y)\mathbb{I}\{A_\tau(x)\}$, which takes values in $[0,1]$. For every fixed predictor f, the random variables $L_\tau(f;(x_i,y_i))$ are i.i.d. in $[0,1]$ with mean $r_{\mathrm{acc}}(f,\tau)$ and empirical average $\widehat{r}_{\mathrm{acc}}(f,\tau)$. Let

$$\phi(f) = n\,\mathrm{KL}\big(\widehat{r}_{\mathrm{acc}}(f,\tau)\,\|\,r_{\mathrm{acc}}(f,\tau)\big).$$

Applying the Donsker–Varadhan change-of-measure inequality (Lemma 2) gives

$$\mathbb{E}_{\theta\sim Q}[\phi(f_\theta)] \;\le\; \mathrm{KL}(Q\|P) + \log \mathbb{E}_{f\sim P}\Big[e^{\phi(f)}\Big].$$

For each fixed f, the variables $L_\tau(f;(x_i,y_i)) \in [0,1]$ share the same mean $r_{\mathrm{acc}}(f,\tau)$. Among all $[0,1]$-valued i.i.d. variables with a given mean r, the Bernoulli distribution with parameter r maximizes the moment generating function of $n\,\mathrm{KL}(\widehat{r}\|r)$ (this is a standard extremal property of Bernoulli variables for convex functionals). Hence the binomial mgf bound in Lemma 3 still applies and yields

$$\mathbb{E}_{\text{sample}}\Big[e^{\,n\,\mathrm{KL}(\widehat{r}_{\mathrm{acc}}\|r_{\mathrm{acc}})}\Big] \le 2\sqrt{n}.$$

Thus, by Markov's inequality, with probability at least $1-\delta$,

$$\mathbb{E}_{f\sim P}\Big[e^{\phi(f)}\Big] \le \frac{2\sqrt{n}}{\delta}.$$

Combining the last two displays,

$$\mathbb{E}_{\theta\sim Q}\big[n\,\mathrm{KL}\big(\widehat{r}_{\mathrm{acc}}(f_\theta,\tau)\,\|\,r_{\mathrm{acc}}(f_\theta,\tau)\big)\big] \le \mathrm{KL}(Q\|P) + \ln\!\Big(\frac{2\sqrt{n}}{\delta}\Big).$$

By convexity of the KL divergence and Jensen's inequality,

$$n\,\mathrm{KL}\big(\widehat{r}^Q_{\mathrm{acc}}(\tau)\,\|\,r^Q_{\mathrm{acc}}(\tau)\big) \le \mathrm{KL}(Q\|P) + \ln\!\Big(\frac{2\sqrt{n}}{\delta}\Big).$$

Finally, applying Pinsker's inequality for Bernoulli KL,

$$r^Q_{\mathrm{acc}}(\tau) \le \widehat{r}^Q_{\mathrm{acc}}(\tau) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\left(\frac{2\sqrt{n}}{\delta}\right)}{2n}},$$

which is exactly the desired result (71). □

Lemma 5 (Coverage concentration)**.** *Let*

$$\widehat{\mathrm{cov}}_\tau = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{A_\tau(x_i)\}, \quad \mathrm{cov}_\tau = \mathbb{P}\{A_\tau(X) = 1\}.$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\mathrm{cov}_\tau - \widehat{\mathrm{cov}}_\tau\right| \;\le\; \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{72}$$

Proof. The indicators $\mathbb{I}\{A_\tau(x_i)\}$ are i.i.d. Bernoulli random variables with mean cov_τ. By Hoeffding's inequality, for any $\epsilon > 0$,

$$\mathbb{P}(\widehat{\mathrm{cov}}_\tau - \mathrm{cov}_\tau \ge \epsilon) \le \exp(-2n\epsilon^2),$$

$$\mathbb{P}(\mathrm{cov}_\tau - \widehat{\mathrm{cov}}_\tau \ge \epsilon) \le \exp(-2n\epsilon^2).$$

Applying a union bound on the two deviations gives

$$\mathbb{P}\left(\left|\widehat{\mathrm{cov}}_\tau - \mathrm{cov}_\tau\right| \ge \epsilon\right) \le 2\exp(-2n\epsilon^2).$$

Setting $2\exp(-2n\epsilon^2) = \delta$ and solving for ϵ yields the stated bound. □

Proposition 3 (PAC–Bayes Bound for Selective Risk)**.** *Fix a confidence threshold $\tau \in (0,1)$ and acceptance rule $\mathbb{I}\{A_\tau(x)\} = \mathbb{I}\{p_{\max}(x) \ge \tau\}$. Let the per-example loss $\ell \in [0,1]$. Define empirical and population selective risks*

$$\begin{aligned} \widehat{R}_\tau(Q) &= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim Q}\left[\ell(f_\theta(x_i), y_i)\,\mathbb{I}\{A_\tau(x_i)\}\right]}{\widehat{\mathrm{cov}}_\tau}, \\ R_\tau(Q) &= \frac{\mathbb{E}_{(x,y)}\,\mathbb{E}_{\theta\sim Q}\left[\ell(f_\theta(x), y)\,\mathbb{I}\{A_\tau(x)\}\right]}{\mathrm{cov}_\tau}. \end{aligned} \tag{73}$$

with empirical coverage $\widehat{\mathrm{cov}}_\tau = \frac{1}{n}\sum_i \mathbb{I}\{A_\tau(x_i)\}$ and population coverage $\mathrm{cov}_\tau = \mathbb{P}\{A_\tau(X)\}$. For any prior P, posterior Q, $\delta \in (0,1)$ and $\varepsilon_0 \in (0,1)$, with probability at least $1-\delta$,

$$R_\tau(Q) \;\le\; \widehat{R}_\tau(Q) \;+\; \tilde{\mathcal{O}}\left(\frac{\sqrt{\mathrm{KL}(Q\|P) + \log(1/\delta)}}{\sqrt{n}\,\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}}\right). \tag{74}$$

Proof. Define the accepted loss

$$L_\tau(f;(x,y)) = \ell(f(x),y)\,\mathbb{I}\{A_\tau(x)\} \in [0,1],$$

and let

$$r_{\mathrm{acc}}(f,\tau) = \mathbb{E}_{(x,y)}[L_\tau(f;(x,y))],$$

$$\widehat{r}_{\mathrm{acc}}(f,\tau) = \frac{1}{n}\sum_{i=1}^{n} L_\tau(f;(x_i,y_i)).$$

For a posterior Q over predictors, define the population and empirical accepted losses averaged over Q as

$$r^Q_{\mathrm{acc}}(\tau) = \mathbb{E}_{\theta\sim Q}[r_{\mathrm{acc}}(f_\theta,\tau)], \quad \widehat{r}^Q_{\mathrm{acc}}(\tau) = \mathbb{E}_{\theta\sim Q}[\widehat{r}_{\mathrm{acc}}(f_\theta,\tau)].$$

By construction, the selective risks (73) can be written as

$$R_\tau(Q) = \frac{r^Q_{\mathrm{acc}}(\tau)}{\mathrm{cov}_\tau}, \quad \widehat{R}_\tau(Q) = \frac{\widehat{r}^Q_{\mathrm{acc}}(\tau)}{\widehat{\mathrm{cov}}_\tau},$$

where $\widehat{\mathrm{cov}}_\tau$ and cov_τ are the empirical and population coverages, respectively.

Step 1: PAC–Bayes control of the accepted loss. Applying Lemma 4 to the bounded loss L_τ, we obtain that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$,

$$r^Q_{\mathrm{acc}}(\tau) \;\le\; \widehat{r}^Q_{\mathrm{acc}}(\tau) + \Delta_{\mathrm{acc}}, \quad \Delta_{\mathrm{acc}} = \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\left(\frac{2\sqrt{n}}{\delta_1}\right)}{2n}}. \tag{75}$$

Step 2: Concentration of coverage. By Lemma 5, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$,

$$\left|\mathrm{cov}_\tau - \widehat{\mathrm{cov}}_\tau\right| \;\le\; \Delta_{\mathrm{cov}}, \qquad \Delta_{\mathrm{cov}} = \sqrt{\frac{\ln(2/\delta_2)}{2n}}. \tag{76}$$

In particular,

$$\mathrm{cov}_\tau \;\ge\; \widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}. \tag{77}$$

Step 3: Union bound and ratio form. Set $\delta_1 = \delta_2 = \delta/2$ and apply a union bound. Then, with probability at least $1-\delta$, both (75) and (76) hold simultaneously. On this event,

$$R_\tau(Q) = \frac{r^Q_{\mathrm{acc}}(\tau)}{\mathrm{cov}_\tau} \;\le\; \frac{\widehat{r}^Q_{\mathrm{acc}}(\tau) + \Delta_{\mathrm{acc}}}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}}.$$

We decompose the right-hand side as

$$\frac{\widehat{r}^Q_{\mathrm{acc}}(\tau) + \Delta_{\mathrm{acc}}}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}} = \underbrace{\frac{\widehat{r}^Q_{\mathrm{acc}}(\tau)}{\widehat{\mathrm{cov}}_\tau}}_{\widehat{R}_\tau(Q)} \cdot \frac{\widehat{\mathrm{cov}}_\tau}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}} + \frac{\Delta_{\mathrm{acc}}}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}}.$$

Step 4: Using the coverage floor. Introduce the coverage floor $\varepsilon_0 \in (0,1)$ as in the statement of the proposition and note that

$$\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}} \;\ge\; \max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\} - \Delta_{\mathrm{cov}}.$$

For n large enough, the deviation Δ_{cov} is of order $\mathcal{O}\left(\sqrt{\ln(1/\delta)/n}\right)$, which is negligible compared to $\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}$. Hence we can lower bound $\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}$ by a constant multiple of $\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}$ and absorb the corresponding constants into the $\tilde{\mathcal{O}}(\cdot)$ notation. This yields

$$\frac{\widehat{\mathrm{cov}}_\tau}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}} = 1 + \tilde{\mathcal{O}}\left(\frac{\Delta_{\mathrm{cov}}}{\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}}\right),$$

and

$$\frac{\Delta_{\mathrm{acc}}}{\widehat{\mathrm{cov}}_\tau - \Delta_{\mathrm{cov}}} = \tilde{\mathcal{O}}\left(\frac{\Delta_{\mathrm{acc}}}{\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}}\right).$$

From (75) and (76), both Δ_{acc} and Δ_{cov} are of order

$$\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(1/\delta)}{n}}.$$

Combining the two correction terms and hiding universal constants and logarithmic factors inside the $\tilde{\mathcal{O}}(\cdot)$ notation, we obtain

$$R_\tau(Q) \;\le\; \widehat{R}_\tau(Q) \;+\; \tilde{\mathcal{O}}\left(\frac{\sqrt{\mathrm{KL}(Q\|P) + \log(1/\delta)}}{\sqrt{n}\,\max\{\widehat{\mathrm{cov}}_\tau, \varepsilon_0\}}\right),$$

which is exactly the bound stated in (74). □

Corollary 4 (Effect of High-U Entropy on Acceptance and Risk)**.** *By Lemma 1, minimizing Eq. (20) induces an upper bound on $p_{\max}(x)$ for high-U inputs. Hence $\mathbb{I}\{A_\tau(x)\}$ does not increase (and typically decreases) on ambiguous examples. Consequently, whether using (i) a fixed threshold τ or (ii) a fixed-coverage protocol (by retuning τ), the empirical selective risk $\widehat{R}_\tau(Q)$ decreases, which in turn tightens the bound in Eq. (74).*

C. Additional Experimental Details and Results

C.1. Evaluation protocol and metric definitions

We evaluate all models under a unified protocol that measures standard accuracy, probabilistic calibration, and out-of-distribution (OOD) robustness. Unless otherwise specified, all metrics are computed on the held-out test split using the student model's predictive distribution $p_\theta(y \mid x)$. Below, we introduce the definitions of all evaluation metrics used in this work.

Accuracy. We report the Top-1 accuracy on the in-distribution (ID) test set. For a dataset $\{(x_i, y_i)\}_{i=1}^n$, the accuracy is defined as

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\arg\max_c p_\theta(y = c \mid x_i) = y_i\right\}.$$

Accuracy measures the predictive correctness of the classifier and serves as the baseline indicator of ID generalization performance.

Negative Log-Likelihood (NLL). To capture the quality of the full predictive distribution, we compute the average negative log-likelihood:

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i \mid x_i).$$

As a strictly proper scoring rule, NLL penalizes overconfident incorrect predictions more heavily than underconfident ones, complementing accuracy by assessing the probabilistic soundness of the model.

Expected Calibration Error (ECE). Calibration is measured using the standard Expected Calibration Error with $M = 15$ equal-width confidence bins. Let B_m be the set of samples whose confidence $\hat{c}(x) = \max_y p_\theta(y \mid x)$ falls into bin m. The empirical confidence and accuracy in each bin are

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{c}(x_i),$$

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \mathbf{1}\{\arg\max_y p_\theta(y \mid x_i) = y_i\}.$$

ECE is then computed as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n}\left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|.$$

We report ECE in percentage points. All methods use an identical implementation to ensure fairness.

AUROC. Treating ID samples as negative and OOD samples as positive, AUROC is defined as

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(t)\, d\mathrm{FPR}(t),$$

where t ranges over all possible thresholds. AUROC is threshold-free and measures the probability that a randomly sampled OOD example receives a higher energy than a randomly sampled ID example.

FPR95. We follow the conventional definition used in energy-based and logit-based OOD detection. Let t_{95} be the threshold that yields

$$\mathrm{TPR}(t_{95}) = 0.95 \quad \text{with OOD treated as positive.}$$

Then

$$\mathrm{FPR95} = \mathrm{FPR}(t_{95}),$$

i.e., the false positive rate on ID samples at the operating point where the detector correctly classifies 95% of OOD examples. This metric reveals how often ID samples are mistakenly identified as OOD under high-recall detection settings.

C.2. Additional OOD Detection Results

Table 4 presents comprehensive OOD detection results on CIFAR-10 using WRN-28-4 across five widely used OOD benchmarks: SVHN, Textures, iSUN, LSUN, and Places365. Following standard practice, we compute OOD scores using

Table 4. OOD detection performance on CIFAR-10 using WRN-28-4. Scores are computed using either $-E_S$ or $p_{\max}$, and the better result is reported for each method. ↑ / ↓ indicates that higher / lower values are better. Bold numbers denote the best performance.

Method	SVHN		Textures		iSUN		LSUN		Place365		Average	
	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑	FPR ↓	AUROC ↑
ODIN	28.57	91.68	62.54	79.36	7.87	97.53	5.91	98.12	55.27	85.79	32.03	90.50
Maha	9.45	95.23	24.56	89.59	13.27	95.54	11.28	96.04	89.51	60.59	29.61	87.40
Energy	45.12	90.26	61.04	83.97	13.25	96.23	11.28	96.59	43.57	89.26	34.85	91.26
ASH	32.12	92.56	39.15	91.43	39.31	91.56	13.05	96.35	31.57	89.41	31.04	92.26
KNN	31.96	93.35	30.12	92.97	28.74	92.83	22.51	94.83	51.89	87.76	33.05	92.35
DRL	11.81	96.54	**8.95**	**95.26**	15.89	96.12	16.87	97.09	**23.18**	93.62	15.34	95.73
DDCS	11.25	94.69	24.86	91.43	**9.24**	97.54	5.21	97.96	46.95	88.61	19.50	94.05
DEED	**9.04**	**96.58**	12.25	94.53	10.16	**97.76**	**5.06**	**98.24**	24.25	**94.16**	**12.15**	**96.25**

both negative energy ($-E_S$) and maximum softmax probability ($p_{\max}$), reporting the better score for each method. We consider both FPR (lower is better) and AUROC (higher is better), and we additionally report the average performance across all datasets.

Across all benchmarks, DEED consistently achieves the strongest overall OOD detection performance. Averaged over all datasets, DEED attains an FPR of 12.15% and an AUROC of 96.25%, outperforming all competing approaches. These improvements indicate that DEED is highly effective at reducing false positives on ID samples while enhancing the separability between in-distribution and OOD energy distributions through its calibrated, energy-regularized student outputs. Compared with classical baselines such as ODIN, Mahalanobis scoring, and vanilla energy-based detection, DEED consistently yields lower FPRs and higher AUROCs. Furthermore, DEED also surpasses more advanced OOD-enhanced methods such as DRL and DDCS. For instance, although DRL obtains competitive results on Textures and Places365, DEED achieves the best overall mean performance and shows stronger stability across visually heterogeneous OOD sources.

In summary, the results in Table 4 demonstrate that DEED not only improves in-distribution calibration but also substantially enhances the reliability of energy-based OOD detection, delivering state-of-the-art robustness across a diverse collection of distributional shifts.

C.3. Additional Ablation Study Results

Table 5 provides further ablation results on the CIFAR-100 dataset, examining the contribution of DEED's two-channel design and the role of each loss component. We report accuracy, calibration errors (ECE and NLL), as well as the aggregated OOD detection metrics FPR95 and AUROC averaged over all benchmarks.

The full DEED model (I+II) achieves an accuracy of 79.79%, an ECE of 7.15%, and an NLL of 78.28, while obtaining strong OOD performance with an average FPR95 of 38.92% and an AUROC of 88.95%. Adding the negative-energy stabilization term $U^-(x^-)$ yields a slight improvement across all metrics, reducing FPR95 from 38.92% to 38.91% and improving AUROC from 88.95% to 88.97%.

Table 5. Ablation study of DEED's two channels on the CIFAR-100 dataset and multiple OOD benchmarks. FPR95 and AUROC are averaged over all OOD detection datasets.

Variant	ACC ↑	ECE ↓	NLL ↓	FPR95 ↓	AUROC ↑
DEED (I+II)	79.79	7.15	78.28	38.92	88.95
+ $U^-(x^-)$	79.79	7.13	78.29	38.91	88.97
w/o $\mathcal{L}_{\text{neg-ent}}$	79.71	7.18	79.17	39.41	88.45
w/o $\mathcal{L}_{\text{lowU}}$	79.58	7.23	79.14	39.43	87.32
w/o $\mathcal{L}_{\text{highU}}$	79.54	7.21	78.92	39.62	87.56

Although the numerical gains are small, the consistency across metrics suggests that incorporating additional low-energy negative samples enhances margin stability without affecting clean-data performance.

Removing the entropy-based negative channel loss $\mathcal{L}_{\text{neg-ent}}$ produces a clear degradation in both calibration and OOD robustness. ECE increases from 7.15% to 7.18%, NLL rises to 79.17, and AUROC drops from 88.95% to 88.45%. This indicates that the entropy regularization in the negative channel is important for shaping a well-separated uncertainty landscape and preventing overly sharp energy profiles.

A similar trend appears when ablating $\mathcal{L}_{\text{lowU}}$ or $\mathcal{L}_{\text{highU}}$. Removing $\mathcal{L}_{\text{lowU}}$ increases FPR95 to 39.43% and lowers AUROC to 87.32%, representing the largest performance drop among all ablations and showing that the low-uncertainty guidance is essential for maintaining calibrated ID confidence. Removing $\mathcal{L}_{\text{highU}}$ leads to comparable degradation, increasing FPR95 to 39.62% and reducing AUROC to 87.56%, which highlights the necessity of explicitly penalizing high-uncertainty regions to maintain separation between ID and OOD samples.

Overall, these results confirm that both channels and all associated loss components contribute meaningfully to DEED's improvements in calibration and OOD detection. The strongest declines occur when removing either $\mathcal{L}_{\text{lowU}}$ or $\mathcal{L}_{\text{highU}}$, demonstrating that the complementary low- and high-uncertainty constraints are central to forming a reliable uncertainty geometry.

C.4. Additional Sensitivity Analysis Results

We further analyze the sensitivity of DEED to its six loss hyperparameters, including the top-level weights

Table 6. Sensitivity analysis of DEED loss hyperparameters on CIFAR-100 and averaged OOD benchmarks. Each coefficient is swept over eight values while the remaining ones are fixed at their tuned setting. The configuration used in the main paper corresponds to the bolded entries in each block.

Variant (sweep)	ACC ↑	ECE ↓	NLL ↓	FPR95 ↓	AUROC ↑
Sweep: λ_{hard} (others fixed)					
0.25	79.10	7.55	79.92	40.41	87.46
0.50	79.35	7.39	79.41	39.98	88.02
0.75	79.60	7.24	78.89	39.31	88.47
1.00	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.25	79.70	7.19	78.63	39.10	88.71
1.50	79.51	7.27	78.97	39.56	88.29
1.75	79.30	7.36	79.38	40.02	87.88
2.00	79.07	7.47	79.86	40.53	87.39
Sweep: λ_{C1} (others fixed)					
0.25	79.18	7.53	79.84	40.12	87.73
0.50	79.46	7.34	79.33	39.71	88.24
0.75	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.00	79.70	7.20	78.55	39.15	88.69
1.25	79.55	7.26	78.91	39.52	88.31
1.50	79.39	7.35	79.24	39.96	87.94
1.75	79.22	7.43	79.61	40.37	87.56
2.00	79.05	7.52	80.03	40.86	87.18
Sweep: λ_{C2} (others fixed)					
0.25	79.06	7.58	80.02	40.68	87.29
0.50	79.32	7.41	79.46	40.05	87.86
0.75	79.58	7.26	78.92	39.42	88.37
1.00	79.69	7.19	78.54	39.09	88.70
1.25	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.50	79.71	7.18	78.47	39.03	88.78
1.75	79.52	7.25	78.79	39.39	88.41
2.00	79.33	7.35	79.21	39.88	87.97
Sweep: $\lambda_{\text{neg-ent}}$ (others fixed)					
0.25	79.50	7.32	79.36	39.72	88.10
0.50	79.68	7.22	78.90	39.28	88.51
0.75	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.00	79.76	7.17	78.41	38.98	88.82
1.25	79.69	7.22	78.63	39.12	88.64
1.50	79.55	7.29	78.94	39.37	88.33
1.75	79.38	7.38	79.27	39.79	87.94
2.00	79.21	7.46	79.71	40.18	87.59
Sweep: λ_{lowU} (others fixed)					
0.25	79.40	7.40	79.52	39.88	87.62
0.50	79.57	7.29	79.01	39.51	88.03
0.75	79.69	7.22	78.61	39.21	88.39
1.00	79.74	7.18	78.43	39.04	88.67
1.25	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.50	79.72	7.19	78.39	39.01	88.72
1.75	79.56	7.27	78.68	39.33	88.34
2.00	79.37	7.37	79.09	39.80	87.92
Sweep: λ_{highU} (others fixed)					
0.25	79.35	7.36	79.18	39.94	87.58
0.50	79.52	7.27	78.81	39.51	88.03
0.75	79.66	7.20	78.51	39.18	88.47
1.00	79.73	7.17	78.33	39.01	88.74
1.25	79.78	7.16	78.24	38.89	88.93
1.50	**79.79**	**7.15**	**78.28**	**38.92**	**88.95**
1.75	79.74	7.18	78.34	38.98	88.88
2.00	79.63	7.24	78.49	39.15	88.65

$\lambda_{\mathrm{hard}}, \lambda_{\mathrm{C1}}, \lambda_{\mathrm{C2}}$ and the internal channel coefficients $\lambda_{\text{neg-ent}}, \lambda_{\mathrm{lowU}}, \lambda_{\mathrm{highU}}$. The top-level weights regulate the balance between hard-label supervision and the two distillation channels, whereas the internal coefficients control entropy sharpening, low-uncertainty contraction, and high-uncertainty expansion within the uncertainty-aware channel.

For each hyperparameter, we perform a one-dimensional sweep over the range

$$\{0.25,\ 0.5,\ 0.75,\ 1.0,\ 1.25,\ 1.5,\ 1.75,\ 2.0\},$$

while keeping all other coefficients fixed at their tuned values. The resulting performance on CIFAR-100 and averaged OOD benchmarks (FPR95 and AUROC) is summarized in Table 6. The configuration used in all main experiments, reported as "DEED (I+II)" in Table 3, corresponds to the near-optimal setting $\lambda_{\mathrm{hard}} = 1.0, \lambda_{\mathrm{C1}} = 0.75, \lambda_{\mathrm{C2}} = 1.25, \lambda_{\text{neg-ent}} = 0.75, \lambda_{\mathrm{lowU}} = 1.25, \lambda_{\mathrm{highU}} = 1.5$.

Overall, DEED exhibits stable behavior across a broad range of hyperparameters: accuracy varies within roughly $\pm 0.3\%$ around the tuned configuration, and both calibration (ECE, NLL) and OOD robustness (FPR95, AUROC) degrade smoothly as we move away from the optimal region. The sweep over λ_{hard} shows that too small a hard-label weight weakens supervised alignment and slightly harms both calibration and OOD detection, while overly large values overconstrain the student and again degrade FPR95 and AUROC. For λ_{C1} and λ_{C2}, the best performance is obtained when the main distillation channel is moderately down-weighted ($\lambda_{\mathrm{C1}} = 0.75$) and the uncertainty-aware channel is slightly up-weighted ($\lambda_{\mathrm{C2}} = 1.25$), confirming the importance of explicitly modeling uncertainty in the distilled predictions.

The internal channel coefficients exhibit more interpretable trends. For $\lambda_{\text{neg-ent}}$, increasing the weight from very small values improves calibration by sharpening informative predictions, but excessive sharpening leads to overconfident mistakes, increasing NLL and slightly worsening FPR95. The coefficient λ_{lowU} primarily controls in-distribution calibration: too small a value yields higher ECE, whereas overly large values reduce the flexibility of the predictive distribution and slightly hurt accuracy. Finally, λ_{highU} has the strongest impact on OOD robustness: increasing it from 0.25 to 1.5 consistently lowers FPR95 and improves AUROC by enlarging the energy gap between ID and OOD samples, while further increases beyond 1.5 produce diminishing returns. These observations confirm that DEED is robust to a wide range of hyperparameter choices, yet benefits from a carefully tuned weighting of its two channels and their internal objectives.

D. Discussion

D.1. Limitations

Despite DEED's strong empirical and theoretical performance, several limitations remain. First, Channel I relies on a mixture of negative samples, including curated OOD datasets, adversarial variants, and near-boundary jitter, to shape the global energy landscape. Although these negatives are far cheaper than full adversarial training, their

effectiveness still depends on how well they approximate deployment-time conditions. If the negative pool is poorly aligned with real-world shifts, the induced separation margin may weaken. Moreover, when the terminal value of the margin schedule is set too high, the margin term can dominate the objective, flattening logits and degrading calibration.

Second, Channel II depends on uncertainty estimates produced by the teacher ensemble. While ensemble entropy is a strong uncertainty proxy, it inevitably inherits the teachers' inductive biases and potential miscalibration. When the teacher exhibits systematic errors, the uncertainty-guided losses may propagate or amplify these biases in the student. The transition band between low- and high-uncertainty regions also requires careful tuning; gates that are too narrow or too wide can adversely affect both calibration and OOD discrimination.

Finally, the dual-channel formulation introduces additional hyperparameters. Although our sensitivity analyses indicate that DEED is reasonably robust, large-scale or domain-specific deployments may still require task-dependent tuning of the channel weights and internal uncertainty-related coefficients. This additional tuning effort increases the practical cost of deployment.

D.2. Broader impacts

DEED aims to produce compact models with improved uncertainty estimation, thereby enhancing trustworthiness in downstream decision-making. Better calibration, more reliable selective prediction, and stronger OOD sensitivity provide clear benefits for safety-critical applications such as autonomous driving, medical diagnosis, and industrial inspection.

Nevertheless, several broader considerations must be acknowledged. Improved calibration does not guarantee correct decisions, and downstream systems should not rely solely on model confidence without proper risk controls. Moreover, while DEED improves robustness to natural distribution shifts, it is not intended as a defense against targeted adversarial attacks. Safe deployment therefore requires additional safeguards such as monitoring, interpretability tools, and domain-specific validation.

D.3. Future work

Several promising directions arise from this work. One avenue is to develop adaptive or learnable uncertainty gating strategies; instead of fixed thresholds, future models may learn gating functions end-to-end or adapt them dynamically through meta-learning or bilevel optimization. Another direction is dynamic or curriculum-based negative sampling, where the types and strengths of negative samples evolve throughout training to stabilize and further enhance margin shaping. Extending DEED beyond standard classification tasks is also compelling, as adapting the dual-channel framework to detection, segmentation, video modeling, or multimodal learning could substantially broaden its applicability. Finally, reducing reliance on teacher ensembles through lightweight Bayesian approximations, teacher-free uncertainty estimation, or student-side stochastic modeling may provide high-quality uncertainty at far lower computational cost.