CAN WE BREAK THE LOOP? ACTIVE BOUNDARY RESTORATION FOR LEARNING WITH NOISY LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most methods for learning with noisy labels identify reliable samples through training loss, predictive confidence, or model agreement. These reliability estimates depend on an evolving representation: as noisy supervision distorts class geometry, hard clean examples and mislabeled examples can become increasingly difficult to distinguish near class boundaries. Robust learning therefore requires maintaining a representation in which sample reliability remains identifiable. We propose HamBR, a framework that couples a dynamically filtered support bank with active boundary regularization under noisy supervision. Hamiltonian exploration serves as one structured proposal mechanism within this intervention, while an explicit density-retention rule controls anchor support relative to inter-class initialization. Geometric regularization then draws reliable features toward their class cores and away from the anchors. Theoretical analysis establishes a relative support-density bound, characterizes optimization of prototype–outlier margins, and provides an anchor-induced boundary-separation guarantee. Our evaluation examines classification performance, representation geometry, and alternative candidate-generation strategies under noisy supervision. Across synthetic- and real-noise benchmarks, the reported results support the benefits of active geometric intervention for robust learning. These findings motivate considering sample selection and representation geometry jointly, with active boundary restoration complementing advances in selection rules.

1 INTRODUCTION

Label noise in large-scale visual datasets can lead deep networks to memorize incorrect supervision, making reliable-sample identification a central problem in learning with noisy labels (Arpit et al., 2017). Many methods exploit small training losses, predictive confidence, model agreement, or statistical distributions to estimate sample reliability. Selected examples then support supervised learning, label correction, or semi-supervised training (Han et al., 2018; Li et al., 2020; Xiao et al., 2022). Sample selection has consequently become a major paradigm for robust learning.

This paradigm relies on a fundamental condition: the current representation must support meaningful reliability estimates. Losses, confidence scores, and relationships between examples all depend on features learned by the model. When those features preserve clear class structure, clean and mislabeled examples can often be distinguished statistically. As noisy supervision distorts the representation, the geometric basis of this distinction can weaken. Representation learning is therefore central to robust sample selection, as also emphasized by methods that improve feature structure under label noise (Li et al., 2022; Karim et al., 2022).

The difficulty is particularly acute near class boundaries. Noisy supervision can disrupt compact class structure, narrow low-support inter-class regions, and move hard clean examples toward ambiguous boundaries. At the same time, some mislabeled examples can acquire low loss or high confidence, complicating reliability estimation (Northcutt et al., 2021; Xia et al., 2020). A selector that successfully identifies easy examples may consequently omit informative hard clean examples and admit mislabeled ones near these regions. Those decisions shape subsequent training and the representation used by the next selection step. We refer to this coupling as *selection–boundary feed-*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

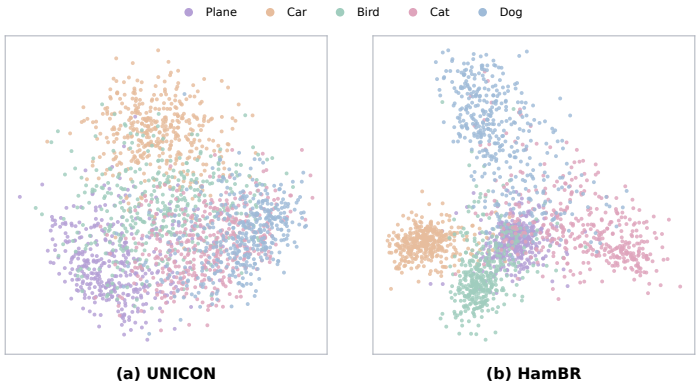

Figure 1: **Feature distributions for five classes.** (a) UNICON. (b) HamBR. Colors identify the plane, car, bird, cat, and dog classes; each panel contains 400 displayed points per class. The two-dimensional projections visualize class-wise feature organization.

back. The feature distributions in Figure 1 motivate examining class geometry alongside selection behavior.

This perspective broadens the question of sample selection. Every selection rule depends on the separability available in the current representation. Once class geometry has deteriorated, improving the selection rule alone leaves the underlying source of ambiguity in place. This raises a more fundamental question: *Can we actively maintain and restore the class boundaries on which sample selection depends, making hard clean examples and incorrect supervision distinguishable again?*

We propose HamBR, an active boundary restoration framework that couples a dynamically filtered support bank with active boundary regularization under noisy supervision. The bank estimates class structure from currently reliable features and is updated as reliability estimates evolve. Hamiltonian exploration is used as one structured proposal mechanism within this intervention, and an explicit density-retention rule constrains anchor support relative to inter-class initialization. Geometric regularization then draws reliable features toward their class prototypes and away from the retained anchors, enlarging the separation between class cores and ambiguous inter-class regions. HamBR thus intervenes directly on the representation used by subsequent sample selection (Figure 2).

The components have complementary roles. Building on Hamiltonian outlier synthesis for OOD detection (Li & Zhang, 2025), finite Hamiltonian trajectories explore candidate positions along the current feature geometry. Inter-class initialization and explicit density retention control the support of the resulting anchors. Our theoretical analysis quantifies this relative density constraint and shows how the HamBR objective favors larger prototype–outlier similarity margins, yielding a separation certificate for anchor-induced boundaries. A local transfer analysis further relates proxy-margin improvement to classifier geometry under alignment and bounded head drift.

Our evaluation studies how active geometric intervention affects representation structure and robust classification. We compare alternative candidate-generation mechanisms and analyze feature geometry alongside performance on synthetic- and real-noise benchmarks. Hard clean examples, actual classifier margins, and contamination of the selected set provide complementary criteria for assessing whether the intervention improves the local structure most relevant to reliability estimation.

Our contributions are threefold:

- **A representation-based view of sample-selection failure.** We formulate selection–boundary feedback: reliability estimation depends on evolving class geometry, while selected supervision shapes that geometry. This expands the focus from designing selectors to maintaining representations that support reliable selection.
- **An active boundary restoration framework.** HamBR couples a dynamically filtered support bank with active boundary regularization under noisy supervision. Hamiltonian exploration is used as one structured proposal mechanism within this intervention.

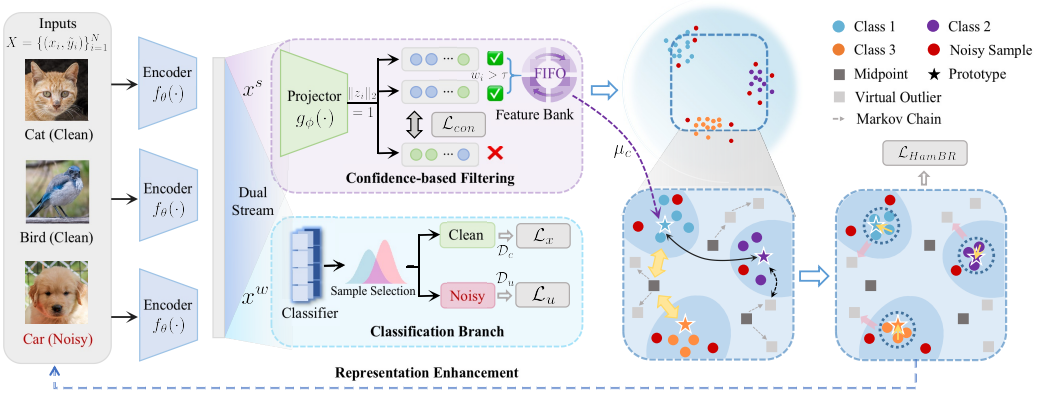

Figure 2: **HamBR framework.** The classification branch partitions examples, while a confidence-filtered bank supports prototype construction and candidate exploration. Retained virtual anchors regularize selected features relative to their class prototypes.

- **Analysis connecting geometry to robust learning.** We establish relative density and prototype–outlier margin properties, derive an anchor-induced separation certificate, and evaluate candidate-generation strategies, feature geometry, and performance on noisy-label benchmarks.

2 Related Work

Sample selection and robust representations. Small-loss selection exploits early learning (Arpit et al., 2017); Co-teaching cross-filters examples (Yu et al., 2019), while DivideMix and later variants partition or expand the reliable set (Li et al., 2020; Cordeiro et al., 2023; Zhang et al., 2025; Xiao et al., 2022). However, boundary noise can retain low loss or high confidence (Northcutt et al., 2021; Xia et al., 2020). Representation methods improve feature geometry (Li et al., 2022; Karim et al., 2022), but conventional contrastive objectives do not explicitly constrain low-density decision regions (Chen et al., 2020; He et al., 2020). HamBR couples selection with boundary-oriented regularization.

Relationship to HamOS. VOS and HamOS regularize OOD recognition through synthesized outliers (Du et al., 2022; Li & Zhang, 2025). HamOS introduces inter-class midpoint initialization, hyperspherical Hamiltonian exploration, and a density-based barrier. HamOS synthesizes virtual outliers for OOD detection from labeled in-distribution data. In noisy-label learning, however, membership in the clean support set is latent, dynamically estimated, and potentially contaminated. HamBR therefore couples boundary-anchor construction with reliability estimation and subsequent representation updates.

3 Method

Let $h_\theta(x)$ denote a backbone feature, $f_W(h)$ the classification head, and $z = P_\phi(h)/\|P_\phi(h)\|$ the projected unit feature used by HamBR. The classifier and geometric regularizer act in different spaces. Selection at epoch t produces $S_t = \mathcal{A}(\{\ell_i(\theta_t, W_t)\}_i)$, and training on S_t changes the representation used to form S_{t+1}. HamBR regularizes this update while retaining the sample-selection rule $\mathcal{A}$.

3.1 Representation Space Initialization

We warm up two networks with Generalized Cross Entropy (GCE) and fit a GMM to sample losses to estimate clean probabilities. The support bank retains only samples whose probabilities exceed γ for T_{filter} consecutive epochs:

$$\Omega_{ref}^{t} = \bigcap_{\tau=t-T_{filter}+1}^{t} i \mid P(w_i = 1 \mid \mathcal{L}_i^{(\tau)}) > \gamma. \tag{1}$$

3.2 Boundary-Oriented Virtual Outlier Probing

HamBR generates candidates with finite spherical Hamiltonian proposals and retains them using an explicit support-density constraint.

Support-density energy. Let $\mathcal{B} = \{(\mathbf{k}_i, w_i)\}_{i=1}^{N}$ be the confidence-weighted feature bank, with normalized features $\mathbf{k}_i \in \mathbb{S}^{d-1}$ and weights $w_i \geq 0$. For a query $\mathbf{z} \in \mathbb{S}^{d-1}$, we define the weighted top-K support score

$$\widehat{\rho}_K(\mathbf{z}) = \sum_{(\mathbf{k}_j, w_j) \in \mathcal{N}_K(\mathbf{z})} (w_j + \varepsilon_w) \exp\left(\frac{\mathbf{z}^\top \mathbf{k}_j}{\tau}\right), \tag{2}$$

where $\mathcal{N}_K(\mathbf{z})$ indexes the K largest weighted kernel terms, equivalently the K largest values of $\mathbf{z}^\top \mathbf{k}_j/\tau + \log(w_j + \varepsilon_w)$. The corresponding potential is

$$U(\mathbf{z}) = -\log \widehat{\rho}_K(\mathbf{z}). \tag{3}$$

Well-supported class interiors have low U, whereas low-support regions have high U.

Candidate generation and retention. Each chain starts at $\mathbf{z}^{(0)} = \mathrm{Normalize}((\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)/2 + 0.01\boldsymbol{\xi})$, where $a \neq b$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, I)$, with Gaussian momentum projected onto the tangent space. For step size ϵ, the normalized, leapfrog-inspired updates are

$$\mathbf{v}^- = \mathbf{v} - \tfrac{\epsilon}{2}\nabla U(\mathbf{z}), \tag{4}$$

$$\mathbf{z}^+ = \mathrm{Normalize}\left(\mathbf{z}\cos(\omega\epsilon) + \frac{\mathbf{v}^-}{\omega}\sin(\omega\epsilon)\right), \quad \omega = \|\mathbf{v}^-\|_2, \tag{5}$$

followed by momentum rotation and a second half-gradient update. These updates use Euclidean gradients and position normalization to generate finite candidates; the support criterion below determines candidate retention.

Candidates are evaluated using the unweighted all-bank log-support score

$$s(\mathbf{z}) = \log \sum_{i=1}^{N} \exp\left(\kappa \mathbf{z}^\top \mathbf{k}_i\right). \tag{6}$$

After a proposal round, the candidate $\widetilde{\mathbf{z}}^{(r)}$ is retained only if

$$s\left(\widetilde{\mathbf{z}}^{(r)}\right) \leq s\left(\mathbf{z}^{(0)}\right) + \delta; \tag{7}$$

otherwise, the position rolls back to its state before that round. This bounds support relative to initialization; absolute low support additionally requires a low-support initial midpoint. Algorithm 1 and Appendix F provide the full procedure and proof.

3.3 MLE-based Manifold Regularization

For a clean unit-normalized feature $\mathbf{x}$, its class prototype $\boldsymbol{\mu}_c$, and M accepted virtual outliers $\mathbf{v}_j$, we maximize the class posterior by treating virtual outliers as competing anchors. The resulting negative log-likelihood is

$$\mathcal{L}_{\text{HamBR}} = -\log \frac{\exp(\mathbf{x}^\top \boldsymbol{\mu}_c/\tau)}{\exp(\mathbf{x}^\top \boldsymbol{\mu}_c/\tau) + \sum_{j=1}^{M} \exp(\mathbf{x}^\top \mathbf{v}_j/\tau)}. \tag{8}$$

At fixed anchors, minimizing this loss favors larger prototype–outlier similarity margins.

Dataset	CIFAR-10					
Noise Mode	Sym.				Asym.	
Method	20%	50%	80%	90%	40%	Avg.
Standard CE	73.4	58.9	49.2	30.8	60.9	54.6
GCE	77.2	75.7	66.7	44.1	67.7	66.3
DivideMix	79.5	74.6	70.7	44.4	68.1	67.5
RRL	88.4	83.3	74.7	45.2	76.5	73.6
UNICON	72.8	71.5	68.1	41.8	76.4	66.1
ProMix	87.4	83.7	68.7	49.3	77.0	73.2
LongReMix	81.4	79.0	67.6	37.3	67.7	66.6
L2B	86.7	82.1	59.4	37.0	76.7	68.4
PSSCL	80.7	82.3	77.6	46.1	71.0	71.5
DCD	87.6	78.7	78.5	49.6	75.8	74.0
MEEs	87.9	83.7	79.3	51.7	76.6	75.8
Ours	**88.8**	**85.1**	**80.3**	**62.3**	**77.9**	**78.9**

Table 1: Experimental results on CIFAR-10 dataset.

Boundary-restoration certificate. Let $m_j(\mathbf{x}) = \mathbf{x}^\top(\boldsymbol{\mu}_c - \mathbf{v}_j)$ and $\Gamma_c(\mathbf{x}) = \min_j m_j(\mathbf{x})$. If $\mathcal{L}_{\mathrm{HamBR}}(\mathbf{x}) \le \ell$, then

$$\Gamma_c(\mathbf{x}) \ge -\tau \log(e^\ell - 1). \tag{9}$$

For unit-normalized features, $\ell < \log 2$ places $\mathbf{x}$ strictly on the prototype side of every prototype–outlier boundary. This certificate concerns anchor-induced proxy boundaries; the derivation and margin-monotonicity proof are in Appendix F.

From proxy to classifier. Transfer to the classifier requires alignment between projected-feature updates and classifier geometry. Appendix F.3 defines the affine-head boundary distance and derives a sufficient local transfer bound under directional alignment and bounded head drift.

3.4 Joint Optimization Objective

HamBR is a modular geometric component of semi-supervised noisy label learning. The GMM-selected labeled set uses label co-correction and a cross-entropy loss $\mathcal{L}_x$; the unlabeled set uses sharpened dual-network pseudo-labels and a mean-squared consistency loss $\mathcal{L}_u$. A batch-prior entropy term $\mathcal{L}_{reg}$ discourages class collapse, and a self-supervised contrastive loss $\mathcal{L}_{con}$ aligns augmented views. The combined objective is

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_x + \lambda_u \mathcal{L}_u + \lambda_{reg}\mathcal{L}_{reg}}_{\mathcal{L}_{SSL}} + \lambda_c \mathcal{L}_{con} + \lambda_{hambr}\mathcal{L}_{HamBR}. \tag{10}$$

Each iteration updates the support bank before generating anchors and optimizing the joint loss. Appendix D gives the individual loss definitions.

4 Experiments

We evaluate classification, OOD robustness, and mechanism behavior on CIFAR-10/100 (Wang et al., 2022), Animal-10N, and Food-101. Baselines include DivideMix (Li et al., 2020), UNICON (Karim et al., 2022), LongReMix (Cordeiro et al., 2023), PSSCL (Zhang et al., 2025), ProMix (Xiao et al., 2022), DCD (Mu et al., 2025), and MEEs (Yuan et al., 2025).

Shared evaluation protocol. We target practical fine-tuning scenarios with limited training budgets and therefore adopt a 30-epoch training budget for all methods. Within each dataset, all methods use the same backbone and training duration to ensure a controlled comparison. Dataset descriptions, backbone configurations, and optimization parameters are given in Appendix E.

Dataset	CIFAR-100			
Noise Mode	Sym.		Asym.	
Method	20%	50%	40%	Avg.
DivideMix	61.6	47.9	53.1	54.2
RRL	61.0	58.6	53.5	57.7
UNICON	60.7	57.1	54.2	57.3
ProMix	45.2	42.8	40.6	42.9
LongReMix	61.3	56.9	47.8	55.3
L2B	42.9	46.1	39.8	42.9
PSSCL	56.5	45.6	46.0	49.4
DCD	63.3	54.9	53.3	57.2
MEEs	63.7	59.6	52.0	58.4
Ours	**64.2**	**63.5**	**56.1**	**61.3**

Table 2: Experimental results on CIFAR-100 dataset.

Method	Ref.	Accuracy
DivideMix	ICLR20	71.5
RRL	ICCV21	73.8
UNICON	CVPR22	71.1
ProMix	IJCAI23	76.2
LongReMix	PR23	71.6
L2B	CVPR24	71.4
PSSCL	PR25	72.7
DCD	ICCV25	77.3
MEEs	NeurIPS25	76.9
Ours	**-**	**79.7**

Table 3: Experimental results on Animal-10N dataset using VGG-19.

Method	Ref.	Accuracy
DivideMix	ICLR20	67.2
RRL	ICCV21	69.7
UNICON	CVPR22	70.4
ProMix	IJCAI23	70.1
LongReMix	PR23	68.7
L2B	CVPR24	68.9
PSSCL	PR25	70.7
DCD	ICCV25	71.2
MEEs	NeurIPS25	71.0
Ours	**-**	**72.5**

Table 4: Experimental results on Food-101 dataset using pre-trained ResNet-50.

4.1 Classification under Synthetic and Real-World Noise

We evaluate symmetric and asymmetric label noise at the dataset-specific rates listed in Tables 1 and 2.

CIFAR-10: HamBR achieves 78.9% average accuracy, compared with 74.0% for DCD and 75.8% for MEEs (Table 1). Under 90% symmetric noise, it retains 62.3%, versus 41.8% for UNICON and 37.3% for LongReMix.

CIFAR-100: HamBR reaches 61.3% average accuracy, compared with 57.2% for DCD and 58.4% for MEEs (Table 2), supporting its effectiveness on fine-grained classes under label noise.

Real-world noise: HamBR achieves 79.7% on Animal-10N, exceeding DCD and MEEs by 2.4 and 2.8 points, and 72.5% on Food-101, exceeding them by 1.3 and 1.5 points (Tables 3 and 4). These gains extend the comparison to naturally noisy labels.

4.2 Noise Robustness in OOD Detection

Following OpenOOD protocols, we evaluate CIFAR-10 models against SVHN, LSUN, Places365, Textures, and MNIST using AUROC and FPR95. Table 5 shows that HamBR achieves average AUROC/FPR95 of 95.8/18.3, 95.9/19.6, 91.0/26.3, and 93.7/22.3 under Sym0.2, Sym0.5, Sym0.8, and Asym0.4, respectively. The consistent separation from unseen data provides complementary evidence that the learned in-distribution manifold is compact. Dataset-level results are provided in the appendix.

Dataset	CIFAR-10				
Noise Mode	Sym0.2	Sym0.5	Sym0.8	Asym0.4	Avg.
Method	AUROC ↑/FPR95 ↓				
DivideMix	69.7/64.1	77.2/51.2	77.9/67.0	49.2/93.1	68.5/68.9
RRL	66.3/84.9	51.1/98.8	30.8/99.1	41.7/98.0	47.5/95.2
UNICON	84.7/38.4	92.9/19.9	83.9/35.3	82.5/33.9	86.0/31.9
ProMix	76.6/45.6	82.0/36.0	79.4/40.5	81.2/39.2	79.8/40.3
LongReMix	70.8/64.9	67.6/58.6	62.6/62.5	70.5/68.6	67.9/63.7
L2B	74.7/61.3	75.5/64.9	68.1/82.8	78.9/66.4	74.3/68.9
PSSCL	93.4/24.0	92.3/26.4	87.8/31.1	87.5/33.8	90.3/28.8
DCD	89.6/29.8	90.7/27.4	85.3/37.9	87.6/33.1	88.3/32.1
MEEs	86.1/42.7	87.4/38.6	81.8/51.2	84.9/43.8	85.1/44.1
Ours	**95.8/18.3**	**95.9/19.6**	**91.0/26.3**	**93.7/22.3**	**94.1/21.6**

Table 5: Comparison of OOD detection performance on CIFAR-10. Avg. is the arithmetic mean across the four noise settings for each metric.

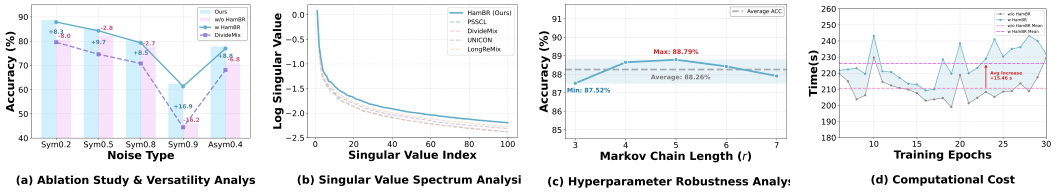

Figure 3: (a) HamBR ablation and integration into DivideMix. (b) Feature singular-value spectra. (c) Sensitivity to proposal-chain length r. (d) Training time per epoch.

4.3 Mechanism Analysis

Candidate-generation strategies. We compare Hamiltonian candidate generation with prototype midpoints, random spherical perturbations, spherical random walks, and gradient search on CIFAR-10 under the same training and evaluation settings. Figure 4 shows that HamBR achieves the highest accuracy at all three symmetric noise rates. Its gains over the strongest alternative, random spherical perturbations, are 1.70, 0.45, and 0.50 percentage points at 20%, 50%, and 80% noise, respectively. The consistent advantage supports Hamiltonian exploration as an effective candidate-generation choice among the evaluated strategies. Additional comparisons with Gaussian, Mixup, and feature-scale perturbation sampling are reported in Appendix Table 11.

Hard-clean samples and selection quality. Table 6 evaluates fixed hard-clean subsets: the lowest 20% of true-class probability margins among uncorrupted training examples, ranked by the 20%-noise baseline at epoch 4 (8,205 and 2,792 examples at 20% and 80% noise). Accuracy uses the sum of two classifiers' logits. AUC and recall use fresh single-pass GMM clean probabilities, averaged over branches; recall thresholds are determined post hoc from diagnostic labels. At the same 1% noise-misacceptance cap, HamBR retains more hard-clean examples in both settings, alongside higher hard-clean accuracy and separation AUC.

Module contribution and integration. Figure 3(a) shows that HamBR improves accuracy by 7.32 percentage points on average and by 16.24 points under 0.9 symmetric noise, from 46.10% to 62.34%. Consistent gains under both symmetric and asymmetric noise confirm the module's contribution to robustness.

We further integrate HamBR into DivideMix, obtaining an average gain of 10.46 percentage points across the evaluated noise settings. Accuracy increases from 44.43% to 61.36% under 0.9 symmetric noise and by 8.86 points under 0.4 asymmetric noise, validating HamBR as a modular enhancement component.

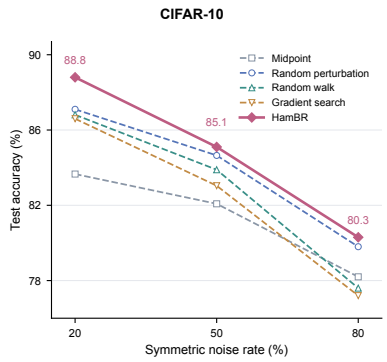

Figure 4: Candidate-generation comparison on CIFAR-10.

Noise	Method	Acc.	AUC	Recall
20%	Baseline	51.32	0.8928	35.83
	HamBR	**69.63**	**0.9459**	**64.11**
80%	Baseline	48.07	0.8904	27.97
	HamBR	**54.30**	**0.9251**	**36.46**

Table 6: **Hard-clean diagnostics.** Acc.: hard-clean accuracy (%); AUC: hard-clean/noisy separation; Recall: hard-clean recall (%) at noise misacceptance $\leq 1\%$.

Metric	w/o HamBR	w/ HamBR
Silhouette Score ↑	0.43	**0.75**
D_{pos} ↓	0.68	**0.37**
D_{neg} ↑	0.86	**1.29**

Table 7: Feature-space evidence of decision-boundary restoration.

Singular Value Spectrum Analysis. Figure 3(b) shows that HamBR has a slower singular-value decay than the baselines, particularly in the mid-tail region. The higher-amplitude spectrum indicates a larger effective feature rank and less noise-induced representation collapse.

Sensitivity and computational cost. Figure 3(c) shows that accuracy peaks at 88.79% for $r = 5$, compared with 87.52%, 88.43%, and 87.91% for $r = 3, 6, 7$, respectively. A moderate chain therefore provides the best exploration–efficiency balance; additional sensitivity results for K, L, and ϵ are reported in the appendix.

Figure 3(d) shows a stable average overhead of 15.46 seconds per epoch, mainly from virtual-sample synthesis and energy-based constraints. This moderate cost accompanies the improvements in representation structure and robustness.

Accuracy attainment during training. On CIFAR-10 with 20% symmetric noise, DivideMix obtains 79.55% at epoch 30, while HamBR first exceeds this target at epoch 18 with 80.32%. UNICON obtains 72.85% at epoch 30, while HamBR first exceeds it at epoch 7 with 73.46%. These comparisons show that HamBR reaches the corresponding accuracy targets earlier in training. In this setting, although HamBR incurs a higher per-epoch cost, it reaches the corresponding accuracy targets in fewer epochs; the observed benefit is earlier accuracy attainment.

Feature-space geometry. Table 7 shows a higher Silhouette Score, smaller intra-class distance, and larger inter-class distance with HamBR, indicating more compact and separated class representations. Supplementary ablations, parameter sensitivity, computational costs, and dataset-level OOD results are provided in the appendix.

5 Limitations

The absolute density of retained anchors depends on midpoint initialization, selected-feature reliability, and prototype quality. The geometric certificate applies to anchor-induced proxy boundaries; transfer to classifier margins requires the alignment and drift conditions in Appendix F.3. The coupled training process also calls for controlled interventions to distinguish individual causal effects.

6 CONCLUSION

HamBR reframes noisy-label selection as part of a dynamic learning loop: the representation determines reliability estimates, and selected supervision shapes the next representation. Active boundary anchors provide one intervention on this coupling. Finite Hamiltonian proposals play a supporting role as structured exploration; density retention and proxy-margin analysis have explicit, conditional scope. The candidate-generation comparison supports Hamiltonian exploration across the three evaluated CIFAR-10 noise rates.

AI USE STATEMENT

We used generative AI tools to assist with writing and language polishing, experimental design, and code implementation.

REFERENCES

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Filipe R Cordeiro, Ragav Sachdeva, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Longremix: Robust learning with high confidence samples in a noisy label environment. *Pattern recognition*, 133:109013, 2023.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9676–9686, 2022.

Hengzhuang Li and Teng Zhang. Outlier synthesis via hamiltonian monte carlo for out-of-distribution detection. In *International Conference on Learning Representations*, 2025.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 316–325, 2022.

Chenyu Mu, Yijun Qu, Jiexi Yan, Erkun Yang, and Cheng Deng. Meta-learning dynamic center distance: Hard sample mining for learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 415–425, October 2025.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pp. 5907–5915. PMLR, 2019.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. *ICLR*, 1(2):5, 2022.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in neural information processing systems*, 33:7597–7610, 2020.

Ruixuan Xiao, Yiwen Dong, Haobo Wang, Lei Feng, Runze Wu, Gang Chen, and Junbo Zhao. Promix: Combating label noise via maximizing clean sample utility. *arXiv preprint arXiv:2207.10276*, 2022.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pp. 7164–7173. PMLR, 2019.

Suqin Yuan, Lei Feng, Bo Han, and Tongliang Liu. Enhancing sample selection against label noise by cutting mislabeled easy examples. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Qian Zhang, Yi Zhu, Filipe R Cordeiro, and Qiu Chen. Psscl: A progressive sample selection framework with contrastive loss designed for noisy labels. *Pattern Recognition*, 161:111284, 2025.

Can We Break the Loop?

Active Boundary Restoration for Learning with Noisy Labels

Supplementary Material

CONTENTS

A Full Related Work 12
A.1 Sample Selection and Semi-Supervised Noisy Learning 12
A.2 Robust Representation Learning 12
A.3 Virtual Outlier Synthesis and Manifold Regularization 12

B Notation and Objective Recap 13

C Additional Empirical Analyses 13
C.1 Ablations and Diagnostics 13

D Optimization and Algorithm Details 15
D.1 Posterior and Partition Function 15
D.2 Detailed Joint Optimization Objective 15

E Implementation Details 17

F Theoretical Analysis and Proofs of HamBR 17
F.1 Boundary Locality of Accepted Proposals 18
F.2 Local Margin Effect of Virtual Outliers 19
F.3 Conditional Transfer to the Classifier Margin 20

G Noise Robustness in OOD Detection 20

A Full Related Work

A.1 Sample Selection and Semi-Supervised Noisy Learning

Sample selection is one of the most prominent paradigms for handling noisy labels. Its core assumption relies on the small-loss trick, which suggests that deep networks tend to prioritize the learning of simple and clean samples (Arpit et al. 2017). Early works, such as Co-teaching and its variants, use dual networks to cross-filter samples and mitigate accumulated errors (Yu et al. 2019). DivideMix (Li, Socher, and Hoi 2020) extends this idea to a Semi-Supervised Learning (SSL) framework by using a Gaussian Mixture Model (GMM) to dynamically divide data into a labeled set (clean) and an unlabeled set (noisy), and applying the MixMatch strategy (Berthelot et al. 2019) for joint training. Subsequent studies, such as LongReMix (Cordeiro et al. 2023) and PSSCL (Zhang et al. 2025), have introduced more refined selection strategies or two-stage training processes on this basis to address the challenges of high noise rates or long-tailed distributions. ProMix (Xiao et al. 2022) attempts to maximize the utility of clean samples by expanding the base clean set through matched high-confidence selection.

Although these methods have made significant progress, they rely primarily on prediction confidence for selection. When noisy samples are located near decision boundaries or have misleadingly high confidence (Northcutt, Jiang, and Chuang 2021), GMMs often struggle to distinguish them (Xia et al. 2020). This results in the erroneous categorization of noise into the clean set, thereby contaminating the training of the model. Unlike these methods that rely solely on statistical selection, our approach introduces an active geometric defense mechanism, which reinforces the ability to identify boundary noise through physical constraints (Bouveyron and Girard 2009).

A.2 Robust Representation Learning

In addition to label cleaning, another strategy focuses on learning robust feature representations that are insensitive to label noise (Zhong et al. 2019). RRL (Li et al. 2022) proposes the learning of robust representations by regularizing the geometric structure of the feature space without any label cleaning. UNICON (Karim et al. 2022) notes that traditional sample selection tends to select classes that are easy to learn, which leads to an unbalanced feature space. Therefore, it proposes the use of contrastive learning and Jensen-Shannon divergence to enhance the discriminability of features.

However, existing contrastive learning methods focus mainly on pairwise relationships between samples and lack explicit constraints on the boundaries of the feature manifold (Chen et al. 2020; He et al. 2020). Noisy samples often drift at the edges of class clusters or in inter-class voids; relying solely on the minimization of intra-class distances makes it difficult to isolate them completely. Our work establishes explicit energy barriers in the feature space by introducing virtual outliers as negative pivots, thereby geometrically compressing high-confidence clean features and indirectly reducing boundary interference from noisy samples.

A.3 Virtual Outlier Synthesis and Manifold Regularization

Virtual Outlier Synthesis (VOS) (Du et al. 2022) has been extensively investigated in OOD detection (Hendrycks and Gimpel 2016), with the aim of regularizing decision boundaries by generating virtual samples in low-density regions. However, the direct transfer of traditional VOS techniques, typically based on GANs or simple Gaussian sampling, to LNL faces a dual challenge: excessive computational overhead for generation or the inability to capture complex manifold topologies precisely (Goodfellow et al. 2014; Lee et al. 2018). To bridge this gap, HamBR studies boundary intervention in noisy-label learning, building on Hamiltonian outlier synthesis introduced by HamOS (Li & Zhang, 2025). Under noisy supervision, membership in the clean support set is latent and must be estimated as the representation evolves. HamBR couples this dynamically filtered, potentially contaminated support bank with anchor construction and subsequent representation updates; Hamiltonian exploration supplies one structured proposal mechanism within that coupling.

Method	Sym0.2	Sym0.5	Asym0.4	Avg.
		CIFAR-10		
DCD	87.6	78.7	75.8	80.7
MEEs	87.9	83.7	76.6	82.7
Ours	**88.8**	**85.1**	**77.9**	**83.9**
		CIFAR-100		
DCD	63.3	54.9	53.3	57.2
MEEs	63.7	59.6	52.0	58.4
Ours	**64.2**	**63.5**	**56.1**	**61.3**

Table 8: Comparison with recent baselines under the shared restricted training setting.

Variant	S0.2	S0.5	S0.8	A0.4	Avg.
Original	**88.8**	**85.1**	**80.3**	**77.9**	**83.0**
+ Other prototypes	89.0	84.6	75.1	72.6	80.3

Table 9: Ablation of adding other-class prototypes to Eq. (14).

B Notation and Objective Recap

For completeness, we restate the quantities needed by the proofs. For a normalized query $\mathbf{z}$ and support bank $\mathcal{B} = \{(\mathbf{k}_i, w_i)\}_{i=1}^{N}$, HamBR uses

$$U(\mathbf{z}) = -\log \sum_{(\mathbf{k}_j, w_j) \in \mathcal{N}_K(\mathbf{z})} (w_j + \varepsilon_w) \exp(\mathbf{z}^\top \mathbf{k}_j / \tau), \tag{11}$$

$$s(\mathbf{z}) = \log \sum_{i=1}^{N} \exp(\kappa \mathbf{z}^\top \mathbf{k}_i). \tag{12}$$

A finite spherical Hamiltonian proposal $\widetilde{\mathbf{z}}^{(r)}$ initialized at a prototype midpoint $\mathbf{z}^{(0)}$ is retained only when

$$s(\widetilde{\mathbf{z}}^{(r)}) \le s(\mathbf{z}^{(0)}) + \delta. \tag{13}$$

For a clean feature $\mathbf{x}$ of class c and accepted virtual outliers $\{\mathbf{v}_j\}_{j=1}^{M}$, the geometric regularizer is

$$\mathcal{L}_{\mathrm{HamBR}} = -\log \frac{\exp(\mathbf{x}^\top \boldsymbol{\mu}_c / \tau)}{\exp(\mathbf{x}^\top \boldsymbol{\mu}_c / \tau) + \sum_{j=1}^{M} \exp(\mathbf{x}^\top \mathbf{v}_j / \tau)}. \tag{14}$$

C Additional Empirical Analyses

C.1 Ablations and Diagnostics

C.1.1 Ablation Study of HamBR Module

As shown in Figure 3(a) of the main paper, the HamBR module consistently enhances model robustness across all noise configurations, yielding an average accuracy improvement of approximately 7.32 percentage points. This advantage is particularly pronounced under extreme noise conditions: at a 0.9 symmetric noise ratio, the module boosts accuracy from 46.10% to 62.34%, a margin of 16.24 percentage points. Furthermore, robust gains are consistently observed under asymmetric noise, supporting HamBR's role in mitigating boundary collapse under severe noise.

We also examine whether prototypes from other classes should be added as extra negative competitors in Eq. (14). As shown in Table 9, this variant slightly improves the easiest 20% symmetric setting, but it degrades sharply under high symmetric and asymmetric noise. The reason is prototype corruption: when class prototypes become unreliable under severe noise, pushing a sample away from all other prototypes can introduce conflicting gradients against the supervised objective. By contrast, HamBR uses virtual outliers as class-agnostic boundary anchors, which preserves compact manifolds without relying on potentially corrupted class prototypes.

Metric	w/o HamBR	w/ HamBR
Silhouette Score ↑	0.43	**0.75**
D_{pos} ↓	0.68	**0.37**
D_{neg} ↑	0.86	**1.29**

Table 10: Quantitative evidence of decision-boundary sharpening.

C.1.2 Integration Capability of HamBR as an Enhancement Module

To validate the modular integration capability of HamBR, we integrate it into DivideMix. As illustrated in Figure 3(a) of the main paper, incorporating HamBR yields consistent gains across all evaluated noise settings, improving performance by approximately 10.46 percentage points on average. Under 0.9 symmetric noise, accuracy increases from 44.43% to 61.36%, while under 0.4 asymmetric noise it improves by 8.86 percentage points. These results validate the effectiveness of HamBR as an enhancement module and demonstrate its potential for integration with existing noisy label learning frameworks.

C.1.3 Singular Value Spectrum Analysis

As evidenced by the singular value spectrum analysis in Figure 3(b) of the main paper, HamBR exhibits slower singular-value decay than the other methods, indicating that active feature-space reshaping mitigates feature collapse. The difference is most visible in the mid-tail region (indices 20–100); at index 40, HamBR maintains a log singular value of approximately -1.9, whereas competing methods drop below -2.0. This spectrum indicates a higher effective feature rank and richer discriminative structure under label noise.

To directly quantify boundary restoration, Table 10 reports feature-space geometry metrics. HamBR improves the Silhouette Score from 0.43 to 0.75, indicating stronger cluster separation. It also reduces the average intra-class distance D_{pos} from 0.68 to 0.37 while increasing the inter-class distance D_{neg} from 0.86 to 1.29. These dataset-level measurements agree with the boundary-restoration certificate in Proposition 3: HamBR simultaneously contracts clean features toward their prototypes and increases their separation from low-density boundary anchors, alongside the classification gains.

C.1.4 Ablation Study of Moderate Proposal-Chain Length

As illustrated in Figure 3(c) of the main paper, we analyze the impact of the proposal-chain length r on classification accuracy. Increasing r from 3 to 5 yields a steady improvement from 87.52% to 88.79%, suggesting that a moderate number of rounds provides more diverse finite trajectories within the permitted low-density band. When r is further increased to 6 and 7, accuracy drops to 88.43% and 87.91%, respectively. Excessive rounds can therefore add redundant proposals that are either rolled back by the retention rule or contribute little new boundary information. The best trade-off between exploration and computation is obtained at $r = 5$.

C.1.5 Per-Epoch Training Cost Analysis

As shown in Figure 3(d) of the main paper, the training time with HamBR is consistently higher than that without HamBR. This indicates that the additional overhead introduced by HamBR is systematic, arising mainly from virtual sample synthesis and the associated energy-based constraint computation, rather than from occasional fluctuations caused by specific parameter settings. Without HamBR, the training time is roughly between 199 s and 229 s, whereas the version with HamBR ranges from approximately 209 s to 243 s. Overall, although HamBR consistently increases the training time, the magnitude of this increase remains moderate and within a controllable range for most settings, reflecting a stable trade-off between the benefits of enhanced representation structure and computational efficiency.

C.1.6 Comparison of Different Sampling Strategies

Table 11 compares HamBR with Gaussian sampling based on low-likelihood GMM regions, Mixup interpolation between clean features, and perturbations scaled by feature standard deviation. HamBR achieves 79.1% average accuracy across the four symmetric-noise settings, compared with 65.0%, 63.5%, and 64.2%, respectively. At 20% noise, its advantage over Mixup is 10.2 percentage points. At 80% noise, Gaussian and Mixup obtain 65.8% and 57.0%; at 90% noise, all three alternatives obtain approximately 47%, compared with 62.3% for HamBR. These results complement the candidate-exploration comparison in Figure 4. A direct comparison with a conventional VOS-style boundary generation baseline further supports this observation: on CIFAR-10 Sym0.2, the VOS-style variant obtains 86.51%, whereas HamBR reaches 88.8%. This comparison supports the combined candidate-generation and retention design used by HamBR.

Method	S0.2	S0.5	S0.8	S0.9	Avg.
Gaussian	75.9	71.1	65.8	47.0	65.0
Mixup	78.6	71.2	57.0	47.1	63.5
Perturbation	76.4	72.0	61.4	46.9	64.2
Ours	**88.8**	**85.1**	**80.3**	**62.3**	**79.1**

Table 11: Comparison of different sampling strategies.

D Optimization and Algorithm Details

D.1 Posterior and Partition Function

After obtaining the virtual outliers $\mathcal{V}$, we aim to reshape the feature manifold using geometric constraints. Assuming a clean sample $\mathbf{x}$ and its corresponding class prototype $\boldsymbol{\mu}_c$, we seek to maximize the posterior probability $P(c|\mathbf{x})$ of belonging to the correct class c. Under the open-space assumption containing M virtual outliers (acting as negative samples/background class), this probability can be expressed in the form of a Boltzmann distribution:

$$P(c|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{x}, \boldsymbol{\mu}_c)/\tau)}{Z(\mathbf{x})}, \tag{15}$$

where the partition function $Z(\mathbf{x})$ must account for both the positive sample (class prototype) and all potential distractors (i.e., synthesized virtual outliers $\mathbf{v}_j \in \mathcal{V}$):

$$Z(\mathbf{x}) = \exp(\text{sim}(\mathbf{x}, \boldsymbol{\mu}_c)/\tau) + \sum_{j=1}^{M} \exp(\text{sim}(\mathbf{x}, \mathbf{v}_j)/\tau). \tag{16}$$

In this context, we ignore prototypes of other classes because virtual outliers are designed to populate the gaps between the current class and confusing classes, constituting the primary competitors for distinguishing the correct category.

Our optimization objective is to maximize the log-likelihood function $\mathcal{L} = \sum \log P(c|\mathbf{x})$. By taking the negative, we obtain the final **Geometric Perception Regularization Loss**, which aligns with the form of the InfoNCE loss:

The resulting objective is Eq. (8).

D.2 Detailed Joint Optimization Objective

To utilize useful information in noisy data fully and achieve robust learning, HamBR is designed as a modular geometric defense component that can be integrated into semi-supervised noisy label learning frameworks. Following the mainstream paradigm of sample selection, we use collaborative label correction and consistency regularization strategies for the current labeled set $\mathcal{X}_t$ and noisy

unlabeled set $\mathcal{U}$. For samples in $\mathcal{X}_{support}$, we combine the clean probability predicted by the GMM with model predictions for label co-correction and apply data augmentation strategies to compute the cross-entropy loss:

$$\mathcal{L}_x = -\frac{1}{|\mathcal{X}'|} \sum_{x_i \in \mathcal{X}'} \sum_{c=1}^{C} y_i^c \log(p_i^c(\theta)), \tag{17}$$

where $\mathcal{X}'$ is the augmented labeled dataset, y_i is the corrected pseudo-label, and $p_i(\theta)$ is the model prediction. For samples in $\mathcal{U}$, we use feature-distribution information through consistency training. We generate pseudo-labels by sharpening the predictions of the dual-network and compute the Mean Squared Error (MSE) as a consistency loss to enhance the robustness of the model against perturbations:

$$\mathcal{L}_u = \frac{1}{|\mathcal{U}'|} \sum_{u_i \in \mathcal{U}'} \|\hat{q}_i - p_i(\theta)\|_2^2, \tag{18}$$

where $\hat{q}_i$ is the sharpened guess label. Furthermore, to prevent the collapse of the model into a single class, we introduce a regularization term $\mathcal{L}_{reg}$. Thus, the basic semi-supervised noisy learning loss is defined as:

$$\mathcal{L}_{SSL} = \mathcal{L}_x + \lambda_u \mathcal{L}_u + \lambda_{reg} \mathcal{L}_{reg}. \tag{19}$$

Following DivideMix, the regularization term discourages degenerate class collapse by encouraging the average prediction distribution in a mini-batch to remain high-entropy:

$$\mathcal{L}_{reg} = \sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c, \quad \bar{p}_c = \frac{1}{B} \sum_{i=1}^{B} p(y = c \mid x_i). \tag{20}$$

Because this term is minimized together with the other losses, it maximizes the entropy of the batch-level prediction prior and prevents all samples from being assigned to a single dominant class. In each training iteration, Algorithm 1 is invoked after the feature bank is updated by the GMM-based filtering branch; the generated virtual outliers are then used in Eq. (8) together with the SSL losses above.

To further enhance the discriminability of features for unlabeled data, we introduce a self-supervised contrastive loss $\mathcal{L}_{con}$. By pulling together different augmented views of the same image and pushing away views of different images, this loss mitigates the over-reliance of the model on potentially incorrect pseudo-labels:

$$\mathcal{L}_{con} = -\sum_{i \in \mathcal{U}_{batch}} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i')/\tau_{con})}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau_{con})}, \tag{21}$$

where $\mathbf{z}_i, \mathbf{z}_i'$ are features of two augmented views of sample i, and $\text{sim}(\cdot)$ denotes cosine similarity. Finally, we integrate the aforementioned perceptual losses with the proposed active geometric defense loss $\mathcal{L}_{HamBR}$:

$$\mathcal{L}_{total} = \mathcal{L}_{SSL} + \lambda_c \mathcal{L}_{con} + \lambda_{hambr} \mathcal{L}_{HamBR}, \tag{22}$$

where λ_c and λ_{hambr} are balancing coefficients. By minimizing $\mathcal{L}_{total}$, this joint optimization ensures that while the model uses label information, it is constrained by a strong geometric inductive bias. HamBR directly constrains high-confidence clean samples to move toward their class prototypes and away from virtual outliers, thereby restoring decision boundaries and indirectly reducing interference from noisy samples.

Algorithm 1 Boundary-Constrained Spherical Hamiltonian Proposals

Require: Bank $\mathcal{B}$, prototypes $\{\boldsymbol{\mu}_c\}$, proposal count M, step size ϵ, rounds R, steps L, margin δ.
Ensure: Virtual outlier set $\mathcal{V}$.

$\mathcal{V} \leftarrow \varnothing$.
for $m \leftarrow 1$ to M **do**
  Sample $a \neq b$ and set $\mathbf{z}^{(0)} \leftarrow \mathrm{Normalize}((\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)/2 + 0.01\boldsymbol{\xi}),\ \boldsymbol{\xi} \sim \mathcal{N}(0, I)$.
  $b_m \leftarrow s(\mathbf{z}^{(0)}) + \delta$; $\mathbf{z} \leftarrow \mathbf{z}^{(0)}$.
  for $r \leftarrow 1$ to R **do**
    $\mathbf{z}_{\mathrm{prev}} \leftarrow \mathbf{z}$; sample $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and project it onto $\mathcal{T}_{\mathbf{z}}\mathbb{S}^{d-1}$.
    for $\ell \leftarrow 1$ to L **do**
      $\mathbf{v} \leftarrow \mathbf{v} - \frac{\epsilon}{2}\nabla U(\mathbf{z})$.
      $\omega \leftarrow \|\mathbf{v}\|_2$; $\mathbf{z}^{+} \leftarrow \mathbf{z}\cos(\omega\epsilon) + \frac{\mathbf{v}}{\omega}\sin(\omega\epsilon)$.
      $\mathbf{v} \leftarrow -\mathbf{z}\omega\sin(\omega\epsilon) + \mathbf{v}\cos(\omega\epsilon)$.
      $\mathbf{z} \leftarrow \mathrm{Normalize}(\mathbf{z}^{+})$; $\mathbf{v} \leftarrow \mathbf{v} - \frac{\epsilon}{2}\nabla U(\mathbf{z})$.
    end for
    if $s(\mathbf{z}) > b_m$ **then**
      $\mathbf{z} \leftarrow \mathbf{z}_{\mathrm{prev}}$.
    end if
  end for
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{\mathbf{z}\}$.
end for
return $\mathcal{V}$.

E Implementation Details

Datasets. CIFAR-10/100: Each benchmark contains 50,000 training and 10,000 test images. We evaluate symmetric label flips and standard asymmetric corruption: semantically similar-class mappings for CIFAR-10 and circular within-superclass flips for CIFAR-100.

Animal-10N (Song et al., 2019): This is a benchmark dataset that contains real-world label noise, primarily composed of five pairs of animal species with highly confusing visual features, with an estimated intrinsic noise rate of approximately 8%.

Food-101 (Bossard et al., 2014): This dataset consists of real-world food images crawled from the web, which contains naturally occurring label noise.

Shared training protocol. All methods are evaluated with the same dataset-specific backbone and a 30-epoch training budget. We use a pre-trained ResNet-18 for CIFAR-10/100, VGG-19 for Animal-10N, and an ImageNet-pre-trained ResNet-50 for Food-101. The comparison therefore measures performance under this common training horizon; equal epoch counts do not imply equal wall-clock cost.

HamBR configuration. HamBR is optimized using stochastic gradient descent (SGD) with momentum 0.9 and weight decay 1×10^{-3}. Its warm-up lasts 5 epochs. The candidate generator uses $M = 64$ parallel proposals, $R = 5$ rounds, and $L = 3$ leapfrog-inspired steps per round. Table 12 lists the optimization and geometric regularization parameters, which are fixed across the main experiments unless otherwise specified.

Table 13 further analyzes the stability of Hamiltonian proposals. The best result is obtained with $K = 20$, $L = 3$, and $\epsilon = 0.1$, which balances local energy smoothing with sufficient inter-class exploration. Too few neighbors make the energy surface sensitive to isolated noisy samples, whereas too many neighbors over-smooth class geometry. Similarly, overly long or aggressive trajectories may produce redundant candidates, while short traces provide limited exploration.

F Theoretical Analysis and Proofs of HamBR

In this appendix, we establish three claims used by HamBR: accepted proposals remain in a controlled low-support-density band around their prototype-midpoint initialization; the HamBR objec-

Hyperparameter	Value
Training budget (epochs)	30
Optimizer	SGD
Momentum	0.9
Weight decay	1×10^{-3}
Warm-up epochs	5
Temperature τ	0.1
Parallel proposals M	64
Leapfrog steps L	3
Step size ϵ	0.1
Sampling rounds R	5
KNN neighbors K	20
Density margin δ	0.1
λ_{hambr}	0.1
λ_c	0.025

Table 12: Hyperparameter settings shared by HamBR across the main experiments.

	KNN neighbors K			
K	10	15	20	25
Accuracy	87.8	88.3	**88.8**	86.6
	Leapfrog steps and step size ϵ			
(L, ϵ)	(2,0.1)	(3,0.05)	(3,0.1)	(4,0.1)
Accuracy	86.5	85.6	**88.8**	83.7

Table 13: Sensitivity of Hamiltonian proposal settings on CIFAR-10 Sym0.2.

tive is strictly decreasing in every prototype-to-outlier similarity margin; and a small HamBR loss certifies a positive minimum distance from the resulting boundary surfaces. Together, these claims connect the finite-step proposal mechanism, the optimization objective, and anchor-induced geometric separation.

F.1 Boundary Locality of Accepted Proposals

Under Eq. (11), class interiors with large support density have low energy, whereas inter-class low-density regions have high energy. HamBR uses finite spherical Hamiltonian trajectories to explore candidates and controls their support through prototype-midpoint initialization together with Eq. (13).

Define the positive all-bank support density and its associated energy as

$$\rho_{\mathrm{all}}(\mathbf{z}) = \sum_{i=1}^{N} \exp\left(\kappa \mathbf{z}^\top \mathbf{k}_i\right), \qquad U_{\mathrm{all}}(\mathbf{z}) = -\log \rho_{\mathrm{all}}(\mathbf{z}). \tag{23}$$

Equation (12) gives $s(\mathbf{z}) = \log \rho_{\mathrm{all}}(\mathbf{z}) = -U_{\mathrm{all}}(\mathbf{z})$.

Proposition 1 (density-band guarantee). Let $\mathbf{z}^{(0)}$ be a prototype-midpoint initialization and let $\widetilde{\mathbf{z}}$ be any proposal retained by Eq. (13). Then

$$\frac{\rho_{\mathrm{all}}(\widetilde{\mathbf{z}})}{\rho_{\mathrm{all}}(\mathbf{z}^{(0)})} \le e^{\delta}, \qquad U_{\mathrm{all}}(\widetilde{\mathbf{z}}) \ge U_{\mathrm{all}}(\mathbf{z}^{(0)}) - \delta. \tag{24}$$

Proof. The acceptance condition implies $\log \rho_{\mathrm{all}}(\widetilde{\mathbf{z}}) \le \log \rho_{\mathrm{all}}(\mathbf{z}^{(0)}) + \delta$. Exponentiating both sides yields the density-ratio bound. Multiplying the log inequality by -1 yields the energy bound. □

If the selected prototype midpoint satisfies $\rho_{\mathrm{all}}(\mathbf{z}^{(0)}) \le \rho_b$ for a boundary-density level ρ_b, Proposition 1 immediately gives $\rho_{\mathrm{all}}(\widetilde{\mathbf{z}}) \le e^{\delta}\rho_b$. Thus, the finite Hamiltonian trajectory may explore different directions, but every retained state has support at most e^{δ} times the initial support. The absolute support level is therefore determined jointly by initialization and the retention margin.

F.2 Local Margin Effect of Virtual Outliers

For a clean feature $\mathbf{x}$ with prototype $\boldsymbol{\mu}_c$ and accepted virtual outliers $\{\mathbf{v}_j\}_{j=1}^{M}$, define the prototype-to-outlier similarity margin

$$m_j(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\mu}_c - \mathbf{x}^\top \mathbf{v}_j. \tag{25}$$

The loss in Eq. (14) can be rewritten as

$$\mathcal{L}_{\mathrm{HamBR}}(\mathbf{x}) = \log\left(1 + \sum_{j=1}^{M} \exp\left[-\frac{m_j(\mathbf{x})}{\tau}\right]\right). \tag{26}$$

Proposition 2 (margin monotonicity). For every virtual outlier j,

$$\frac{\partial \mathcal{L}_{\mathrm{HamBR}}}{\partial m_j} = -\frac{1}{\tau}\frac{\exp(-m_j/\tau)}{1 + \sum_{k=1}^{M}\exp(-m_k/\tau)} < 0. \tag{27}$$

Proof. The result follows by differentiating Eq. (26); positivity of the exponential and $\tau > 0$ gives the strict sign. □

Therefore, minimizing $\mathcal{L}_{\mathrm{HamBR}}$ locally favors larger similarity margins between the correct prototype and every active virtual outlier. Equivalently, its feature gradient is

$$\nabla_{\mathbf{x}} \mathcal{L}_{\mathrm{HamBR}} = \frac{1}{\tau}\left[-(1 - p_c)\boldsymbol{\mu}_c + \sum_{j=1}^{M} p_j \mathbf{v}_j\right], \tag{28}$$

where p_c and p_j are the normalized softmax probabilities in Eq. (14). The first term attracts a clean feature toward its prototype, while the second repels it from low-density accepted outliers. Proposition 2 establishes the local optimization direction. We next turn this monotonicity into an explicit boundary-separation certificate.

For every accepted outlier $\mathbf{v}_j$, define the prototype–outlier boundary surface and the minimum signed score margin as

$$\mathcal{H}_{cj} = \{\mathbf{z} : \mathbf{z}^\top \boldsymbol{\mu}_c = \mathbf{z}^\top \mathbf{v}_j\}, \qquad \Gamma_c(\mathbf{x}) = \min_{1 \le j \le M} m_j(\mathbf{x}). \tag{29}$$

The surface $\mathcal{H}_{cj}$ is the decision surface induced by treating $\boldsymbol{\mu}_c$ as the positive anchor and $\mathbf{v}_j$ as a competing low-density anchor, exactly as in Eq. (14).

Proposition 3 (boundary-restoration certificate). Assume $\|\mathbf{x}\|_2 = \|\boldsymbol{\mu}_c\|_2 = \|\mathbf{v}_j\|_2 = 1$, $\mathbf{v}_j \neq \boldsymbol{\mu}_c$, and $\ell > 0$. If $\mathcal{L}_{\mathrm{HamBR}}(\mathbf{x}) \le \ell$, then

$$\Gamma_c(\mathbf{x}) \ge -\tau \log\left(e^{\ell} - 1\right) \equiv g(\ell). \tag{30}$$

Moreover, for every j,

$$\|\mathbf{x} - \mathbf{v}_j\|_2^2 - \|\mathbf{x} - \boldsymbol{\mu}_c\|_2^2 \ge 2g(\ell). \tag{31}$$

When $\ell < \log 2$, $g(\ell) > 0$: the feature lies strictly on the prototype side of every $\mathcal{H}_{cj}$, and its minimum Euclidean signed distance to these hyperplanes is at least $g(\ell)/2$.

Proof. Let $j^* = \arg\min_j m_j(\mathbf{x})$. From Eq. (26),

$$\mathcal{L}_{\mathrm{HamBR}}(\mathbf{x}) \ge \log(1 + \exp[-\Gamma_c(\mathbf{x})/\tau]). \tag{32}$$

Combining this inequality with $\mathcal{L}_{\mathrm{HamBR}}(\mathbf{x}) \le \ell$ and rearranging gives Eq. (30). Unit normalization yields $2m_j(\mathbf{x}) = \|\mathbf{x} - \mathbf{v}_j\|_2^2 - \|\mathbf{x} - \boldsymbol{\mu}_c\|_2^2$, proving Eq. (31). Finally, $\|\boldsymbol{\mu}_c - \mathbf{v}_j\|_2 \le 2$, so a positive score margin m_j gives signed distance $m_j/\|\boldsymbol{\mu}_c - \mathbf{v}_j\|_2 \ge g(\ell)/2$. □

The certificate $g(\ell)$ increases strictly as the HamBR loss decreases. Proposition 1 supplies a relative support upper bound for the competing anchors, Proposition 2 explains the optimization direction, and Proposition 3 proves the resulting geometric separation. This establishes a geometric certificate for the anchor-induced proxy boundaries optimized by HamBR.

F.3 Conditional Transfer to the Classifier Margin

The classifier acts on $h \in \mathbb{R}^d$, whereas HamBR acts on a projected feature. Consider a fixed class c and normalized affine constraints $d_{cj,t}(h) = a_{cj,t}^\top h + \beta_{cj,t}$, where $a_{cj,t} = (w_{c,t} - w_{j,t})/\|w_{c,t} - w_{j,t}\|$. Let $D_t(h) = \min_{j \neq c} d_{cj,t}(h)$. Assume nonzero pairwise weight differences. When h is correctly classified, $D_t(h)$ is the exact Euclidean distance to the boundary of its class region; otherwise it is a signed constraint margin, not a nearest-boundary distance.

Proposition 4 (sufficient local transfer). Fix the projection and anchors over a feature displacement $h \mapsto h + u$. Let $p(h)$ be the minimum proxy anchor margin. Suppose its active anchor stays fixed along this segment and ∇p is L-Lipschitz there. If, for $\alpha, \varepsilon, \eta \geq 0$ and every $j \neq c$,

$$a_{cj,t}^\top u \geq \alpha \nabla p(h)^\top u - \varepsilon \|u\|, \tag{33}$$

$$d_{cj,t+1}(h + u) \geq d_{cj,t}(h + u) - \eta, \tag{34}$$

then

$$D_{t+1}(h + u) - D_t(h) \geq \alpha[p(h + u) - p(h)] - \varepsilon \|u\| - \frac{\alpha L}{2}\|u\|^2 - \eta. \tag{35}$$

Proof. The inequality $\min_j(A_j + B_j) \geq \min_j A_j + \min_j B_j$ yields $D_t(h + u) - D_t(h) \geq \min_{j \neq c} a_{cj,t}^\top u$. Apply directional alignment and $p(h + u) - p(h) \leq \nabla p(h)^\top u + L\|u\|^2/2$. The head-drift condition decreases the bound by at most η. □

The transfer bound applies over segments with fixed projection and active anchors, directional alignment, and bounded head drift. Its right-hand side quantifies the proxy gain required for a positive classifier-margin change.

G Noise Robustness in OOD Detection

For OOD evaluation, we sum the logits from the two networks and use the temperature-scaled energy score $S(\mathbf{x}) = T \log \sum_c \exp((f_{1,c}(\mathbf{x}) + f_{2,c}(\mathbf{x}))/T)$, with $T = 1$. A higher score indicates that a sample is more likely to belong to the in-distribution data.

As shown in Tables 14, 15, and 16, the experimental results demonstrate that the proposed method consistently achieves the strongest OOD detection performance among the evaluated methods on the CIFAR-10 dataset across various symmetric label noise levels.

At 20% symmetric noise, HamBR achieves average AUROC/FPR95 of 95.76%/18.30%, compared with 93.40%/24.04% for PSSCL, the strongest baseline on these average metrics. At 50% noise, HamBR obtains 95.93%/19.59%, compared with 92.88%/19.86% for UNICON, corresponding to a 3.05-percentage-point AUROC gain and a 0.27-point FPR95 reduction. At 80% noise, HamBR obtains 91.04%/26.25%, compared with 87.82%/31.12% for PSSCL.

Performance varies across OOD datasets. At 50% symmetric noise, ProMix achieves better LSUN AUROC/FPR95 (98.31%/5.61%) than HamBR (97.88%/9.11%), and UNICON obtains lower FPR95 on Textures and MNIST. HamBR achieves the best average AUROC and FPR95 across the evaluated OOD datasets in each reported noise setting.

Furthermore, as evidenced in Table 17, the proposed method excels under asymmetric noise conditions. At a 0.4 asymmetric noise ratio, the proposed method achieves an average AUROC of 93.71% and an FPR95 of 22.27%. This represents an improvement over PSSCL of 6.23 percentage points in AUROC and 11.49 percentage points in FPR95. On SVHN under asymmetric noise, the proposed method achieves an FPR95 of 3.80%. The average results support improved OOD detection across the evaluated noise settings.

Dataset	SVHN	LSUN	Places365	Textures	MNIST	Average
			AUROC ↑/FPR95 ↓			
DivideMix	81.09/71.95	79.66/48.43	81.40/56.12	87.19/47.15	19.16/96.68	69.70/64.07
RRL	82.82/60.01	55.69/96.65	53.62/92.36	56.08/94.26	83.26/81.42	66.29/84.94
UNICON	94.66/20.67	95.59/16.22	86.03/42.04	89.39/31.84	57.68/80.99	84.67/38.35
ProMix	90.83/32.92	97.22/10.05	80.81/45.15	79.67/47.41	34.30/92.47	76.57/45.60
LongReMix	69.35/79.11	88.69/41.12	79.56/59.24	82.63/54.70	33.98/90.10	70.84/64.85
L2B	59.06/79.77	75.18/63.33	81.65/55.70	71.30/68.09	86.30/39.40	74.70/61.26
PSSCL	97.62/11.55	97.33/10.77	89.76/35.13	92.74/25.98	89.55/36.75	93.40/24.04
Ours	**97.71/5.82**	**98.58/6.42**	**94.33/26.14**	**96.15/17.70**	**92.04/35.42**	**95.76/18.30**

Table 14: OOD Detection Performance on CIFAR-10 under Symmetric Noise (0.2). Bold denotes the best value for each metric.

Method	SVHN	LSUN	Places365	Textures	MNIST	Average
			AUROC ↑/FPR95 ↓			
DivideMix	72.39/72.93	94.30/21.28	84.30/42.76	85.38/40.55	49.38/78.48	77.15/51.20
RRL	51.43/99.83	53.12/98.60	55.38/97.51	60.04/98.21	35.58/99.97	51.11/98.82
UNICON	94.31/18.72	96.45/10.72	87.85/33.49	90.31/**24.38**	95.48/**12.01**	92.88/19.86
ProMix	89.65/29.07	**98.31/5.61**	84.73/35.85	82.96/39.57	54.42/69.98	82.01/36.02
LongReMix	79.12/53.88	89.95/25.73	74.20/67.87	79.72/48.88	15.16/96.78	67.63/58.63
L2B	61.28/90.37	85.86/50.76	74.20/67.87	67.57/76.15	88.34/39.21	75.45/64.87
PSSCL	96.88/14.13	95.86/17.27	87.30/41.08	91.98/28.99	89.43/30.43	92.29/26.38
Ours	**97.77/12.88**	97.88/9.11	**93.80/27.29**	**94.04**/25.30	**96.17**/23.37	**95.93/19.59**

Table 15: OOD Detection Performance on CIFAR-10 under Symmetric Noise (0.5). Bold denotes the best value for each metric.

Method	SVHN	LSUN	Places365	Textures	MNIST	Average
			AUROC ↑/FPR95 ↓			
DivideMix	63.69/96.17	95.55/24.67	78.78/63.45	74.76/71.97	76.52/78.71	77.86/66.99
RRL	27.85/98.96	29.43/99.69	41.93/99.13	30.94/98.30	23.95/99.56	30.82/99.13
UNICON	94.32/22.10	**96.84**/11.89	86.58/40.90	90.87/27.48	50.70/73.90	83.86/35.25
ProMix	87.77/39.40	96.34/14.19	**87.94**/31.70	79.86/42.64	45.10/74.66	79.40/40.52
LongReMix	58.95/76.57	88.79/28.18	76.74/53.34	74.30/56.95	14.36/97.25	62.63/62.46
L2B	78.12/88.91	43.26/96.51	66.70/86.52	58.70/91.42	**93.76**/50.80	68.11/82.83
PSSCL	93.45/20.83	95.98/12.44	85.91/44.53	91.39/29.88	72.37/47.90	87.82/31.12
Ours	**95.44/17.70**	96.66/**10.83**	86.68/**31.43**	**92.12/25.57**	84.32/**45.71**	**91.04/26.25**

Table 16: OOD Detection Performance on CIFAR-10 under Symmetric Noise (0.8). Bold denotes the best value for each metric.

Method	SVHN	LSUN	Places365	Textures	MNIST	Average
			AUROC ↑/FPR95 ↓			
DivideMix	29.80/99.54	54.91/91.82	53.63/95.12	37.64/96.17	70.23/82.73	49.24/93.08
RRL	49.58/98.04	49.49/95.84	44.73/98.17	38.16/97.93	26.47/99.92	41.69/97.98
UNICON	95.95/14.69	96.83/**9.50**	87.09/**34.55**	92.02/23.21	40.43/87.51	82.46/33.89
ProMix	95.80/18.28	95.64/19.25	84.85/38.59	83.92/40.50	45.70/79.53	81.18/39.23
LongReMix	71.62/69.63	89.12/42.90	74.53/72.51	81.46/61.51	35.51/96.29	70.45/68.57
L2B	79.90/74.21	81.80/60.42	75.45/69.56	71.47/72.55	**85.63**/55.20	78.85/66.39
PSSCL	96.58/14.82	**97.47**/11.02	86.66/48.38	93.25/28.60	63.46/65.98	87.48/33.76
Ours	**98.99/3.80**	97.17/10.01	**91.65**/34.89	**95.70/21.24**	85.02/**41.41**	**93.71/22.27**

Table 17: OOD Detection Performance on CIFAR-10 under Asymmetric Noise (0.4). Bold denotes the best value for each metric.