Radial-Angular Geometry for Reliable Update Diagnosis in Noisy-Label Learning

Ningkang Peng[*]
Nanjing Normal University

Jingyang Mao[*]
Nanjing Normal University

Xiaoqian Peng
Nanjing University of Chinese Medicine

Weiguang Qu
Nanjing Normal University

Yanhui Gu[†]
Nanjing Normal University

Abstract

Noisy-label methods often estimate sample reliability from forward-space signals such as loss, confidence, or entropy. These signals indicate whether a sample is difficult to predict, but they do not directly test whether its observed label induces a reliable parameter update. This gap matters because hard clean samples and mislabeled samples can have similar loss while inducing different updates. We recast reliability estimation as diagnosis of the observed-label update. The sample-wise empirical Fisher trace gives a backward-space measure of update energy: for the classifier layer, it factorizes into a prediction-residual term and a feature-sensitivity term, so it captures information beyond scalar loss. Trace, however, is still a radial magnitude signal and cannot decide whether a large update is useful or harmful. We therefore propose Relative Geometric Conflict (RGC), which compares the observed-label gradient with a reference gradient induced by an EMA teacher. The conflict term helps distinguish large but aligned hard-clean updates from large conflicting updates caused by corrupted labels. Across synthetic and real-world noisy-label benchmarks, RGC improves hard-clean preservation and accuracy under our evaluation protocol.

1 Introduction

Learning with noisy labels (LNL) can be viewed as a problem of unreliable supervision [27, 11, 22]. Most existing methods estimate sample reliability from forward-space quantities such as loss, confidence, or entropy [25, 22]. This design also appears in more recent sample-selection and label-refinement frameworks such as ProMix and PSSCL [36, 40]. These signals are convenient because they are available during training, but they primarily answer how difficult the current prediction is under the observed label. They leave open the optimization question of whether the update induced by that label is trustworthy. Hard clean samples and mislabeled samples can therefore overlap in forward space, since both may exhibit high loss or low confidence, especially under instance-dependent or feature-dependent noise [35, 41].

This observation motivates us to shift the reliability criterion from forward difficulty to parameter-space update behavior. For a sample with observed label $\tilde{y}_i$, the actual training effect is determined by the gradient $\mathbf{g}_i^{\mathrm{obs}} = \nabla_\theta \ell(f_\theta(x_i), \tilde{y}_i)$. The sample-wise empirical Fisher trace, which reduces to the squared gradient norm for $\mathbf{g}_i \mathbf{g}_i^\top$, gives a backward-space diagnostic of this effect [1, 19]. For

[*]These authors contributed equally to this work.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the classifier layer, the trace factorizes as $\tau_i^{\text{last}} = \|p_i - e_{\tilde{y}_i}\|_2^2 \|h_i\|_2^2$, where $e_{\tilde{y}_i}$ is the one-hot vector corresponding to the observed label $\tilde{y}_i$. This factorization shows that the trace captures not only the prediction residual but also how strongly this residual is transmitted into parameter updates through the feature representation. Thus, trace provides information beyond the scalar forward loss and can distinguish samples that have similar loss but different update energy.

However, update energy alone is still not reliability. A hard clean sample may induce a large but useful update, while a mislabeled sample may induce a large and misleading update. Therefore, Fisher trace reveals how strongly a sample acts on the model, but not whether the induced update points in a trustworthy direction. This radial ambiguity is one reason magnitude-based criteria can still suppress informative hard clean samples.

To address this ambiguity, we introduce relative geometric conflict (RGC), a direction-aware reliability principle. Instead of judging the observed-label update in isolation, RGC compares it with a reference update induced by an EMA teacher [34, 20]. If the observed label is correct but the sample is hard, its update can be large while remaining directionally aligned with the reference. If the observed label is corrupted, the update is more likely to conflict with the reference direction. We quantify this inconsistency by $C_i = 1 - \cos(\mathbf{g}_i^{\text{obs}}, \mathbf{g}_i^{\text{ref}})$, which turns sample reliability estimation into a relative test of update consistency rather than an absolute test of difficulty or magnitude. Our contributions are:

- We reformulate noisy-label reliability estimation as the problem of assessing the reliability of observed-label-induced parameter updates, rather than measuring forward prediction difficulty.
- We use the sample-wise empirical Fisher trace as a backward-space diagnostic of update energy and make explicit why magnitude-only reliability criteria remain ambiguous.
- We propose relative geometric conflict, which compares observed-label and reference update directions to distinguish conflicting noisy updates from large but aligned hard-clean updates.

2 Related Work

Learning with noisy labels. Learning with noisy labels has been addressed by loss correction and robust objectives [27, 10, 2, 42], memorization-based sample selection or exchange [15, 11, 37], and semi-supervised or mixture-based label refinement [22, 7, 36, 40]. Despite their differences, many methods estimate reliability from forward-space cues such as loss, confidence, entropy, prediction history, or small-loss selection. These cues are useful but primarily measure learning difficulty rather than label reliability, and may therefore confuse hard clean samples with corrupted samples, especially under instance-dependent or feature-dependent noise [35, 41].

Representation and gradient geometry. Several methods improve noisy-label robustness by shaping representations rather than only correcting labels. Contrastive learning provides a general tool for representation separation [6, 12, 17], and noisy-label variants use selective contrastive objectives, uniform selection, or progressive selection to improve feature discriminability [16, 40]. These methods use representation geometry through feature similarity, class compactness, or selection confidence, whereas RGC tests whether the observed-label update is directionally consistent with a reference update in parameter space. Recent approaches also use active decision-boundary restoration and geometry-aware manifold regularization to improve learning with noisy labels [29, 28]. Information geometry views the Fisher matrix as a local metric, and Fisher/Hessian spectra are widely used to study curvature, sharpness, optimization, and generalization [1, 9, 14], though empirical Fisher approximations require care [19]. Our sample-wise empirical trace is a local rank-one convention only for the individual matrix $\mathbf{g}_i\mathbf{g}_i^\top$, whose trace equals $\|\mathbf{g}_i\|_2^2$; it is used as an update-energy diagnostic rather than as a true or population Fisher estimate [13, 31].

Teacher and reference targets. Teacher-based learning and self-ensembling average predictions or weights to provide stable targets [20, 34], and are widely used in noisy-label learning for pseudo-labeling, label refinement, and consistency regularization [30, 5, 22]. Here, the EMA teacher is not merely a replacement label source; it defines a reference gradient for testing the geometric reliability of the observed-label update, and we explicitly treat teacher drift as a boundary condition.

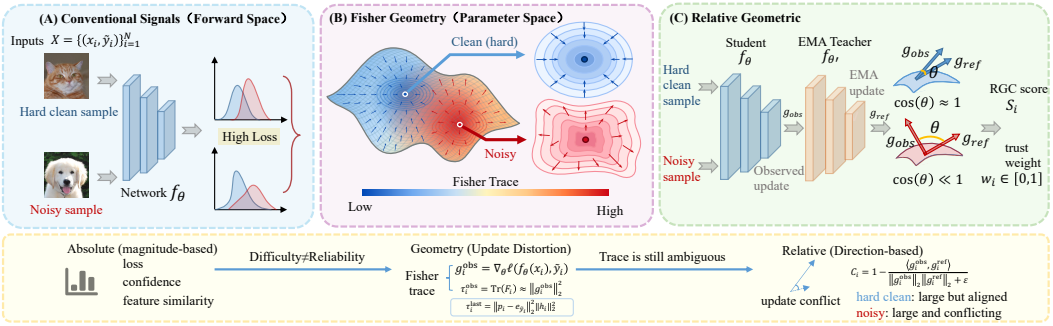


Figure 1: Overall motivation of RGC. Conventional forward-space signals mainly measure sample difficulty and confuse hard clean samples with noisy samples. Fisher trace moves the signal into parameter space but remains magnitude-based. RGC further compares update directions and helps identify unreliable samples by their directional conflict with a reference update.

3 Method

We propose a geometry-aware noisy-label learning framework that estimates whether the update induced by an observed label is reliable. The method has three ingredients: a temporally smoothed EMA reference target, a Fisher-trace proxy for update magnitude, and relative geometric conflict (RGC) for update direction. Figure 1 outlines the framework. Magnitude measures how strongly a sample updates the model, while direction tests whether this update agrees with the reference.

3.1 Problem Setup and Model Definition

Given a noisy training set $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^N$, the observed label $\tilde{y}_i \in \{1, \ldots, K\}$ may differ from the latent clean label y_i. We use a student network f_θ as the main model to be optimized, where $\theta = (\psi, W)$ consists of a representation encoder ϕ_ψ and a classifier head W. For each input x_i, the encoder extracts a representation $h_i = \phi_\psi(x_i)$, and the classifier produces logits and class probabilities by

$$z_i = Wh_i, \qquad p_i = \mathrm{softmax}(z_i).$$

Our framework contains four functional modules. First, the *student prediction module* produces the observed-label update induced by the noisy annotation $\tilde{y}_i$. Second, the *EMA reference module* maintains a temporally smoothed teacher model, which provides the reference target for measuring whether the observed-label update is reliable. Third, the *geometric reliability module* computes sample-wise update statistics, including a lightweight Fisher-trace proxy for update magnitude and relative geometric conflict for update direction. Finally, the *trust-weighted optimization module* converts the reliability score into a sample weight $w_i \in [0, 1]$, so that high-trust samples are weighted toward their observed labels and low-trust samples toward the reference target.

The reference target is produced by an EMA teacher $f_{\theta'}$ with the same architecture as the student:

$$q_i = \mathrm{Sharpen}(\mathrm{softmax}(f_{\theta'}(x_i)); T)\,, \tag{1}$$

$$\theta'_t \leftarrow m\theta'_{t-1} + (1-m)\theta_t. \tag{2}$$

Here, $T > 0$ denotes the sharpening temperature and m is the EMA momentum. For a probability vector u, sharpening is defined component-wise as

$$\mathrm{Sharpen}(u; T)_k = \frac{u_k^{1/T}}{\sum_{j=1}^K u_j^{1/T}}, \qquad k = 1, \ldots, K. \tag{3}$$

The teacher is not assumed to be an oracle. Instead, it acts as a lightweight temporal ensemble of historical student models, providing a smoother reference update without training multiple independent networks. Compared with explicit ensemble or prototype-based references, the EMA teacher introduces one additional forward pass and keeps the extra computation limited.

Practical trace proxy. The geometric reliability module requires a sample-wise trace signal to measure the update magnitude induced by each example. Throughout the paper, "sample-wise Fisher trace" means the empirical rank-one matrix formed from an individual sample gradient,

$\mathbf{F}_i^{\text{emp}} = \mathbf{g}_i \mathbf{g}_i^\top$. Its trace is exactly $\text{Tr}(\mathbf{F}_i^{\text{emp}}) = \|\mathbf{g}_i\|_2^2$; we do not use it as a population Fisher estimate. For cross-entropy loss with observed label $\tilde{y}_i$, denote

$$\delta_i = p_i - e_{\tilde{y}_i}.$$

The full-model sample-wise empirical Fisher trace is the squared gradient norm over all trainable parameters:

$$\tau_i^{\text{full}} = \|\nabla_\theta \ell_i\|_2^2 = \|\nabla_W \ell_i\|_F^2 + \|\nabla_\psi \ell_i\|_2^2. \tag{4}$$

Computing Eq. 4 for every sample is expensive. We therefore use the classifier-layer contribution as a lightweight surrogate. Since

$$\nabla_W \ell_i = \delta_i h_i^\top, \tag{5}$$

the last-layer trace admits the closed form

$$\tau_i^{\text{last}} = \|\nabla_W \ell_i\|_F^2 = \|\delta_i h_i^\top\|_F^2 = \|\delta_i\|_2^2 \|h_i\|_2^2. \tag{6}$$

This surrogate avoids full per-sample backpropagation through the encoder and can be computed directly from the prediction residual and the feature norm.

Importantly, τ_i^{last} is not an unconditional replacement for the full trace. The approximation error is exactly the omitted representation-gradient energy:

$$\tau_i^{\text{full}} = \tau_i^{\text{last}} + \epsilon_i, \qquad \epsilon_i = \|\nabla_\psi \ell_i\|_2^2 \geq 0. \tag{7}$$

Thus, the last-layer trace is a lower-bound surrogate whose fidelity depends on whether ϵ_i is small relative to the classifier-layer contribution and to the trace margin between samples. In particular, if $\epsilon_i \leq \kappa_i \tau_i^{\text{last}}$, then

$$\tau_i^{\text{last}} \leq \tau_i^{\text{full}} \leq (1 + \kappa_i)\tau_i^{\text{last}}. \tag{8}$$

Proposition 1. For samples i and j, suppose $\epsilon_k \leq \kappa_k \tau_k^{\text{last}}$ for $k \in \{i, j\}$. If

$$\tau_i^{\text{last}} > (1 + \kappa_j)\tau_j^{\text{last}}, \tag{9}$$

then $\tau_i^{\text{full}} > \tau_j^{\text{full}}$. This is a sufficient condition, not a necessary one: smaller last-layer gaps may still preserve the full-trace ordering when the omitted residuals are benign. The complete error bound, ordering-fidelity condition, and residual analysis across easy-clean, hard-clean, and noisy samples are provided in Appendix C.

3.2 From Forward Difficulty to Trace Geometry

Forward-space criteria such as loss, confidence, or entropy are widely used to estimate sample reliability, but they primarily measure learning difficulty rather than label correctness. A hard clean sample and a mislabeled sample can both produce high observed-label loss or low confidence, making them difficult to distinguish in prediction space. To move beyond such forward-space ambiguity, we inspect the update induced by each training example:

$$\mathbf{g}_i^{\text{obs}} = \nabla_\theta \ell(f_\theta(x_i), \tilde{y}_i). \tag{10}$$

The sample-wise empirical Fisher trace provides a parameter-space measure of this update magnitude:

$$\mathbf{F}_i^{\text{emp}} = \mathbf{g}_i^{\text{obs}} (\mathbf{g}_i^{\text{obs}})^\top, \qquad \tau_i^{\text{obs}} = \text{Tr}(\mathbf{F}_i^{\text{emp}}) = \|\mathbf{g}_i^{\text{obs}}\|_2^2. \tag{11}$$

This equality is an empirical rank-one convention for the individual sample gradient, not a claim about estimating the population Fisher. This trace signal captures how strongly a sample acts on the model parameters, and therefore provides information that is not visible from loss or confidence alone. In practice, we use the classifier-layer surrogate derived in Sec. 3.1.

However, trace remains a magnitude-based criterion. A hard clean sample may induce a large but useful update, while a noisy sample may induce a large and misleading update. Thus, update strength alone cannot determine whether the observed-label supervision is reliable. This motivates the next step: comparing the *direction* of the observed-label update with a reference update, which leads to relative geometric conflict.

3.3 Relative Geometric Conflict

RGC addresses the remaining ambiguity by comparing the observed-label update with a reference update. To remain consistent with the last-layer trace surrogate above, we compute this comparison on the classifier head W rather than on the full parameter set θ. Given q_i from Eq. 1, define

$$\mathbf{g}_{i,W}^{\mathrm{obs}} = \nabla_W \ell(f_\theta(x_i), \tilde{y}_i), \qquad \mathbf{g}_{i,W}^{\mathrm{ref}} = \nabla_W \ell(f_\theta(x_i), q_i), \qquad C_i^\epsilon = 1 - \frac{\langle \mathbf{g}_{i,W}^{\mathrm{obs}}, \mathbf{g}_{i,W}^{\mathrm{ref}} \rangle}{\|\mathbf{g}_{i,W}^{\mathrm{obs}}\|_F \|\mathbf{g}_{i,W}^{\mathrm{ref}}\|_F + \epsilon}. \tag{12}$$

Small C_i^ϵ means that the observed-label update is aligned with the reference; large C_i^ϵ indicates directional conflict. This distinction targets the case trace alone can mix: hard clean samples can have large gradient norm but small conflict, whereas mislabeled samples tend to induce directions that disagree with the reference.

For the softmax classifier, these gradients admit the closed forms

$$\mathbf{g}_{i,W}^{\mathrm{obs}} = (p_i - e_{\tilde{y}_i}) h_i^\top, \qquad \mathbf{g}_{i,W}^{\mathrm{ref}} = (p_i - q_i) h_i^\top. \tag{13}$$

The score uses $\tau_i^{\mathrm{obs}} = \|p_i - e_{\tilde{y}_i}\|_2^2 \|h_i\|_2^2$ and $\tau_i^{\mathrm{ref}} = \|p_i - q_i\|_2^2 \|h_i\|_2^2$; when the feature and residual norms are nonzero, C_i^ϵ is the stabilized implementation of the resulting logit-residual angular discrepancy. For a near-zero reference residual $p_i - q_i$, the direction is ill-defined. At $p_i = q_i$, Eq. 12 gives $C_i^\epsilon = 1$; ϵ ensures computability but cannot restore directional meaning or establish genuine conflict. We use classifier-space conflict for the same reason we use last-layer trace: it is the local interface where the observed label acts on the representation, providing a lightweight proxy for whether supervision is geometrically consistent at the decision boundary.

Let $q_i^\star$ be the ideal clean target and assume the logit-to-parameter Jacobian satisfies $\|J_i\|_2 \le L_i$. For soft-label cross-entropy, when the EMA target is closer to the clean target than the corrupted observed label, the resulting upper bound on its gradient error is smaller:

$$\|\nabla_\theta \ell(p_\theta(x_i), q_i) - \nabla_\theta \ell(p_\theta(x_i), q_i^\star)\|_2 \le L_i \|q_i - q_i^\star\|_2. \tag{14}$$

When the teacher drifts toward noisy labels, this condition can fail and the reference can become harmful. In practice, we mitigate this failure mode with warm-up, per-sample score smoothing, and a gated objective that applies teacher supervision mainly to low-trust samples.

This observed–reference comparison also specifies the roles of trace and RGC. For the nondegenerate theoretical case with $a_i = \|\mathbf{g}_{i,W}^{\mathrm{obs}}\|_F > 0$ and $b_i = \|\mathbf{g}_{i,W}^{\mathrm{ref}}\|_F > 0$, let $r_i = a_i / b_i$ and define the unstabilized angular conflict $C_i^\circ = 1 - \cos\theta_i$. The normalized gradient discrepancy decomposes as

$$\frac{\|\mathbf{g}_{i,W}^{\mathrm{obs}} - \mathbf{g}_{i,W}^{\mathrm{ref}}\|_F^2}{b_i^2} = (r_i - 1)^2 + 2 r_i C_i^\circ. \tag{15}$$

This identity is exact for the unstabilized geometry; trace captures radial mismatch, while RGC uses C_i^ϵ as a numerically robust angular signal. This decomposition reframes the score from update strength alone to whether the observed-label update is geometrically consistent with the reference.

3.4 Reliability Score and Overall Objective

We convert the observed–reference update discrepancy into a sample-level unreliability score. Given the observed-label trace τ_i^{obs}, the reference trace τ_i^{ref}, and the stabilized relative geometric conflict C_i^ϵ in Eq. 12, we define

$$R_i^\tau = \log \frac{\tau_i^{\mathrm{obs}} + \epsilon}{\tau_i^{\mathrm{ref}} + \epsilon}, \qquad s_i = \beta \widehat{R}_i^\tau + (1 - \beta) \widehat{C}_i^\epsilon, \tag{16}$$

where $\widehat{R}_i^\tau$ and $\widehat{C}_i^\epsilon$ are standardized within each fine-tuning epoch. Here, R_i^τ measures radial mismatch in update magnitude, while C_i^ϵ measures angular mismatch in update direction. We use a direction-dominant default because trace can be large for both noisy samples and hard clean samples; directional conflict is meant to test agreement with the reference rather than provide an optimal mixing rule.

Proposition 2. For a noisy sample n and a hard clean sample h, define

$$\Delta_C = \widehat{C}_n - \widehat{C}_h, \qquad \Delta_R = \widehat{R}^{\tau}_n - \widehat{R}^{\tau}_h.$$

If the directional margin satisfies $\Delta_C \geq m_C > 0$ and the radial term is no worse than $\Delta_R \geq -m_R$ for $m_R \geq 0$, then the fused score preserves the noisy-over-hard-clean ordering, $s_n > s_h$, for any

$$0 \leq \beta < \frac{m_C}{m_C + m_R}. \tag{17}$$

Indeed, $s_n - s_h \geq (1-\beta)m_C - \beta m_R > 0$, so this sufficient interval keeps the angular term dominant when radial trace alone is ambiguous.

The raw unreliability score is smoothed by a per-sample EMA and mapped to the observed-label trust weight:

$$\tilde{s}_{i,t} \leftarrow \mu \tilde{s}_{i,t-1} + (1-\mu)s_{i,t}, \qquad w_i = \sigma\left(-\alpha \frac{\tilde{s}_{i,t} - \mathrm{mean}(\tilde{s}_t)}{\mathrm{std}(\tilde{s}_t) + \epsilon}\right). \tag{18}$$

A larger s_i yields a smaller w_i, indicating lower trust in the observed label.

The student is trained with a trust-weighted objective:

$$\mathcal{L}_{\mathrm{RGC}} = \frac{1}{N}\sum_{i=1}^{N}\left[w_i \ell(f_\theta(x_i), \tilde{y}_i) + \lambda(1-w_i)\ell(f_\theta(x_i), q_i)\right]. \tag{19}$$

Thus, high-trust samples remain supervised mainly by their observed labels, while low-trust samples receive more reference-target supervision. This continuous gate avoids a hard clean/noisy split: trust in the observed label decreases as its induced update becomes less consistent with the reference. Appendix D.1 gives the conditional reference-fidelity statement, the EMA lag bound, and the score-smoothing stability remarks used to interpret these gates.

4 Experiments

Setup. We evaluate Trace and RGC on CIFAR-10 and CIFAR-100 [18] under symmetric and asymmetric noise, and on Animal-10N [33], Food-101N [21], and WebVision-50 [24] under real-world label noise. For WebVision-50, we report accuracy on both its validation set and the corresponding ILSVRC12-50 validation set [8, 32]. CIFAR noise rates range from 20% to 80%.

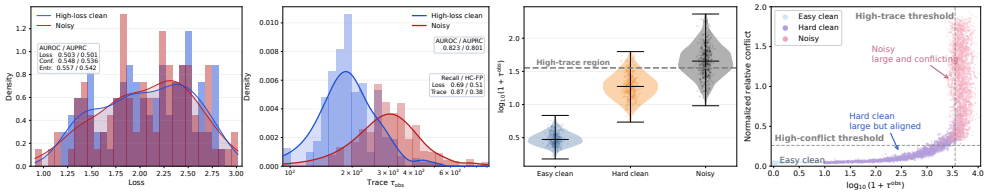
Within each experimental setting, we match the external configurations across methods, including the backbone, initialization, data and noise splits, optimizer, augmentation, learning-rate schedule, and checkpoint selection rule, while retaining their method-specific training structures. Under the submitted setting used in the main tables, single-stage methods are trained for 30 epochs, whereas methods with warm-up or two-stage training use 5 warm-up epochs followed by 25 training epochs. EarlyCut retains its native iterative structure of three selection rounds and one final training round, with 30 epochs in the final training stage.

For RGC in the main-table setting, the model is first warmed up for $E_w = 5$ epochs using observed-label cross-entropy, followed by 25 training epochs with an EMA teacher ($m = 0.999$), sharpened targets ($T = 0.7$), score smoothing ($\mu = 0.9$), trust sharpness $\alpha = 0.6$, and radial–angular fusion coefficient $\beta = 0.2$. The rationale for this short warm-up and its interaction with the EMA reference is discussed in Appendix D.1.

In addition, we conduct an EarlyCut-compatible comparison on CIFAR-10 using ResNet-18, from-scratch initialization, and matched external training configurations. RGC performs a single 300-epoch training run, while EarlyCut retains three selection rounds followed by a final 300-epoch retraining stage. This comparison covers Sym.20%, Sym.80%, IDN-20%, and IDN-40%. The methods match the external configurations and final-stage training schedule, but their total training compute differs because EarlyCut includes additional selection stages.

We report classification accuracy, noisy-sample detection, and hard-clean preservation. For diagnostic evaluation on synthetic-noise benchmarks, noisy samples are $\mathcal{N} = \{i : \tilde{y}_i \neq y_i\}$, and hard clean samples $\mathcal{H}$ are defined as the top-20% high-loss subset among clean samples. Given a threshold η, samples with $s_i > \eta$ are predicted as unreliable, and we report

$$\mathrm{NoisyRec.} = \frac{|\{i \in \mathcal{N} : s_i > \eta\}|}{|\mathcal{N}|}, \quad \mathrm{HC\text{-}FP} = \frac{|\{i \in \mathcal{H} : s_i > \eta\}|}{|\mathcal{H}|}, \quad \mathrm{HCPR} = 1 - \mathrm{HC\text{-}FP}.$$



(a) Loss overlap in the high-loss clean region (b) Trace distribution (c) Trace magnitude across groups (d) Relative geometric conflict

Figure 2: Diagnostic analysis of reliability signals. Hard-clean samples are the top 20% of clean samples ranked by loss after warm-up. In this studied high-loss clean region, forward-space signals show substantial overlap between hard-clean and noisy samples. Trace improves discrimination but remains susceptible to hard-clean false positives; relative geometric conflict adds directional separation. These observations are specific to this loss-based diagnostic subset.

These metrics test whether a reliability signal detects corrupted labels while preserving hard but correctly labeled examples.

4.1 Diagnostic analysis of reliability signals.

We select the top 20% of clean samples ranked by loss after warm-up as hard-clean samples and compare them with noisy samples to analyze reliability signals in the high-loss clean region. As shown in Fig. 2(a), the observed-label loss distributions of hard-clean and noisy samples overlap substantially in this diagnostic subset: Loss achieves an AUROC/AUPRC of only 0.503/0.501, while Confidence and Entropy are also close to chance. Fig. 2(b) shows that Trace partially reduces this ambiguity, reaching an AUROC/AUPRC of 0.823/0.801 and indicating that update magnitude contains information beyond scalar loss.

However, Fig. 2(c) shows that Trace remains a radial signal: hard-clean and noisy samples can still overlap in the high-Trace region. Fig. 2(d) shows that, among the samples studied, some hard-clean samples induce large updates whose directions remain relatively aligned with the reference, whereas noisy samples tend to exhibit greater directional conflict. To directly evaluate the discriminative ability of the full score, we further compare reliability signals under CIFAR-10 Sym.20% using a fixed Student/EMA checkpoint and the same evaluation samples. Full RGC achieves an AUROC/AUPRC of 0.872/0.854, outperforming JS disagreement and Conflict-only, consistent with radial and directional information playing complementary roles. The complete comparison is provided in Appendix E.1.

We further apply predetermined visual perturbations to class-balanced clean samples to construct a difficult-sample set independently of loss and all other evaluated scores. Under CIFAR-10 Sym.20%, Full RGC achieves an AUROC/AUPRC of 0.839/0.817, outperforming Loss at 0.632/0.611, indicating that its advantage is maintained under the tested controlled visual-difficulty setting. Construction details and results are provided in Appendix E.2.

At thresholds chosen separately for each signal to achieve 80% noisy-sample recall, we audit hard-clean samples additionally retained by RGC, commonly retained by Loss, Trace, and RGC, and incorrectly rejected by RGC. These cases illustrate the potential benefits and false-rejection risks of directional information; they do not replace aggregate quantitative evaluation. Details are provided in Appendix E.3.

4.2 Main Results

Tables 1 and 2 report the main comparison results. Under this evaluation protocol, RGC is the top entry among the evaluated methods in each CIFAR-10 noise setting. The largest gaps appear in the 80% symmetric and 40% asymmetric settings, with improvements of +6.2 and +3.3 points over the next-highest non-RGC entries in those columns. These results are consistent with RGC combining update magnitude and EMA-referenced directional conflict to distinguish large aligned hard-clean updates from conflicting noisy updates, as examined in Sec. 4.1 and the component ablations below.

To examine the dependence of the comparison results on the training protocol, we additionally conduct a comparison under the EarlyCut-compatible protocol. On CIFAR-10 under Sym.20%, Sym.80%, IDN-20%, and IDN-40%, RGC achieves accuracies of 95.4%, 88.3%, 94.1%, and 91.4%, exceeding EarlyCut by 0.6, 2.2, 0.7, and 0.6 percentage points, respectively. These results show that RGC maintains a modest and consistent performance advantage when the external configurations and final-stage training schedules are matched. The complete cross-protocol results are provided

Table 1: Test accuracy (%) on the CIFAR-10 and CIFAR-100 datasets under symmetric and asymmetric noise. Bold indicates the highest value in each column.

Method	CIFAR-10					CIFAR-100			
	Sym.			Asym.	Avg.	Sym.		Asym.	Avg.
	20%	50%	80%	40%		20%	50%	40%	
Standard CE	73.4	58.9	49.2	60.9	60.6	44.6	32.8	38.7	38.7
GCE [42]	77.2	75.7	66.7	67.7	71.8	50.1	42.3	45.7	46.0
DivideMix [22]	79.5	74.6	70.7	68.1	73.2	61.6	47.9	53.1	54.2
RRL [23]	88.4	83.3	74.7	76.5	80.7	61.0	58.6	53.5	57.7
UNICON [16]	72.8	71.5	68.1	76.4	72.2	60.7	57.1	54.2	57.3
ProMix [36]	87.4	83.7	68.7	77.0	79.2	55.2	52.8	44.6	50.9
LongReMix [7]	81.4	79.0	67.6	67.7	73.9	61.3	56.9	47.8	55.3
L2B [43]	86.7	82.1	59.4	76.7	76.2	42.9	46.1	39.8	42.9
PSSCL [40]	82.3	80.7	77.6	71.0	77.9	56.5	45.6	46.0	49.4
DCD [26]	87.6	78.7	72.9	75.8	78.8	63.3	54.9	53.3	57.2
EarlyCut [38]	87.9	83.7	78.4	76.6	81.7	63.7	59.6	52.0	58.4
IDO [39]	90.4	86.9	80.7	84.3	85.6	70.8	63.2	60.5	64.8
RGC (Ours)	**93.5**	**92.4**	**86.9**	**87.6**	**90.1**	**73.7**	**68.6**	**64.6**	**69.0**

Table 2: Classification accuracy (%) on three real-world noisy-label benchmarks, reported across four evaluation sets. Bold indicates the best result in each column.

Method	Ref.	Animal-10N	Food-101N	WebVision-50 val.	ILSVRC12-50 val.
DivideMix	ICLR20	71.5	67.2	76.0	75.4
RRL	ICCV21	73.8	69.7	81.6	79.0
UNICON	CVPR22	71.1	70.4	78.6	77.2
ProMix	IJCAI23	76.2	70.1	78.5	78.0
LongReMix	PR23	71.6	68.7	84.3	81.6
L2B	CVPR24	71.4	68.9	80.4	78.9
PSSCL	PR25	72.7	70.7	76.6	74.3
DCD	ICCV25	74.1	71.2	80.8	78.4
EarlyCut	NeurIPS25	75.0	72.1	81.7	80.8
IDO	NeurIPS25	77.2	74.4	81.8	78.7
RGC (Ours)	–	**81.4**	**78.5**	**85.9**	**82.8**

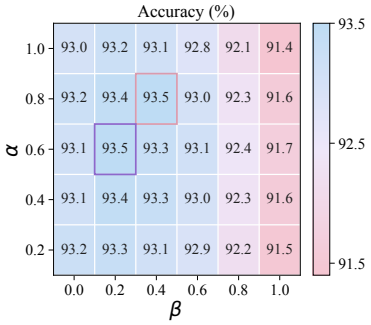


Figure 3: Hyperparameter sensitivity to α and β.

in Appendix B.1; because EarlyCut includes additional selection stages, this comparison does not imply equal total compute budgets.

4.3 Ablation Studies

4.3.1 Ablation on Geometric Reliability Components

Table 3 presents an ablation study of the three components in geometric reliability modeling. Teacher denotes EMA-based reference supervision, Trace-only uses absolute last-layer Trace, and Trace+Teacher combines Trace magnitude with teacher supervision. Full RGC further incorporates directional conflict modeling. Although Trace and Teacher exhibit varying performance when used individually, their combination improves results across all seven settings, with gains of up to 2.7 percentage points, demonstrating their complementary effects. Adding Conflict further yields improvements of 0.2–1.8 percentage points, highlighting the complementary value of directional information.

4.3.2 Hyperparameter Sensitivity

Figure 3 examines sensitivity to α and β, which control trust-weight sharpness and radial–angular balance, respectively. Accuracy varies little across the tested α values, consistent with α changing the weight mapping rather than sample ordering. With small-to-moderate β, directional conflict remains the primary cue and trace provides auxiliary magnitude information. Larger β overemphasizes magnitude and reduces accuracy, supporting a balance that preserves directional information.

4.3.3 Teacher-Related Ablations

Reference Design. Figure 4(a) compares reference designs. Without temporal smoothing ($m = 0$), the teacher provides only limited gains over Trace-only. Increasing m stabilizes the EMA reference in the tested settings, with accuracy and HCPR supporting its role in hard-clean preservation. Class-prototype references depend on noisy representations, and combining them with EMA does

Table 3: Ablation on the geometric reliability components on CIFAR-10 and CIFAR-100. Trace denotes radial update-magnitude scoring, Teacher denotes the EMA-based reference target, and Conflict denotes the proposed relative geometric conflict.

Variant	Module			CIFAR-10				CIFAR-100		
	Trace	Teacher	Conflict	Sym. 20%	Sym. 50%	Sym. 80%	Asym. 40%	Sym. 20%	Sym. 50%	Asym. 40%
Trace	✓	✗	✗	91.5	88.4	83.0	85.8	71.2	64.4	61.7
Teacher	✗	✓	✗	91.2	88.7	82.5	85.4	71.4	64.9	62.5
Trace+Teacher	✓	✓	✗	92.6	90.8	85.1	86.8	72.2	66.8	64.4
RGC (Ours)	✓	✓	✓	**93.5**	**92.4**	**86.9**	**87.6**	**73.7**	**68.6**	**64.6**

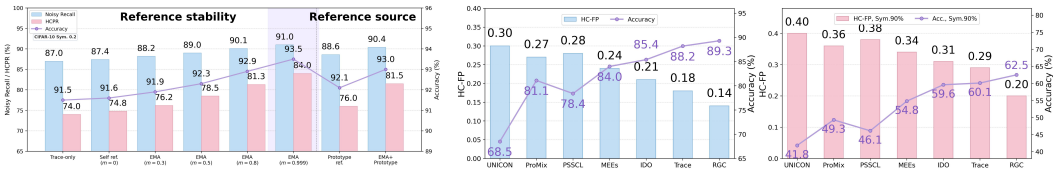


Figure 4: Ablation and boundary analysis of RGC. The figure summarizes the reference design ablation and the stress tests under weak model capacity and extreme noise.

not consistently outperform pure EMA. These results support using EMA to define the reference direction.

Necessity of the Teacher-Target Term. To isolate the benefit of teacher supervision beyond down-weighting, we retain the EMA-teacher mechanism, radial–angular score construction, and trust-weight mapping, removing only the teacher-target loss. The down-weighting-only variant optimizes $w_i \ell(f_\theta(x_i), \tilde{y}_i)$; full RGC additionally includes $\lambda(1 - w_i)\ell(f_\theta(x_i), q_i)$, providing alternative supervision for low-trust observed labels. As shown in Fig. 5, the full objective improves over down-weighting only across the tested settings, averaging approximately 1.4 percentage points. Thus, in these settings, down-weighting unreliable samples alone does not recover the full objective's benefits: the teacher-target term provides additional value through alternative supervision.

Teacher-Related Hyperparameters. We examine EMA momentum m, sharpening temperature T, and teacher-loss coefficient λ, varying one parameter at a time while keeping the others at their defaults. The maximum accuracy variation across the tested ranges is 1.08 percentage points, indicating relatively stable performance within these ranges. Detailed test ranges and accuracy and HCPR results are provided in Appendix D.2.

4.4 Boundary and Robustness Validation

We evaluate RGC under low-data, small-backbone, and extreme-noise settings to examine when the reference direction becomes unreliable and how this limits the gains of direction-aware reliability estimation.

Low-Data Boundary Table 4 shows that all methods degrade as the amount of training data decreases, while HC-FP generally rises, confirming that low-data training weakens reliability estimation. RGC remains above IDO across all data ratios, but the hard-clean false-positive rate worsens as the data ratio decreases. At the extreme 20% setting, Trace slightly surpasses RGC in accuracy (84.7 vs. 84.3) and HC-FP (0.46 vs. 0.49), while RGC remains above IDO in accuracy. This marks a boundary case: with very little data, the EMA reference becomes higher-variance, so conflict may no longer distinguish useful hard-clean updates from harmful noisy ones.

Teacher Quality and Error Propagation. With less training data, teacher accuracy and calibration deteriorate alongside Conflict AUROC; this association does not establish causality. In the audited standard setting, 83.3% of the total teacher-supervision weight is assigned to samples with correct teacher predictions. Incorrect teacher targets still receive weight, so EMA smoothing and trust weighting cannot eliminate error propagation, highlighting RGC's dependence on reference quality. Complete teacher-quality diagnostics are provided in Appendix D.3.

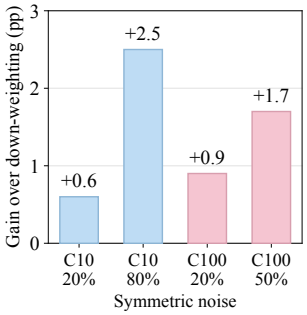


Figure 5: Teacher-target ablation.

Table 4: Low-data robustness under different training-data ratios. Higher Acc. and lower HC-FP are better; $\Delta = \mathrm{RGC} - \mathrm{IDO}$.

Data ratio	Metric	ProMix	PSSCL	EarlyCut	IDO	Trace	**RGC**	Δ
100%	Acc. ↑	87.4	82.3	87.9	90.4	91.5	**93.5**	+3.1
	HC-FP ↓	0.34	0.32	0.29	0.25	0.22	**0.16**	-0.09
80%	Acc. ↑	86.1	80.9	86.4	88.7	90.7	**92.1**	+3.4
	HC-FP ↓	0.36	0.39	0.31	0.32	0.27	**0.23**	-0.09
50%	Acc. ↑	82.7	77.2	83.9	85.4	89.2	**90.8**	+5.4
	HC-FP ↓	0.46	0.47	0.40	0.38	0.34	**0.28**	-0.10
20%	Acc. ↑	76.9	71.0	79.1	79.0	**84.7**	84.3	+5.3
	HC-FP ↓	0.58	0.60	0.54	0.51	**0.46**	0.49	-0.02

Table 5: Runtime and memory on CIFAR-10 with 20% symmetric noise.

Method	Sec./Epoch ↓	Memory ↓
CE	52	3.2 GB
DivideMix	138	8.4 GB
PSSCL	165	8.6 GB
ProMix	121	6.3 GB
DCD	114	6.0 GB
IDO	98	4.9 GB
RGC	110	5.3 GB

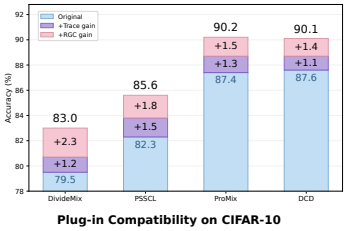


Figure 6: Plug-in compatibility of Trace and RGC on representative noisy-label learning frameworks.

Model Capacity and Extreme Noise Model capacity and extreme noise stress the geometric reference. A smaller backbone weakens representations and reduces the gain over Trace (Fig. 4(b)). Extreme symmetric noise can drift the teacher toward corrupted labels (Fig. 4(c)), while extreme asymmetric noise can make wrong-class updates partly aligned; in both cases, RGC becomes more reference-limited.

5 Efficiency and Plug-in Compatibility

Table 5 reports per-epoch runtime and peak GPU memory on CIFAR-10 with 20% symmetric noise. RGC requires 110 s and 5.3 GB, compared with 98 s and 4.9 GB for IDO. Its additional trace and conflict calculations operate at the classifier layer, avoiding full-model per-sample Fisher computation.

Figure 6 shows accuracy gains from Trace and RGC across the tested host frameworks while retaining each host's training recipe and changing its reliability module. These results support RGC's practical efficiency and plug-in compatibility in the evaluated settings.

6 Limitations

RGC relies on the reference direction being more stable than the observed-label update, which can weaken when the EMA teacher has high variance or drifts toward corrupted labels. The last-layer trace and conflict proxies focus on the classifier interface rather than full-model geometry; the appendix gives sufficient conditions and residual diagnostics, not a general guarantee. The β interval is likewise sufficient, not an optimal tuning rule. Near-zero reference residuals also make the angular score directionally uninformative. Larger architectures and non-classification tasks remain future work.

7 Conclusion

We presented RGC, a direction-aware reliability criterion for noisy-label learning that compares the observed-label update with a teacher-induced reference update. It combines radial update energy with observed–reference conflict to score trust in noisy supervision. In the evaluated protocols, RGC improves over magnitude-only trace criteria when the reference is reliable, with gains shrinking when the teacher can drift. Code is available at `https://github.com/maojingyangmao-cloud/RGC`.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (22673056 and 92370127). The authors declare no competing interests.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the International Conference on Machine Learning*, 2019.

[3] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, and Yoshua Bengio. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

[4] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

[7] 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.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the International Conference on Machine Learning*, 2019.

[10] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017.

[11] 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. In *Advances in Neural Information Processing Systems*, 2018.

[12] 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*, 2020.

[13] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[14] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR, 2021.

[15] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning*, 2018.

[16] 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*, pages 9676–9686, 2022.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[21] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456. IEEE, 2018.

[22] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[23] Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9485–9494, 2021.

[24] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[25] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, 2020.

[26] 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*, pages 415–425, 2025. doi: 10.1109/ICCV51701.2025.00046.

[27] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] Ningkang Peng, Jingyang Mao, Xiaoqian Peng, Peirong Ma, Xichen Yang, Weiguang Qu, and Yanhui Gu. Gamr: Geometric-aware manifold regularization with virtual outlier synthesis for learning with noisy labels, 2026. URL `https://arxiv.org/abs/2605.20727`.

[29] Ningkang Peng, Jingyang Mao, Qianfeng Yu, Xiaoqian Peng, Peirong Ma, and Yanhui Gu. Hambr: Active decision boundary restoration based on hamiltonian dynamics for learning with noisy labels, 2026. URL `https://arxiv.org/abs/2605.11383`.

[30] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations Workshop*, 2015.

[31] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, 2007.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

[35] 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.

[36] Ruixuan Xiao, Yiwen Dong, Haobo Wang, Lei Feng, Runze Wu, Gang Chen, and Junbo Zhao. ProMix: Combating label noise via maximizing clean sample utility. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4442–4450, 2023. doi: 10.24963/ijcai.2023/494.

[37] 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*, pages 7164–7173. PMLR, 2019.

[38] 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*, 2025.

[39] Kuan Zhang, Chengliang Chai, Jingzhe Xu, Chi Zhang, Han Han, Ye Yuan, Guoren Wang, and Lei Cao. Handling label noise via instance-level difficulty modeling and dynamic optimization. In *Advances in Neural Information Processing Systems*, 2025.

[40] 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.

[41] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021.

[42] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

[43] Yuyin Zhou, Xianhang Li, Fengze Liu, Qingyue Wei, Xuxi Chen, Lequan Yu, Cihang Xie, Matthew P Lungren, and Lei Xing. L2b: Learning to bootstrap robust models for combating label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23523–23533, 2024.

A Training Procedure

Algorithm 1 summarizes the training procedure used by RGC. The released implementation follows an epoch-based pipeline; here, a round denotes one fine-tuning epoch.

Algorithm 1 RGC training with last-layer trace and relative geometric conflict

Require: Noisy training set $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, student f_θ, total rounds R, warm-up rounds E_w, EMA momentum m, sharpening temperature T, score-smoothing momentum μ, trust sharpness α, radial weight β, reference loss weight λ, numerical constant ϵ.

Ensure: Trained student parameters θ.

Initialize smoothed scores $\tilde{s}_{i,0} \leftarrow 0$ for all samples.
for $t = 1$ **to** E_w **do**
  for each mini-batch $B \subset \mathcal{D}$ **do**
    Update θ by minimizing observed-label cross-entropy $\frac{1}{|B|}\sum_{i\in B} \ell(f_\theta(x_i), \tilde{y}_i)$.
  end for
end for
Initialize the teacher from the warmed-up student, $\theta' \leftarrow \theta$.
for $t = E_w + 1$ **to** R **do**
  for each mini-batch $B \subset \mathcal{D}$ **do**
    for each sample $(x_i, \tilde{y}_i) \in B$ **do**
      Compute $h_i = \phi_\psi(x_i)$ and $p_i = \mathrm{softmax}(f_\theta(x_i))$.
      Compute the teacher target $q_i = \mathrm{Sharpen}(\mathrm{softmax}(f_{\theta'}(x_i)); T)$.
      Compute

$$\tau_i^{\mathrm{obs}} = \|p_i - e_{\tilde{y}_i}\|_2^2 \|h_i\|_2^2, \qquad \tau_i^{\mathrm{ref}} = \|p_i - q_i\|_2^2 \|h_i\|_2^2.$$

      Compute stabilized conflict C_i^ϵ by Eq. 12 and radial mismatch $R_i^\tau = \log((\tau_i^{\mathrm{obs}} + \epsilon)/(\tau_i^{\mathrm{ref}} + \epsilon))$.
      Standardize R_i^τ and C_i^ϵ with the current epoch statistics, then set $s_i = \beta \widehat{R}_i^\tau + (1-\beta)\widehat{C}_i^\epsilon$.
      Update the stored score $\tilde{s}_{i,t} \leftarrow \mu \tilde{s}_{i,t-1} + (1-\mu) s_i$ and compute

$$w_i = \sigma\left(-\alpha \frac{\tilde{s}_{i,t} - \mathrm{mean}(\tilde{s}_t)}{\mathrm{std}(\tilde{s}_t) + \epsilon}\right).$$

    end for
    Update θ by minimizing

$$\frac{1}{|B|}\sum_{i\in B} \left[w_i \ell(f_\theta(x_i), \tilde{y}_i) + \lambda(1 - w_i)\ell(f_\theta(x_i), q_i) \right].$$

    Update the teacher by EMA: $\theta' \leftarrow m\theta' + (1-m)\theta$.
  end for
end for
return θ.

Full RGC retains the warm-up initialization and basic training scaffold of IDO [39], while rebuilding the reliability-estimation pipeline and downstream supervision objective. Specifically, RGC uses an EMA-referenced radial–angular score in place of IDO's cumulative wrong-event statistics, class-wise Beta mixture models, and associated posterior reliability and difficulty weights. The score is mapped to an observed-label trust weight w_i: the observed-label loss is weighted by w_i, while the EMA-teacher-target loss is weighted by $\lambda(1 - w_i)$.

B EarlyCut Comparison and Compute Resources

B.1 Cross-Protocol Comparison with EarlyCut

Cross-protocol comparison. Table B.1 presents the results of RGC and EarlyCut under the Submitted and EarlyCut-compatible protocols. The EarlyCut-compatible setting uses ResNet-18 with

Table B.1: Cross-protocol test accuracy (%) on CIFAR-10.

Method	Protocol	Backbone	Training structure	Sym. 20%	Sym. 80%	IDN 20%	IDN 40%
RGC	Submitted	ViT-B	5-epoch warm-up + 25-epoch training	93.5	86.9	90.8	86.3
EarlyCut	Submitted	ViT-B	3 selection rounds + 30-epoch final training	87.9	78.4	87.2	84.3
RGC	EarlyCut-compatible	ResNet-18	Single 300-epoch run	95.4	88.3	94.1	91.4
EarlyCut	EarlyCut-compatible	ResNet-18	3 selection rounds + 300-epoch final retraining	94.8	86.1	93.4	90.8

from-scratch initialization; RGC performs a single 300-epoch training run, while EarlyCut retains three selection rounds followed by a final 300-epoch retraining stage.

Under this compatible protocol, RGC improves over EarlyCut by 0.6, 2.2, 0.7, and 0.6 percentage points on Sym.20%, Sym.80%, IDN-20%, and IDN-40%, respectively, showing that its performance advantage is maintained in this setting. Because the two protocols differ in both backbone and training schedule, cross-protocol performance changes cannot be attributed to training duration alone.

Within each protocol, the two methods use matched external configurations while retaining their method-specific stage structures. In the Submitted setting, RGC uses 5 warm-up epochs followed by 25 training epochs, whereas EarlyCut retains three selection rounds (at most 30 epochs per round) and a 30-epoch final stage. In the EarlyCut-compatible setting, RGC performs one 300-epoch run and EarlyCut retains its selection rounds followed by 300-epoch final retraining; the 300-epoch schedules are matched only in this compatible comparison. However, EarlyCut includes additional selection stages, so the comparison does not imply equal total training compute budgets.

B.2 Compute Resources

Hardware and software environment. Experiments were run on a Linux server equipped with NVIDIA L20 GPUs. Each individual training run used one L20 GPU with 46 GB memory. The server contains eight such GPUs, which were used to run multiple experiments in parallel. The machine also has two AMD EPYC 9334 processors, with 128 logical CPUs in total, and 503 GB system RAM. The software environment is Python 3.13.5, PyTorch 2.9.1+cu128, CUDA 12.8, and cuDNN 91002.

C Last-Layer Trace Approximation

Computing the full-model Fisher trace for each sample requires gradients with respect to all trainable parameters, which is computationally expensive for routine noisy-label training. In the main paper, we therefore use the classifier-layer trace as a lightweight surrogate. This approximation is not intended to be an unconditional replacement for the full trace. Instead, it is a conditional surrogate whose error is exactly the omitted representation-gradient energy. As in the main text, all sample-wise Fisher traces in this appendix refer to empirical rank-one matrices $\mathbf{g}_i\mathbf{g}_i^\top$, whose traces equal squared gradient norms exactly. Here we provide the full derivation, the deterministic error bound, the ordering-fidelity condition, and the residual analysis used to verify whether the approximation systematically fails on easy-clean, hard-clean, or noisy samples.

Full-model trace decomposition. Let the student model be decomposed as

$$h_i = \phi_\psi(x_i), \qquad z_i = W h_i, \qquad p_i = \mathrm{softmax}(z_i), \tag{20}$$

where $\theta = (W, \psi)$ consists of the classifier parameters W and the representation parameters ψ. For cross-entropy loss with observed label $\tilde{y}_i$, denote

$$\delta_i = p_i - e_{\tilde{y}_i}. \tag{21}$$

The sample-wise full-model empirical Fisher trace is the squared norm of the sample gradient:

$$\tau_i^{\mathrm{full}} = \|\nabla_\theta \ell_i\|_2^2 = \|\nabla_W \ell_i\|_F^2 + \|\nabla_\psi \ell_i\|_2^2. \tag{22}$$

For the classifier layer, the gradient has the closed form

$$\nabla_W \ell_i = \delta_i h_i^\top. \tag{23}$$

Therefore, the last-layer trace is

$$\tau_i^{\mathrm{last}} = \|\nabla_W \ell_i\|_F^2 = \|\delta_i h_i^\top\|_F^2 = \|\delta_i\|_2^2 \|h_i\|_2^2. \tag{24}$$

This quantity can be computed from the prediction residual and the feature norm without full per-sample backpropagation through the encoder.

Representation-gradient residual. The representation gradient can be written as

$$\nabla_\psi \ell_i = J_\psi(x_i)^\top W^\top \delta_i, \tag{25}$$

where $J_\psi(x_i)$ is the Jacobian of the representation encoder with respect to its parameters. For a classifier with a bias term, the same derivation applies by augmenting h_i with a constant coordinate.

Define the omitted residual as

$$\epsilon_i = \|\nabla_\psi \ell_i\|_2^2. \tag{26}$$

Then

$$\tau_i^{\text{full}} = \tau_i^{\text{last}} + \epsilon_i, \qquad \epsilon_i \geq 0. \tag{27}$$

Thus, the last-layer trace is always a lower-bound surrogate of the full trace. The approximation is faithful when the residual ϵ_i is small relative to the classifier-layer contribution or, more importantly, when it does not change the ordering of samples used for reliability estimation.

Deterministic error bound. By sub-multiplicativity of the spectral norm,

$$0 \leq \epsilon_i = \|J_\psi(x_i)^\top W^\top \delta_i\|_2^2 \leq \|J_\psi(x_i)\|_2^2 \|W\|_2^2 \|\delta_i\|_2^2. \tag{28}$$

When $\tau_i^{\text{last}} > 0$, using Eq. 24, we obtain the relative bound

$$\tau_i^{\text{full}} = \tau_i^{\text{last}}(1 + r_i), \qquad 0 \leq r_i \leq \kappa_i, \qquad \kappa_i = \frac{\|J_\psi(x_i)\|_2^2 \|W\|_2^2}{\|h_i\|_2^2}. \tag{29}$$

Equivalently,

$$\tau_i^{\text{last}} \leq \tau_i^{\text{full}} \leq (1 + \kappa_i)\tau_i^{\text{last}}, \qquad \frac{\tau_i^{\text{full}} - \tau_i^{\text{last}}}{\tau_i^{\text{full}}} \leq \frac{\kappa_i}{1 + \kappa_i}. \tag{30}$$

If the representation is locally Lipschitz with $\|J_\psi(x_i)\|_2 \leq L_\psi$, the classifier norm is bounded by $\|W\|_2 \leq B_W$, and the feature norm is non-degenerate with $\|h_i\|_2 \geq m_h > 0$, then all samples obey the uniform distortion bound

$$\tau_i^{\text{last}} \leq \tau_i^{\text{full}} \leq \left(1 + \frac{L_\psi^2 B_W^2}{m_h^2}\right) \tau_i^{\text{last}}. \tag{31}$$

If $\delta_i = 0$, both classifier and representation gradients vanish and the approximation is exact.

Ordering-fidelity condition. The learning algorithm does not require the last-layer trace to exactly match the full trace numerically. Instead, it mainly requires the trace proxy to preserve the ordering of samples. For two samples i and j, suppose

$$\tau_i^{\text{last}} > (1 + \kappa_j)\tau_j^{\text{last}}. \tag{32}$$

Then, by Eq. 30,

$$\tau_i^{\text{full}} \geq \tau_i^{\text{last}} > (1 + \kappa_j)\tau_j^{\text{last}} \geq \tau_j^{\text{full}}. \tag{33}$$

Therefore,

$$\tau_i^{\text{full}} > \tau_j^{\text{full}}. \tag{34}$$

This gives a sufficient condition for preserving the ordering induced by the full trace. With a global bound $\kappa_i \leq \kappa$, any pair whose last-layer traces differ by a multiplicative margin larger than $1 + \kappa$ cannot be reversed by the omitted representation-gradient residual. Thus, rank reversals are expected mainly for near-tie samples or for samples with unusually large residual factors.

Correlation-fidelity condition. We can also state a perturbation-style condition for correlation fidelity. Let $t^{\text{last}} \in \mathbb{R}^n$ be the vector of last-layer traces over an evaluation set, let $\epsilon \in \mathbb{R}^n$ be the residual vector, and let

$$P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \tag{35}$$

be the centering matrix. Since

$$t^{\text{full}} = t^{\text{last}} + \epsilon, \tag{36}$$

define

$$\gamma = \frac{\|P\epsilon\|_2}{\|Pt^{\text{last}}\|_2}. \tag{37}$$

When $\gamma < 1$, the Pearson correlation between the last-layer trace and the full trace satisfies the perturbation lower bound

$$\rho\left(t^{\text{last}}, t^{\text{full}}\right) \geq \frac{1-\gamma}{1+\gamma}. \tag{38}$$

This condition shows that high empirical correlation is expected when the centered residual energy is small relative to the variation of the last-layer trace. In our experiments, correlation and residual analyses are used to verify this condition rather than to serve as the sole justification of the approximation.

Residual analysis across sample groups. We further examine whether the last-layer trace approximation introduces systematic bias across sample groups. We compute full per-sample traces on a stratified subset of CIFAR-10 with 20% symmetric noise after warm-up. Noisy samples are those with $\tilde{y}_i \neq y_i$. Among clean samples with $\tilde{y}_i = y_i$, hard clean samples are defined as the top 20% highest observed-label loss subset, while easy clean samples are defined as the bottom 30% lowest observed-label loss subset. Rank reversal is computed within each group by comparing pairwise orderings induced by τ_i^{last} and τ_i^{full}.

Table C.1 reports the residual ratio $\epsilon_i/\tau_i^{\text{last}}$ and the rank reversal rate. The residual ratio is lowest for easy clean samples, indicating that their full-model trace is well captured by the classifier-layer contribution. Hard clean and noisy samples exhibit moderately larger residual ratios, which is expected because ambiguous or corrupted samples tend to induce stronger representation-level updates. However, their residual ratios remain below one, suggesting that the omitted representation-gradient component does not dominate the full trace. More importantly, the rank reversal rates remain low across all groups: 4.6% for easy clean samples, 8.9% for hard clean samples, and 8.1% for noisy samples. This indicates that although the last-layer trace is not numerically identical to the full trace, it largely preserves the sample ordering induced by the full-model trace. Therefore, the approximation does not systematically fail on the critical hard-clean or noisy groups, supporting its use as a lightweight surrogate for reliability estimation.

D EMA Teacher: Reference Design and Hyperparameter Sensitivity

D.1 EMA Teacher as a Reference

We use an EMA teacher as the default reference model:

$$\theta'_t = m\theta'_{t-1} + (1-m)\theta_t. \tag{39}$$

Unrolling the recursion gives:

$$\theta'_t = (1-m)\sum_{s=1}^{t} m^{t-s}\theta_s + m^t\theta'_0. \tag{40}$$

Thus the teacher is an exponentially weighted temporal ensemble of historical students. This explains why it is a natural default: it provides the stabilizing effect of ensembling without training or evaluating multiple independent networks.

Variance reduction. To illustrate temporal smoothing, consider a scalar component of the student prediction at a fixed sample, decomposed as

$$z_t(x_i) = \bar{z}(x_i) + \xi_t, \tag{41}$$

where ξ_t denotes zero-mean stochastic fluctuation caused by mini-batch sampling and noisy labels. For independent, identically distributed fluctuations, applying EMA directly to this scalar prediction gives

$$\operatorname{Var}\left[(1-m)\sum_{k=0}^{\infty} m^k \xi_{t-k}\right] = \frac{1-m}{1+m}\operatorname{Var}[\xi_t]. \tag{42}$$

Table C.1: Residual analysis of last-layer trace approximation. $\epsilon_i/\tau_i^{\text{last}}$ measures the omitted representation-gradient residual relative to the classifier-layer trace. Rank reversal rate measures how often the ordering induced by τ_i^{last} disagrees with that induced by τ_i^{full}.

Group	$\epsilon_i/\tau_i^{\text{last}}$	Rank reversal rate	Conclusion
Easy clean	0.18	4.6%	reliable
Hard clean	0.31	8.9%	no systematic failure
Noisy	0.28	8.1%	no systematic failure

This identity describes prediction-space averaging under the stated scalar-noise assumptions. In RGC, however, EMA is applied to network parameters; for a nonlinear network, $p_{\text{EMA}(\theta_t)}(x_i)$ need not equal the EMA of $p_{\theta_t}(x_i)$. Thus, the variance factor is an explanatory approximation for temporal smoothing, not a strict variance guarantee for the actual parameter-EMA teacher. It motivates why a larger momentum can smooth transient fluctuations, while the behavior of the network predictions depends on the model and training trajectory.

Why short warm-up does not immediately poison the teacher. The teacher is not used during the first E_w warm-up epochs; it is activated only after the student has acquired a coarse representation and the loss distribution has become more structured. This matters because the purpose of warm-up is not to obtain a reliable classifier, but to avoid initializing the EMA reference from a completely random predictor. In the early phase of deep training, even under substantial label noise, networks tend to fit easier and more consistent patterns before fully memorizing corrupted labels [3, 25, 4]. Our warm-up is intentionally short relative to the onset of heavy memorization, so the teacher is seeded from a partially organized student rather than from a late-stage noise-fitting model.

The EMA update further reduces the risk of severe drift. With $m = 0.999$, the teacher changes much more slowly than the student, so transient noisy gradients from the end of warm-up cannot immediately rotate the reference into a corrupted direction. After warm-up, the teacher is also sharpened, the reliability score is smoothed, and the final objective gates teacher supervision through $(1 - w_i)$. Consequently, even if the initial teacher is imperfect under extreme noise, it is not treated as an oracle and mainly influences samples whose observed-label supervision has already become low-trust. This is why RGC should be interpreted as a reference-based reliability test rather than as a pure pseudo-labeling method.

Conditional reference-gradient fidelity. For cross-entropy with soft target q_i, the gradient with respect to logits is $p_\theta(x_i) - q_i$. Let $q_i^\star$ be the ideal clean target and assume the logit-to-parameter Jacobian satisfies $\|J_i\|_2 \le L_i$. Then:

$$\|\nabla_\theta \ell(p_\theta(x_i), q_i) - \nabla_\theta \ell(p_\theta(x_i), q_i^\star)\|_2 \le L_i \|q_i - q_i^\star\|_2. \tag{43}$$

The same bound holds for the observed one-hot label $e_{\tilde{y}_i}$. Therefore, the teacher target yields a smaller upper bound on gradient error whenever its distribution is closer to the clean target than the observed noisy label is:

$$\|q_i - q_i^\star\|_2 < \|e_{\tilde{y}_i} - q_i^\star\|_2. \tag{44}$$

This comparison concerns the upper bounds; under the Jacobian norm assumption alone, it does not establish an ordering of the actual full-model gradient errors.

Drift and failure modes. The EMA teacher can fail if it accumulates systematic wrong predictions. Let $u_t = \theta_t - \theta_{t-1}$ be the student update. The student-teacher lag satisfies:

$$\theta_t - \theta'_t = m(\theta_{t-1} - \theta'_{t-1}) + m u_t. \tag{45}$$

Unrolling Eq. 45 yields the deterministic lag bound

$$\|\theta_t - \theta'_t\|_2 \le m^t \|\theta_0 - \theta'_0\|_2 + \sum_{s=1}^{t} m^{t-s+1} \|u_s\|_2. \tag{46}$$

If $\|u_s\|_2 \le U$ for all $s \le t$, this further gives

$$\|\theta_t - \theta'_t\|_2 \le m^t \|\theta_0 - \theta'_0\|_2 + \frac{m(1-m^t)}{1-m} U. \tag{47}$$

Thus rapid parameter motion, poor early predictions, or severe class-dependent noise can increase teacher drift. In this regime the reference error $\|q_i - q_i^\star\|_2$ becomes large and the conditional fidelity statement above weakens.

Score smoothing and sigmoid stability. The score smoother in Eq. 18 is also an EMA. If the raw score can be written as $s_{i,t} = s_i^\star + \zeta_{i,t}$ with independent zero-mean noise of variance σ_i^2, then in the stationary scalar case

$$\mathrm{Var}[\tilde{s}_{i,t}] = \frac{1-\mu}{1+\mu}\sigma_i^2. \tag{48}$$

Thus larger μ reduces high-frequency score variance before the score affects the training objective. The mapping from normalized score to trust weight is locally stable because the sigmoid is $1/4$-Lipschitz:

$$|\sigma(a) - \sigma(b)| \le \frac{1}{4}|a-b|. \tag{49}$$

When the normalization statistics are fixed or vary slowly, a perturbation $\Delta\tilde{s}_{i,t}$ changes w_i by at most approximately $\alpha|\Delta\tilde{s}_{i,t}|/[4(\mathrm{std}(\tilde{s}_t)+\epsilon)]$. This is a stability remark for the weighting map, not a convergence statement.

Implementation safeguards. The training code uses several safeguards against teacher failure: the teacher is updated by stop-gradient EMA, the teacher distribution is sharpened with temperature T, the first few epochs are used as warm-up, and per-sample reliability scores are smoothed before they affect the loss. The final objective gates teacher supervision by $(1 - w_i)$, so teacher targets mainly affect samples whose observed-label updates are already deemed unreliable.

The teacher itself is not the contribution. The contribution is to use a reference target to evaluate whether the observed-label update is geometrically reliable. EMA is chosen because the reference ablation favors it among the implemented choices while keeping the method lightweight.

D.2 Teacher-Related Hyperparameter Sensitivity

We examine the effects of EMA momentum m, sharpening temperature T, and teacher-loss coefficient λ. The default settings are $m = 0.999$, $T = 0.7$, and $\lambda = 1.0$. We vary one parameter at a time while keeping all other parameters and training settings fixed. Table D.1 reports classification accuracy and hard-clean preservation rate (HCPR), both of which are higher-is-better metrics.

Across the tested ranges, accuracy varies from 92.42% to 93.50%, a maximum difference of 1.08 percentage points, while HCPR ranges from 0.78 to 0.84. The observed trends are consistent with the intended roles of these hyperparameters: a smaller EMA momentum can yield a less stable reference, overly aggressive sharpening can amplify incorrect teacher predictions, and a large λ may overemphasize teacher supervision when the reference is imperfect. Nevertheless, the maximum accuracy variation is only 1.08 percentage points, and all tested configurations remain substantially above the corresponding baseline.

D.3 Teacher Quality and Target-Routing Diagnostics

We examine RGC's dependence on teacher-reference quality from two perspectives: how teacher prediction quality and angular-conflict diagnostic ability co-vary as training data decrease, and whether teacher-supervision weight is primarily assigned to samples with correct teacher predictions under the standard setting. All clean labels are used solely for post-training diagnostics and are not accessed during RGC training, reliability estimation, or trust-weight computation.

Teacher Quality under Limited Data. We reduce the available training data under CIFAR-10 Sym.20% while keeping all other training settings unchanged. In addition to classification performance, we report teacher accuracy, expected calibration error (ECE), and Conflict AUROC. Teacher accuracy measures agreement between teacher predictions and clean labels. ECE measures the discrepancy between prediction confidence and empirical accuracy, with lower values indicating better calibration. Conflict AUROC measures how well angular-conflict scores distinguish noisy samples from the hard clean samples defined in this paper, treating noisy samples as the positive class and larger scores as indicating greater unreliability. Table D.2 summarizes the endpoint results.

As the training-data ratio decreases from 100% to 20%, teacher accuracy drops from 92.6% to 79.4%, ECE rises from 0.041 to 0.126, and Conflict AUROC falls from 0.846 to 0.682. Degradation in teacher prediction and calibration quality accompanies a decline in the discriminative ability of angular conflict, while RGC's gain over Trace decreases from $+2.0$ to -0.4 percentage points. These

Table D.1: Sensitivity to teacher-related hyperparameters. Accuracy is reported in %; bold parameter values indicate the default settings.

Parameter	Value	Accuracy ↑	HCPR ↑
EMA momentum m	0.9	92.42	0.78
	0.99	93.08	0.82
	0.999	93.50	0.84
Sharpening temperature T	0.25	92.71	0.79
	0.5	93.24	0.82
	0.7	93.50	0.84
	1.0	92.93	0.80
Teacher-loss coefficient λ	0.5	93.02	0.81
	1.0	93.50	0.84
	2.0	92.76	0.78

Table D.2: Teacher quality and angular-conflict diagnostic ability under limited training data. Accuracies are reported in %, and performance differences are in percentage points.

Training data ratio	Teacher accuracy ↑	ECE ↓	Conflict AUROC ↑	Trace accuracy ↑	RGC accuracy ↑	RGC − Trace
100%	92.6	0.041	0.846	91.5	93.5	+2.0
20%	79.4	0.126	0.682	84.7	84.3	−0.4

observations reflect co-varying trends and do not, by themselves, establish that declining teacher quality causes the loss of diagnostic ability. As reported in Table 4, RGC still outperforms Trace by 1.4 and 1.6 percentage points when retaining 80% and 50% of the training data, respectively. Under the extreme 20% setting, however, directional information provides no additional gain, illustrating the method's dependence on reference quality.

Teacher-Target Routing Audit. We audit teacher supervision under the standard CIFAR-10 Sym.20% setting. Let $\hat{y}_i^{\mathrm{EMA}} = \arg\max_k q_{i,k}$ and $a_i = \lambda(1 - w_i)$ be the coefficient of the teacher-target loss. We partition the audited samples according to whether the observed label and teacher prediction agree with the clean label. For a group $\mathcal{G}$ within the audited set $\mathcal{S}$, its share of total teacher-supervision weight is

$$\mathrm{Share}(\mathcal{G}) = \frac{\sum_{i\in\mathcal{G}} a_i}{\sum_{i\in\mathcal{S}} a_i}. \tag{50}$$

This quantity measures weight allocation, not sample frequency or the fraction of the loss or gradient contributed by a group. Clean labels are used only for this post-hoc audit.

Table D.3: Teacher-supervision weight allocation under CIFAR-10 Sym.20%. The two rows describe disjoint groups; the remaining groups are not shown.

Observed label	Teacher prediction	Share of total teacher weight
Incorrect	Correct	47.3%
Correct	Incorrect	9.3%

The teacher correctly predicts 92.6% of the training samples in this setting. As shown in Table D.3, the potentially corrective group receives 47.3% of total teacher-supervision weight, compared with 9.3% for the potentially harmful overwriting group, a ratio of approximately 5.1. Across both groups with correct teacher predictions, the combined share is 83.3%. This is a weight share, not the fraction of samples predicted correctly. Incorrect teacher targets still receive nonzero weight, and the audit does not establish the causal effect of teacher supervision on student predictions. In particular, the 5.1 ratio is not a ratio of realized correction benefits to error-propagation harm.

E Reliability Diagnostics and Qualitative Analysis

E.1 Same-Checkpoint Reliability Diagnostics

To directly compare the discriminative ability of reliability signals, we compute the diagnostic scores of JS disagreement, Conflict-only, and Full RGC under CIFAR-10 Sym.20%, using a fixed Student/EMA checkpoint and the same noisy and hard-clean evaluation samples. Noisy samples are

treated as the positive class, all scores are oriented so that larger values indicate greater unreliability, and we report AUROC and AUPRC.

JS disagreement measures the difference between the student and EMA-teacher predictive distributions. Conflict-only compares the direction of the observed-label update with that of the teacher-reference update, while Full RGC further combines radial information with directional conflict.

Conflict-only achieves higher AUROC and AUPRC than JS disagreement, indicating that label-conditioned directional information provides additional discriminative value in this setting. Full RGC further outperforms Conflict-only, consistent with radial and directional information playing complementary roles. These results apply to this fixed checkpoint and evaluation set and do not establish a general guarantee across all noise conditions or training stages.

Table E.1(a): Same-checkpoint raw-score diagnostics on CIFAR-10. Noisy samples are the positive class. Dashes indicate unreported results.

	Sym.20%		IDN-20%	
Reliability signal	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
JS disagreement	0.679	0.657	–	–
Teacher observed-label confidence	0.748	0.725	0.714	0.690
Teacher observed-label margin	0.779	0.756	0.747	0.723
Conflict-only	0.846	0.827	**0.818**	**0.794**
Full RGC	**0.872**	**0.854**	–	–

Label-Conditioned Teacher Baselines. To distinguish the contribution of angular conflict from label conditioning alone, Table E.1(a) additionally compares teacher observed-label confidence and teacher observed-label margin under CIFAR-10 Sym.20% and IDN-20%. Both criteria condition the EMA prediction on the observed label. All scores are oriented so that larger values indicate greater unreliability and are evaluated on the same samples within each setting. Conflict-only outperforms both teacher criteria in these tested settings; this standalone comparison is complemented by the incremental analysis below.

Incremental Diagnostic Information. Without retraining RGC, we extract teacher observed-label margin and angular conflict from the same fixed Student/EMA checkpoint. We compare three L_2-regularized logistic diagnostic probes using teacher margin only, Conflict only, and teacher margin plus Conflict. All probes use identical stratified five-fold out-of-fold splits, preprocessing, regularization, and class-weight settings. Feature standardization and probe coefficients are fitted only on each fold's training split; predictions are evaluated on its held-out split. Clean labels are used only to define the post-hoc diagnostic task, not to train RGC.

Table E.1(b): Incremental diagnostic information of angular conflict beyond teacher observed-label margin. Results use the same out-of-fold logistic-probe protocol.

	Sym.20%		IDN-20%	
Diagnostic input	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Teacher margin only	0.777	0.753	0.745	0.721
Conflict only	0.842	0.826	0.816	0.791
Teacher margin + Conflict	**0.861**	**0.842**	**0.834**	**0.811**
Gain over teacher margin	+0.084	+0.089	+0.089	+0.090

The single-feature probes in Table E.1(b) are re-estimated through the common out-of-fold pipeline rather than copied from the raw-score comparison. Adding Conflict to the teacher-margin probe improves AUROC/AUPRC from 0.777/0.753 to 0.861/0.842 under Sym.20% and from 0.745/0.721 to 0.834/0.811 under IDN-20%. Since the combined probe adds only Conflict to the same label-conditioned teacher feature, these results support incremental discriminative information within this controlled linear diagnostic setting, rather than a general guarantee across noise conditions or training stages.

E.2 Loss-Independent Controlled Visual Difficulty

The original diagnostic defines hard-clean samples as the top 20% of clean samples ranked by loss. This construction is useful for analyzing the high-loss region, but it couples sample selection with the

Table E.2: Diagnostics on the controlled visual-difficulty set under CIFAR-10 Sym.20%.

Reliability signal	AUROC ↑	AUPRC ↑
Loss	0.632	0.611
Confidence	0.646	0.624
Entropy	0.651	0.631
Trace-only	0.781	0.758
Conflict-only	0.812	0.789
Full RGC	**0.839**	**0.817**

Table E.3: Sample-level audit of hard-clean samples. Score columns are percentiles relative to the corresponding score distributions over known noisy samples. Types 1–4 denote the visual-difficulty categories defined in the text. The final three columns give the decisions of Loss, Trace, and RGC.

ID	Clean class	Type	Loss pct.	Trace pct.	Conflict pct.	RGC pct.	Loss	Trace	RGC
L1	cat	4	76	12	9	11	Reject	Retain	Retain
L2	bird	1	68	16	7	13	Reject	Retain	Retain
L3	dog	3	83	14	12	16	Reject	Retain	Retain
T1	ship	3	11	74	6	9	Retain	Reject	Retain
T2	cat	1	15	88	8	13	Retain	Reject	Retain
T3	bird	2	8	69	11	15	Retain	Reject	Retain
B1	dog	4	71	82	3	17	Reject	Reject	Retain
B2	cat	1	79	91	2	18	Reject	Reject	Retain
C1	airplane	2	7	11	8	9	Retain	Retain	Retain
C2	ship	1	13	16	10	12	Retain	Retain	Retain
F1	dog	4	73	87	89	84	Reject	Reject	Reject
F2	cat	3	12	17	76	58	Retain	Retain	Reject
F3	airplane	1	14	72	67	55	Retain	Reject	Reject
F4	horse	3	62	79	81	73	Reject	Reject	Reject

Loss criterion. To examine whether the diagnostic results depend on this construction, we conduct an additional controlled visual-difficulty experiment under CIFAR-10 Sym.20%.

Using the known label-corruption mask, we first identify clean samples whose observed labels agree with their clean labels. With a fixed random seed, we perform class-balanced random sampling to select 100 samples per class, yielding 1,000 clean samples. Each sample receives exactly one predetermined visual perturbation intended to preserve its class identity: Gaussian blur, local occlusion, rotation, or color/contrast perturbation. Within each class, the 100 samples are evenly divided across the four perturbations, with 25 samples per perturbation. Sample indices, perturbation categories, strengths, spatial parameters, and random seeds are fixed before inspecting any reliability score. Loss, Confidence, Entropy, EMA predictions, Trace, Conflict, and RGC scores are not used for selection.

We also draw 1,000 noisy samples whose observed labels differ from their clean labels, using class-balanced sampling with 100 samples per clean class. These noisy samples do not receive the controlled perturbations. The resulting evaluation set contains 1,000 controlled visually difficult clean samples and 1,000 original noisy samples. All signals use the same fixed Student/EMA checkpoint and deterministic evaluation preprocessing, with stochastic training augmentation disabled. Noisy samples are the positive class and controlled visually difficult clean samples are the negative class; larger scores consistently indicate greater unreliability.

Full RGC outperforms the evaluated forward-space signals, Trace-only, and Conflict-only on this set, indicating that its diagnostic advantage is maintained under the tested controlled visual-difficulty setting. Because difficult samples are selected independently of the current model's loss and all other evaluated scores, this experiment complements the original high-loss diagnostic.

This result is limited to the controlled visual perturbations studied here. It does not directly generalize to all naturally occurring difficult samples or establish that forward-space signals lack discriminative ability in every hard-clean region.

E.3 Qualitative Audit of Hard-Clean Preservation

Under CIFAR-10 Sym.20%, we inspect sample-retention behavior using the same fixed hard-clean subset and Student/EMA checkpoint as in the original diagnostic experiment.

Thresholds and Sampling. For each reliability signal, we independently determine the operating threshold corresponding to 80% noisy-sample recall. Since larger scores indicate greater unreliability, the threshold is the 20th percentile of the scores over the known noisy subset. Samples with scores no greater than the threshold are classified as reliable, and the remaining samples as unreliable. Clean labels and the corruption mask are used only for post-hoc diagnostics, not for training or weight computation.

Using these fixed thresholds, we construct three candidate groups:

1. Hard-clean samples classified as reliable by RGC but unreliable by Loss or Trace;
2. Hard-clean samples classified as reliable by Loss, Trace, and RGC;
3. Hard-clean samples classified as unreliable by RGC.

We sample 8, 2, and 4 cases from these groups, respectively, yielding 14 audited cases covering additional preservation, common preservation, and failures.

Before inspecting the cases, we define four primary types of visual difficulty: (1) occlusion, truncation, or a small foreground object; (2) atypical viewpoint or pose, or background interference; (3) blur, low contrast, or insufficient visual evidence; and (4) fine-grained class ambiguity or similarity to a neighboring class. Each case is assigned one primary difficulty type.

Additional Preservation Cases. Loss rejects L1–L3, whereas their Trace, Conflict, and full RGC scores all fall in the reliable region, illustrating that high loss does not necessarily indicate an incorrect label. T1–T3 have high Trace percentiles but low Conflict percentiles, illustrating large updates whose directions remain consistent with the reference. Both Loss and Trace reject B1–B2, but their directional conflict is low and RGC retains them. These cases are consistent with the motivation for jointly using magnitude and directional information.

Common Preservation and Failure Cases. All three methods retain C1–C2. F1–F4 are hard-clean samples rejected by RGC. In particular, Loss and Trace retain F2, but its relatively high directional conflict leads RGC to classify it as unreliable, showing that the directional reference can also introduce additional false rejections. The qualitative audit therefore illustrates both the potential benefits and failure boundaries of RGC; these cases do not replace aggregate quantitative evaluation.

F Dataset Sources and Terms of Use

Table F.1 summarizes the official sources and stated terms of use for the datasets used in our experiments.

The absence of a named standard license on an inspected project page does not imply unrestricted reuse or redistribution. Access to each dataset remains subject to the original provider's applicable terms.

Table F.1: Dataset sources and terms of use.

Dataset	Official source	License or terms of use
CIFAR-10 / CIFAR-100	CIFAR project page	The project page requests citation of Krizhevsky's technical report (2009) [18]. No named standard dataset license is specified on the inspected page.
Animal-10N	Animal-10N project page	The project page provides access information and requests citation of SELFIE (ICML 2019) [33]. No named standard image license is specified on the inspected page.
Food-101N	Food-101N project page	Restricted to non-commercial research use. The provider requests citation of CleanNet (CVPR 2018) [21] and states that access does not grant intellectual-property rights to the underlying images.
WebVision	WebVision download page	The provider requests citation of the original dataset paper [24] and prohibits redistribution of the images. These conditions also apply to the WebVision-50 subset used here.
ILSVRC 2012 / ImageNet	ImageNet access page	Access is provided for non-commercial research and educational use, subject to the provider's terms. ImageNet does not own the copyright of the underlying images. We use the corresponding 50-class validation subset [8, 32].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the method, its reliability criterion, and the benchmark scope. The main text and limitations section qualify the dependence on reference quality.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses low-data, model-capacity, and extreme-noise boundary cases in Sec. 4.4 and summarizes reference-quality limitations in the Limitations section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The theoretical content consists of gradient identities, sufficient propositions, and approximation bounds. Assumptions and derivations are stated in Sec. 3, Appendix C, and Appendix D.1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper provides pseudocode for RGC training, a unified evaluation protocol, implementation details, hardware/software information, and a public code repository.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The code repository is available at `https://github.com/maojingyangmao-cloud/RGC`. It provides the released code and will host the accompanying instructions for reproducing the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The core paper summarizes the protocol; the appendix and supplement list data splits, all hyperparameters (including RGC, EMA, and baselines) with selection criteria, optimizer and learning-rate schedule, batch size, epochs, augmentation, hardware context, and random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: The current submission reports single-run results for the main tables and does not report error bars or significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix B reports the hardware and software environment, including that each individual run uses one NVIDIA L20 GPU with 46 GB memory; Sec. 5 reports representative runtime and memory measurements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses standard image-classification benchmarks and algorithmic training procedures, and we are not aware of deviations from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A].

Justification: The paper studies robust learning under label noise and does not introduce a deployed system or application-specific societal intervention. Potential impacts are limited to improving reliability of supervised learning pipelines.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release high-risk models, generative models, or newly scraped datasets requiring misuse safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The appendix lists each benchmark and third-party asset with citation, version or split where applicable, and license or terms of use (including links where standard). The released code repository includes a LICENSE file consistent with redistribution of our implementation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The public code repository documents installation, configuration, training, evaluation, and licensing in the repository README and supplemental material. The scientific focus remains the RGC method; the codebase is the accompanying reproducibility artifact.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The work does not involve crowdsourcing experiments or human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The work does not involve human-subject research, so IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core method does not use LLMs as an original or non-standard component. Any writing or formatting assistance, if used, does not affect the method or experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.