FROZEN FOUNDATION MODELS AND NOISY LABELS: ACCURACY HEADROOM AND CALIBRATION GAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation-model encoders change the starting point for learning from noisy downstream labels: a visual representation is available before the corrupted supervision is observed. We ask how much accuracy headroom remains for noise-specific training when the encoder is immutable, and whether accurate readouts also yield reliable probabilities. Linear probes serve as matched diagnostics over the same frozen representation, not as substitutes for representation learning. Across the evaluated benchmarks, cross-entropy probing often approaches a matched clean-label readout and specialized noisy-label learning rules under moderate noise, while severe or structured corruption leaves meaningful accuracy headroom. Absolute accuracy also varies across encoders, so this finding is conditional on the tested representations. Yet similar top-1 accuracy can coexist with markedly different calibration errors. A population analysis separates changes in Bayes decisions from changes in posterior probabilities. We then evaluate Anchor-based Efficient Calibration (AEC), a post-hoc temperature-matching intervention that uses training logits rather than clean calibration labels. AEC often improves ECE and representative proper scores without changing predictions, but its fixed target can also worsen calibration. These results position decision quality and probability reliability as distinct objectives in noisy-label learning with frozen foundation models.

1 INTRODUCTION

Foundation models such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) supply transferable visual representations before noisy labels for a downstream task are observed. This changes the starting point of *Learning with Noisy Labels (LNL)*. Robust losses (Zhang & Sabuncu, 2018; Wang et al., 2019), sample selection (Han et al., 2018; Li et al., 2020), label correction (Song et al., 2019; Patrini et al., 2017), and semi-supervised training (Li et al., 2020; Liu et al., 2020; Li et al., 2022) were often studied while visual features and decision rules were learned together. Once a representation is supplied, their marginal value for downstream classifier learning becomes a separate question: what remains to be learned, and what remains unreliable?

We study an immutable-encoder setting in which features are available but the encoder's parameters and inference path cannot be adapted. This covers restricted model access and workflows that reuse extracted representations; it does not presume that fine-tuning, LoRA, or prompt tuning are impossible. A linear probe is a controlled readout in this setting, not a replacement for deep representation learning. Holding the encoder and readout family fixed lets us compare cross-entropy with noise-aware training rules, while a matched clean-label probe provides a within-family accuracy reference. Our question is twofold: **how much decision improvement remains for noise-aware training, and do accurate readouts also yield reliable probabilities?**

The answer depends on the representation and corruption. With strong frozen representations and moderate noise, ordinary cross-entropy can approach both specialized LNL readouts and the matched clean reference. Severe or structured corruption still leaves meaningful gains for robust training. Absolute accuracy also varies across the evaluated encoders, so a small noisy-versus-clean gap does not imply that every representation is adequate. The linear readout makes these differences measurable without conflating a change in the training rule with a change in the encoder.

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

Even when accuracy headroom is small, probability quality can remain poor. Classification accuracy tests the selected class; calibration tests whether confidence agrees with the frequency of correct predictions, a distinction that matters when confidence controls acceptance or further review. On CIFAR-10 with 50% symmetric noise, CE probing trails its matched clean reference by only 0.23 accuracy points, yet raw ECE is 51.59 rather than 8.34. Our population analysis separates changes in Bayes decisions from changes in posterior probabilities. A frozen encoder does not alter these input-level posteriors; it can make useful decisions accessible to a simple readout without guaranteeing reliable probabilities.

This residual gap motivates a calibration signal that does not require clean validation labels. Standard temperature scaling uses labeled validation data (Guo et al., 2017). We therefore examine **Anchor-based Efficient Calibration (AEC)** as a minimal post-hoc diagnostic: it selects high-confidence examples within each predicted class and fits positive temperatures to a fixed confidence target using training logits. The transformation preserves predicted classes, separating probability changes from accuracy changes. AEC can improve calibration without encoder retraining, but unreliable anchors or a mismatched target can also worsen it.

The study makes three contributions:

- **A foundation-model regime assessment.** With representations held fixed, we measure the accuracy headroom available to noise-aware readout training relative to a matched clean-label reference, identifying both limited-headroom and hard-noise regimes.
- **An accuracy–probability distinction.** We separate the population effects of label corruption on Bayes decisions and posterior probabilities, clarifying why strong representations and competitive accuracy do not by themselves establish reliable confidence.
- **A post-hoc calibration diagnostic.** We evaluate how anchor-based temperature matching changes ECE and proper scoring rules without changing decisions, including cases where its fixed target is unsuitable.

2 Related Work and Scope

LNL methods use robust objectives, sample selection, label correction, and semi-supervised training to limit corrupted supervision (Zhang & Sabuncu, 2018; Han et al., 2018; Li et al., 2020; Song et al., 2019). They are usually evaluated end to end, so noise affects both representation and classifier learning. We instead hold the representation fixed and compare representative principles only in the linear-readout stage; this measures the marginal value of noise-aware *classifier learning*, not the overall ranking of full LNL systems. Closest to our setting, Di Salvo et al. (2025) compare robust linear losses and neighborhood inference over fixed foundation-model embeddings, while Xu et al. (2025) analyze how compact self-supervised clusters and the induced noise structure favor noisy-label learning. Our contribution is complementary: we use a matched clean readout to quantify the remaining accuracy headroom and then examine the distinct calibration gap.

Pretrained models have also been used to detect or clean noisy labels and adapted through fine-tuning or prompts (Feng et al., 2024; Wei et al., 2024; Chen et al., 2025; Pan et al., 2025). Those methods change the supervision or representation, whereas our immutable-encoder regime isolates what remains after features are supplied. Consequently, our results neither compare against representation adaptation nor apply automatically to training from scratch. For calibration under noisy labels, Penso et al. (2024) estimate noise and model accuracy before calibration; AEC assumes neither is available and instead probes fixed logits with a heuristic target. Expanded positioning appears in Appendix C.

3 Noisy-Label Learning with Frozen Foundation Models

Noisy supervision. Let $(X, Y) \sim \mathcal{D}$ denote the clean distribution over inputs and K classes, while the learner observes $S_{\text{noisy}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ from a corrupted distribution $\widetilde{\mathcal{D}}$. With clean and noisy posteriors $\eta_i(x) = P(Y = i \mid X = x)$ and $\tilde{\eta}_j(x) = P(\widetilde{Y} = j \mid X = x)$, a general instance-

dependent corruption process satisfies

$$T_{ij}(x) = P(\widetilde{Y} = j \mid Y = i, X = x), \tag{1}$$

$$\tilde{\eta}_j(x) = \sum_{i=1}^{K} T_{ij}(x)\eta_i(x). \tag{2}$$

and includes symmetric, asymmetric, class-dependent, and instance-dependent noise. The learner receives no clean training or validation subset; clean test labels are used only for evaluation.

Frozen-access model and permitted interventions. We query an immutable pretrained foundation-model encoder $\phi : \mathcal{X} \to \mathbb{R}^d$ for features $z = \phi(x)$ and train only a linear readout $g_W(z) = Wz + b$. Feature vectors are observable, but no gradient or trainable component enters the encoder; deterministic features may therefore be cached, while DivideMix queries the same frozen encoder for its prescribed augmented views. Methods may change the readout loss, weighting, sample selection, optimization, or post-hoc logits, but not ϕ or the learned readout during post-processing. Full fine-tuning, LoRA, adapters, and prompt tuning are consequently outside scope. This access boundary does not claim that adaptation is infeasible; it isolates noisy-label learning after a representation has been supplied.

Matched reference. A clean-label probe with the same encoder, readout family, and optimization protocol provides a within-family reference for the effect of noisy supervision. It is an evaluation reference, not an available training procedure or a mathematical upper bound.

4 How Label Noise Separates Decisions from Probabilities

We first establish the population distinction that motivates the empirical study. Label noise may change the Bayes-optimal decision, and it may change posterior probabilities even where that decision is preserved. Foundation-model representations and low-capacity readouts concern which decisions are accessible to downstream learning; they do not change the data-generating posteriors or eliminate probability distortion. The analysis therefore motivates examining calibration separately from accuracy, but does not imply that every frozen probe is robust or that a particular post-hoc calibrator must succeed.

4.1 Accuracy Preservation under Noisy Labels

We first separate corruption of the population decision target from error introduced when that target is estimated by a finite-sample learner. Let $h^*(x) = \arg\max_c \eta_c(x)$ and $\tilde{h}^*(x) = \arg\max_c \tilde{\eta}_c(x)$ denote the clean and noisy Bayes classifiers, respectively. These induce the partition

$$\mathcal{X}_{\mathrm{BC}} = \{x \in \mathcal{X} : h^*(x) = \tilde{h}^*(x)\}, \quad \mathcal{X}_{\mathrm{BI}} = \mathcal{X} \setminus \mathcal{X}_{\mathrm{BC}}, \tag{3}$$

where $\mathcal{X}_{\mathrm{BC}}$ and $\mathcal{X}_{\mathrm{BI}}$ are the Bayes-consistent and Bayes-inconsistent regions. For any deterministic classifier $h : \mathcal{X} \to \mathcal{Y}$, define $\mathrm{Acc}_{\mathrm{clean}}(h) = \mathbb{E}_X[\eta_{h(X)}(X)]$, the usual expected clean accuracy. This definition allows stochastic clean labels; under Assumption 1, it reduces to $P(h(X) = Y)$. Its clean Bayes regret admits the exact decomposition

$$\mathrm{Acc}_{\mathrm{clean}}(h^*) - \mathrm{Acc}_{\mathrm{clean}}(h) = \underbrace{\Delta_{\mathrm{BI}}}_{\text{Bayes decision shift}} + \underbrace{\Delta_{\mathrm{L}}(h)}_{\text{learner gap}}, \tag{4}$$

where $\Delta_{\mathrm{BI}} = \mathrm{Acc}_{\mathrm{clean}}(h^*) - \mathrm{Acc}_{\mathrm{clean}}(\tilde{h}^*)$ and $\Delta_{\mathrm{L}}(h) = \mathrm{Acc}_{\mathrm{clean}}(\tilde{h}^*) - \mathrm{Acc}_{\mathrm{clean}}(h)$. The first term is nonnegative and intrinsic to the corruption process. The second is deliberately called a *gap*, not a penalty: it can be negative because a finite-sample learner optimized on noisy labels may, by chance or inductive bias, achieve higher clean accuracy than $\tilde{h}^*$. This distinction also cautions against treating a matched clean probe as a deterministic upper bound.

Bayes decision shift (Δ_{BI}). The intrinsic accuracy cost of pursuing the noisy rather than the clean Bayes decision is supported only on the Bayes-inconsistent region:

$$\begin{aligned}\Delta_{\mathrm{BI}} &\triangleq \mathrm{Acc}_{\mathrm{clean}}(h^*) - \mathrm{Acc}_{\mathrm{clean}}(\tilde{h}^*) \\ &= \mathbb{E}_X\Big[\mathbf{1}\{X \in \mathcal{X}_{\mathrm{BI}}\} \\ &\qquad \cdot \Big(\eta_{h^*(X)}(X) - \eta_{\tilde{h}^*(X)}(X)\Big)\Big].\end{aligned} \tag{5}$$

If $\varepsilon_{\mathrm{BI}} = P_X(X \in \mathcal{X}_{\mathrm{BI}})$, then $0 \le \Delta_{\mathrm{BI}} \le \varepsilon_{\mathrm{BI}}$. The following familiar sufficient conditions make this term vanish.

Assumption 1 (Deterministic Clean Posterior)**.** *For every x in the data support, the clean posterior is deterministic: for $k = h^*(x)$, $\eta_k(x) = 1$ and $\eta_i(x) = 0$ for all $i \neq k$ (Liu & Tao, 2015; Cheng et al., 2020).*

Assumption 2 (Diagonal Dominance / Non-overwhelming Noise)**.** *For every x in the support, the label corruption process is diagonally dominant for the clean Bayes class $k = h^*(x)$: $T_{kk}(x) > \max_{j \neq k} T_{kj}(x)$ (Zhang et al., 2021; Tanno et al., 2019).*

Corollary 1 (Global Bayes Consistency)**.** *Under Assumptions 1 and 2, the noisy and clean Bayes classifiers agree everywhere on the data support. Equivalently, $\varepsilon_{\mathrm{BI}} = 0$ and $\Delta_{\mathrm{BI}} = 0$.*

These assumptions are sufficient, not necessary. More importantly, they concern population posteriors and say nothing about whether a finite-sample probe reaches $\tilde{h}^*$. The frozen encoder ϕ does not alter η, $\tilde{\eta}$, or Δ_{BI}; it only changes which decision functions are accessible to the readout. We therefore do not derive accuracy saturation from a generic surrogate-risk bound. Instead, the matched-probe experiments assess a finite-sample accuracy difference under controlled features and training; they do not identify the population learner gap. The decomposition is an interpretive separation: severe or structured noise can enlarge $\mathcal{X}_{\mathrm{BI}}$, while an inadequately aligned representation can keep the learner far from either Bayes rule.

4.2 Posterior Distortion and Population Calibration

Decision agreement requires only equality of two argmax operations; probability agreement requires substantially more. Consider the idealized predictor that matches the noisy posterior, predicts $k(x) = \tilde{h}^*(x)$, and reports top-label confidence $C(x) = \tilde{\eta}_{k(x)}(x)$. Define its pointwise clean–noisy confidence discrepancy by

$$d_{\mathrm{post}}(x) \triangleq |\eta_{k(x)}(x) - \tilde{\eta}_{k(x)}(x)|, \qquad \text{where } k(x) = \tilde{h}^*(x). \tag{6}$$

To avoid conflating a population quantity with the binned estimator used in experiments, define the population top-label calibration error

$$\mathrm{CE}_{\mathrm{pop}} = \mathbb{E}_C[|P(Y = k(X) \mid C) - C|]\,. \tag{7}$$

The 15-bin ECE reported in our experiments is a finite-sample, bin-dependent estimate of calibration behavior and is not identical to $\mathrm{CE}_{\mathrm{pop}}$.

Proposition 1 (Posterior-Distortion Decomposition)**.** *For the idealized noisy-posterior predictor above, let $k = k(x)$,*

$$\begin{aligned}\mathrm{CE}_{\mathrm{pop}} &\le \mathbb{E}_X\left[d_{\mathrm{post}}(X)\right] \\ &= \mathbb{E}_X\left[\Big|\underbrace{(1 - T_{kk}(x))\eta_k(x)}_{\textit{Leakage (Out-flow)}} - \underbrace{\sum_{j \neq k} T_{jk}(x)\eta_j(x)}_{\textit{Injection (In-flow)}}\Big|\right],\end{aligned} \tag{8}$$

The two terms describe probability mass leaving the predicted clean class and entering it from competing classes. They can be nonzero even at $x \in \mathcal{X}_{\mathrm{BC}}$, because preserving an argmax does not require preserving its magnitude. In the deterministic case with $k = h^*(x)$, the injection term vanishes and $d_{\mathrm{post}}(x) = 1 - T_{kk}(x)$. Conversely, leakage and injection may cancel, and aggregation

over equal-confidence examples may further reduce $\mathrm{CE}_{\mathrm{pop}}$. The proposition therefore supplies an upper bound and a posterior-distortion diagnostic, not a positive calibration floor.

The analysis establishes a separation, not a universal failure theorem: a small observed clean-accuracy gap can coexist with nonzero posterior distortion. Whether that distortion produces large empirical ECE, and whether a post-hoc transformation improves it, remain empirical questions. Accordingly, the analysis motivates the calibration study but does not prove the effectiveness of AEC.

5 Anchor-Based Temperature Matching

The population analysis motivates examining probability quality independently of decisions. Standard temperature scaling (TS) fits a positive scalar on labeled validation data and preserves the predicted class (Guo et al., 2017). ATS augments class-conditional validation subsets (Mozafari et al., 2019); Penso et al. (2024) estimate label noise and model accuracy before calibration; and TransTS adapts scaling to an estimated noise level (Wu et al., 2026). These approaches use labeled-validation, training-time, or noise-estimation machinery. Our narrower question is how much a fixed predictor's probabilities can be moved using cached logits alone.

We use **Anchor-based Efficient Calibration (AEC)** to test a minimal fitting signal for this intervention. AEC is a recalibration procedure; we use it diagnostically because its prediction-preserving transformation isolates changes to probability quality. AEC consists of only two operations: class-balanced anchor mining and positive temperature rescaling. It consumes cached logits, uses no clean validation labels, and leaves every top-1 prediction unchanged.

The simple anchor rule is a design choice, not a restriction of the framework. More elaborate LNL selectors—including small-loss ranking, neighbourhood agreement, temporal consistency, or co-training—can be substituted without changing the target-matching stage. Confidence ranking is especially close to the small-loss principle: for an example whose supplied label agrees with the predicted class, cross-entropy is the negative logarithm of the corresponding predicted probability and is therefore a monotone re-expression of confidence. The two rankings need not coincide when the supplied and predicted labels disagree, which is why AEC uses predicted-class confidence and avoids depending directly on noisy annotations. Appendix B.4 evaluates an SSR-inspired nearest-neighbour consistency selector as a second anchor construction.

5.1 Confidence-Based Anchor Construction

Let $s(x) = W\phi(x) + b \in \mathbb{R}^K$ be the logits of a trained noisy-label probe, $p(x) = \sigma(s(x))$, $\hat{y}(x) = \arg\max_j s_j(x)$, and $c(x) = \max_j p_j(x)$. AEC constructs anchors separately within each predicted class to avoid allowing frequent or systematically easier classes to dominate the calibration signal. For class c, define

$$\mathcal{D}_c = \{x_i : \hat{y}(x_i) = c\}, \qquad \mathcal{A}_\kappa^{(c)} = \mathrm{Top}_\kappa(\mathcal{D}_c; c(x)), \tag{9}$$

where Top_κ retains the $m_c = \lceil \kappa |\mathcal{D}_c| \rceil$ largest-confidence samples for each nonempty class (at least one and at most $|\mathcal{D}_c|$). The noisy annotations $\tilde{y}_i$ are not used to rank these samples. If $\mathcal{D}_c$ is empty, the implementation uses the identity fallback $\tau_c = 1$.

This construction is motivated by a heuristic, not an identification assumption: high-confidence samples are plausible candidates for low-ambiguity examples, but confidence ranking alone does not guarantee correctness, clean annotation, or a near-deterministic clean posterior. The selected set may be unreliable under systematic model error, domain shift, class imbalance, or severe instance-dependent noise. We therefore refer to *anchor reliability* as a condition to be examined empirically rather than assuming anchor purity by construction. The SSR-inspired alternative and its clean-label diagnostic purity are reported in Appendix B.4; neither selector requires clean labels when constructing anchors.

5.2 Anchor-Guided Target Matching

Standard supervised temperature scaling chooses a temperature by minimizing a proper loss, typically NLL, on a clean labeled validation set (Guo et al., 2017). AEC has no such labels. It instead

treats a target anchor confidence ξ as a diagnostic prior and chooses one positive temperature per predicted class:

$$\tau_c^* = \arg\min_{\tau>0} \left(\frac{1}{|\mathcal{A}_\kappa^{(c)}|} \sum_{x \in \mathcal{A}_\kappa^{(c)}} \max_j \sigma_j\left(\frac{s(x)}{\tau}\right) - \xi \right)^2 . \tag{10}$$

We use $\kappa = 0.1$ and $\xi = 0.995$ as fixed defaults. Optimization is performed over $\log \tau_c$, initialized at zero ($\tau_c = 1$), using at most 100 L-BFGS iterations and no explicit temperature clipping. The method is therefore *anchor-guided target-matching temperature scaling*, not conventional supervised TS.

This distinction also prevents a pseudo-label degeneracy. If the model's own argmax were treated as a hard calibration label and NLL were minimized on it, the objective would generally favor $\tau \to 0^+$. The target-matching objective in equation 10 instead stops sharpening when the mean anchor confidence reaches the finite target ξ. The target remains heuristic: it need not equal anchor accuracy and is not estimated as a clean-posterior parameter.

5.3 Inference

For a new input, AEC first computes the original predicted class $\hat{y}(x)$ and then applies

$$s^{\mathrm{AEC}}(x) = \frac{s(x)}{\tau^*_{\hat{y}(x)}}. \tag{11}$$

Because every $\tau_c^* > 0$ is a scalar applied to all logits of a sample, $\arg\max_j s_j^{\mathrm{AEC}}(x) = \arg\max_j s_j(x)$. AEC can therefore change confidence-based metrics but cannot repair an incorrect class prediction, alter ranking-based accuracy, or compensate for an inadequate representation.

Interpretation and limitations. AEC does not estimate whether a predictor requires calibration, nor does it infer ξ from clean outcomes. Its effectiveness therefore remains dependent on confidence ordering and the suitability of the fixed target. Although AEC improves calibration in most evaluated settings, these methodological limitations remain because the current method provides no clean-label-free mechanism for selecting (κ, ξ). The main configuration uses confidence ranking because it adds no neighbourhood index or iterative selector; Appendix B.4 uses SSR-style nearest-neighbour agreement to test whether a richer selector yields purer anchors. We therefore use AEC to characterize a correctable component of the observed calibration gap, not as evidence of universal clean-posterior recovery.

6 Experiments

6.1 Experimental Protocol

We compare cross-entropy ERM (**CE Noisy-LP**) with GCE-LP, SSR-LP, and DivideMix-LP over the same frozen representation. GCE changes the loss, SSR applies selection and relabelling, and DivideMix retains its original two-model pipeline and published hyperparameters, with linear readouts as its only trainable models. Because no clean validation labels are permitted, we do not retune methods per dataset; the comparison therefore evaluates fixed published recipes under a shared access constraint rather than claiming each method's best achievable frozen-readout performance. A **Matched Clean-LP Reference** uses the same encoder and readout protocol with clean training labels; it measures a within-family noise gap, not a mathematical upper bound. We report top-1 accuracy, 15-bin ECE, NLL, and the multiclass Brier score $n^{-1} \sum_i \sum_{c=1}^{K} (p_{ic} - \mathbf{1}[y_i = c])^2$ on clean test labels. Unless stated otherwise, entries are means $\pm$ standard deviations over three independent seeds. Clean-validation TS is an unavailable oracle; noisy-validation TS uses noisy held-out labels; AEC uses training logits only.

We use DINOv2 ViT-B/14 by default and examine ImageNet-pretrained ViT and CLIP in the backbone comparison. Synthetic CIFAR noise, human-annotated CIFAR-N, Animal-10N, and Clothing1M cover the evaluated corruption settings. Dataset and optimization details appear in Appendix E.

Table 1: **Accuracy headroom under moderate and severe synthetic noise.** CIFAR-10 with frozen DINOv2 features; accuracy (%) is mean ± standard deviation over three seeds. Bold marks the best noisy-label readout. Complete settings appear in Appendix Table A.3.

Readout training rule	Sym 50%	Sym 80%	Asym 40%
CE Noisy-LP	97.51 ± 0.11	94.78 ± 0.34	92.69 ± 0.42
GCE-LP	97.57 ± 0.09	96.87 ± 0.15	96.21 ± 0.19
SSR-LP	97.27 ± 0.11	97.04 ± 0.13	97.14 ± 0.12
DivideMix-LP	97.17 ± 0.14	96.86 ± 0.16	96.81 ± 0.15
Matched Clean-LP Reference	97.74 ± 0.05	97.74 ± 0.05	97.74 ± 0.05

Table 2: **Top-1 accuracy under human-annotated, instance-dependent, and real-world noise.** This table presents the main non-synthetic benchmarks; more results under synthetic-noise settings are reported in Appendix Table A.3. All methods use the same frozen encoder and linear-readout family. A matched clean reference is reported only where corresponding clean training labels are available. Bold marks the best noisy-label readout in each column.

Readout Training Rule	**C-10N Aggregate**	**C-10N Worse**	**CIFAR-100N**	**Animal-10N**	**Clothing1M**
CE Noisy-LP	96.86 ± 0.12	93.80 ± 0.25	80.10 ± 0.35	91.24 ± 0.18	69.88 ± 0.55
GCE-LP (Zhang & Sabuncu, 2018)	97.24 ± 0.08	94.96 ± 0.18	80.72 ± 0.28	91.02 ± 0.20	70.20 ± 0.48
SSR-LP (Feng et al., 2022)	97.40 ± 0.06	97.24 ± 0.09	81.06 ± 0.22	90.50 ± 0.25	67.69 ± 0.62
DivideMix-LP (Li et al., 2020)	96.94 ± 0.15	97.03 ± 0.11	75.46 ± 0.45	86.22 ± 0.42	62.10 ± 0.75
Matched Clean-LP Reference	97.74 ± 0.05	97.74 ± 0.05	87.45 ± 0.15	—	—

6.2 Accuracy Headroom under Noisy Supervision

We first compare the raw predictors, before applying AEC. We use *limited accuracy headroom* to describe settings in which CE probing is close to its matched clean reference and the evaluated noise-aware readouts offer only small gains. This is a description of the observed comparisons, without a universal numerical threshold.

Controlled synthetic noise. Table 1 contrasts moderate and severe corruption. At 50% symmetric noise, CE Noisy-LP is only 0.23 accuracy points below the matched clean reference, and the best noise-aware readout improves on CE by 0.06 points. The complete results in Appendix Table A.3 show similarly limited headroom at 20% symmetric and 10% asymmetric noise.

The conclusion changes under severe corruption. At 80% symmetric noise, CE falls 2.96 points below the matched clean reference, while GCE-LP and SSR-LP recover more than two points relative to CE. Under 40% asymmetric noise, SSR-LP improves over CE by 4.45 points and approaches the clean reference. Thus, frozen representations reduce—but do not eliminate—the value of noise-aware readout learning. Accuracy saturation is an empirical regime associated with sufficiently informative features and non-overwhelming corruption, not a universal property of linear probing.

Human-annotated and real-world noise. Table 2 gives a similarly conditional picture. On CIFAR-10N Aggregate, all noisy-label readouts fall within 0.54 points of one another. On Animal-10N, simple CE is strongest; on Clothing1M, GCE improves over CE by only 0.32 points, whereas DivideMix is lower. Conversely, CIFAR-10N Worse benefits substantially from selection-based learning: SSR-LP improves over CE by 3.44 points. CIFAR-100N retains a large gap to the matched clean reference, and DivideMix is lower than the simpler readouts in this frozen-feature instantiation.

Dependence on the frozen representation. The matched reference is specific to the encoder and readout family. Repeating CE probing with ImageNet-pretrained ViT, CLIP, and DINOv2 features under 50% symmetric noise shows that absolute accuracy varies substantially with representation, while noisy and clean probes remain close within each matched pair (Table 3). In contrast, the same noisy-label predictors have raw ECE near 50 points across all three representations, and AEC lowers it below six points without changing accuracy. The table also reports the matched clean-probe ECE;

Table 3: **Accuracy and calibration across frozen representations** on CIFAR-10 with 50% symmetric noise. Clean-LP and Noisy-LP share the encoder and readout protocol. AEC preserves Noisy-LP accuracy. Values are percentages; variability is not displayed in this summary.

	Accuracy ↑		ECE ↓		
Backbone	Clean-LP	Noisy-LP	Clean-LP	Noisy-LP	+AEC
ViT (ImageNet)	93.07	93.16	10.05	50.43	5.58
CLIP	89.53	90.48	24.18	50.68	4.96
DINOv2	97.74	97.51	8.34	51.59	2.49

Rule	C10N Agg.	C10N Worse	C100N	Animal10N	Clothing1M
CE	+9.99	+31.02	+24.04	+8.24	+1.53
GCE	+2.78	+1.27	−1.34	−0.73	+2.56
SSR	+6.62	+4.39	+5.12	+1.75	+3.97
DivideMix	+6.29	+7.15	+11.38	+1.88	+4.95

Figure 1: **Signed ECE changes after AEC on human-annotated and web-collected noise.** Values are raw minus post-AEC ECE (percentage points): positive blue cells indicate improvement; negative orange cells indicate degradation. Rows are frozen-DINOv2 training rules; C10N/C100N abbreviate CIFAR-10N/CIFAR-100N. Top-1 predictions are unchanged. Full results with standard deviations appear in Appendix Table A.1.

further interpretation appears in Appendix B.3. A small matched accuracy gap therefore does not imply that the representation itself is adequate or that its probabilities are reliable, and finite-sample variation can even place the noisy probe slightly above the clean probe; this is why we do not call the latter an upper bound.

Finding and boundary. Taken together, the results support a qualified accuracy-saturation finding. When the frozen representation is strong for the downstream task and the corruption is moderate, CE Noisy-LP can already approach both specialized readout-level alternatives and its matched clean reference. Under severe asymmetric or human-annotation noise, or when the matched clean probe itself leaves substantial error, specialized LNL rules may still provide meaningful gains and accuracy remains unresolved. We next examine whether the remaining probability errors follow the same pattern.

6.3 Probability Reliability and Post-hoc Calibration

The empirical reassessment above showed that accuracy can saturate in several—but not all—frozen-representation regimes. We now hold each learned predictor fixed and examine its probability outputs. Because AEC applies a positive scalar to all logits of a sample, Raw and +AEC rows have identical top-1 predictions; accuracy is therefore not repeated below. ECE is reported in percentage points with 15 equal-width bins. We complement this bin-dependent metric with NLL and Brier score, and treat every degradation as part of the result rather than excluding it as a failed setting.

We report calibration for CE Noisy-LP and the same representative readout-level LNL variants used in the accuracy study. This comparison asks two separate questions: whether a training rule that improves accuracy also produces reliable probabilities, and whether target-matching rescaling changes the reliability of the resulting fixed predictor.

Controlled synthetic label noise. Appendix B.2 reports all synthetic-noise ECE results; the matched clean probe is a within-representation reference.

Table 4: **Temperature-scaling references on CIFAR-10N Aggregate.** Top-1 accuracy remains 96.86% in every row. Clean-Val TS is an unavailable oracle. ECE is in percentage points; run-to-run variability was not reported for this comparison.

Method	**Fitting Signal**	**ECE ↓**
Raw predictor	None	13.13
Noisy-Val TS	Noisy validation labels	8.14
AEC	Training-logit anchors	3.14
Clean-Val TS	Clean validation labels	0.41

At 50% symmetric noise, CE Noisy-LP trails its matched clean reference by only 0.23 accuracy points (Table 1), yet raw ECE is 51.59 versus 8.34. AEC lowers CE ECE to 2.49 without changing predictions. It also reduces ECE for every rule and synthetic corruption evaluated, including DivideMix-LP from 23.61 to 3.14 at 80% symmetric noise. Thus accuracy saturation can coexist with a large, correctable calibration gap.

Real-world and instance-dependent noise. Figure 1 shows the change in ECE on human-annotated and web-collected noise; Appendix Table A.1 provides the underlying values. AEC reduces ECE for CE Noisy-LP on all five datasets, with predictor-dependent gains, and improves every evaluated training rule on Clothing1M. The two exceptions overall are GCE-LP on CIFAR-100N and Animal-10N, reinforcing that a fixed target is not uniformly beneficial.

Comparison with Validation-Based Temperature Scaling Table 4 compares fitting signals on CIFAR-10N Aggregate. Noisy-Val TS improves on the raw CE probe but remains worse than AEC; an unavailable clean-validation oracle is better than both. This single-dataset result underscores the value of clean calibration labels without establishing a general method ranking.

Proper scoring rules. ECE is bin-dependent and only considers top-label confidence. Table 5 adds representative NLL and multiclass Brier results for CE Noisy-LP; the full set is in Appendix Table A.2. Both scores improve here, though not necessarily for every predictor.

Table 5: **Proper scoring rules for CE Noisy-LP.** The same predictor is evaluated before and after AEC. Entries are mean ± standard deviation over three seeds; lower is better.

Dataset / noise	State	NLL	Brier
CIFAR-10 / Sym 50%	Raw	0.82 ± 0.02	0.35 ± 0.01
	+AEC	0.12 ± 0.01	0.05 ± 0.01
Clothing1M	Raw	0.90 ± 0.02	0.43 ± 0.01
	+AEC	0.63 ± 0.02	0.35 ± 0.01

7 Discussion and Conclusion

With strong frozen foundation-model encoders, CE probing can approach matched clean-label and LNL readouts under moderate noise while remaining miscalibrated. Severe or structured noise leaves accuracy headroom, so a noise-aware rule should be judged by its incremental accuracy over a matched CE probe. Comparing decisions and probabilities within the same encoder and readout family isolates this contribution. Our fixed published recipes are not method optima; encoder adaptation and possible pretraining-data overlap limit generalization.

Although AEC preserves predictions, erroneous anchors or a mismatched target can worsen ECE. It offers no clean-label-free rule for whether to calibrate or how to select (κ, ξ). Better anchor discovery cannot alone determine an appropriate confidence target. Decision quality and probability reliability therefore merit separate reporting; future work should test calibration signals under encoder adaptation.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript restructuring, language editing, LaTeX formatting, literature discovery, and code scaffolding and debugging. The authors independently checked cited sources, verified the experimental implementation and reported measurements, and reviewed all AI-assisted changes. No generative AI system was used as part of the proposed method or to generate experimental measurements. The authors take responsibility for all content.

REPRODUCIBILITY STATEMENT

Experimental settings and implementation details are provided in Appendix E. We intend to release the experimental code upon acceptance.

ETHICS STATEMENT

Improved calibration should not be interpreted as evidence of fairness or safe deployment. AEC preserves predicted classes and therefore cannot correct misclassifications; consequential applications require separate evaluation of data provenance, subgroup performance, and downstream harms.

REFERENCES

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Guanyu Chen, Ruihao Li, Defu Liu, and Wei Yi. Robust-clip: End-to-end fine-tuning of vision-language models under noisy labels. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE, 2025.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

Francesco Di Salvo, Sebastian Doerrich, Ines Rieger, and Christian Ledig. An embedding is worth a thousand noisy labels. *Transactions on Machine Learning Research*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Ssr: An efficient and robust framework for learning with unknown label noise. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Clipcleaner: Cleaning noisy labels with clip. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 876–885, 2024.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2703–2708, 2021.

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

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. *arXiv preprint arXiv:1804.06872*, 2018.

Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 926–934, 2023a.

Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *CVPR*, 2023b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9676–9686, 2022.

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

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*, 2020.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pp. 21475–21496. PMLR, 2023.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. Attended temperature scaling: A practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bikang Pan, Qun Li, Xiaoying Tang, Wei Huang, Zhen Fang, Feng Liu, Jingya Wang, Jingyi Yu, and Ye Shi. Nlprompt: Noise-label prompt learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19963–19973, 2025.

Deep Patel and PS Sastry. Adaptive sample selection for robust learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3932–3942, 2023.

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*, pp. 1944–1952, 2017.

Coby Penso, Lior Frenkel, and Jacob Goldberger. Confidence calibration of a medical imaging classification system that is robust to label noise. *IEEE Transactions on Medical Imaging*, 43(6): 2050–2060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Mengmeng Sheng, Zeren Sun, Tao Chen, Shuchao Pang, Yucheng Wang, and Yazhou Yao. Foster adaptivity and balance in learning with noisy labels. In *European Conference on Computer Vision*, pp. 217–235. Springer, 2024.

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

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11244–11253, 2019.

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

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. *Advances in Neural Information Processing Systems*, 37:58154–58173, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Yuefei Wu, Bin Shi, Bo Dong, and Qinghua Zheng. Transts: An adaptive post-hoc method for probability calibration under label noise. *Science China Information Sciences*, 69:122105, 2026. doi: 10.1007/s11432-025-4576-9.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Pengcheng Xu, Li Yi, Gezheng Xu, Xi Chen, Ian McLeod, Charles Ling, and Boyu Wang. Uniform noise distribution and compact clusters: Unveiling the success of self-supervised learning in label noise. *Transactions on Machine Learning Research*, 2025.

Xichen Ye, Yifan Wu, Weizhong Zhang, Xiaoqiang Li, Yifan Chen, and Cheng Jin. Optimized gradient clipping for noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9463–9471, 2025.

Jingfeng Zhang, Bo Song, Haohan Wang, Bo Han, Tongliang Liu, Lei Liu, and Masashi Sugiyama. Badlabel: A robust perspective on evaluating and enhancing label-noise learning. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4398–4409, 2024.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *International Conference on Machine Learning*, pp. 12501–12512. PMLR, 2021.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

Rui Zhao, Bin Shi, Jianfei Ruan, Tianze Pan, and Bo Dong. Estimating noisy class posterior with part-level labels for noisy label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22809–22819, June 2024.

Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. *arXiv preprint arXiv:2103.13646*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Chen-Chen Zong, Ye-Wen Wang, Ming-Kun Xie, and Sheng-Jun Huang. Dirichlet-based prediction calibration for learning with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17254–17262, 2024.

Frozen Foundation Models and Noisy Labels: Accuracy Headroom and Calibration Gaps

— Appendix —

OVERVIEW

The appendix provides additional theoretical proofs, expanded related work, and comprehensive experimental details. It is organized as follows:

- **Additional results**: probabilistic metrics, backbone diagnostics, AEC sensitivity, and efficiency.
- **Extended related work**: additional positioning relative to LNL, foundation-model adaptation, and calibration.
- **Theory proofs**: Bayes consistency and the posterior-distortion decomposition.
- **Experimental details**: datasets, hyperparameters, metrics, and reproducibility information.

A ADDITIONAL AEC DETAILS

Algorithm 1 Class-Conditional AEC (Default Variant)

Require: Training inputs $\{x_i\}_{i=1}^N$, fixed logits $s(x)$
Require: Hyperparameters: anchor ratio κ (default: 0.1), target confidence ξ (default: 0.995)
for each class $c \in \{1, \ldots, K\}$ **do**
  $\mathcal{D}_c \leftarrow \{x_i : \arg\max_j s_j(x_i) = c\}$
  if $\mathcal{D}_c = \emptyset$ **then**
    $\tau_c^* \leftarrow 1$; **continue**
  end if
  $c_i \leftarrow \max_j \sigma_j(s(x_i))$ for $x_i \in \mathcal{D}_c$
  $m_c \leftarrow \min\{|\mathcal{D}_c|, \max\{1, \lceil \kappa |\mathcal{D}_c| \rceil\}\}$
  $\mathcal{A}_\kappa^{(c)} \leftarrow$ the m_c highest-confidence samples in $\mathcal{D}_c$
  Fit $\tau_c^* > 0$ with the target-matching objective
end for
For a query x, set $\hat{y}(x) \leftarrow \arg\max_j s_j(x)$
Output: $s^{\mathrm{AEC}}(x) = s(x)/\tau_{\hat{y}(x)}^*$

B ADDITIONAL PROBABILISTIC RESULTS AND DIAGNOSTICS

This appendix provides complete synthetic-noise results, proper-score references, backbone diagnostics, anchor-selection diagnostics, and efficiency details that complement the main text.

B.1 PROPER SCORES

B.2 COMPLETE RESULTS UNDER CONTROLLED SYNTHETIC NOISE

Tables A.3 and A.4 provide the complete accuracy and ECE results under the controlled symmetric- and asymmetric-noise settings summarized in the main text.

B.3 IMPACT OF PRE-TRAINED REPRESENTATIONS

The matched clean linear-probing reference depends on the frozen representation. We therefore report the complete accuracy/ECE comparison across ViT-ImageNet, CLIP, and DINOv2. Table 3

Table A.1: **Complete ECE results under human-annotated and real-world noise.** All methods use frozen DINOv2 features. +AEC preserves top-1 predictions. ECE is in percentage points; entries are mean ± standard deviation over three seeds.

Training Rule	State	C-10N Aggregate	C-10N Worse	CIFAR-100N	Animal-10N	Clothing1M
CE Noisy-LP	Raw	13.13 ± 0.34	34.85 ± 0.72	27.66 ± 0.58	13.36 ± 0.29	10.37 ± 0.14
	+AEC	3.14 ± 0.15	3.83 ± 0.18	3.62 ± 0.12	5.12 ± 0.16	8.84 ± 0.22
GCE-LP	Raw	5.54 ± 0.18	5.92 ± 0.21	3.19 ± 0.11	5.34 ± 0.19	12.24 ± 0.31
	+AEC	2.76 ± 0.16	4.65 ± 0.14	4.53 ± 0.15	6.07 ± 0.17	9.68 ± 0.25
SSR-LP	Raw	9.22 ± 0.25	7.15 ± 0.28	10.25 ± 0.32	7.23 ± 0.22	16.58 ± 0.42
	+AEC	2.60 ± 0.19	2.76 ± 0.20	5.13 ± 0.17	5.48 ± 0.15	12.61 ± 0.29
DivideMix-LP	Raw	9.35 ± 0.31	10.12 ± 0.38	17.88 ± 0.45	6.05 ± 0.24	19.27 ± 0.48
	+AEC	3.06 ± 0.14	2.97 ± 0.19	6.50 ± 0.21	4.17 ± 0.08	14.32 ± 0.52
Matched Clean-LP Reference	Raw	8.34 ± 0.22	8.34 ± 0.22	16.77 ± 0.35	—	—

Table A.2: **Representative proper scoring rules for CE Noisy-LP.** Raw and +AEC evaluate the same top-1 predictor. The Brier score is the unnormalized sum over classes defined in Section 6.1. Lower is better.

Dataset / Noise	State	NLL	Brier
CIFAR-10 / Sym 50%	Raw	0.82 ± 0.02	0.35 ± 0.01
	+AEC	0.12 ± 0.01	0.05 ± 0.01
CIFAR-10 / Sym 80%	Raw	1.69 ± 0.04	0.74 ± 0.02
	+AEC	0.24 ± 0.01	0.09 ± 0.01
CIFAR-10N Worse	Raw	0.60 ± 0.01	0.25 ± 0.01
	+AEC	0.20 ± 0.01	0.10 ± 0.01
CIFAR-100N	Raw	1.05 ± 0.02	0.38 ± 0.01
	+AEC	0.89 ± 0.02	0.30 ± 0.01
Animal-10N	Raw	0.37 ± 0.01	0.16 ± 0.01
	+AEC	0.29 ± 0.01	0.13 ± 0.02
Clothing1M	Raw	0.90 ± 0.02	0.43 ± 0.01
	+AEC	0.63 ± 0.02	0.35 ± 0.01

Table A.3: **Top-1 accuracy under synthetic noise on CIFAR-10.** All methods use the same frozen DINOv2 encoder and linear-classifier family; DivideMix retains its prescribed augmented views while keeping the encoder fixed. The matched clean reference uses the corresponding clean training labels and is not an upper bound. Bold marks the best noisy-label readout in each column.

Readout Training Rule	Sym 20%	Sym 50%	Sym 80%	Asym 10%	Asym 40%
CE Noisy-LP	97.69 ± 0.08	97.51 ± 0.11	94.78 ± 0.34	97.37 ± 0.09	92.69 ± 0.42
GCE-LP (Zhang & Sabuncu, 2018)	97.64 ± 0.07	97.57 ± 0.09	96.87 ± 0.15	97.58 ± 0.08	96.21 ± 0.19
SSR-LP (Feng et al., 2022)	97.44 ± 0.09	97.27 ± 0.11	97.04 ± 0.13	97.39 ± 0.09	97.14 ± 0.12
DivideMix-LP (Li et al., 2020)	97.07 ± 0.12	97.17 ± 0.14	96.86 ± 0.16	96.85 ± 0.13	96.81 ± 0.15
Matched Clean-LP Reference	97.74 ± 0.05	97.74 ± 0.05	97.74 ± 0.05	97.74 ± 0.05	97.74 ± 0.05

Table A.4: **ECE under synthetic noise on CIFAR-10.** All predictors use the frozen DINOv2 encoder. +AEC changes logits but not predicted classes.

Training Rule	State	Sym 20%	Sym 50%	Sym 80%	Asym 10%	Asym 40%
CE Noisy-LP	Raw	24.47 ± 0.48	51.59 ± 0.85	75.92 ± 1.12	14.96 ± 0.35	27.94 ± 0.54
	+AEC	2.31 ± 0.08	2.49 ± 0.11	5.22 ± 0.34	2.63 ± 0.09	3.72 ± 0.15
GCE-LP	Raw	5.85 ± 0.18	7.46 ± 0.24	43.62 ± 0.76	5.64 ± 0.16	13.67 ± 0.31
	+AEC	2.36 ± 0.07	2.43 ± 0.09	3.13 ± 0.15	2.42 ± 0.08	3.79 ± 0.19
SSR-LP	Raw	8.52 ± 0.22	7.52 ± 0.21	8.71 ± 0.26	9.12 ± 0.23	8.88 ± 0.25
	+AEC	2.56 ± 0.09	2.73 ± 0.11	2.96 ± 0.13	2.61 ± 0.09	2.86 ± 0.12
DivideMix-LP	Raw	9.88 ± 0.26	11.76 ± 0.32	23.61 ± 0.51	9.55 ± 0.24	12.73 ± 0.34
	+AEC	2.93 ± 0.12	2.83 ± 0.14	3.14 ± 0.16	3.15 ± 0.13	3.19 ± 0.15
Matched Clean-LP Reference	Raw	8.34 ± 0.22	8.34 ± 0.22	8.34 ± 0.22	8.34 ± 0.22	8.34 ± 0.22

separates representation-dependent absolute accuracy from post-hoc changes to a fixed noisy-label predictor.

The noisy-label probe should be interpreted relative to the matched clean probe for the same backbone rather than relative to training-from-scratch LNL pipelines. If a weaker encoder has large clean-probe error, AEC is not expected to repair it because AEC does not modify the decision rule.

Analysis. DINOv2 provides the highest absolute clean and noisy probing accuracy, while ViT-ImageNet and CLIP are weaker under the same linear protocol. Within each matched pair, noisy and clean accuracy remain close, including cases where finite-sample variation puts the noisy probe slightly above the clean probe. Raw ECE is high for all three noisy probes, and AEC lowers it below six points without changing accuracy. This experiment is suggestive rather than exhaustive: broader encoder, noise, and domain coverage remains necessary before attributing saturation causally to representation quality.

B.4 SSR-Inspired Neighbourhood-Consistent Anchors

AEC intentionally uses the least expensive selector: confidence ranking within each predicted class. To test whether the calibration stage can reuse a richer LNL signal, we additionally construct anchors using the non-parametric neighbourhood principle of SSR (Feng et al., 2022). Let $\bar{z}_i = \phi(x_i)/\|\phi(x_i)\|_2$ and let $\mathcal{N}_k(i)$ contain the k nearest training features under cosine similarity, excluding i. We define the neighbourhood label distribution and its majority label as

$$q_i(c) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathbf{1}\{\tilde{y}_j = c\}, \qquad y_i^{\mathrm{NPK}} = \arg\max_c q_i(c). \tag{12}$$

For predicted class c, the SSR-inspired candidate set contains examples satisfying $\hat{y}(x_i) = y_i^{\mathrm{NPK}} = c$; among them, the top κ fraction ranked by $q_i(c)$ are retained. This selector changes only anchor construction. The target-matching temperatures, target confidence ξ, and inference rule remain identical to AEC. We use the SSR neighbourhood size $k = 500$ and do not tune it using clean outcomes.

We consider below anchor "purity":

$$\mathrm{Purity}_{\mathrm{label}}(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbf{1}\{\tilde{y}_i = y_i\}. \tag{13}$$

Clean labels are used only to evaluate these diagnostics, never to construct either anchor set. Please refer to Table A.5 for results.

The richer selector is not uniformly purer: it slightly improves purity at 50% symmetric noise but is lower in the two harder settings. This supports the modularity of anchor construction while showing that additional selection complexity does not automatically provide a better target-matching signal. Designing selectors that improve both anchor reliability and downstream calibration remains future work.

Table A.5: **Purity of confidence-ranked and SSR-inspired anchors.** Both selectors use the same anchor ratio $\kappa = 0.1$ and the same frozen predictor. Values are reported as the mean over three seeds.

Dataset / Noise	Anchor Selector	Annotation Purity
CIFAR-10 / Sym 50%	Confidence-ranked AEC	0.963
	SSR-neighbour anchors	0.968
CIFAR-10 / Sym 80%	Confidence-ranked AEC	0.942
	SSR-neighbour anchors	0.937
CIFAR-10 / Asym 40%	Confidence-ranked AEC	0.956
	SSR-neighbour anchors	0.944

B.5 Computational Complexity and Efficiency

AEC is designed to be orthogonal to training-time LNL methods and lightweight enough to apply after any fixed predictor (Table A.6). It operates on cached logits: anchor construction requires a confidence quantile computation, and class-wise temperature fitting optimizes scalar parameters over the selected anchor sets.

Table A.6: **Illustrative wall-clock measurements on CIFAR-10.** Measurements were taken on an NVIDIA A100 GPU. The pipelines perform different workloads, so the end-to-end rows are contextual references rather than a controlled runtime ranking.

Method Framework	Pipeline Components	Approx. Time
Standard End-to-End LNL (e.g., DivideMix)	Full Network Training (e.g., 300 Epochs)	∼3–4 hours
Standard End-to-End LNL (e.g., SSR)	Multi-stage Sample Selection	∼2–3 hours
Feature-level LNL (DivideMix-LP)	Frozen Encoder + Full DivideMix Readout Training	∼8–10 mins
CE Noisy-LP + AEC	Linear Training (standard CE)	∼10 secs
	AEC Calibration (post-hoc optimization)	∼2 secs

Analysis. Once logits have been cached, AEC adds approximately two seconds to the reported CE probing pipeline, whereas the full two-readout, multi-view DivideMix-LP procedure takes approximately 8–10 minutes in the recorded setup. These figures support only the narrow claim that the post-hoc operation itself is lightweight. The end-to-end rows include different representation-learning workloads and should not be interpreted as a hardware-normalized efficiency comparison.

C Extended Related Work

Problem formulations and surveys. Label noise has been studied across statistical learning, deep representation learning, and weak supervision, with recent surveys emphasizing that conclusions depend strongly on the corruption mechanism (Song et al., 2022). Benchmarks likewise cover synthetic flips, human annotation noise, and web supervision (Wei et al., 2022; Xiao et al., 2015). These distinctions matter for our scope: a frozen representation can reduce the need to relearn visual structure, but it cannot make different corruption mechanisms statistically equivalent.

Risk correction, robust objectives, and transition modelling. One family modifies empirical risk through corrected or noise-robust objectives (Patrini et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019); another estimates a class- or instance-dependent transition mechanism (Xia et al., 2019). Recent theory clarifies that transition estimation from noisy observations requires identifiability conditions or additional structure (Liu et al., 2023). These methods target recovery of the clean decision rule or noise process; AEC instead leaves the learned decision rule fixed and makes a narrower claim. Recent alternatives also study posterior estimation with auxiliary part-level labels and optimized gradient clipping (Zhao et al., 2024; Ye et al., 2025).

Sample selection, correction, and collaborative training. Selection-based methods exploit memorization dynamics, disagreement, loss statistics, neighbourhood structure, or cross-model

agreement to identify likely clean examples (Han et al., 2018; Song et al., 2019). Recent methods add adaptive thresholds, uniform selection, and greater reuse of uncertain samples (Patel & Sastry, 2023; Karim et al., 2022; Wang et al., 2022). This literature provides a library of possible anchor selectors. AEC adopts a simpler confidence-ranked selector; the SSR-inspired experiment illustrates how a richer selector can be substituted without changing target matching. Recent studies further emphasize balanced adaptive training and robust evaluation across noise regimes (Sheng et al., 2024; Zhang et al., 2024).

Semi-supervised and representation-learning pipelines. Modern LNL systems frequently treat selected and unselected examples as labelled and unlabelled subsets, combining label refinement with consistency regularization or pseudo-labelling (Berthelot et al., 2019; Li et al., 2020; Liu et al., 2020). Contrastive and self-supervised objectives can further reduce reliance on corrupted class labels by learning structure before or alongside robust classification (Ghosh & Lan, 2021; Li et al., 2022; Huang et al., 2023b; Zheltonozhskii et al., 2021). Such pipelines remain valuable when representation learning is part of the task. Our experiments remove that degree of freedom and ask how much classifier-level benefit remains after a representation has already been supplied.

Pretrained representations and adaptation. Large-scale image and image–text pretraining provides transferable features under clean, weak, or naturally noisy supervision (Radford et al., 2021; Jia et al., 2021; Oquab et al., 2023). Downstream adaptation ranges from full fine-tuning to adapters and learned prompts (Wortsman et al., 2022; Gao et al., 2021; Zhou et al., 2022; Chen et al., 2022). In noisy-label settings, pretrained models have also been used for label detection, cleaning, noise-aware pretraining, and parameter-efficient adaptation (Feng et al., 2024; Wei et al., 2024; Huang et al., 2023a; Chen et al., 2025; Pan et al., 2025). Our frozen-access protocol holds the encoder and its inference path fixed, so it complements rather than ranks these directions.

Most directly related, Di Salvo et al. (2025) evaluate robust linear losses and neighbourhood inference on fixed foundation-model embeddings, and Xu et al. (2025) connect the success of self-supervised representations under label noise to compact feature clusters and induced noise structure. Our study asks a different but adjacent question: after fixing both the representation family and linear readout family, how much accuracy headroom remains relative to a matched clean probe, and does probability reliability saturate at the same time?

Calibration, proper scores, and unavailable clean labels. Post-hoc temperature scaling is typically fitted on clean validation labels (Guo et al., 2017), while recent extensions consider richer parameterizations or noisy-label-specific fitting (Zong et al., 2024; Wu et al., 2026). Calibration evaluation is sensitive to binning and finite-sample bias, motivating debiased estimators and complementary proper scoring rules (Roelofs et al., 2022; Gruber & Buettner, 2022); calibration under label noise is also being studied in applications such as medical imaging (Penso et al., 2024). AEC uses cached logits, no clean validation labels, and no retraining, at the cost of a heuristic target and no clean-label-free parameter-selection guarantee.

D Proofs of Theoretical Analysis

This appendix provides the mathematical derivations for the results in the main theoretical analysis.

D.1 Proof of Corollary 1 (Global Bayes Consistency)

Problem Setup. Let x be an input instance. Let $\eta(x) = [\eta_1(x), \ldots, \eta_K(x)]^\top$ be the clean posterior probability vector, where $\eta_i(x) = P(y = i|x)$. Let $\tilde{\eta}(x) = [\tilde{\eta}_1(x), \ldots, \tilde{\eta}_K(x)]^\top$ be the noisy posterior probability vector, where $\tilde{\eta}_j(x) = P(\tilde{y} = j|x)$. The relationship between the clean and noisy posteriors is governed by the Law of Total Probability via the instance-dependent transition matrix $T(x)$, where $T_{ij}(x) = P(\tilde{y} = j|y = i, x)$:

$$\tilde{\eta}_j(x) = \sum_{i=1}^{K} T_{ij}(x)\eta_i(x). \tag{14}$$

Goal. Let $k = h^*(x) = \arg\max_c \eta_c(x)$ be the clean Bayes-optimal label. We show that under Assumptions 1 and 2, the noisy Bayes classifier yields the same prediction, i.e.,

$$\tilde{\eta}_k(x) > \tilde{\eta}_j(x), \quad \forall j \neq k. \tag{15}$$

Proof. By Assumption 1, the clean posterior is deterministic in the idealized limit: $\eta_k(x) = 1$ and $\eta_i(x) = 0$ for every $i \neq k$. Therefore, for the clean Bayes class k,

$$\tilde{\eta}_k(x) = \sum_{i=1}^{K} T_{ik}(x)\eta_i(x) = T_{kk}(x), \tag{16}$$

and for any competing class $j \neq k$,

$$\tilde{\eta}_j(x) = \sum_{i=1}^{K} T_{ij}(x)\eta_i(x) = T_{kj}(x). \tag{17}$$

Assumption 2 gives $T_{kk}(x) > T_{kj}(x)$ for every $j \neq k$. Hence $\tilde{\eta}_k(x) > \tilde{\eta}_j(x)$ for all competing classes, and $\tilde{h}^*(x) = h^*(x)$. Since this holds for every x in the data support, $\mathcal{X}_{\mathrm{BI}}$ has probability mass zero and $\varepsilon_{\mathrm{BI}} = 0$. □

D.2 Proof of Proposition 1 (Posterior-Distortion Decomposition)

Problem Setup. Let (X, Y) denote the clean data and $(X, \tilde{Y})$ denote the noisy data. Let $\eta_c(x) = P(Y = c|x)$ be the clean posterior and $\tilde{\eta}_c(x) = P(\tilde{Y} = c|x)$ be the noisy posterior. We assume the classifier f has fitted the noisy distribution for its predicted class. Write $k(X)$ for the predicted class at input X, so its confidence output is $\hat{p}(X) = \tilde{\eta}_{k(X)}(X)$ and $\hat{y} = k(X)$.

Proof. **Step 1: Population calibration bound.** The population top-label calibration error defined in the main theoretical analysis is

$$\mathrm{CE}_{\mathrm{pop}} = \mathbb{E}_{\hat{p}}\left[|\mathbb{P}(Y = \hat{y} \mid \hat{p}) - \hat{p}|\right]. \tag{18}$$

Since the model predicts $k(X)$ and fits the noisy posterior for that predicted class, we have $\hat{y} = k(X)$ and $\hat{p} = \tilde{\eta}_{k(X)}(X)$. The term $\mathbb{P}(Y = \hat{y}|\hat{p})$ represents the true conditional accuracy given the confidence score. By the law of iterated expectations, this is the expectation of the clean posterior $\eta_{k(X)}(X)$ conditioned on the confidence:

$$\mathbb{P}(Y = \hat{y}|\hat{p}) = \mathbb{E}_X[\eta_{k(X)}(X)|\hat{p}]. \tag{19}$$

Similarly, the confidence $\hat{p}$ itself is trivially $\mathbb{E}_X[\tilde{\eta}_{k(X)}(X)|\hat{p}]$. Substituting these quantities into the population definition gives

$$\mathrm{CE}_{\mathrm{pop}} = \mathbb{E}_{\hat{p}}\left[\left|\mathbb{E}_X[\eta_{k(X)}(X) - \tilde{\eta}_{k(X)}(X) \mid \hat{p}]\right|\right]. \tag{20}$$

Since the absolute value function $|\cdot|$ is convex, we apply **Jensen's Inequality** ($|\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$):

$$\mathrm{CE}_{\mathrm{pop}} \leq \mathbb{E}_{\hat{p}}\left[\mathbb{E}_X[|\eta_{k(X)}(X) - \tilde{\eta}_{k(X)}(X)| \mid \hat{p}]\right]. \tag{21}$$

By the Law of Iterated Expectations, the outer expectation over $\hat{p}$ and inner expectation over $X|\hat{p}$ simplify to the expectation over the input space X:

$$\mathrm{CE}_{\mathrm{pop}} \leq \mathbb{E}_X\left[|\eta_{k(X)}(X) - \tilde{\eta}_{k(X)}(X)|\right]. \tag{22}$$

This establishes the first line of the proposition. It concerns the population quantity, not the finite-sample binned ECE used in experiments.

Step 2: Posterior-Distortion Expansion. Now we expand the term inside the expectation. By the Law of Total Probability, the noisy posterior $\tilde{\eta}_k(x)$ relates to the clean posterior via the transition matrix T:

$$\begin{aligned}\tilde{\eta}_k(x) &= \sum_{j=1}^{K} P(\tilde{Y} = k \mid Y = j, x)P(Y = j \mid x) \\ &= \sum_{j=1}^{K} T_{jk}(x)\eta_j(x).\end{aligned} \tag{23}$$

We split the summation into the predicted class ($j = k$) and the other classes ($j \neq k$):

$$\tilde{\eta}_k(x) = T_{kk}(x)\eta_k(x) + \sum_{j \neq k} T_{jk}(x)\eta_j(x). \tag{24}$$

Now, we calculate the difference $\Delta(x) = \eta_k(x) - \tilde{\eta}_k(x)$:

$$\Delta(x) = \eta_k(x) - \left(T_{kk}(x)\eta_k(x) + \sum_{j \neq k} T_{jk}(x)\eta_j(x) \right) \tag{25}$$

$$= (1 - T_{kk}(x))\eta_k(x) - \sum_{j \neq k} T_{jk}(x)\eta_j(x). \tag{26}$$

Here, we identify the physical meaning of the terms:

- **Mass Leakage (Out-flow):** $(1 - T_{kk}(x))\eta_k(x)$ is the portion of clean probability mass for the predicted class k that is corrupted (flipped away) to other classes.
- **Noise Injection (In-flow):** $\sum_{j \neq k} T_{jk}(x)\eta_j(x)$ is the probability mass from other classes j that is incorrectly flipped into class k.

Step 3: Sign Analysis in the High-Confidence Regime. The theorem keeps the absolute value, so no sign assumption is needed for the bound. To interpret the direction of miscalibration, we additionally consider the high-confidence regime where the leakage term dominates the injection term:

$$(1 - T_{kk})\eta_k \geq \sum_{j \neq k} T_{jk}\eta_j. \tag{27}$$

Consequently, $\Delta(x) \geq 0$, implying $\eta_k(x) \geq \tilde{\eta}_k(x)$. In this additional regime, $|\Delta(x)| = \Delta(x)$, and the bound can be interpreted as:

$$\mathrm{CE}_{\mathrm{pop}} \leq \mathbb{E}_x \left[\underbrace{(1 - T_{kk}(x))\eta_k(x)}_{\text{Leakage}} - \underbrace{\sum_{j \neq k} T_{jk}(x)\eta_j(x)}_{\text{Injection}} \right]. \tag{28}$$

Since the difference is nonnegative in this additional regime, the clean correctness probability η_k is at least as high as the noisy-posterior confidence $\tilde{\eta}_k$. This is a sufficient condition for pointwise under-confidence, not a claim that every noisy-label predictor or confidence group is under-confident. □

E Experiment Details

In this appendix, we provide the complete details regarding the datasets, implementation protocols, and evaluation metrics used in our main experiments.

E.1 Datasets

We evaluate our method on both synthetic and real-world noisy datasets to ensure comprehensive validation:

- **Synthetic Noise (CIFAR-10/100):** Following standard protocols (Li et al., 2020), we inject label noise into the clean CIFAR-10 and CIFAR-100 datasets. We consider symmetric noise, where labels are flipped uniformly to other classes with rates $\epsilon \in \{20\%, 50\%, 80\%\}$, and asymmetric noise, which simulates semantic confusion (e.g., Cat $\leftrightarrow$ Dog) at rates of 10% and 40%.
- **Real-world Noise:** To assess robustness against human and web-collected noise, we use **CIFAR-10N/100N** (Wei et al., 2022), **Animal-10N** (Song et al., 2019), and **Clothing1M** (Xiao et al., 2015).

E.2 Implementation Details

To demonstrate the generality of our method across different representation learning paradigms, we employ three representative pre-trained models as frozen feature extractors:

- **CLIP ViT-B/16** (Radford et al., 2021): A multi-modal model trained via language-image contrastive learning.
- **DINOv2 ViT-B/14** (Oquab et al., 2023): A self-supervised vision transformer used as the default frozen encoder.
- **ViT-B/16 (ImageNet)** (Dosovitskiy et al., 2021): A standard vision transformer pre-trained on ImageNet-21k with supervised learning.

Training Protocol. For all backbones, a simple linear classification layer is trained on top of the frozen encoder using the AdamW optimizer for 50 epochs, with a learning rate of 0.001 and a batch size of 2048. No data augmentation is applied to the cached features for CE, GCE, SSR, or the Matched Clean-LP Reference. The latter uses the identical protocol and features, replacing only the noisy training labels with the corresponding clean labels. Final results are defined as the mean and standard deviation over three independent seeds shared across methods.

Feature-Level LNL Baselines. GCE-LP (Zhang & Sabuncu, 2018) replaces cross-entropy while keeping the same linear head and optimizer. SSR-LP (Feng et al., 2022) applies the original non-parametric KNN selection and parametric relabelling logic to the frozen features. DivideMix-LP (Li et al., 2020) retains the complete DivideMix procedure: two independently initialized linear heads; per-example cross-entropy modelling by a two-component GMM; cross-head co-division at clean-probability threshold $p_{\text{th}} = 0.5$; label co-refinement and co-guessing; sharpening temperature $T = 0.5$; and MixMatch with MixUp. Augmented image views are passed through the frozen encoder without gradient propagation, so the augmentation and semi-supervised machinery are retained while representation parameters remain unchanged.

For CIFAR-10 and CIFAR-100, we use the original DivideMix settings without retuning: Beta MixUp strength $\alpha = 4$, unsupervised weight $\lambda_u = 25$ with the original ramp-up, batch size 64, learning rate 0.02, 300 epochs, and warm-up lengths 10 and 30 epochs, respectively. For Clothing1M, we use the original dataset-specific settings $\alpha = 0.5$, $\lambda_u = 0$, $p_{\text{th}} = 0.5$, $T = 0.5$, batch size 32, learning rate 0.002, 80 epochs, and one warm-up epoch. CIFAR-10N and Animal-10N inherit the CIFAR-10 settings, while CIFAR-100N inherits the CIFAR-100 settings; no dataset-specific tuning is performed.

Calibration Settings. For our AEC calibration, we use the default anchor ratio $\kappa = 0.1$ and target confidence $\xi = 0.995$ unless otherwise specified. The optimization is solved via L-BFGS, typically converging within 100 iterations. We compare against standard temperature scaling (TS) fit either on noisy validation labels or on clean validation labels. The clean-validation setting is an oracle reference rather than an available LNL protocol. Since AEC and TS apply positive temperatures after the predicted class is selected, they are evaluated primarily as calibration modules rather than as mechanisms for changing the classifier's top-1 decision.

E.3 Evaluation Metrics

We report accuracy to measure decision quality and ECE to measure confidence-accuracy alignment. Because ECE depends on binning, we additionally include strictly proper scoring rules such as NLL and Brier score:

- **Test Accuracy (Acc):** The percentage of correctly classified samples on the clean test set.
- **Expected Calibration Error (ECE):** The weighted average difference between prediction confidence and actual accuracy across $M = 15$ bins. Lower ECE indicates better calibration reliability.
- **Negative Log-Likelihood (NLL):** A strictly proper scoring rule that penalizes confident incorrect predictions and evaluates the full predictive distribution.
- **Brier Score:** A strictly proper squared-error scoring rule for probabilistic classification, complementing ECE by being insensitive to binning choices.

F SOCIETAL IMPACT

The work aims to improve reliability assessment for classifiers trained with noisy labels. Better calibration can support safer downstream decision making by making confidence estimates more informative. A potential negative impact is that calibrated-looking probabilities may encourage inappropriate deployment when the underlying decision boundary is biased, especially in high-stakes domains or under distribution shift. AEC should therefore be used as a reliability tool for fixed predictors, not as evidence that a model is fair, robust, or suitable for deployment.