When Better Models Score Lower: Test-Set Label Noise Inverts Method Rankings in Learning with Noisy Labels

Anonymous authors
Paper under double-blind review

Abstract

Label noise in fine-grained recognition is not random: annotators confuse *semantically adjacent* classes, so errors concentrate on twin class pairs. This affects both the methods of learning with noisy labels (LNL) and the protocol used to compare them.

(i) Evaluation. Test sets are annotated by the same procedure as training sets, so they carry the same structured noise, and a noisy test set does not merely lower scores: it *inverts rankings*. Measured accuracy obeys an exact two-term law, $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$, where NF ("noise fitting") is a method-intrinsic propensity to agree with the noise direction and is *anti-correlated* with true quality. Across three fine-grained benchmarks and seven methods, 58 of 63 method pairs change order, with a median crossover at $r^\star = 0.522$; a symmetric-noise control shows NF at chance level and no inversion. The inversion is not an artefact of the injected noise: it also appears on real crowd re-annotation of CIFAR-100N, where no CLIP output enters the noise at any point and a method the clean test set rejects by a wide margin comes out ahead on the human-annotated test set of the same images. The closed-form crossover $r^\star = \Delta A/(\Delta A + \Delta\mathrm{NF})$ recovers, as its small-ΔA limit, the earlier observation that test-set errors can flip model selection. Nor is the law tied to our method, backbone or modality: it holds on a 1000-class human re-annotation of ImageNet val, where with no clean labels the correction still recovers the human-verified accuracy to within a point, and the rankings invert again on 20 Newsgroups with a bag-of-words model and no pretrained network. Given any class-level confusion partner map it predicts the accuracy curve of *any* model, with NF the only free quantity; and the threshold grows as $r^\star(\rho) = \Delta A/(\Delta A + \rho\,\Delta\mathrm{NF})$ when only a fraction ρ of test errors shares the training direction.

(ii) Methods. To measure this we need a method that is actually strong under semantic-confusion noise, so we build SASR — a vehicle for the measurement, not a claimed advance. Every remedy based on data-internal statistics fails here: kNN label propagation *inverts* and lands below the CLIP zero-shot baseline. SASR beats cross-entropy by 14–18 points on all three benchmarks. Against contemporary CLIP-based remedies the result is less flattering, and we report it: a plain CLIP hard-selection baseline matches SASR on CUB-200 and beats it on Cars196.

(iii) Protocol. Correcting an accuracy estimate for an imperfect evaluator is classical, and the standard correction assumes that model errors and annotation errors are independent. LNL methods violate that assumption by construction, being optimised on the same noise that reappears in the test set. PAD (Pair-Agreement Debiasing) replaces the assumption with a measurement: using class-name semantics to obtain the noise direction, a single model's clean accuracy is recovered in closed form from two agreement rates on one noisy test set (error ≤ 0.01 for $r \leq 0.2$). PAD reduces *exactly* to the classical correction when independence holds, and as $r \to 0.5$ the two unknowns become unidentifiable — the distortion is worst exactly where it cannot be repaired. PAD needs no clean reference, no retraining and no model internals, so it can be applied post hoc to an existing leaderboard.

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

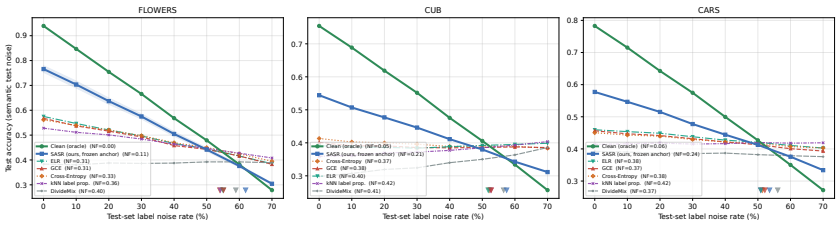

Figure 1: **Test-set label noise inverts method rankings.** Same trained models, same predictions; only the test labels are perturbed, with structured (twin-class) noise at rate r (independent seed, i.e. a second set of annotators). Measured accuracy follows $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$ with a method-specific NF (legend). Because NF anti-correlates with true quality, curves cross: 58 of 63 method pairs reverse order, median crossover $r^\star = 0.522$. The law holds for any model; the only ingredient needed is the twin map τ.

1 Introduction

Fine-grained recognition is annotation-bottlenecked. Distinguishing a *Setophaga striata* from a *Setophaga caerulescens*, or one BMW 3-series year from the next, is hard even for a careful expert, so the label noise in these datasets is not the uniform random flip that dominates the LNL literature: the wrong label is nearly always the semantically and visually nearest alternative. We call this **semantic-confusion noise** and formalise it by a twin map τ that names, for each class, its nearest neighbour in CLIP text embedding space.

The evaluation protocol is broken. Because test sets are annotated by the same procedure as training sets, they carry the same structure, so the field's protocol — train on noisy data, rank methods on a *clean* test set — is unrepresentative of deployment; and when the test set is noisy the ranking does not merely degrade, it *inverts* (Figure 1): the same trained models on the same images change order purely because of the annotation quality, and the strongest method is the biggest victim.

This is not a hypothetical. On CIFAR-100N, whose test labels are crowd re-annotations with no synthetic noise and no CLIP output anywhere in the pipeline, the verdict of the standard protocol reverses: a model that trails by 13.4 points on the clean test set leads by 3.8 points on the human-annotated test set of the same images, in the ten noisiest classes. The clean-test deficit grows monotonically with the annotation error rate, and the reversal survives both a random-subset control and a second train/test partition (Section 7.4).

That test-set label errors exist, and that they can perturb model selection, is known (Northcutt et al., 2021); what has not been established is what happens to *method* rankings, and by how much, when the noise is *structured* and shares its direction with the training noise. Section 3 derives an exact law, $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$, whose second term is a method-intrinsic propensity to agree with the noise direction and which holds for *any* model given the twin map τ: across three fine-grained benchmarks and seven methods, **58 of 63 method pairs reverse order** (median crossover $r^\star = 0.522$), and a symmetric-noise control yields NF at chance with zero inversions. The per-dataset crossovers and the remaining checks are in Appendix F.

Why the ranking inverts, and why we need a method that does not. An entire family of LNL remedies is structurally mis-specified here: small-loss selection and neighbour consistency both rely on a signal that semantic twins destroy, so kNN label propagation assigns *higher* weight to mislabelled samples — in our ablation it lands 24.7 points below our method, 7 below plain CE, and below the CLIP zero-shot baseline (0.523 vs 0.634). To *measure* the inversion we therefore need a method genuinely strong under this noise, or "the best method is the biggest victim" is vacuous; SASR (Section 4) recovers 14–18 points over CE.

We are explicit about the status of that component: our claim on SASR is the *composition* on an already crowded axis, and Section 6.1 tests that rather than asserting it — the reimplemented CLIP baselines match SASR on CUB-200 and beat it on Cars196.

Repair and its limits. We finally ask whether the distortion can be undone without clean labels. Exploiting the involution structure of twin confusion we derive PAD, which recovers clean accuracy

in closed form from two observable agreement rates, using only class-name semantics and a noise-rate estimate. The estimator is exact in its regime (error ≤ 0.01 for $r \leq 0.2$) and comes with an exact sensitivity law. We also prove an identifiability limit: as $r \to 0.5$, A and NF cannot be separated from noisy labels alone — precisely where the distortion is most severe, so it is worst exactly where it is unrepairable, which is itself an argument for validating test-set cleanliness rather than correcting scores after the fact.

Contributions.

1. Semantic-confusion noise as the realistic fine-grained noise model, with a diagnosis of why data-internal LNL methods must fail under it.
2. SASR, a method strong enough under semantic-confusion noise for the measurement to be meaningful (14–18 points over CE on all three benchmarks); we claim the composition rather than a new method.
3. The accuracy-decomposition law $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$, its closed-form crossover, and the demonstration that noisy test sets invert method rankings (58/63 pairs), with a negative symmetric-noise control.
4. PAD, a label-free closed-form debiasing estimator with an exact sensitivity law, and the $r \to 1/2$ identifiability limit.

2 Related Work

Label errors in test sets. Northcutt et al. (2021) established that test sets of common benchmarks carry systematic label errors ($\geq 3.3\%$ on average), that *correcting* them leaves relative rankings largely unchanged, and—most relevant to us—that higher-capacity models track those errors more closely, so that selection would flip if the mislabeled prevalence rose by only 5–6%; Karakatič et al. (2026) reports the analogous effect in algorithmic fairness. We therefore do **not** claim to be the first to notice that test-set errors perturb benchmark conclusions: we turn these observations into a law that predicts their magnitude, and extend them from model capacity to arbitrary training objectives.

VLM-based remedies. CLIPCleaner (Feng et al., 2024), NLPrompt (Pan et al., 2025), DSPT (Li et al., 2026a) and VisPrompt (Geng et al., 2026), alongside CRoF (Deng et al., 2024), WTS (Li et al., 2026b) and LANE (Sosea & Caragea, 2026), all call on a frozen vision-language model as an external arbiter. They confirm semantic structure as the right prior, but each either gates its correction by a threshold or modulates a signal computed from the model being trained. We borrow the same paradigm—a frozen anchor as a purely external judge—without claiming novelty for it.

Estimating accuracy when labels are imperfect. The latent-class formulation (Hui & Walter, 1980) and the agreement-rate family (Platanios et al., 2014; 2016; 2017) correct measured accuracy under an *assumed* conditional independence; prediction-powered inference (Angelopoulos et al., 2023) calibrates residuals against a small set of gold-standard labels. Our setting violates independence by construction, and PAD *measures* the violation rather than assuming it, using no gold labels at all. REVEAL (Pang et al., 2025), HLE-Verified (Zhai et al., 2026) and Fault Lines (Jackson et al., 2025) treat the test set as something to be cleaned; we treat its noise as something to be measured.

Appendix D gives the per-work comparison, the remaining literature, and the adjacent settings where the same phenomenon has been reported.

3 Theory

3.1 Semantic-confusion noise and the twin map

Let the class set be $\{1..K\}$ with names $c_1, \ldots, c_K$; encoding each name with a frozen text encoder, let

$$\tau(k) = \arg\max_{j \neq k} \cos\big(\phi(c_k), \phi(c_j)\big) \tag{1}$$

be the **twin** of class k, and model the noisy label $\tilde{y}$ of an image with true label y as

$$\tilde{y} = \begin{cases} \tau(y) & \text{with probability } r, \\ y & \text{otherwise,} \end{cases} \tag{2}$$

with r the noise rate. Unlike symmetric noise this transition is *anisotropic* and can be *shared* between training and test sets, because both come from the same annotation process.

3.2 The accuracy decomposition

With $A = \mathbb{P}(\hat{y} = y)$ the clean accuracy and $\mathrm{NF} = \mathbb{P}(\hat{y} = \tau(y))$ the **noise-fitting coefficient**,

Proposition 1 (Accuracy decomposition)**.** *On a test set corrupted at rate r, the measured accuracy is*

$$\mathrm{Acc}(r) \;=\; (1-r)\,A \;+\; r\,\mathrm{NF}. \tag{3}$$

Two consequences. **(a)** NF is empirically nearly constant across $r \in [0.1, 0.7]$, so it is a property of the method, not of the noise level. **(b)** Comparing two methods a, b via (3) gives a closed-form crossover

$$r^{\star}_{ab} \;=\; \frac{A_a - A_b}{(A_a - A_b) + (\mathrm{NF}_b - \mathrm{NF}_a)}, \tag{4}$$

at which the ranking reverses; since NF is anti-correlated with A, a noisy test set systematically *rewards the worse method*.

3.3 Why symmetric noise cannot do this

Under symmetric noise the flip target is uniform over the remaining $K-1$ classes, so $\mathrm{NF} \to 1/K$ for every method and $\mathrm{Acc}(r) = (1-r)A + r/K$ — a monotone transform of A, so rankings are *invariant by construction*. That is the falsifiable prediction we use as a control.

3.4 An identifiability limit

With $u = \mathbb{P}(\hat{y} = \tilde{y})$ and $v = \mathbb{P}(\hat{y} = \tau(\tilde{y}))$, on mutual twins ($\tau(\tau(k)) = k$)

$$u = (1-r)A + r\,\mathrm{NF}, \qquad v = rA + (1-r)\,\mathrm{NF}, \tag{5}$$

which solves to

$$\hat{A} = \frac{(1-r)u - rv}{1-2r}, \qquad \hat{\mathrm{NF}} = \frac{(1-r)v - ru}{1-2r}. \tag{6}$$

The system is singular at $r = 1/2$: A and NF become unidentifiable. This is not an artefact of our estimator — it says that when the test set is about half mislabelled, no method that sees only test labels can separate true quality from noise fitting, so the regime where ranking distortion is worst is exactly the regime where it cannot be corrected. PAD *contains* the classical correction (Hui & Walter, 1980) as the independent-error special case (Appendix B); the derivations are in Appendix J.

4 Method used as a vehicle (SASR)

To measure the inversion we need a method that is genuinely strong under semantic-confusion noise, otherwise "the best method is the biggest victim" is vacuous. Any method with that property will do, and Section 6.1 shows that a much simpler one also qualifies, so we treat this section as infrastructure rather than as our methodological claim. SASR keeps a *frozen* CLIP text anchor as an external judge — the only quantity in the pipeline that the noisy labels cannot move — and uses it three ways. For each sample with given label $\tilde{y}$ it computes the relative-confidence ratio $\mathrm{ratio}_i = p_{\tilde{y},i} / \max_k p_{k,i}$, where $p = \mathrm{softmax}(f_0^{\top} c_{\text{text}}/\tau)$ is the frozen anchor's posterior. (i) The ratio becomes a steep soft weight $w_i = w_{\min} + (1 - w_{\min})\,\mathrm{ratio}_i^{p}$, with $w_{\min} = 0.05$, $p = 20$. (ii) The loss is weighted cross-entropy: soft, so no sample is discarded and none is relabelled. (iii) The same weights purify the prototype EMA, so the classifier is not dragged by mislabelled features. On Flowers102 the raw ratios of the two populations are much closer together (means 0.89 against 0.57) than the weights they induce; the steep exponent is what turns that into a clean separation, $w(\text{clean}) = 0.71$ against $w(\text{mislabelled}) = 0.12$. Full specification, component ablations and rejected alternatives are in Appendix A.

Table 1: Test accuracy under 50% semantic-confusion noise. All entries are peak accuracies over training; SASR is averaged over three seeds. The two CLIP-based rows are our own reimplementations under this protocol (Section 6.1), also averaged over three seeds. The kNN entry for Cars196 is the final-epoch accuracy of the checkpoint used in Section 7 (no peak log exists for that run); CUB-200's kNN entry is omitted for the same reason, so the kNN-on-CUB statements below rest on Table 4 rather than on this column.

Method	Flowers102	CUB-200	Cars196
Clean oracle	0.939	0.755	0.785
SASR (ours)	**0.770**	0.557	0.593
CLIP selection (ours)	0.710	0.557	**0.604**
CLIP pseudo-labels	0.667	**0.559**	0.585
CLIP zero-shot	0.634	0.538	0.568
ELR	0.593	0.391	0.460
GCE	0.574	0.407	0.456
CE	0.592	0.413	0.450
kNN label prop.	0.523	—	0.415
DivideMix	0.385	0.307	0.394

5 PROTOCOL

Flowers102, CUB-200-2011 and Stanford Cars with the official splits; class names used verbatim for the prototypes and the twin map τ; a frozen CLIP ViT-B/32 with only a residual image-side adapter and the class prototypes trained; test noise applied to the *same* predictions at $r \in \{0, 0.1, \ldots, 0.7\}$ with an independent seed, so no retraining is involved. The full protocol — datasets, optimiser and schedule, the methods compared, and the official-split leakage measurement of Section 6.2 — is in Appendix A (Table 3).

6 RESULTS I: SEMANTIC-CONFUSION NOISE BREAKS DATA-INTERNAL METHODS

Main comparison (50% semantic noise).

SASR leads CE by 14–18 points on all three datasets (Flowers 17.8, CUB-200 14.4, Cars196 14.3) and leads the strongest *dedicated* LNL baseline—ELR on Flowers and Cars, GCE on CUB-200—by 13–18 points. DivideMix, the strongest method on standard symmetric-noise benchmarks, falls *below* plain CE on all three datasets; kNN label propagation falls below CLIP zero-shot.

Baseline budget, stated plainly. Every baseline shares SASR's backbone, data pipeline, classifier parameterisation, batch size, learning rate and epoch budget (20 epochs, batch 64, $\mathrm{lr} = 5 \times 10^{-4}$, one seed); the pre-CLIP methods use their published default hyperparameters and we did not search them. We therefore report these as *published-default* baselines, not as tuned competitors, and we do not claim they are the best each method can reach: a per-method search under a matched budget could move them. What the table supports is the comparison we actually ran under a shared budget, not the claim that no LNL method can beat ours; the per-method settings are in Appendix A.

6.1 SASR AGAINST CONTEMPORARY VLM-BASED REMEDIES

Those baselines are all pre-CLIP, so the comparison above is not the one a 2026 reviewer would run. We therefore reimplement two contemporary remedies under the *identical* protocol — same backbone, same data pipeline, same classifier parameterisation, same epochs, no clean reference and no extra data — and report the outcome as measured, including where it goes against us: *CLIP selection* keeps the training samples whose frozen CLIP zero-shot prediction agrees with the given label and trains on that subset with plain CE, the offline hard-selection mechanism of Feng et al. (2024), and *CLIP pseudo-labels* replaces every training label with the zero-shot prediction, the trivial member of the prompt/teacher family (Wei et al., 2024; Pan et al., 2025). Averaged over three seeds, both are far stronger than every pre-CLIP baseline on every dataset, which is the point: on this noise

family the deciding ingredient is consulting an external semantic anchor at all, not how elaborately one consults it. Against them, SASR's advantage survives only on Flowers102, where it leads CLIP selection by 6.0 points (5.2 once near-duplicate test images are removed, Section 6.2). On CUB-200 the two are tied to within 0.002, and on Cars196 the simpler baseline is *ahead* by 1.1 points, on all three seeds. We state the consequence plainly: the method contribution of Section 4 is a composition, as its own scope statement says, and the experiments in Sections 7–8 do not depend on SASR being state of the art — they need only a method that does not collapse under semantic-confusion noise, and both CLIP selection and SASR satisfy that.

Remark 1 (Protocol note)**.** *Table 1 reports peak accuracy over training (SASR averaged over three seeds); Section 7 evaluates the final-epoch checkpoint, since the ranking study requires one fixed set of predictions to perturb. The two differ by at most 1.5 points on any method (SASR on CUB-200: 0.557 peak versus 0.543 final epoch; SASR on Cars196: 0.593 three-seed peak versus 0.581 single checkpoint), and the seed spread of the first is of the same order as that gap. All conclusions are identical under either protocol. The three quantities are defined separately in Appendix J.*

Noise-rate ablation (Flowers102). SASR 0.886 / 0.770 / 0.558 and CE 0.818 / 0.592 / 0.287 at 30 / 50 / 70%; the advantage grows with the noise rate (+7, +18, +27).

Noise-type control. With symmetric noise at the same rate the ranking collapses to "everyone similar": GCE reaches 0.886 under symmetric noise but 0.574 under semantic confusion (-31 points), whereas SASR loses only 10. This is the decisive evidence that the *type* of noise, not its rate, determines which methods work. The component ablation (prototype purification -2.7; hard subset selection -4.7; anchor-argmax correction -8.1; kNN -24.7) and the clean-label-weight ceiling are in Appendix A.

6.2 Control: near-duplicate leakage in the test splits

Fine-grained splits can contain test images that are near-duplicates of training images, and that matters here because SASR classifies by comparing a test feature against class prototypes, so a leaked image can be answered by retrieval rather than by generalisation. Taking each test image's frozen CLIP feature and its maximum cosine similarity against the *training pool actually used*, counted as leaked at $\cos \geq 0.95$, the rate varies by an order of magnitude across benchmarks all called "fine-grained": 31.1% on Flowers102 against 3.2% on CUB-200 and 2.8% on Cars196. We therefore re-scored the *identical* predictions on the leak-free subset and report both columns. The absolute numbers are inflated, and unevenly so: leakage is worth 3.8 points to CE on Flowers102 but only 1.4 to SASR. But *leakage does not manufacture the ranking*: leaked images are simply easier, and easier for everyone, so because the weakest classifier gains most from an easier test set, leakage *understates* the advantage of a strong one. On Flowers102 the SASR$-$CE gap therefore *widens* from 17.4 to 19.8 points once leakage is removed (17.4 from the control's own full-split run; Table 1 reports 17.8 for the same quantity and this table's rounded cells give 17.3, the spread being the recomputation drift its caption notes), while the SASR$-$CLIP-selection gap narrows only from 6.0 to 5.2. On CUB-200 and Cars196 every method moves by at most 0.4 points and no ordering changes. The control thus closes the obvious alternative explanation of our headline comparison, at the cost of a disclosure: on Flowers102 the gap we report is smaller than the leaked split makes it look, not larger. The full detail, including why the correction is a lower bound, is in Appendix B.

6.3 The external dependency: what happens when the anchor is wrong

SASR's only external input comes from CLIP's text space. Damaging the anchor in a controlled way, the gain over cross-entropy decays monotonically to zero at $q \approx 0.86$, is $+2.8$ points at $q = 0.75$, and a completely broken anchor leaves SASR 4.8 points *below* plain cross-entropy; a frozen *random* anchor is no better than a fully permuted text anchor, so an anchor must be informative and not merely fixed. SASR therefore needs a class-name lexicon whose embeddings separate the classes, and a τ defined by that same lexicon; where neither holds, the method degrades to — and can fall below — cross-entropy. That is the honest boundary of the contribution. What survives such a domain is not the method but the *diagnostic*: NF and the crossover law need only *some* class-level confusion partner map, and Section 9 shows that they also need no vision.

7 RESULTS II: NOISY TEST SETS INVERT METHOD RANKINGS

Figure 1 already showed the effect; here we quantify it.

7.1 THE NOISE-FITTING COEFFICIENT ANTICORRELATES WITH QUALITY

Fitting (3) to each method's accuracy curve gives (A, NF) per method and dataset: Spearman(A, NF) is -1.000 on Flowers102, -0.964 on CUB-200 and -0.714 on Cars196 (DivideMix on Cars is the single discordant point); **pooled over all 21 method–dataset pairs,** -0.914. The per-method NF values are stable across noise levels (single-seed run), confirming that NF is intrinsic rather than a curve-fitting artefact. A control rules out an architectural explanation: every method, including the clean oracle, initialises its prototypes from the *same* CLIP text prototypes and the same τ, yet the oracle has NF near chance $(1/K)$, so NF is learned from the noisy labels, not inherited from the initialisation.

7.2 RANKINGS INVERT, WITH A PREDICTABLE CROSSOVER

Using (4) on the fitted (A, NF): Flowers102 $21/21$ pairs reverse within $r < 1$; CUB-200 $20/21$; Cars196 $17/21$. **Pooled** $58/63$**, median** $r^\star = 0.522$, i.e. once a test set is about half mislabelled the standard protocol's ranking is no longer meaningful. Observed rankings confirm the predictions: at $r = 0$ the order is oracle $>$ SASR $>$ CE $>$ kNN, while at $r = 0.7$ it becomes oracle $> \cdots >$ CE $>$ SASR in the Flowers case, and kNN overtakes both CE and SASR on CUB. The individual crossovers, and the "some methods improve" case (a model can be made to look 9.3 points better, on the same weights, purely by degrading the test-set annotation quality — the sharpest demonstration of the failure mode), are in Appendix F.

7.3 CONTROL: SYMMETRIC TEST NOISE PRODUCES NO INVERSION

Same trained models, symmetric test noise: NF collapses to chance for every method (mean $|\mathrm{NF}| \leq 0.03$ against $1/102 = 0.0098$), the accuracy ratio between methods is preserved to three decimals (SASR/oracle $= 0.795$ at both $r = 0$ and $r = 0.7$), and zero pairs invert. This reproduces the prediction of Section 3 and rules out a "the test set just got harder" explanation.

7.4 REAL HUMAN ANNOTATION NOISE, WITH NO SYNTHETIC COMPONENT

Everything above injects a known transition τ, which invites the objection that a CLIP-based remedy fits that noise by construction. We therefore repeat the experiment on *human* annotation only. CIFAR-100N (Wei et al., 2022) re-annotates the $50{,}000$ training images by crowd workers; we use the majority-vote labels as the noisy training set (error 40.2%) and, separately, as the test labels of a held-out pool, so no CLIP output enters the noise at any point. Ranking every class by its *training-pool* error rate and keeping the worst N yields real, human-annotated test sets at $r = 0.538$ $(N = 50)$ down to 0.763 (10), selected without looking at the evaluation pool. In Table 12 the oracle leads on the clean test set increasingly so, its margin growing from 2.6 points in the full pool to 13.4 in the ten noisiest classes, while on the human-annotated test set of those *same* images SASR leads in every stratum, most widely in the noisiest one ($+3.8$ points), and significantly so throughout ($p \leq 0.009$). Both controls hold (random class draws; a second train/test partition, $p = 0.0001$). With nine methods none of the individual $\mathrm{Spearman}(A, \mathrm{NF})$ correlations is significant ($p \geq 0.33$), so we read this as establishing that the inversion occurs with no synthetic noise, not as independent confirmation of the mechanism; annotation error is structured in its *direction* but not reproducible *per image*. That is the boundary of the phenomenon. The four stratum rates, the granularity check and the ImageNet ReaL validation of $\hat{A}$ are in Appendix I.

8 PAD: DEBIASING WITHOUT CLEAN LABELS

Mutual-twin subset. τ is an involution only on mutual twins: $30/102$ such classes on Flowers102, $88/200$ on CUB-200 and $102/196$ on Cars196, covering 1952 / 2487 / 4186 test samples. **A selection trap that must be handled.** Selecting the samples whose *given* label lies in the mutual set lets

in samples whose true class is outside it but whose flip lands inside, which inflates the realised flip rate (0.306 at $r = 0.20$) so that (6) returns errors of -0.108 and -0.161; restricting the *injection* to the mutual classes makes the subset closed under flipping and the realised rate then matches r.

Estimator accuracy. Mean $|\hat{A} - A|$ ranges from 0.0039 to 0.0458 across the three benchmarks over $r = 0.1 \ldots 0.4$, growing with r as the $1/(1-2r)$ amplification predicts, and is ≤ 0.01 for $r \leq 0.2$.

PAD on real human annotation noise. The same estimator, with no re-tuning, applied to CIFAR-10N crowd labels whose ten class names induce four classes in mutual pairs, covering 40% of the evaluation pool, across three real label sets that differ in rate: the three-annotator majority vote ($r = 0.105$), a single *other* annotator ($r = 0.207$, so training and test noise come from different people), and the worst annotator ($r = 0.444$). The uncorrected accuracy understates the best of five trained models by 8.8 and 18.0 points; PAD removes 89% and 84% of that bias. Real noise is a harder test: only two twin pairs survive on ten classes and the per-class rate is heterogeneous (0.085 to 0.863 on CIFAR-100N); the deployable variant costs a further 1.5 points at $r = 0.207$. At $r = 0.444$ the estimator does not merely lose accuracy, it returns a value above one — the identifiability limit of Section 3.4, reached by real annotation rather than by injected noise. PAD is usable on real crowd labels exactly where the theory says it is: below $r \approx 0.2$, and not near one half.

Against the classical correction, on the same logs. Even granted the correct multi-class formula, the independence assumption costs about 10 points of accuracy and a systematic $+0.10$ overestimate, whereas PAD has a mean bias of $+0.007$ and a mean absolute error of 0.014; the reason is that LNL methods deliberately fit the noise and thereby break independence. This is the empirical counterpart of the exact reduction proved in Appendix B.

Honest assessment of the ranking benefit. On the same subset, ranking by $\hat{A}$ is comparable to ranking by the raw agreement rate u, not uniformly better (e.g. Flowers102 at $r = 0.4$ both give Spearman $+0.83$; Cars196 at $r = 0.4$ PAD gives $+0.43$ while u gives $+1.00$). PAD's value is a *calibrated, noise-rate-independent estimate of* A; it is not a magic ranking fix.

9 Scope and limitations

What bounds the claim. Four things. (i) *Identifiability:* by (6) the problem is singular at $r = 1/2$ and ill-conditioned nearby, so no test-label-only method can cure the worst case. (ii) *The noise rate must be known:* we take r from re-annotation, dual annotation or a label-free agreement estimator (Platanios et al., 2014), and supply the resulting cost in Table 6. (iii) *Twin-map coverage:* A, NF and PAD are computed on mutual-twin classes, 29–52% of classes, so absolute accuracies on that subset differ from full-test accuracies. (iv) *Anchor dependence:* SASR's advantage over cross-entropy vanishes once roughly three-quarters of the class anchors are wrong and falls *below* plain cross-entropy when the anchor is fully broken; the diagnostic does not require a lexicon whose embeddings separate the classes.

The law also holds without vision: on 20 Newsgroups, with τ built from class names alone, it holds to within 0.003 for every method and both noise types, and the rankings still invert. The field noise rates and the bounds on that real-noise inversion are in Appendix E; our claim remains conditional on shared annotation structure, which Section G.2 quantifies.

10 Discussion: the decomposition beyond annotation error

The law needs only a test set that splits into a clean part $(1-r)$ and a *compromised* part r: averaging the measured accuracy over the two parts gives the same two-term form, $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$, with no assumption about how the compromise arose. Benchmark contamination is the instance of this that the LLM literature cares about: r is the fraction of leaked items and NF is how often the model returns the leaked answer, and a contaminated benchmark *rewards* contamination, so pairs of models can invert as r grows. Whether the identifiability limit does is less clear, since it is derived from a two-equation system that the twin map makes available and contamination supplies no such map. We therefore state this as a conjecture about a setting we have not tested, because the contribution is the decomposition, not the fine-grained application.

11 Conclusion

Fine-grained label noise is structured, and because test sets are annotated by the same procedure as training sets, the test set is a second noisy sample of the same annotator population rather than a ground truth. Under that condition ranking methods on a clean test set is not merely optimistic, it is misordered: the methods that fit the noise are rewarded, and the strongest method is the largest victim. The effect is described exactly by $\mathrm{Acc}(r) = (1 - r)A + r\,\mathrm{NF}$, and the law is not specific to our method, our backbone or our noise family: given a map τ that names each class's confusion partner, it predicts the accuracy curve of *any* model, and NF—its propensity to agree with the noise direction—is the quantity that anti-correlates with true quality and produces the crossovers. On three fine-grained benchmarks 58 of 63 method pairs reverse order, with a median crossover at $r = 0.52$; a symmetric-noise control shows no inversion at all, so the phenomenon is a property of the structure of the noise rather than of its prevalence.

Two consequences are practical. First, a benchmark that cannot be re-annotated should report NF, or accuracies corrected by PAD, which needs only noisy labels and class names and no clean reference. Second, the noise rate must be estimated, and (15) with Table 6 gives the budget: under a hundred re-annotated images at the 3–6% error rates typical of fine-grained benchmarks, but about fifteen thousand at $r = 0.39$, because the estimator's sensitivity scales as $(1 - 2r)^{-1}$. We do not claim that existing leaderboards are wrong: at the 3–6% unstructured error rates reported in the field the predicted distortion is mild. What we claim is narrower and more useful: when the noise is structured and shares its direction with the training noise, it is known in advance which methods will be rewarded, by how much, and from which noise rate onwards the ranking can no longer be trusted.

12 LLM usage

Large language models were used as a writing and editing aid: to improve the language, to condense sections of the manuscript to meet the page limit, and to check that notation and numbers are used consistently across the paper. They were not used to generate the research idea, to design the method, to run or analyse experiments, or to produce any reported number: every experiment, analysis and quantitative result is the authors' own, and every number in the paper is traceable to a script and log through Appendix C. The authors reviewed and take responsibility for all text.

Reproducibility Statement

All experiments, analyses and quantitative results in this paper are traceable to specific scripts and log files, listed in Appendix C (Number Provenance). Hyperparameters and the training protocol are detailed in Appendix A (Table 3). The complete code, checkpoints, and figure-generation scripts are provided as supplementary material.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. *arXiv preprint arXiv:2601.05420*, 2026.

John Collins and Minh Huynh. Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Statistics in Medicine*, 33(24):4141–4169, 2014. doi: 10.1002/sim.6218.

Shizhuo Deng, Bowen Han, Jiaqi Chen, Hao Wang, Dongyue Chen, and Tong Jia. CRoF: CLIP-based robust few-shot learning on noisy labels. *arXiv preprint arXiv:2412.12793*, 2024.

Florian E. Dorner and Moritz Hardt. Don't label twice: Quantity beats quality when comparing binary classifiers on a budget. In *International Conference on Machine Learning (ICML)*, 2024.

Rui Duan, Ming Cao, Yang Ning, Mingfu Zhu, Bin Zhang, Aidan McDermott, Haitao Chu, Xiaohua Zhou, Jason H. Moore, Joseph G. Ibrahim, Daniel O. Scharfstein, and Yong Chen. Global identifiability of latent class models with applications to diagnostic test accuracy studies: A gröbner basis approach. *Biometrics*, 76(1):98–108, 2019. doi: 10.1111/biom.13133.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. CLIPCleaner: Cleaning noisy labels with CLIP. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, 2024. doi: 10.1145/3664647.3680664.

Zibin Geng, Xuefeng Jiang, Jia Li, Zheng Li, Tian Wen, Lvhua Wu, Sheng Sun, Yuwei Wang, and Min Liu. Seeing is believing: Robust vision-guided cross-modal prompt learning under label noise. *arXiv preprint arXiv:2604.09532*, 2026.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980. doi: 10.2307/2530508.

David Jackson, Paul Groth, and Hazar Harmouch. Fault lines: Benchmarking the impact of label data quality on ML robustness and fairness. *Proceedings of the VLDB Endowment*, 19(4):670–683, 2025. doi: 10.14778/3785297.3785308.

Sašo Karakatič, Ivona Colakovic, and Tjaša Heričko. Fairness evaluation paradox: How biased test data masks true group fairness assessment. *Mathematics*, 14(16):2894, 2026. doi: 10.3390/math14162894.

Ken Lang. NewsWeeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pp. 331–339, 1995. The `20news-bydate` split used here is the standard `20news-bydate.tar.gz` distribution.

Jia-yu Li, Jiaxin Qi, Sheng Zhou, Jianqiang Huang, and Xian-Sheng Hua. Intrinsic gradient suppression for label-noise prompt tuning in vision-language models. In *International Conference on Machine Learning (ICML)*, 2026a. OpenReview: 9W17xHFtdQ.

Mengke Li, Haiquan Ling, Yiqun Zhang, Yang Lu, and Hui Huang. Learning from imperfect text guidance: Robust long-tail visual recognition with high-noise labels. *arXiv preprint arXiv:2604.23125*, 2026b. Accepted at CVM 2026.

Shuai Li, Xinyan Liang, Yuhua Qian, and Li Lv. GFD-EMVC: Evolutionary multi-view classification with label noise via gradient and feature dual-perception. In *International Conference on Machine Learning (ICML)*, 2026c. OpenReview: aWUlNpaZs1.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2021.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19963–19973, 2025.

Zirui Pang, Haosheng Tan, Yuhan Pu, Zhijie Deng, Zhouan Shen, Keyu Hu, and Jiaheng Wei. When VLMs meet image classification: Test sets renovation via missing label identification. *arXiv preprint arXiv:2505.16149*, 2025.

Henrique Pickler, Jorge K. S. Kamassury, and Danilo Silva. Benchmarking noisy label detection methods. *arXiv preprint arXiv:2510.16211*, 2025.

Emmanouil Antonios Platanios, Avrim Blum, and Tom Mitchell. Estimating accuracy from unlabeled data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

Emmanouil Antonios Platanios, Avinava Dubey, and Tom Mitchell. Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning (ICML)*, volume 48 of *PMLR*, pp. 1416–1425, 2016.

Emmanouil Antonios Platanios, Hoifung Poon, Tom M. Mitchell, and Eric Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tiberiu Sosea and Cornelia Caragea. LANE: Label-aware noise elimination for fine-grained text classification. In *International Conference on Learning Representations (ICLR)*, 2026. OpenReview: PuayLKdrFz.

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 (ICLR)*, 2022.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. OpenReview: haUnEiXgQ7.

Weiqi Zhai, Zhihai Wang, Jinghang Wang, Boyu Yang, et al. HLE-Verified: A systematic verification and structured revision of humanity's last exam. *arXiv preprint arXiv:2602.13964*, 2026.

A Method specification, ablations, and rejected alternatives

A.1 Setup

A frozen CLIP image encoder produces f_0; a residual adapter g produces $f_t = g(f_0)$, L2-normalised. Class prototypes c_k are initialised from CLIP text prototypes for the class names and used as a cosine classifier with temperature τ.

A.2 Component 1: frozen-anchor confidence ratio

For each training sample with given label $\tilde{y}$ compute the anchor posterior $p = \mathrm{softmax}(f_0^\top c_{\text{text}}/\tau)$ and the relative-confidence ratio

$$\mathrm{ratio}_i = \frac{p_{\tilde{y},i}}{\max_k p_{k,i}}, \tag{7}$$

which asks how confident the *frozen, noise-free* anchor is that this image belongs to the label it was given, relative to its most likely class. In effect the anchor acts as an external judge that is not updated by the noisy supervision. *Why not an internal signal?* Every internal candidate is corrupted in the same direction: the loss is high not only for mislabelled samples but for hard-but-clean twins, neighbour statistics are contaminated in both directions, and the model's own confidence is precisely the quantity we are trying to supervise. The frozen anchor is the one quantity in the pipeline that the noisy labels cannot move.

A.3 Component 2: steep soft reweighting

Convert the ratio to a weight

$$w_i = w_{\min} + (1 - w_{\min})\,\mathrm{ratio}_i^{\,p}, \qquad w_{\min} = 0.05,\ p = 20, \tag{8}$$

and train with a weighted cross-entropy, $\mathcal{L} = \mathrm{mean}_i\, w_i\, \mathrm{CE}(f_{t,i}, \tilde{y}_i)$. Weights are *soft*: no sample is discarded and none is relabelled. Measured on Flowers102 this separates the two populations sharply: $w(\text{clean}) = 0.71$ vs $w(\text{mislabelled}) = 0.12$, with $\mathrm{ratio}(\text{clean}) = 0.89$ vs $\mathrm{ratio}(\text{mislabelled}) = 0.57$. *Why not hard selection?* Hard subsetting is the obvious alternative and it costs 4.7 points. On twin pairs a sample sits exactly on the boundary where its weight is most uncertain, and a threshold converts that uncertainty into a definite decision, discarding precisely the samples whose labels are hardest to adjudicate. A steep *soft* weight instead keeps them in the objective at low influence; the isolation diagnostic in §6 confirms this is the better trade. The two CLIP baselines of Section 6.1 instantiate the hard-selection end of this spectrum, which is one reason they are competitive.

Table 2: Training-side imbalance. Macro accuracy on a class-balanced holdout as an exponential long tail of factor R is imposed on the training prior. LA and the SASR soft weights are *not* additive; the gap grows with R.

R	CE	SASR	LA	LA + SASR weights	LA − CE
1	0.762	0.773	**0.781**	—	+1.9
5	0.689	—	**0.715**	0.710	+2.6
10	0.650	—	**0.682**	0.676	+3.2
20	0.605	0.616	**0.640**	0.624	+3.5
100	0.508	0.516	**0.546**	0.526	+3.8

A.4 Component 3: weight-purified prototype EMA

Because the classifier prototypes are EMA-updated from the same data, the prototypes themselves are contaminated by mislabelled features. We therefore update them with the *same* per-sample weights,

$$c_k \leftarrow (1-\eta)\, c_k + \eta\, \frac{\sum_{i \in \text{batch},\, y_i = k} w_i\, f_{t,i}}{\sum_{i \in \text{batch},\, y_i = k} w_i}. \tag{9}$$

This component is worth $+2.7$ points on its own. *Why purify them rather than freeze them?* Leaving the prototypes at their text initialisation avoids contamination but also prevents them from tracking the adapted features f_t, which is where the gains over zero-shot come from. Purifying with the same weights keeps the adaptation while draining the contamination, and it is the only component whose removal does not change the objective itself, which is why it is easy to miss.

A.5 Training-side class imbalance is a separate axis

The noise family this paper is about is a property of the *test* labels. Training-side class imbalance is a different axis and calls for a different remedy, so we ran the control rather than conflate the two. On a 500-class web-image set (101,217 training images, class-balanced holdout) we impose an exponential long tail on the *training* prior by subsampling, with imbalance factor $R = \max_k n_k / \min_k n_k$, and evaluate on the balanced holdout with the deployment head (6-view concatenated kNN, $k = 100$, query-side TTA, column normalisation). The four methods are plain cross-entropy (CE); the frozen-anchor steep soft reweighting of Section 4 (SASR); logit adjustment on the training prior (LA), the classical prior-based remedy; and LA combined with the SASR soft weights.

Two things follow. First, the frozen-anchor signal does survive imbalance—SASR stays $+0.8$ to $+1.2$ points above CE at every R—but it is not the right tool for it: the classical prior-based correction is worth $+1.9$ to $+3.8$ points and its advantage grows monotonically with R. Second, the two mechanisms compete rather than compose: adding the SASR soft weights on top of LA costs 0.5 points at $R = 5$ and 2.0 points at $R = 100$. This is why the contributions are kept separate. Section 4 is addressed to the noise family, not to imbalance, and should not be read as a long-tail method.

A.6 Experimental protocol and hyperparameters

The pre-CLIP methods use their published default hyperparameters and we did not search them: they are *published-default* baselines, not tuned competitors. Official fine-grained splits are not leakage-free, so Section 6.2 measures this and re-scores the competitive rows (SASR, the two CLIP rows, CE and the clean oracle) on a leak-free subset, while the pre-CLIP LNL baselines and the VLM remedies are not re-scored.

Table 3: Protocol and hyperparameters. Every row is shared by all methods unless stated.

Item	Setting
Data	Oxford Flowers102 (102 classes; train+val 2040, test 6149), CUB-200-2011 (200; 5994 / 5794) and Stanford Cars (196; 8144 / 8041), official splits; class names are used verbatim to build the text prototypes and the twin map τ
Backbone	CLIP ViT-B/32 with the image and text encoders frozen; only a residual image-side adapter and the class prototypes c_k are trained
Optimisation	AdamW, adapter lr 5×10^{-4}, prototype lr 5×10^{-5}, 20 epochs, batch size 64, temperature $\tau = 0.07$
Training noise	semantic hard confusion at rate 0.5 (ablation 0.3 / 0.7)
Test noise	applied to the *same* predictions at $r \in \{0, 0.1, \ldots, 0.7\}$ with an independent random seed; no retraining, so every extra noise level is free
Methods	CE; GCE ($q = 0.7$); ELR ($\beta = 3.0$, $\alpha = 0.9$, five warm-up epochs); DivideMix (dual networks, GMM split, co-divide and co-guessing intact, five warm-up epochs, pseudo-label threshold 0.5); kNN label propagation ($k = 10$, $\alpha = 0.9$); a clean oracle trained on ground-truth labels; and SASR

B PAD CONTAINS THE CLASSICAL CORRECTION, AND WHY THAT MATTERS

B.1 SETUP

Let $\tilde{y}$ be the given (noisy) test label with noise rate r and twin map τ, and let $u = \mathbb{P}(\hat{y} = \tilde{y})$, $v = \mathbb{P}(\hat{y} = \tau(\tilde{y}))$ be the two observable agreement rates. On mutual twins, (5)–(6) give

$$u = (1-r)A + r\,\mathrm{NF}, \qquad v = rA + (1-r)\,\mathrm{NF}, \qquad \hat{A} = \frac{(1-r)u - rv}{1-2r}, \qquad \hat{\mathrm{NF}} = \frac{(1-r)v - ru}{1-2r}. \tag{10}$$

B.2 THE CLASSICAL CORRECTION AS A SPECIAL CASE

The classical correction (Hui & Walter, 1980) assumes that model errors and annotation errors are independent. Under that assumption a wrong prediction is uninformative about the annotation, so for K classes $\mathrm{NF}_{\mathrm{ind}} = (1-A)/(K-1)$. Substituting into the first line of (5) gives the familiar closed form, which for $K = 2$ is the textbook $(u-r)/(1-2r)$:

$$A_{\mathrm{cls}}(K) = \frac{u(K-1) - r}{(1-r)(K-1) - r}, \qquad A_{\mathrm{cls}}(2) = \frac{u-r}{1-2r}. \tag{11}$$

Proposition 2. *If independence holds, $\hat{A} = A_{cls}(K) = A$ exactly.*

Proof. Under independence, $\mathrm{NF} = (1-A)/(K-1)$. The second line of (5) then pins the second observable to $v = (1-r)\frac{1-A}{K-1} + rA$. Substituting this v into $\hat{A}$:

$$\begin{aligned} \hat{A} &= \frac{(1-r)\left[(1-r)A + \frac{r(1-A)}{K-1}\right] - r\left[\frac{(1-r)(1-A)}{K-1} + rA\right]}{1-2r} \\ &= \frac{A\left[(1-r)^2 - r^2\right]}{1-2r} \;=\; A. \end{aligned} \tag{12}$$

The same value follows from $A_{\mathrm{cls}}(K)$ by construction. □

So PAD is not an alternative to the classical estimator but its extension: the only difference is that PAD *measures* the conditional dependence through v whereas the classical correction *assumes* it away. When the assumption holds the two coincide; when it fails—as it must when the evaluated method was trained to fit the same noise that corrupts the test set—the classical estimator is biased and PAD is not (Table 11 reports a mean bias of $+0.10$ for the classical estimator versus $+0.007$ for PAD on our logs, using identical inputs).

B.3 Where the novelty is, and where it is not

Recorded for the reviewers' benefit, since several ingredients are classical.

Ingredient	Status
Correcting an accuracy estimate for evaluator error	Classical (Hui & Walter, 1980); review in Collins & Huynh (2014)
Estimating accuracy with no labels from agreement rates	Platanios et al. (2014; 2016); extended with inter-class logical constraints in Platanios et al. (2017)
Noticing that structured errors can flip model selection on test sets	Northcutt et al. (2021), described qualitatively for model capacity
Using class semantics as the noise *direction*	Feng et al. (2024) for selection under "hard noise"; Sosea & Caragea (2026) for a margin criterion
A closed-form estimator that uses the noise direction instead of assuming error independence, for a single model with noisy labels, with an explicit identifiability condition	**This work**
The exact two-term law, its closed-form crossover, the negative symmetric control, and the $r \to 1/2$ limit	**This work**

Two differences from the closest ML line (Platanios et al., 2014; 2016; 2017) are worth stating plainly: they require *several competing classifiers* and *no labels at all* and solve for accuracies by optimisation; PAD requires *one* model, *noisy* labels, and a single class-name lexicon, and is closed-form. Two differences from the closest evaluation line (Northcutt et al., 2021) are stated in Section 2: they study low-rate unstructured natural errors and conclude rankings are largely stable, whereas we study structured noise aligned with the training noise and give the law that quantifies when stability fails.

C Number provenance

Every quantitative claim is traceable to a log file under `logs/` either directly (regex over `[RESULT]`, `[TESTNOISE]` and `[DEBIAS*]` lines) or through one of the analysis scripts, which are deterministic.

The shell drivers above write one log per run under `logs/`; the analysis scripts read them and are deterministic, so every table entry in this paper can be regenerated exactly. All reported numbers were re-derived from those logs.

D Detailed related work

Learning with noisy labels. The field's canonical recipe assumes random or class-dependent noise: robust losses (GCE), regularisation of memorisation (ELR), sample selection by small loss (Co-teaching), and semi-supervised division of clean and noisy subsets (DivideMix). These are the baselines we evaluate. Our finding is that their shared assumption, that *the model's own signal* is informative about which labels are wrong, breaks precisely when the wrong label is a semantic twin.

CLIP/VLM as an external arbiter. CRoF (Deng et al., 2024) fine-tunes CLIP on noisy few-shot data with a multi-label smoothing loss; WTS (Li et al., 2026b) uses a VLM text encoder to correct label-image inconsistency in long-tailed high-noise settings; LANE (Sosea & Caragea, 2026) combines class-semantic relations with the *training dynamics* of the model—a margin criterion—to down-weight suspected noisy text examples. VisPrompt (Geng et al., 2026) shares our frozen-backbone, parameter-efficient form—it injects instance-level visual semantics into the prompt through cross-modal attention—whereas our anchor is class-name semantics read off a frozen text encoder. These works confirm that semantic structure is the right prior, but they either gate corrections by a threshold or modulate a signal computed from the model being trained. We instead

Claim / table	Driver (see repository)	Extraction
Tab. 1, SASR	`run_all_rerun.sh`, seeds 42/43/44	`[RESULT]peak=`, averaged
Tab. 1, CE/Clean/ELR/GCE/DivideMix	`run_all_rerun.sh`, seed 42	`[RESULT]`
Tab. 1, zero-shot	same runs	`[RESULT]zs=`
Tab. 1, kNN on Cars196	`run_testnoise5.sh`	`[TESTNOISE]clean_acc=`
Noise-rate ablation	`run_all_rerun.sh`	`[RESULT]peak=`
Noise-type control	`run_all_rerun.sh` (sym runs)	`[RESULT]`
Component ablation	`run_all_rerun.sh`	`[RESULT]peak=`
Tab. 4, fit statistics	`run_testnoise{,2,3,5}.sh`	`plot_testnoise.py`
Tab. 10	`run_testnoise{3,5}.sh`	`[TESTNOISE]` rows
Symmetric control (NF at chance)	`run_testnoise2.sh`	`[TESTNOISE]` rows
Multi-seed curves	`run_testnoise2.sh`	`[TESTNOISE]` rows
Mutual-twin counts	—	`check_debias.py`
Selection-bias figures	`run_testnoise5.sh`	`[DEBIASr]` rows
Restricted-protocol rates	`run_testnoise5.sh`	`[DEBIASRr]` rows
Estimator-error table	`run_testnoise5.sh`	`plot_pad.py`
Tab. 11	`run_testnoise5.sh`	`classical_vs_pad.py`
Sensitivity law, tolerances	`run_testnoise5.sh` + check	`plot_r_sensitivity.py`
Tab. 6, Fig. 3	`run_testnoise5.sh`	`r_budget.py`
Weights, cleanw	`run_testnoise5.sh`	`w(clean)=` / `w(noisy)=`
Fig. 1–7	—	`plot_testnoise.py`
Fig. 9–10	—	`plot_pad.py`, `plot_r_sensitivity.py`
Fig. 4, 5	—	`plot_paper_figs.py`
Partial-alignment law, §G.2	`run_align.sh`	`analyze_align.py`
Fig. 8	—	`plot_align.py`
Tab. 9, Fig. 6	`run_anchor.sh`	`plot_anchor.py`
Tab. 5, Fig. 2	`src/text_noise.py`	in-script
Real-noise strata, §7.4, Tab. 12	`run_c100_stratum.sh`	`src/c100_stratum.py`
Strata robustness controls, §7.4	`run_c100_stratum2.sh`	`src/c100_stratum2.py`
Real-noise PAD, Tab. 13	`run_c10n_pad.sh`	`src/c10n_pad.py`
VLM baselines, §6.1, Tabs. 1, 7	`run_vlm_baseline.sh` + multiseed	`src/baseline_main.py`
Near-duplicate control, §6.2, Tab. 8	`run_leak_mask.sh`, `run_leak_a_*.sh`	`src/leak_a_summary.py`
Real annotation (ImageNet ReaL): alignment, §7.4	`src/real_fetch.py` (→ `data_real/`)	`src/real_labels_analysis.py`, `src/real_align_arbiter.py` (→ `logs/real_align_arbiter.txt`)
Real annotation (ImageNet ReaL): distortion, §7.4	`src/real_fetch_val_zip.py` (val images)	`src/real_prep.py` (→ `logs/real_prep.txt`), `src/real_feats.py`, `src/real_nf.py`
Training-side imbalance, App. A.5, Tab. 2	`src/lt_bench.py`, `src/lt_eval.py`	`logs/lt_eval_R*.txt`
Released tool, §G.3	`check_my_benchmark.py`	`logs/cmb_selftest.txt`

use a *frozen* anchor as a purely external judge, apply continuous steep reweighting, and purify the classifier prototypes with the same weights, and we provide the diagnostic evidence (weight separation between clean and mislabelled samples, and a noise-type control) that distinguishes this route from data-internal statistics. We do not aim to advance this line: §6.1 shows that a plain CLIP hard-selection baseline already matches SASR on CUB-200 and beats it on Cars196. We borrow the external-arbiter paradigm instead to expose why data-internal strategies collapse under semantic-confusion noise, and thereby why it is the evaluation protocol, not the denoiser, that needs fixing.

Label errors in test sets. Northcutt et al. (2021) established that test sets of ten common benchmarks contain pervasive label errors ($\geq 3.3\%$ on average; 5.83% in ImageNet val; 5.85% in CIFAR-100), that the noise is *systematic* rather than random ("a tiger is more likely to be mislabeled cheetah than CD player"), and that *removing or correcting* these errors leaves relative rankings of models largely unchanged. Most relevant to us, they also report the mechanism we formalise: results are *unstable* because higher-capacity models "reflect the distribution of systematic label errors in their predictions to a greater degree", so a model can score *worse* on corrected labels while scoring *better* on the originally given ones, and they show that model selection would flip if the mislabeled prevalence increased by only 5–6% (ResNet-18 over ResNet-50 on ImageNet, VGG-11 over VGG-19 on CIFAR-10).

We therefore do **not** claim to be the first to notice that test-set label errors can perturb the conclusions drawn from a benchmark. They have been seen to destabilise capacity comparisons (Northcutt et al., 2021) and to reverse fairness conclusions on biased test data (Karakatič et al., 2026). Our contribution is to turn these observations into a law that also explains their magnitude, and to extend it from model capacity to *training objectives*:

1. **A quantitative law.** Northcutt et al. (2021) describe noise fitting qualitatively and show that ranking flips are *possible* at small incremental noise. Equation (3) makes this exact, and (4) shows that their observation is the small-ΔA limit of a single expression: $r^\star = \Delta A/(\Delta A + \Delta\mathrm{NF})$. Capacity variants differ mainly in NF, not in A, so their $r^\star$ is small; the LNL methods we compare differ in A by up to 20 points and therefore cross later (median 0.52). One law, two regimes.
2. **Arbitrary methods, not just capacity.** They compare model families of different sizes under *natural, unstructured* noise at 0.5–6%. We compare *methods*, several of which are explicitly optimised to fit the training noise, and show that the reward for noise fitting is a measurable per-method quantity, NF, anti-correlated with true quality (pooled Spearman -0.914). This is the point of contact with practice: the LNL leaderboard's best entries are the largest victims.
3. **Specificity.** Our symmetric-noise control yields NF at chance and zero inversions, so the effect requires *structured* noise aligned with the training noise, a testable condition absent from prior discussions.
4. **A repair and its limit.** Re-annotation (their recommendation, also REVEAL's) is one remedy; we give a label-free closed-form alternative, PAD, and prove that both run into the same identifiability limit at $r \to 0.5$.

Fault Lines (Jackson et al., 2025) benchmarks robustness to corrupted labels on the training side. REVEAL (Pang et al., 2025) renovates noisy and missing labels in image-classification test sets using VLMs, and it motivates that framework with exactly our starting concern—that such errors "result in misleading model comparisons and unfair evaluations"—but it quantifies the *quality of the datasets* it produces, not any change in method *rankings*, and its remedy is re-annotation. A recent detection benchmark (Pickler et al., 2025) notes that label noise affects validation as well as training, so that clean data are needed for reliable evaluation, but does not quantify its consequences for method rankings. All of these treat the test set as something to be *cleaned*; we treat its noise as something to be *measured and corrected for*, with class names as the only external input.

Estimating accuracy when labels are imperfect. That a classifier's measured accuracy is biased by annotation error, and that the bias can be removed, is a long-standing result outside ML. The latent-class formulation of Hui & Walter (1980) estimates the accuracy of tests *without* a perfect gold standard, assuming that errors are conditionally independent; the same independence assumption underlies the standard closed-form correction and, in ML, the family of methods that estimate classifier accuracy from unlabeled data via agreement rates (Platanios et al., 2014), its Bayesian and nonparametric extensions (Platanios et al., 2016), and the use of logical constraints between classes to sharpen the estimate (Platanios et al., 2017). In the medical literature it has long been acknowledged that conditional independence "is rarely thought to be valid" and that conditional dependence must be modelled (Collins & Huynh, 2014), and the identifiability of latent-class models has been characterised algebraically (Duan et al., 2019).

Our setting violates the independence assumption *by construction*: LNL methods are optimised on the same noise that appears in the test set, so model errors and annotation errors are positively corre-

lated, and the amount of correlation is exactly the quantity NF that drives the ranking inversion. PAD therefore does not assume independence; it *measures* the correlation through a second observable agreement rate $v = \mathbb{P}(\hat{y} = \tau(\tilde{y}))$, where the required direction τ comes from class-name semantics rather than from a second classifier. Appendix B shows that PAD reduces exactly to the classical correction when independence holds, and is exact when it does not. Two further differences are practical: PAD applies to a *single* model with noisy labels (Platanios et al. require several competing classifiers and no labels at all), and it is closed-form, so no EM or optimisation is involved. Its degenerate point at $r = 1/2$ is the LNL counterpart of the identifiability questions studied by Duan et al. (2019).

Two recent lines sharpen the contrast. Prediction-powered inference (Angelopoulos et al., 2023) corrects the bias of machine predictions by calibrating residuals on a *small set of gold-standard labels*, and Chen et al. (2026) extend that machinery, alongside Rogan–Gladen-style measurement-error corrections, to average benchmark scores and pairwise win rates. PAD sits at the opposite corner of that design space, needing no gold labels at all, at the cost of assuming that the noise is a semantic twin map τ that class names can supply. In benchmarking theory, Dorner & Hardt (2024) ask how to spend a budget of noisy labels when the goal is to identify the better of two classifiers; that is the annotation-side design question, whereas we ask how much test noise makes the comparison itself reverse.

Evaluation bias beyond classification. The concern is not confined to our setting. In algorithmic fairness, a 2026 study (Karakatič et al., 2026) shows that biased test data can mask true group fairness and let the same models support opposite conclusions, and notes how rare quantitative studies of the effect are. In LLM benchmarking, HLE-Verified (Zhai et al., 2026) shows that item-level defects depress accuracy and distort cross-model comparisons. In evolutionary multi-view classification, Li et al. (2026c) shows that label noise distorts the fitness ranking that steers the search—a "fitness evaluation bias"—and proposes a detect-then-calibrate purifier; the distortion there arises from training labels inside the search loop, whereas we study how a noisy *test* set reorders methods. These works motivate the problem but do not provide a mechanism, a ranking law or a debiasing estimator for the LNL setting.

Novelty statement. Northcutt et al. (2021) showed that test-set errors can disturb model *selection*, and described noise fitting qualitatively; REVEAL and Northcutt et al. both treat the remedy as re-annotating the test set; fairness and LLM benchmarking have reported evaluation bias without a mechanism. We contribute (i) an exact decomposition of measured accuracy into true quality and a measurable noise-fitting coefficient, (ii) the resulting closed-form crossover, which recovers the regime studied by Northcutt et al. as its small-ΔA limit and extends it to arbitrary methods with much larger $r^\star$, (iii) three fine-grained benchmarks and seven methods showing 58/63 ranking reversals with a negative symmetric-noise control, together with a replication on real crowd re-annotation with no synthetic noise at all (Section 7.4) and a real-noise validation of the estimator (Table 13), and (iv) a label-free closed-form debiasing estimator together with the proof that the worst regime, $r \to 0.5$, is unidentifiable and therefore unfixable by any test-label-only method. We claim no novelty for the method: Section 6.1 shows that a plain CLIP hard-selection baseline matches it on CUB-200 and beats it on Cars196, and we say so in the abstract.

E Scope and limitation details

E.1 Scope: the same law in a non-visual modality

Both the law and the inversion have so far lived in fine-grained images, where the confusion direction is supplied by a vision-language embedding. To separate the phenomenon from the visual encoder we repeat the experiment on 20 Newsgroups (Lang, 1995): 20 topics, 11,314 training and 7,532 test documents, TF-IDF features, and *no* pretrained model of any kind.

A confusion map from class names alone. τ pairs the 20 classes by the cosine similarity of their *names* (word-level plus character-3-gram TF-IDF) under a maximum-weight matching, giving ten twin pairs and an involution that is independent of any classifier. Eight of the ten pairs are genuine near-neighbours—`comp.windows.x` with `comp.os.mswindows.misc`, `comp.sys.ibm.pc.hardware` with `comp.sys.mac.hardware`, `rec.autos` with `rec.motorcycles`, `rec.sport.baseball` with `rec.sport.hockey`, `talk.politics.`-

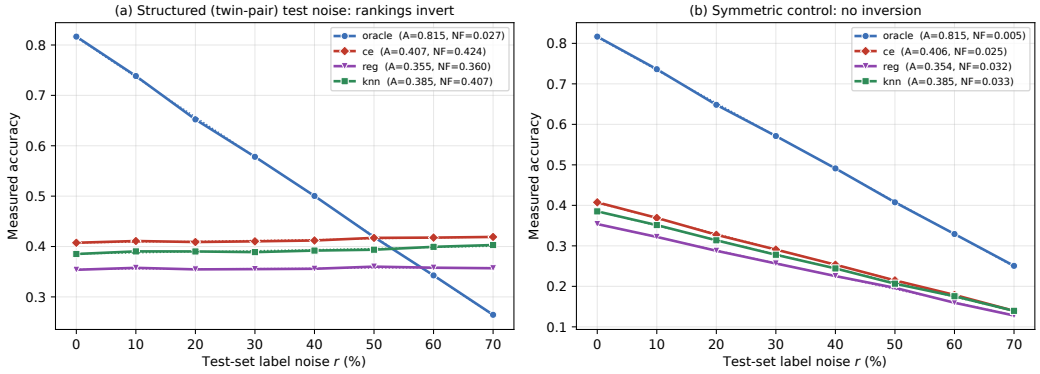

Figure 2: The law without vision: 20 Newsgroups, TF-IDF features, no pretrained model. (a) Structured (twin-pair) test noise: the oracle's curve falls steeply while the noise-fitted methods are nearly flat (NF $\approx A$) and overtake it near $r = 0.5$. (b) Symmetric control with the identical models: every NF collapses towards chance and no ranking changes. Dotted lines are the fits of (3).

`mideast` with `talk.politics.misc`, and `soc.religion.christian` with `talk.religion.misc`; two are forced cross-topic pairings in which the names share no token.

Setting. Training labels are confused along τ at rate 0.5. Four methods are compared: a logistic-regression oracle trained on clean labels; cross-entropy on the noisy labels; a strongly regularised version of the same ($C = 0.05$); and cosine kNN on the noisy labels. Test labels are then corrupted along τ at rate r, exactly as in Section 7, with a symmetric-noise control.

Three things reproduce. First, the two-term law holds to within 0.003 in absolute accuracy for every method and both noise types. Second, the noise-fitting coefficient collapses to chance ($1/20 = 0.05$) under the symmetric control—from 0.424 to 0.025 for cross-entropy, a factor of 17—while the training labels are untouched, exactly as in Figure 7. Third, the rankings invert: under semantic test noise every noisy-trained method overtakes the oracle at $r^\star = 0.51$ (CE), 0.53 (kNN) and 0.58 (regularised CE), whereas under the symmetric control no pair crosses below $r = 0.94$. A model that is 41 points better than cross-entropy on a clean test set loses the comparison entirely past $r \approx 0.5$.

Two qualifications, stated plainly. The anti-correlation between A and NF is weak here (Spearman -0.2 over four methods) because the three noisy-trained methods form one cluster with NF $\geq A$; the coefficient separates *regimes*—clean-trained versus noise-fitted—rather than ranking methods within a regime, which is all the crossover law requires. And because two of the ten twin pairs are forced, the confusion map is less sharp than CLIP's on Flowers102; the effect survives anyway, which is the relevant point.

The law is therefore not a property of images, of CLIP, or of fine-grained recognition: given any class-level confusion map, it predicts the accuracy curve and the crossover of a bag-of-words linear model as well.

E.2 Limitations

1. **Identifiability at $r \to 1/2$.** By (6) the problem is singular at $r = 0.5$ and ill-conditioned nearby. Since the distortion is largest there, there is no test-label-only cure for the worst case. This is a property of the problem.
2. **The noise rate must be known.** PAD requires an estimate of r, and we do not propose an estimator of r itself; we take it from re-annotation, dual annotation, or a label-free agreement estimator (Platanios et al., 2014). What we do provide is the resulting cost: (15) and Table 6 give the number of images that must be re-annotated, which is under a hundred at the 3–6% error rates of curated test sets, but grows to 10^4 by $r \approx 0.4$ — the rate that real crowd re-annotation actually reaches (Section 7.4).
3. **Twin-map coverage.** A, NF and PAD are computed on mutual-twin classes, 29–52% of classes. Absolute accuracies on that subset differ from full-test accuracies, and the

estimator is quantified on that subset only: the mean $|\hat{A} - A|$ over $r = 0.1 \ldots 0.4$ ranges from 0.0039 to 0.0458 (Section 8). We do not measure how the intrinsic difficulty of the mutual-twin subset compares with the full test set, so $\hat{A}$ should be read as an estimate of the subset's A and not of the full-benchmark accuracy.

4. **How far real noise goes.** Section 7.4 measures the field rates directly, and they are higher than the ones usually quoted: crowd re-annotation errs on 40.2% of CIFAR-100N training images, and the per-class rate spans 0.085 to 0.863. The CIFAR-10N figures usually quoted are aggregates over its whole training set (9% for the majority vote, 17–18% per single annotator, 40% for its worst); measured on *our* twin-class evaluation subset the same label sets give $r = 0.105$, 0.207 and 0.444 (Table 13), and it is those we use. Prevalence is therefore not the binding constraint, and neither is alignment: two independent CIFAR-10N annotators agree on the top-1 confusion partner of every class ($r = 0.987$), so ρ is high for real annotators and (17) is close to the fully-aligned threshold. At a thousand classes, however, we measure $\rho = 0.53$–0.77 on ImageNet ReaL (Section 7.4), significant but well below the ten-class value; the crossover should be evaluated at that ρ, not at one, when the class count is large. What does bound the distortion is that human error is structured in *direction* but not reproducible *per image*: the largest NF any method reaches on the full held-out CIFAR-100N pool is 0.128, rising to 0.157 on its noisiest ten-class stratum, against 0.42 under CLIP-aligned synthetic noise, and on the 1000-class ImageNet ReaL set it falls further still, to 0.028–0.050 (Section 7.4), and a $k = 1$ neighbour rule that reproduces its training labels perfectly (0.9997) transfers none of that to unseen images ($\mathrm{NF} = 0.103$). The real-noise inversions are real and significant — a 13.4-point clean-test deficit is reversed into a 3.8-point lead in the noisiest class stratum ($p = 0.006$) — but bounded, against the 9.3 points gained outright under synthetic noise in Table 10. Our claim remains conditional on shared annotation structure, which Section G.2 quantifies.

5. **Scope.** One backbone (CLIP ViT-B/32), three fine-grained image datasets, one noise family and one twin-map construction. Section 9 shows that the law itself also holds for a bag-of-words text model with a confusion map derived from class names alone, so it is not tied to vision; whether other structured confusion families (e.g. hierarchical or attribute-based) behave identically is open. SASR, by contrast, is bounded by its anchor: Section 6.3 shows that its advantage over cross-entropy vanishes once roughly three-quarters of the class anchors are wrong, and that a completely broken anchor leaves it *below* plain cross-entropy. The method needs a lexicon whose embeddings separate the classes; the diagnostic does not. A second bound is that only 18 method pairs sit below their stratum's noise rate on real data, so the real-noise inversions are fewer than the $58/63$ the synthetic setting supplies. Finally, the coarse-class evidence (CIFAR-10N/100N, 10 and 100 classes) is supplemented rather than replaced by a fine-grained set: ImageNet ReaL, a public 1000-class re-annotation, closes both halves of the granularity gap, and the alignment comes out significant but weaker ($\rho = 0.53$–0.77 against 0.987 at ten classes), and on that set the distortion, measured from per-image model predictions, is $\mathrm{NF} = 0.028$–0.050 — above chance, but 2.6–$4.6\times$ below CIFAR-10N and 8–$15\times$ below the synthetic protocol, while the correction still recovers the human-verified accuracy to within 0.013. Two caveats keep this from being a full replication: ReaL annotates ImageNet val, so at 28 training images per class the trained variants underperform a zero-shot head and no method ranking can be read off it; and the upstream annotation procedure is ImageNet's own, not one we control. Training-side class imbalance is likewise out of scope, and deliberately so: Appendix A.5 shows it is a separate axis whose remedy (a prior-based correction, worth $+1.9$ to $+3.8$ points) does not compose with the frozen-anchor weights, so nothing here should be read as a long-tail result.

6. **Relaxing τ from an involution to a distribution.** Both the law and PAD assume a deterministic confusion partner: the class's nearest neighbour. A more faithful model lets a sample flip to any of its K nearest classes with probabilities that depend on the sample. We do not analyse that relaxation, and two of our own measurements indicate both why it matters and how far the top-1 idealisation reaches. Where the map is exact by construction — synthetic confusion along τ — it is exact; on human error it is only approximate, which is what the alignment ρ and the $k = 1$ neighbour check quantify (Section 7.4): at ten classes two annotators agree on the top-1 confusion partner of every class ($r = 0.987$), but at a thousand classes ρ falls to 0.53–0.77, and a $k = 1$ rule that reproduces its training labels

Table 4: Fitted clean accuracy A and noise-fitting coefficient NF per method and dataset (the intercept and slope of (3)).

Method	Flowers102	CUB-200	Cars196
Clean oracle	0.942 / 0.005	0.759 / 0.049	0.788 / 0.059
SASR	0.769 / 0.111	0.543 / 0.213	0.581 / 0.236
ELR	0.575 / 0.308	0.386 / 0.399	0.462 / 0.378
GCE	0.567 / 0.310	0.395 / 0.378	0.458 / 0.366
CE	0.565 / 0.329	0.408 / 0.372	0.451 / 0.381
kNN	0.531 / 0.360	0.356 / 0.418	0.415 / 0.421
DivideMix	0.383 / 0.396	0.294 / 0.414	0.393 / 0.370

Table 5: 20 Newsgroups, 20 classes, TF-IDF, no pretrained model (Section 9). Fitted accuracy curves $\mathrm{Acc}(r) = (1-r)A + r\,\mathrm{NF}$ under semantic (twin-pair) and symmetric test noise. RMSE is the absolute fit error; R^2 is uninformative for the noise-fitting methods because their curves are almost flat.

	Semantic test noise			Symmetric control		
Method	A	NF	RMSE	A	NF	RMSE
Oracle (clean labels)	0.815	0.027	0.0023	0.815	0.005	0.0023
CE (noisy labels)	0.407	0.424	0.0015	0.406	0.025	0.0013
CE, strongly regularised	0.355	0.360	0.0016	0.354	0.032	0.0015
kNN (noisy labels)	0.385	0.407	0.0021	0.385	0.033	0.0014

at 0.9997 transfers only $\mathrm{NF} = 0.103$ to unseen images. Technically, once flips spread over a neighbourhood the involution disappears and the two-equation system (5)–(6) no longer closes, so the estimator would have to be rebuilt; we leave that to future work.

F Additional tables and figures

G Additional analyses

G.1 Sensitivity to the noise-rate estimate

PAD needs r. Differentiating (6) and substituting (5) gives, exactly,

$$\left.\frac{\partial \hat{A}}{\partial \hat{r}}\right|_{\hat{r}=r} = \frac{A - \mathrm{NF}}{1-2r}. \tag{13}$$

The quantity $A - \mathrm{NF}$ is the method's own sensitivity to noise (the slope of (3)), amplified by the debiasing. Stronger methods and noise rates closer to $1/2$ demand a more accurate r. Requiring $|\hat{A} - A| \le \epsilon$ yields the budget

$$|\hat{r} - r| \;\le\; \frac{\epsilon\,(1-2r)}{A - \mathrm{NF}}. \tag{14}$$

From tolerance to annotation budget. The bound becomes a concrete cost when $\hat{r}$ is estimated by re-annotating n randomly chosen test images: then $\hat{r} = k/n$ with $k \sim \mathrm{Bin}(n, r)$, so $|\hat{r} - r| \simeq z\sqrt{r(1-r)/n}$ at confidence level z, and substituting into (14) gives the annotation budget

$$n^\star(r, \epsilon) \;\approx\; \frac{z^2\, r(1-r)}{\epsilon^2} \left(\frac{A - \mathrm{NF}}{1-2r}\right)^2. \tag{15}$$

Equation (15) is the local, linearised law. Because (6) is a *rational* function of $\hat{r}$, its curvature makes the exact budget larger as $r \to 1/2$: we obtain the exact value by solving $\hat{A}(\hat{r}) = A \pm \epsilon$ in closed form and inverting the binomial coverage of $\hat{r}$. Figure 3 reports both, together with a simulation check (exact and Monte-Carlo coverage agree to within 0.004).

Table 6: Images that must be re-annotated so that PAD's accuracy estimate on the mutual-twin subset lies within ϵ of the truth, at 95% confidence (Flowers102). Parentheses give the linearised law (15); the exact figure is the binomial coverage of (6). The cost grows with r, with the strength of the method, and with the curvature that (15) ignores.

Method	ϵ	$r = 0.10$	0.19	0.29	0.39
SASR (ours)	0.02	195 (173)	501 (484)	1776 (1360)	15396 (5659)
	0.05	47 (28)	81 (78)	314 (218)	1584 (906)
Clean oracle	0.02	1260 (1136)	3518 (3193)	12271 (8976)	67224 (37356)
	0.05	195 (182)	563 (511)	1413 (1437)	6211 (5977)

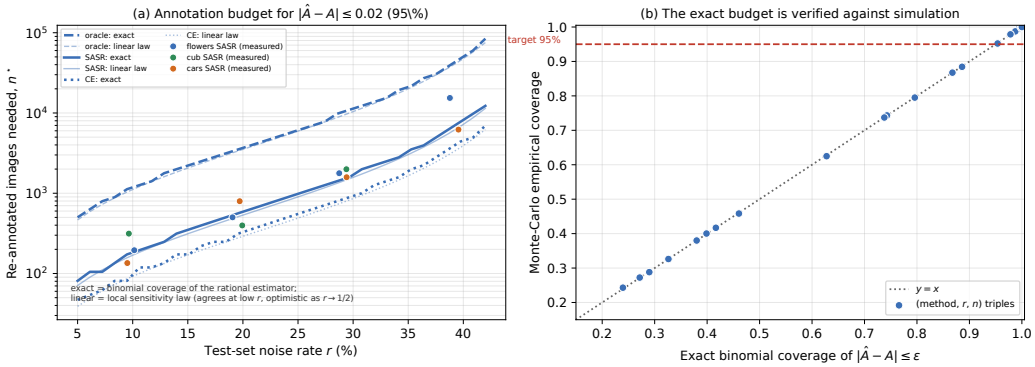

Figure 3: How many images must be re-annotated. (a) Exact annotation budget for $|\hat{A} - A| \le 0.02$ at 95% confidence (thick) against the linearised law (15) (thin), for the three Flowers102 methods; markers are the measured budgets of the SASR models on all three datasets. The two agree at low r and diverge as $r \to 1/2$, where the curvature of (6) dominates. The budget rises with r and with method strength ($A - \mathrm{NF}$). (b) The exact binomial coverage agrees with a Monte-Carlo simulation to within 0.004, validating the inversion used in (a).

Two consequences are practical. First, at the 3–6% error rates reported for natural fine-grained test sets, the budget is trivial: fewer than a hundred images for SASR and about five hundred for the oracle. The low-error-rate regime therefore needs no new infrastructure, which is the regime in which our control experiments also predict a mild distortion. Second, the cost grows sharply with the noise rate—from ∼500 images at $r = 0.19$ to ∼15,000 at $r = 0.39$ for SASR—because the same $(1 - 2r)^{-1}$ that makes debiasing most valuable at high r also makes it most fragile, and because a *stronger* method (larger $A - \mathrm{NF}$) is more expensive to certify. Any estimate of r may be substituted into (15): re-annotation and independent dual annotation both give a binomial estimate, and the label-free agreement estimators of Platanios et al. (Platanios et al., 2014; 2016) may be used when no re-annotation budget exists at all.

G.2 Partial alignment: inversion needs a shared annotation process

The inversion above presumes that test errors point in the same direction as training errors, i.e. that the two sets were annotated by the same procedure. How much alignment is needed? Let a flipped test label be drawn from the same transition τ with probability ρ and uniformly from the other $K-1$ classes otherwise, so that $\rho = 1$ is the setting of this section and $\rho = 0$ reduces to the symmetric control above. For a method whose agreement with the noise is mediated by τ, the second term of (3) then decays linearly,

$$\mathrm{NF}(\rho) \;=\; \rho\,\mathrm{NF}(1) \;+\; (1-\rho)\,\frac{1-A}{K-1}, \tag{16}$$

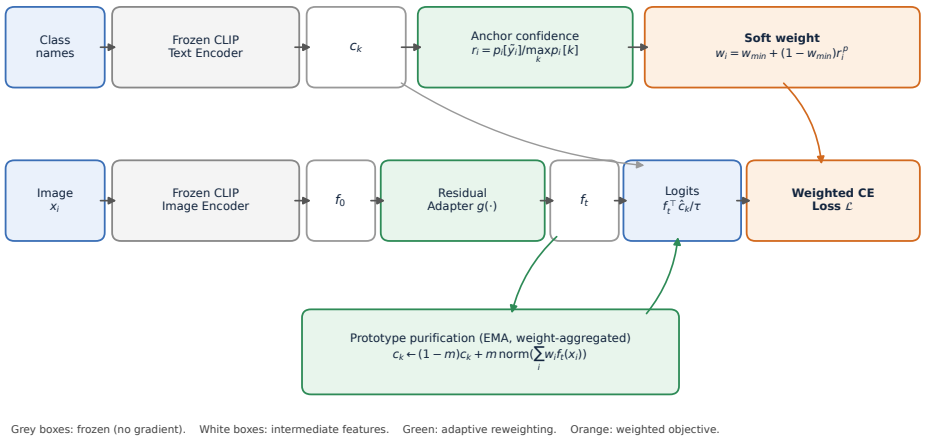

Figure 4: SASR overview: a frozen CLIP text anchor produces a per-sample relative confidence ratio, which is turned into a steep soft weight; the same weights drive both the weighted cross-entropy and the prototype EMA. Full specification in Appendix A.

Table 7: Our reimplementations of two contemporary CLIP-based remedies, under the protocol of Section 5. Three-seed means of peak accuracy.

Method	Flowers102	CUB-200	Cars196
SASR (ours)	**0.770**	0.557	0.593
CLIP selection	0.710	0.557	**0.604**
CLIP pseudo-labels	0.667	**0.559**	0.585

where the last term is the chance agreement with a uniform flip; at $K \geq 102$ it is at most 0.005 and negligible. Substituting into (4) gives the inversion threshold

$$r^\star(\rho) \;=\; \frac{\Delta A}{\Delta A + \rho\,\Delta \mathrm{NF}}. \tag{17}$$

Partial alignment therefore *delays* inversion, and it does so substantially: on Flowers102 the oracle–CE threshold moves from 0.54 at full alignment to 0.70 at $\rho = 1/2$, and at $\rho = 0$ no method pair reverses anywhere in $r \leq 1$.

Figure 8 verifies (16) and (17) on all three datasets, three methods and five alignment levels. Two internal checks validate the implementation, because both have known answers: the fitted clean term $A(\rho)$ is independent of ρ ($|\Delta A| \leq 0.009$, the same checkpoint being evaluated throughout), and the $\rho = 1$ curves reproduce the fits of Table 4 to within 0.027 (single-seed against that table's three-seed means: Flowers102 SASR 0.748 against 0.769, CE 0.538 against 0.565). Normalised by its fully-aligned value, NF tracks the diagonal ρ for every method–dataset pair (averaged over all pairs, $0.76/0.52/0.27$ at $\rho = 0.75/0.5/0.25$, against 0 for a chance-level term), and the measured thresholds track (17) to within 0.03 wherever $r^\star < 1$. A synthetic simulation with a known ground truth reproduces the same linearity of (16) to within 0.002, and endpoints match analytically: at $\rho = 0$ the flip hits a twin class at the chance rate $1/(K-1)$.

Consequence. What triggers the inversion is not a noisy test set, but a test set annotated by the *same* procedure as the training set. Where the two annotation processes are independent, structured test noise is no more harmful than symmetric noise of the same rate. This is why our claim is stated conditionally, and it tells a benchmark maintainer what to measure: the direction of test-set errors relative to the training confusion structure, not only their prevalence.

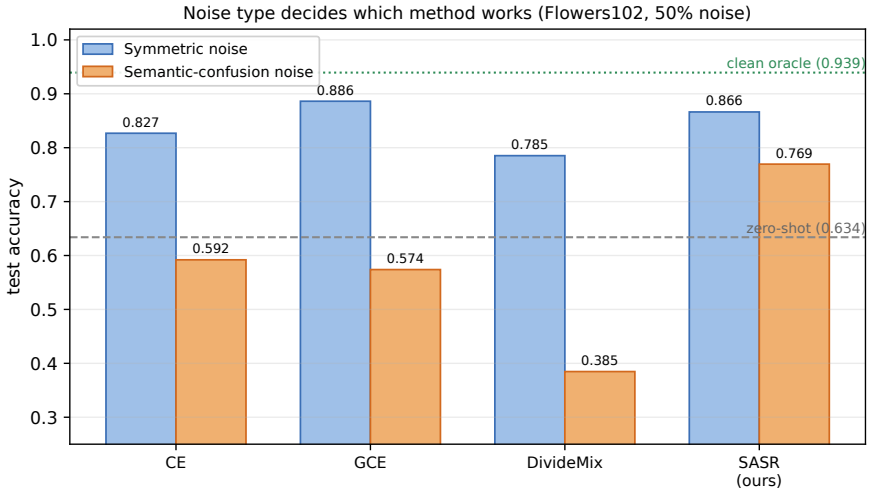

Figure 5: Symmetric versus semantic-confusion noise at the same rate. Under symmetric noise every method is comparable; under semantic confusion the data-internal methods collapse while the frozen-anchor method does not.

Table 8: Near-duplicate control. *full* is the protocol of Table 1 recomputed for this control; being a separate run it differs from that table by up to 1.3 points (CE: +0.4 Flowers102, +0.3 CUB-200, +1.3 Cars196), so every comparison is made *within* this table and the two are never mixed. *leak-free* re-scores the same predictions after removing every test image that has a $\cos \geq 0.95$ neighbour in the training pool; Δ is the accuracy that leakage contributes. Three seeds (SASR, the two CLIP rows, CE); one seed for the oracle.

Dataset	Method	full	leak-free	Δ
Flowers102	SASR	0.769	0.756	+1.4
	CLIP selection	0.710	0.704	+0.6
	CLIP pseudo-labels	0.667	0.657	+1.0
	CE	0.596	0.558	+3.8
	Clean oracle	0.939	0.916	+2.3
CUB-200	SASR	0.557	0.554	+0.3
	CE	0.417	0.414	+0.2
	Clean oracle	0.755	0.752	+0.4
Cars196	SASR	0.593	0.594	−0.1
	CE	0.463	0.463	0.0
	Clean oracle	0.785	0.787	−0.3

G.3 Reproducibility

Appendix C lists, table by table, the script and log that produce every number in this paper; each entry there is a file that exists in the released code.

The Flowers102 curves were reproduced with three seeds for SASR, CE and kNN and are visually indistinguishable (SASR clean accuracy 0.7746 and 0.7759 for seeds 43 and 44; accuracy at $r = 0.7$: 0.3090 and 0.3022).

Data. The ImageNet ReaL experiments use the public ILSVRC-2012 validation set (the devkit's `validation_ground_truth.txt` plus the image archive) together with ReaL's `real.json` and `raters.npz` (Beyer et al., 2020). Every script takes its ImageNet root from the `IMAGENET_ROOT` environment variable and its CLIP weights from `CLIP_ROOT`, so the pipeline runs outside our own filesystem layout. One caveat we hit and record for reproducers: the common

Table 9: Anchor-quality ablation on Flowers102 (50% semantic training noise, seed 42). q is the fraction of class anchors replaced by another class's text prototype; all rows share the cross-entropy baseline of Table 1 (0.592). The last row is a frozen *random* anchor: frozen, but carrying no semantics.

q	0	0.25	0.50	0.75	1.00
SASR peak accuracy	0.756	0.731	0.698	0.620	0.545
gain over CE	+16.4	+13.9	+10.6	+2.8	−4.8
anchor zero-shot	0.634	0.475	0.351	0.138	0.006
frozen random anchor	peak 0.562, gain −3.0, zero-shot 0.003				

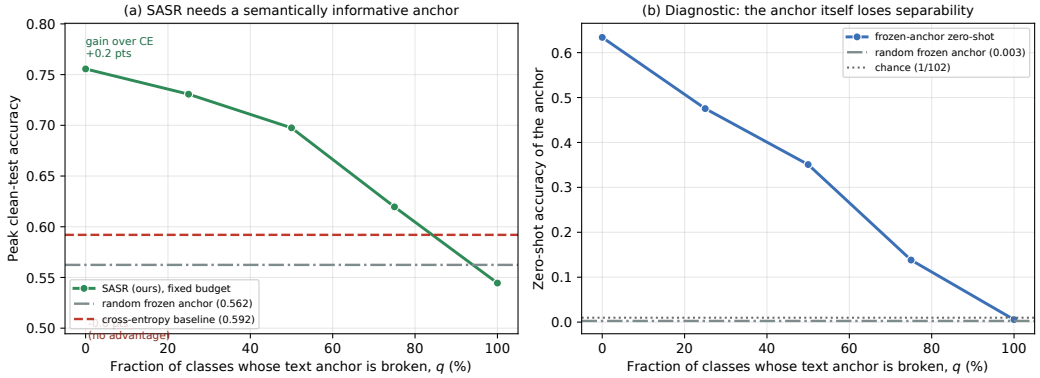

Figure 6: Anchor quality. (a) SASR's peak accuracy as a fraction q of the class anchors is replaced by another class's text prototype (dashed: cross-entropy baseline; dash-dotted: frozen random anchor). The advantage is monotone in anchor quality and vanishes at $q \approx 0.86$. (b) Diagnostic: the damaged anchor's own zero-shot accuracy, which falls from 0.634 to chance.

convention that line i of the devkit ground truth labels the i-th filename does *not* hold on the copy we used. Checked against `raters.npz`, which is keyed by filename and therefore independent of any line-order convention, the synset directory a file sits in agrees with the human votes: the directory label is among the proposed candidates for 100% of sampled files (median vote share 0.50) and is the human *top-1* partner for 0.5660 of them, whereas the filename-index convention reaches a top-1 match rate of 0.0006, i.e. chance. We take the given label from the directory and validate every step in `src/real_prep.py`; a reproducer whose ImageNet copy is laid out differently should re-run that agreement check before trusting any number above.

Released tool. To make the correction usable rather than merely demonstrated, we release `check_my_benchmark.py`. It takes a practitioner's predictions on a noisy test set, the given (noisy) labels, the measured noise rate and the twin map, and returns $(u, v, \hat{A}, \hat{\mathrm{NF}})$ together with the pairwise crossover $r^\star$ of (4) — the PAD estimate and the ranking-inversion threshold, with no clean labels, no retraining and no access to model internals. Its verification is a synthetic self-test with known (A, NF) (`logs/cmb_selftest.txt`), in which the recovery error stays at or below 0.0081 for $r \leq 0.4$; it implements the same estimator whose deployable error Table 13 reports as 0.043 at $r = 0.207$ (that figure is computed by `src/c10n_pad.py`, not by the tool). It computes the deployable estimate, i.e. conditioning on the given label, which is the only one a practitioner without clean labels can run; when clean labels happen to be available it additionally reports the restricted-protocol estimate and the realised error, as a check on the selection trap discussed above. It warns explicitly when $|1 - 2r|$ is small enough that Section 3.4 makes the estimate unusable.

Table 10: Accuracies that cannot happen on a clean test set: four method–dataset combinations whose reported accuracy is higher at 50%/70% test noise than on the clean test set.

Method / dataset	clean	$r = 0.5$	$r = 0.7$
DivideMix / CUB-200	0.296	0.351 (+5.5)	**0.389** (+9.3)
kNN / CUB-200	0.359	0.385 (+2.6)	**0.404** (+4.5)
ELR / CUB-200	0.392	0.392 (+0.1)	0.399 (+0.7)
DivideMix / Flowers102	0.386	0.392 (+0.6)	0.388 (+0.2)

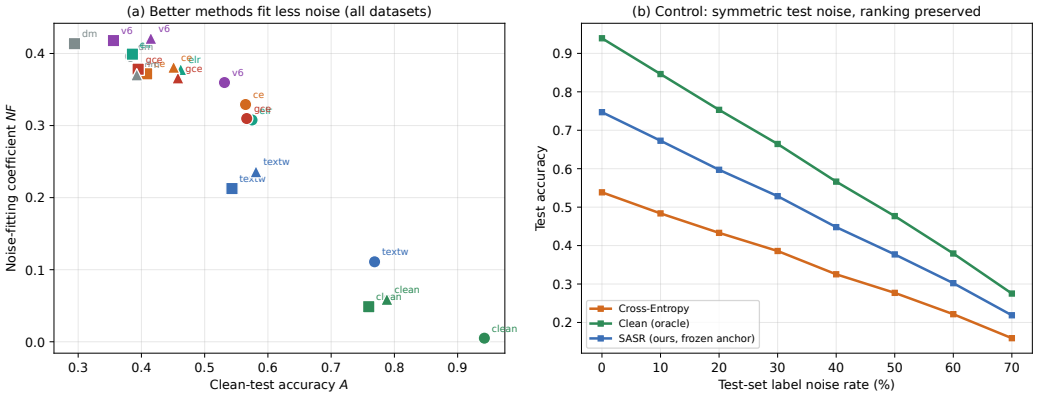

Figure 7: (a) The noise-fitting coefficient NF against clean accuracy A, over all 21 method–dataset pairs: quality and noise-fitting are anti-correlated. (b) Symmetric-noise control: NF is at chance and no ranking changes, so the inversion in Figure 1 is specific to structured noise.

H Near-duplicate control table

	Flowers102	CUB-200	Cars196
leaked test images ($\cos \geq 0.95$)	31.1%	3.2%	2.8%
test images with a $\cos \geq 0.90$ neighbour	93.0%	67.7%	45.2%
mean maximum similarity	0.937	0.909	0.892

I Additional result tables and full 20 Newsgroups detail

The law is not a property of images, of CLIP, or of fine-grained recognition. On 20 Newsgroups (Lang, 1995) (20 topics, TF-IDF, no pretrained model) with τ built from class names alone by maximum-weight matching, the two-term law holds to within 0.003 in absolute accuracy for every method and both noise types, NF collapses to chance under the symmetric control while the training labels are untouched ($0.424 \to 0.025$, a factor of 17), and the rankings invert at $r^\star = 0.51$ (CE), 0.53 (kNN) and 0.58 (regularised CE), whereas no pair crosses below $r = 0.94$ under the control: a model 41 points better on clean labels loses the comparison entirely past $r \approx 0.5$. The anti-correlation between A and NF is weak there (Spearman -0.2 over four methods) because the three noisy-trained methods form one cluster, so the coefficient separates *regimes* rather than ranking methods within a regime — which is all the crossover law requires.

J Full derivation and proofs

This appendix holds the complete derivation of Proposition 1, its proof, and one supporting analysis that the main text does not report.

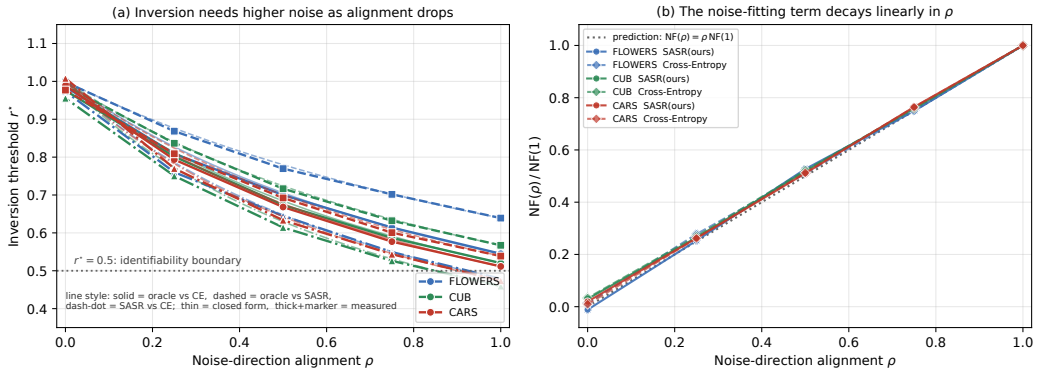

Figure 8: Noise-direction alignment ρ. (a) Inversion threshold $r^\star(\rho)$ for the three method pairs, measured (thick, markers) and closed form (17) (thin). Full alignment ($\rho = 1$) is the regime in which the disagreements of Section 7 appear at moderate noise; independent annotation ($\rho = 0$) leaves the ranking intact up to $r = 1$. (b) The noise-fitting term normalised by its fully-aligned value: the prediction $\mathrm{NF}(\rho) = \rho\,\mathrm{NF}(1)$ (dotted, exact to 0.03 here for the methods with significant NF) holds for every method and dataset.

Table 11: Absolute estimation error of the classical independence-based correction and of PAD, over 72 (method, dataset, noise rate) combinations. The classical estimator is granted its correct multi-class form.

Estimator	mean abs error	median abs error	mean bias
classical, binary form (K mis-set)	0.1345	0.0776	-0.1345
classical, K-class form (correct use)	0.1012	0.0870	$+0.1012$
PAD	**0.0143**	**0.0084**	$+$**0.0068**

J.1 Semantic-confusion noise and the twin map

Let the class set be $\{1..K\}$ with names $c_1, \ldots, c_K$. Encode each name with a frozen text encoder and let

$$\tau(k) = \arg\max_{j \neq k} \cos\big(\phi(c_k), \phi(c_j)\big) \tag{18}$$

be the **twin** of class k. Real annotator errors in fine-grained tasks concentrate on twin pairs, so we model the noisy label $\tilde{y}$ of an image with true label y as

$$\tilde{y} = \begin{cases} \tau(y) & \text{with probability } r, \\ y & \text{otherwise,} \end{cases} \tag{19}$$

where r is the (test- or train-set) noise rate. Unlike symmetric noise, this transition is *anisotropic* and, critically, can be *shared* between the training and test sets because both come from the same annotation process.

J.2 The accuracy decomposition

For any fixed model with prediction $\hat{y}$, define two noise-free quantities:

- $A = \mathbb{P}(\hat{y} = y)$, the clean accuracy (true quality);
- $\mathrm{NF} = \mathbb{P}(\hat{y} = \tau(y))$, the **noise-fitting coefficient**.

Proposition 3 (Accuracy decomposition)**.** *On a test set corrupted at rate r, the measured accuracy is*

$$\mathrm{Acc}(r) \;=\; (1 - r)\,A \;+\; r\,\mathrm{NF}. \tag{20}$$

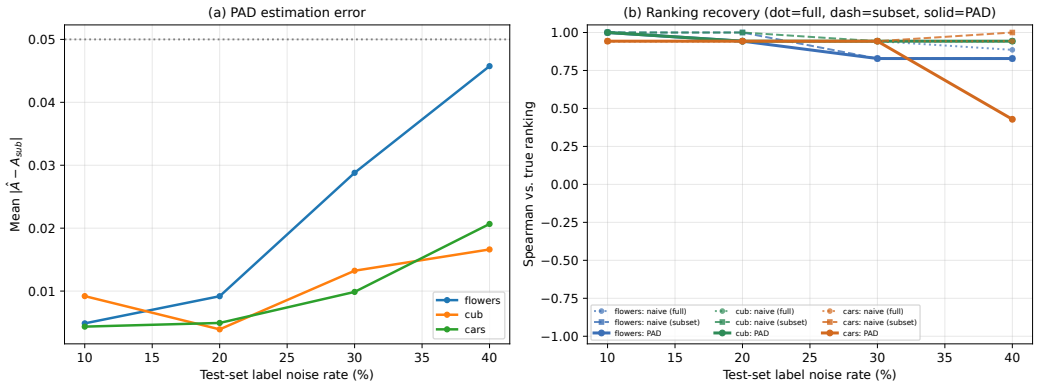

Figure 9: (a) PAD estimation error against the test-set noise rate and (b) ranking recovery (dashed: raw agreement, solid: PAD). PAD's error grows as $1/(1-2r)$ predicts.

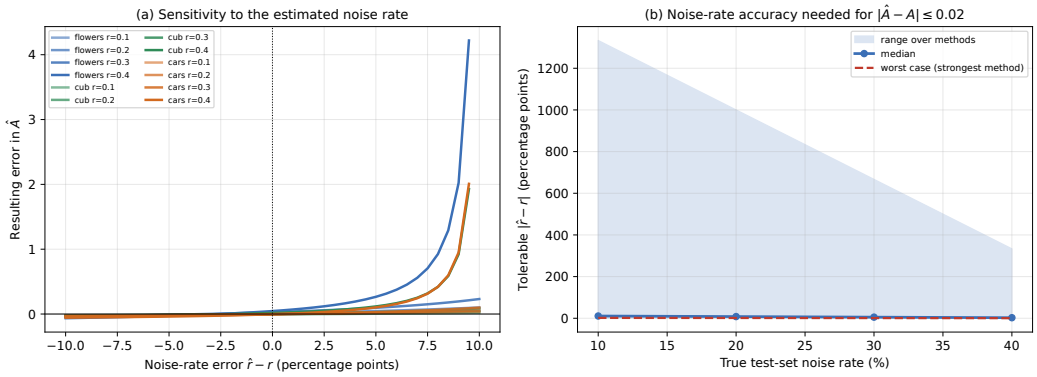

Figure 10: (a) Sensitivity of $\hat{A}$ to the noise-rate estimate; (b) the accuracy to which r must be known for $|\hat{A} - A| \le 0.02$, worst case and median over methods.

Proof. A sample is unflipped with probability $1 - r$; then "prediction equals the given label" is equivalent to $\hat{y} = y$, contributing A. It is flipped with probability r; then the given label is $\tau(y)$, and agreement means $\hat{y} = \tau(y)$, contributing NF. Linearity follows because A and NF do not depend on r. □

Two immediate consequences.

(a) NF **is a method-intrinsic constant.** Equation (3) is linear in r, and we verify empirically that the NF recovered from each noise level is nearly constant across $r \in [0.1, 0.7]$ (Section 7), i.e. NF is a property of the method, not of the noise level.

(b) Closed-form crossover. Comparing two methods a, b via (3):

$$r^{\star}_{ab} = \frac{A_a - A_b}{(A_a - A_b) + (\mathrm{NF}_b - \mathrm{NF}_a)}. \tag{21}$$

The ranking reverses at $r^{\star}_{ab}$. Since good methods learn the truth, their predictions are y rather than $\tau(y)$ and NF tends to chance, whereas methods that memorise noise are pulled towards $\tau(y)$. Empirically NF is anti-correlated with A, so a noisy test set systematically *rewards the worse method*.

J.3 Why symmetric noise cannot do this

If the noise is symmetric, the flip target is uniform over the remaining $K - 1$ classes and no method can anticipate the direction, so $\mathrm{NF} \to 1/K$ for every method and $\mathrm{Acc}(r) = (1 - r)A + r/K$, a

Table 12: Real human noise (CIFAR-100N crowd re-annotation; no synthetic component). Strata are defined by the crowd error rate per class in the *training* pool; accuracies are measured on held-out images of the same classes. Columns 4–5 are clean-test accuracies, columns 6–7 are measured on the human-annotated test set of the same images.

Stratum	Classes	r	oracle (clean)	SASR (clean)	SASR − oracle (noisy / clean)	McNemar p
all	100	0.397	0.703	0.677	$+1.6$ / -2.6	< 0.001
top-50	50	0.538	0.626	0.570	$+2.1$ / -5.6	< 0.001
top-30	30	0.620	0.576	0.487	$+1.9$ / -8.9	0.009
top-20	20	0.690	0.528	0.414	$+2.5$ / -11.4	0.006
top-10	10	0.763	0.431	0.297	$+3.8$ / -13.4	0.006

Table 13: PAD on real human annotation noise (CIFAR-10N). Every cell is the clean-oracle row — the best of the five trained models, and the row the surrounding prose reports. "naive" is the uncorrected accuracy u. The restricted column uses the protocol of this section (conditioning on the true class, an upper bound); the deployable column conditions on the given label and therefore admits the selection trap discussed above.

Label set	r	naive $\lvert u - A\rvert$	PAD $\lvert\hat{A} - A\rvert$	PAD, deployable
clean (reference)	0.000	0.0000	0.0000	0.0000
majority vote	0.105	0.0882	**0.0095**	0.0099
one other annotator	0.207	0.1803	**0.0283**	0.0430
worst annotator	0.444	0.3905	1.66 (undefined)	3.19

monotone transform of A sharing the factor $(1-r)$. Rankings are therefore *invariant by construction*. This yields a sharp falsifiable prediction that we use as a control: under symmetric test noise there should be no inversion, and the accuracy ratio between any two methods should be constant in r.

J.4 An identifiability limit

Suppose we observe only noisy labels and want A. Define the two observable agreement rates $u = \mathbb{P}(\hat{y} = \tilde{y})$ and $v = \mathbb{P}(\hat{y} = \tau(\tilde{y}))$. On classes that form *mutual* twins ($\tau(\tau(k)) = k$, so τ is an involution there) the two rates satisfy

$$u = (1 - r)A + r\,\mathrm{NF}, \qquad v = rA + (1 - r)\,\mathrm{NF}, \tag{22}$$

because a flipped sample's twin is exactly its true class. Solving (5),

$$\hat{A} = \frac{(1 - r)u - rv}{1 - 2r}, \qquad \hat{\mathrm{NF}} = \frac{(1 - r)v - ru}{1 - 2r}. \tag{23}$$

The system is singular at $r = 1/2$: A and NF become unidentifiable. This is not an artefact of our estimator. It says that when the test set is about half mislabelled, no method that sees only test labels can separate true quality from noise fitting. Combining this with (4), the regime where ranking distortion is worst ($r \to 1/2$) is exactly the regime where it cannot be corrected.

Remark 2 (Relation to the classical correction)**.** *The standard correction for an imperfect evaluator, going back to the latent-class model of Hui & Walter (1980), assumes that model errors and annotation errors are* independent*. Under that assumption, in the binary case, the second observable is pinned down to $v = (1 - r)(1 - A) + rA$, and substituting it into* (6) *yields $\hat{A} = A$ exactly; the same value is obtained from the classical closed form $A = (u - r)/(1 - 2r)$. The proof is given in Appendix B. PAD therefore* contains *the classical estimator as the independent-error special case. The difference matters here: LNL methods are trained on the same noise that appears in the test set, so independence fails precisely to the extent that* NF *is elevated, and the classical correction then returns a biased answer while* (6) *remains exact. In this reading,* NF *measures the conditional dependence between model and annotation errors that the classical correction assumes away—the quantity that medical latent-class work models explicitly (Collins & Huynh, 2014) and that we obtain in closed form thanks to the single-direction twin map.*

J.5 Some methods improve as the test set gets noisier

Because $\mathrm{Acc}(r)$ has slope $\mathrm{NF} - A$, a method with $\mathrm{NF} > A$ *gains* from test noise. We observe this for four method–dataset combinations:

A model can therefore be made to look 9.3 points better, on the same weights, purely by degrading the test-set annotation quality. This phenomenon is invisible on clean test sets and is the sharpest possible demonstration of the failure mode.