VALUING DATA BEFORE TRAINING: GEOMETRIC COVERAGE AND TOPOLOGICAL FIDELITY OUTPERFORM GRADIENT-BASED SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the value of training data often requires training the target model, adding cost before useful subsets can be selected. We introduce C-GeoVal (Coverage-Geometric Valuation), which values samples before target-model training using frozen foundation-model embeddings. Sample value is defined as the expected marginal gain of a set utility that rewards density-weighted coverage and penalizes topological distortion. A scalable coreset procedure combines density filtering and greedy coverage maximization with persistence-guided subset repair. At an 80% data budget, C-GeoVal improves CIFAR-10N accuracy by 4.93 percentage points over full-data cross-entropy training and GoEmotions Macro-F1 by 2.06 points over full-data BERT fine-tuning. Its valuation scores detect human label errors with AUROC 0.937. Ablations show gains from topological repair in both tasks. For mathematical instruction tuning, the density-and-coverage variant outperforms DDCF at all three tested budgets on shared embeddings.

1 INTRODUCTION

Training data has become the dominant lever on model quality, and real-world pools are large, redundant, and noisy. Web-collected image labels are wrong at double-digit rates (Wei et al., 2022), crowd-sourced text labels disagree with their aggregated reference on roughly one in ten examples (Demszky et al., 2020), and instruction-tuning corpora contain far more examples than a model needs (Zhou et al., 2023; Xia et al., 2024). Choosing the right subset is therefore a first-order decision, and a mature line of work frames it as *data valuation*: score each training sample by its contribution to the learned model, then keep the most valuable ones (Ghorbani & Zou, 2019; Koh & Liang, 2017; Park et al., 2023).

These valuation tools share a structural requirement that undermines their own purpose. Influence functions differentiate through a trained model (Koh & Liang, 2017); TRAK linearizes an ensemble of trained checkpoints (Park et al., 2023); Shapley-style estimators need repeated training runs or, at best, gradients collected during one full training pass (Ghorbani & Zou, 2019; Wang et al., 2025); recent LLM selectors train warmup models for gradient features (Xia et al., 2024) or fit a correctness predictor from the outputs of 23 language models (Hoang et al., 2026). These methods incur different forms of task-dependent computation before selection, including target-model training, warmup gradients, or predictor fitting. Under label noise, methods that compute gradients against corrupted targets may also inherit the noise they seek to identify.

Our starting point is that the information needed for valuation is already present before training. In the embedding space of a frozen foundation model, local density provides a criterion for filtering noisy or uninformative samples; informative samples cover high-density regions not yet covered by their peers; and rare but genuine subpopulations appear as topological structure of the data manifold that a good subset must preserve. None of these signals requires labels on the target objective, gradients, or a trained target model.

C-GeoVal (Coverage-Geometric Valuation) defines sample value as the expected marginal gain of a set utility $J(A)$ combining density-weighted coverage and a persistent-homology penalty. Coreset construction uses density prefiltering, sketched lazy-greedy coverage maximization, and topology-guided repair. The approximation analysis applies to the coverage stage (Section 3.4). Selection uses frozen embeddings without target-model training or gradients. CIFAR-10N and math selection use no target-task labels, while GoEmotions uses observed labels for density stratification. The math configuration omits topology-guided repair.

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.

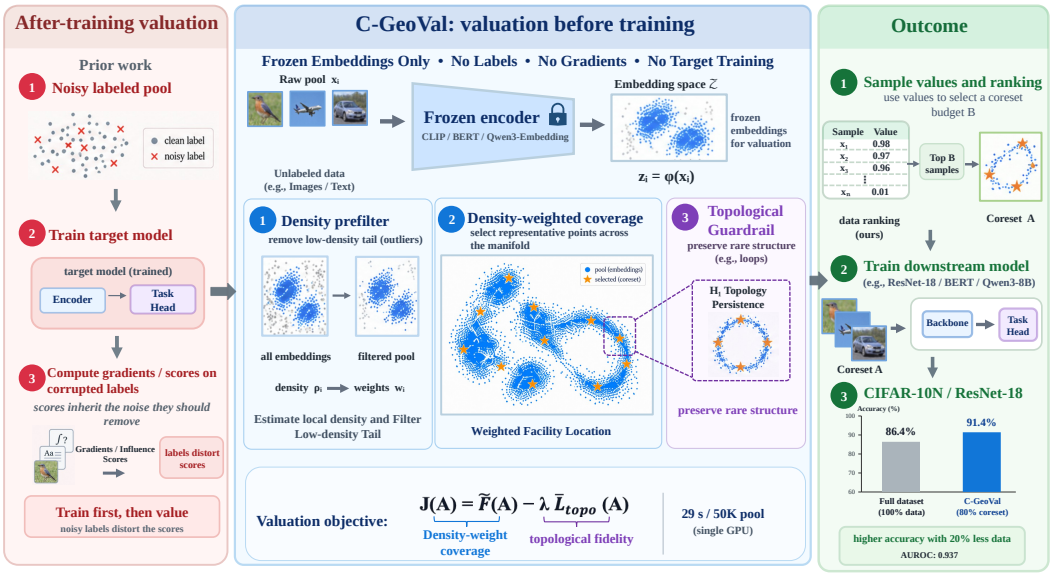


Figure 1: **C-GeoVal values data before target-model training.** Frozen embeddings support density filtering, weighted coverage, and topological repair. The label-free pipeline shown applies to CIFAR-10N; GoEmotions uses observed labels for density stratification, and math selection is label-free with repair disabled. The panels compare valuation workflows; the output ranking illustrates sample scores, while Algorithm 1 constructs coresets through coverage and repair. The 29 s selection time excludes embedding extraction. The displayed utility drops the additive constant in equation 4.

We evaluate C-GeoVal on noisy image and text classification and mathematical instruction tuning. On CIFAR-10N (Wei et al., 2022), C-GeoVal leads the reported selectors at 40% to 80% budgets, and its larger coresets outperform full-data cross-entropy training. Its valuation scores detect human label errors with AUROC 0.937. On GoEmotions (Demszky et al., 2020), the 80% coreset improves Macro-F1 over full-data BERT fine-tuning. For Qwen3-8B (Qwen Team, 2025), density-and-coverage selection outperforms DDCF (Hoang et al., 2026) at every tested budget on identical question embeddings. Ablations quantify the additional gains from topology-guided repair in image and text selection.

Our contributions:

- **A training-free valuation principle.** We define sample value as expected marginal geometric-topological utility on frozen embeddings, derive the marginal decomposition, and connect the valuation to semivalues (Section 3).
- **A scalable coreset procedure.** Density filtering, sketched greedy coverage, and topology-guided repair construct subsets, with an approximation guarantee and explicit sketching-error bound for the coverage stage (Theorem 1).
- **Evidence across three domains.** Coresets improve image accuracy and text Macro-F1 under standard training, and valuation scores detect human label errors. Density-and-coverage selection improves mathematical instruction tuning over the predictor-based baseline on shared embeddings (Section 4).

2 Related Work

Data valuation and attribution. Game-theoretic valuation assigns each sample its Shapley value or a semivalue of a utility defined as trained-model performance (Ghorbani & Zou, 2019; Kwon & Zou, 2022; Wang & Jia, 2023). Influence functions (Koh & Liang, 2017), datamodels (Ilyas et al., 2022), and TRAK (Park et al., 2023) estimate a sample's effect on trained-model predictions from gradients or ensembles of trained checkpoints. In-Run Data Shapley (Wang et al., 2025) reduces the cost to a single training run. A full run over the labeled pool is still required. LAVA (Just et al., 2023) is training-free but requires a clean labeled validation set and computes class-wise optimal transport between labeled distributions, which ties it to classification and label quality. C-GeoVal keeps the semivalue structure, expected marginal utility over background subsets, but replaces the trained-model utility with a geometric-topological one that is computable from frozen embeddings alone. Valuation

therefore precedes training instead of following it, and the scores cannot be contaminated by gradients on mislabeled targets.

Coreset selection and data pruning. Geometry-based selectors include K-Center (Sener & Savarese, 2018), herding (Welling, 2009), and facility-location objectives over gradient or feature spaces (Mirzasoleiman et al., 2020). Training-dynamics scores such as forgetting events (Toneva et al., 2019) and EL2N (Paul et al., 2021) need the target model trained on the full pool, as do FDMat's optimal-transport matching over learned features (Xiao et al., 2024) and moderate coresets (Xia et al., 2023). Self-supervised prototypicality (Sorscher et al., 2022) and ZCore (Griffin et al., 2026) share our premise that foundation-model embeddings suffice for selection, and ZCore is a training-free baseline in our comparisons. C-GeoVal differs in three ways that our experiments isolate: an explicit set utility with an approximation guarantee rather than per-sample heuristic scores, density weighting that actively routes the budget away from the noise-bearing tail, and a topological criterion that no prior selector uses. Under 18.1% real label noise these differences translate into subsets that retain between $2\times$ and $7\times$ less noise than ZCore (Figure 2).

Data selection for LLM fine-tuning. LESS selects instruction data by gradient similarity from LoRA warmup checkpoints (Xia et al., 2024); DEITA scores complexity and quality with trained scorers (Liu et al., 2024); IFD uses the target model's own perplexity ratios (Li et al., 2024); DDCF fits a collaborative-filtering correctness predictor from evaluations of 23 LLMs and greedily balances difficulty against diversity (Hoang et al., 2026). All of these require model forward or backward passes over the pool before selecting. C-GeoVal needs one embedding pass with an off-the-shelf encoder. In a representation-controlled comparison where C-GeoVal and DDCF use identical question embeddings, C-GeoVal outperforms DDCF at every budget without the 23-model correctness matrix (Table 6).

Topological data analysis. Persistent homology summarizes multi-scale topological structure of point clouds (Edelsbrunner et al., 2002; Zomorodian & Carlsson, 2005), admits bottleneck-distance stability under suitable perturbations (Cohen-Steiner et al., 2007), and is computable at scale via landmark constructions (de Silva & Carlsson, 2004) and optimized solvers (Bauer, 2021; Tralie et al., 2018). We use H_1 persistence diagrams as a fidelity criterion for subset selection, which to our knowledge is new in data valuation.

3 METHOD

3.1 A TRAINING-FREE SET UTILITY

Let $V = \{x_1, \ldots, x_n\}$ be the training pool and $z_i = \phi(x_i) \in \mathbb{R}^d$ the ℓ_2-normalized embedding of x_i under a frozen encoder ϕ. Define the local density ρ_i as the mean cosine similarity from z_i to its k nearest neighbors, and the coverage weight

$$w_i \;=\; \rho_i - \min_{j \in V} \rho_j + \varepsilon \;\geq\; 0, \tag{1}$$

so that dense, prototypical regions carry more coverage mass than sparse outlying ones. Let $\tilde{s}_{ui} \in [0, 1]$ denote the similarity between z_u and z_i after a Johnson-Lindenstrauss sketch (Section 3.3), and define the density-weighted coverage of a subset $A \subseteq V$ as the facility-location function

$$\tilde{F}(A) \;=\; \sum_{u \in V} w_u \max_{j \in A} \tilde{s}_{uj}, \qquad \tilde{F}(\varnothing) = 0. \tag{2}$$

Coverage alone can misrepresent the pool: a subset can cover every dense region yet break the connectivity of an elongated or looped subpopulation, precisely the structures that matter for rare classes. We therefore measure topological fidelity. Fix a landmark protocol (density-truncated farthest-point sampling; Section 3.3) and let $D(A)$ be the H_1 persistence diagram computed from the landmarks of A, with D_{ref} the diagram of the full pool. The normalized topological distortion is

$$\bar{\mathcal{L}}_{\mathrm{topo}}(A) \;=\; \frac{1}{C_{\mathrm{topo}}} W_p\big(D(A),\, D_{\mathrm{ref}}\big), \tag{3}$$

where W_p is the p-Wasserstein distance between diagrams and $C_{\mathrm{topo}} > 0$ a scale constant. Degenerate subsets that violate the landmark protocol are assigned a fixed baseline value L_{base}. The C-GeoVal utility is

$$J(A) \;=\; \tilde{F}(A) \;-\; \lambda\big(\bar{\mathcal{L}}_{\mathrm{topo}}(A) - L_{\mathrm{base}}\big). \tag{4}$$

The label-free formulation uses only frozen embeddings. The GoEmotions instantiation additionally uses observed labels to stratify density estimation (Appendix B). Neither requires target-model training or gradients.

Algorithm 1 C-GeoVal selection

Require: embeddings $\{z_i\}_{i=1}^n$, budget b, density order k, removal ratio r, sketch dimension m, landmarks ℓ

1: **Phase 1:** $\rho_i \leftarrow$ mean cosine similarity to k nearest neighbors; drop the bottom r density quantile; $w_i \leftarrow \rho_i - \min_j \rho_j + \varepsilon$

2: **Phase 2:** sketch $\tilde{z}_i \leftarrow \Pi z_i / \|\Pi z_i\|$ with Gaussian $\Pi \in \mathbb{R}^{m \times d}$; build sparse top-$K$ similarity graph $\tilde{s}_{ij} = (1 + \langle \tilde{z}_i, \tilde{z}_j \rangle)/2$; run lazy greedy on $\tilde{F}$ until $|A| = b$

3: **Phase 3:** landmarks of A and of V by density-truncated farthest-point sampling; H_1 diagrams via Vietoris-Rips persistence; flag selected samples whose landmark pairs create H_1 features far from D_{ref}; swap each flagged sample for the densest unselected neighbor, keeping $|A| = b$

Ensure: coreset A; coverage marginal gains collected during Phase 2

3.2 Sample value as expected marginal utility

Write $m_u(A) = \max_{j \in A} \tilde{s}_{uj}$ with $m_u(\varnothing) = 0$. The marginal gain of adding sample x_i to a background subset $A \subseteq V \setminus \{i\}$ has a closed form.

Proposition 1 (Marginal value decomposition)**.** *For every $i \in V$ and $A \subseteq V \setminus \{i\}$,*

$$\Delta_i(A) \;=\; J(A \cup \{i\}) - J(A) \;=\; \sum_{u \in V} w_u \big[\tilde{s}_{ui} - m_u(A)\big]_+ \;-\; \lambda\big(\bar{\mathcal{L}}_{\text{topo}}(A \cup \{i\}) - \bar{\mathcal{L}}_{\text{topo}}(A)\big), \quad (5)$$

where $[a]_+ = \max(a, 0)$.

The decomposition expresses a valuation principle: a sample is worth exactly the weighted coverage mass it newly covers, minus the topological distortion it introduces. Given a family $\mu = \{\mu_i\}_{i=1}^n$ of distributions over background subsets, the *C-Geo value* of x_i is

$$\mathrm{CGV}_\mu(x_i) \;=\; \mathbb{E}_{A \sim \mu_i}\big[\Delta_i(A)\big]. \quad (6)$$

Remark 1. *The valuation in equation 6 recovers semivalues when background subsets of each size are sampled uniformly with a common size distribution across samples (Kwon & Zou, 2022; Ghorbani & Zou, 2019). Uniform weights over sizes recover Shapley values of J, while uniform weights over all subsets recover Banzhaf values (Wang & Jia, 2023). General choices of μ_i define expected marginal scores but need not be semivalues. Unlike trained-model utilities, J uses similarity calculations and persistence-diagram computations without retraining the target model.*

We separate score-based valuation from coreset construction. Monte-Carlo estimates of equation 6 rank samples for the valuation tests. Algorithm 1 instead greedily maximizes coverage, then applies topology-guided swaps. This procedure uses the two terms of J in separate stages and is not greedy maximization of the full utility. Its guarantee applies to the coverage stage.

3.3 Three-phase algorithm

Phase 1: density prefilter and weights. Local density is computed with an exact k-nearest-neighbor search over normalized embeddings ($k = 20$). The bottom $r = 12\%$ density quantile is removed globally. The filter targets low-density embedding outliers, and the density weights equation 1 route the remaining coverage budget toward prototypical regions. Section 4.3 quantifies the effect, and Appendix D.2 shows that accuracy degrades gracefully, by under one point, when the filter is removed entirely.

Phase 2: sketched facility location. A Gaussian JL projection to $m = 128$ dimensions (1024 for math) followed by re-normalization preserves all pairwise similarities within $3\epsilon/2$ in the $[0, 1]$ scale of $\tilde{s}$ (Lemma 1, Appendix A) while reducing the representation dimension by $d/m \approx 6\times$ for the image and text encoders. The graph is sparsified to each sample's top-K neighbors ($K = 200$; $K = 256$ for the 189K LLM pool); the truncated objective remains monotone submodular, so lazy greedy (Minoux, 1978) enjoys the full $(1 - 1/e)$ guarantee of Nemhauser et al. (1978) on the objective it optimizes, at near-linear cost.

Phase 3: topological guardrail. Farthest-point sampling restricted to the top 80% density quantile yields $\ell = 500$ landmarks (600 for math diagnostics) for the coreset and for the pool (de Silva & Carlsson, 2004). Vietoris-Rips persistence on the landmark cosine-distance matrices (Bauer, 2021; Tralie et al., 2018) produces H_1 diagrams; a coreset feature whose distance to the reference diagram exceeds half its own persistence is declared an artifact, the landmark pair realizing its death is traced back to the responsible coreset samples, and each such sample is swapped for the densest unselected sample among its nearest neighbors. The swap preserves $|A| = b$ and acts as a targeted repair step on the penalty term of equation 4. Coverage does not explicitly constrain persistence. The

guardrail targets discrepant loop structure, with its effect on downstream performance and class balance assessed in the ablations.

3.4 GUARANTEE AND COST

Proposition 2 (Monotone submodularity). *$\tilde{F}$ in equation 2 is normalized, monotone, and submodular for any nonnegative weights and similarities, including the JL-sketched and top-K-truncated similarity graph used by Algorithm 1.*

Theorem 1 (Sketched greedy guarantee). *Let V_r be the retained candidates, $1 \le b \le |V_r|$, and $M_{ui} \in \{0,1\}$ the realized top-K graph mask. Define $F_M(A) = \sum_{u \in V} w_u \max_{i \in A} M_{ui} s_{ui}$ using exact similarities $s_{ui} = (1 + \langle z_u, z_i \rangle)/2$ and the same mask as Phase 2, with $F_M(\varnothing) = 0$. For its lazy-greedy output $\hat{A}$, if $m \ge C\epsilon^{-2} \log(n/\delta)$ and $0 < \epsilon \le 1/2$, then with probability at least $1 - \delta$,*

$$F_M(\hat{A}) \;\ge\; \left(1 - \tfrac{1}{e}\right) \max_{\substack{A \subseteq V_r \\ |A| \le b}} F_M(A) \;-\; \tfrac{3}{2}\left(2 - \tfrac{1}{e}\right) \epsilon W, \quad W = \sum_u w_u. \tag{7}$$

Proofs are in Appendix A. This bound controls sketching on the retained candidate set and shared graph support. It does not bound the loss from density filtering, graph truncation, or Phase 3 repair relative to the unrestricted objective. The practical choice $m = 128$ is evaluated empirically in Appendix D.2.

The pipeline uses one frozen-encoder pass, exact kNN search, sparse greedy coverage, and landmark persistence without target-model training or gradients. Phase 2 prefixes share a ranking, while topology-guided repair is applied separately at each budget and can break nesting. Selection takes 29 seconds on CIFAR-10N (Table 4) and 18 minutes on GoEmotions (Appendix C), excluding embedding extraction.

4 EXPERIMENTS

4.1 SETUP

We evaluate three task instantiations with different encoders, label access, and repair settings.

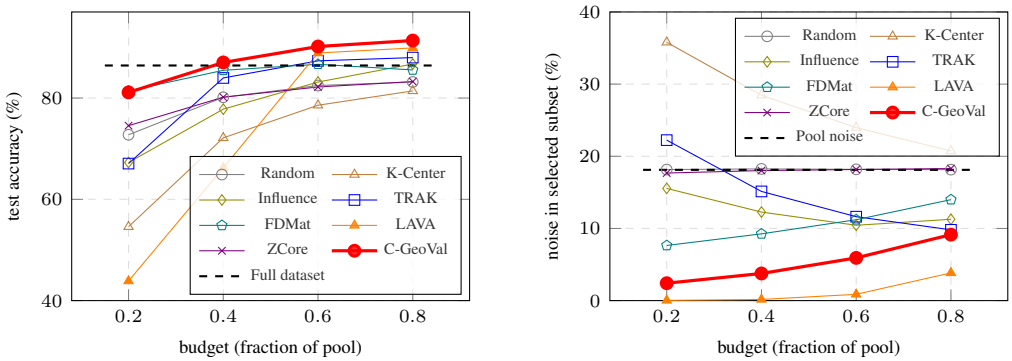
Image classification with real label noise. CIFAR-10N (Wei et al., 2022) pairs CIFAR-10 images with labels from human annotators; we use the `random_label2` set whose measured noise rate is 18.1%. Selection uses CLIP ViT-L/14 image embeddings (Radford et al., 2021); no labels enter the selection. Coresets at budgets from 1% to 80% train ResNet-18 (He et al., 2016) from scratch for a fixed 100-epoch schedule identical across methods and seeds. Baselines: Random, K-Center Greedy (Sener & Savarese, 2018), Prototypicality (Sorscher et al., 2022), Moderate (Xia et al., 2023), Influence Functions (Koh & Liang, 2017), TRAK (Park et al., 2023), FDMat (Xiao et al., 2024), LAVA (Just et al., 2023), and ZCore (Griffin et al., 2026), each run with its official recipe. K-Center, Moderate, IF, TRAK, and FDMat first train the upstream model they require on the full noisy pool; LAVA consumes a clean labeled validation set; C-GeoVal, ZCore, and Prototypicality operate on frozen features only. Image comparisons use the stated protocol and seeds 42/123/456; results are reported as mean $\pm$ standard deviation. Appendix D.6 repeats selection with three further image encoders and two further text encoders.

Text classification with crowd-label noise. GoEmotions (Demszky et al., 2020) provides 28-class emotion labels for Reddit comments; single-rater training labels disagree with the aggregated reference on 9.2% of examples. Selection uses BERT [CLS] embeddings of the pretrained (not fine-tuned) encoder; we use 20%, 60%, and 80% coresets to fine-tune BERT-base (Devlin et al., 2019) for 5 epochs with 3 seeds and report Macro-F1 on the clean test split.

LLM instruction tuning for math reasoning. The pool is the 189,257-question math corpus released with DDCF (Hoang et al., 2026), derived from OpenR1-Math (Hugging Face, 2025). Selection uses Qwen3-Embedding-4B question embeddings (Zhang et al., 2025), pinned to the identical embedding artifact consumed by DDCF, so the comparison is representation-controlled. Subsets of 1K, 10K, and 20K questions are used to fine-tune Qwen3-8B-Base (Qwen Team, 2025); evaluation is AIME24 pass@32 and MATH-500 pass@1 (Lightman et al., 2024) under one shared generation protocol. Baselines: Random, Longest, Difficult, Middle-Perplexity, Diversity, and DDCF. At this pool size the topological term is computed as a diagnostic and Phase 3 swaps are disabled; the density and coverage phases carry the selection. Full protocols, hyperparameters, and the per-task instantiation details are in Appendices B and C.

Table 1: **CIFAR-10N (18.1% real label noise), ResNet-18 test accuracy (%).** Mean ± std over 3 seeds under one unified protocol. LAVA consumes a clean labeled validation set; all other selectors receive no clean labels. C-GeoVal leads at 40% to 80% and exceeds full-data mean accuracy at those budgets. FDMat leads at 20%; budgets of 1% to 10% are in Appendix D.4.

Method	20%	40%	60%	80%
Random	72.75 ±1.74	80.15 ±0.59	82.48 ±0.57	83.17 ±0.78
K-Center (Sener & Savarese, 2018)	54.59 ±0.21	72.14 ±0.57	78.58 ±0.63	81.41 ±0.63
Prototypicality (Sorscher et al., 2022)	66.17 ±2.95	79.61 ±0.71	82.13 ±0.71	83.48 ±0.49
Moderate (Xia et al., 2023)	77.89 ±0.74	83.13 ±0.12	84.69 ±0.25	84.52 ±0.17
Influence Func. (Koh & Liang, 2017)	67.27 ±2.92	77.81 ±2.84	83.16 ±0.92	86.73 ±0.41
TRAK (Park et al., 2023)	67.06 ±3.92	83.97 ±0.29	87.35 ±0.14	87.99 ±0.29
FDMat (Xiao et al., 2024)	**81.41** ±0.11	85.48 ±0.53	86.67 ±0.53	85.55 ±0.21
LAVA (Just et al., 2023)	43.87 ±0.20	66.21 ±0.93	88.90 ±0.07	89.90 ±0.59
ZCore (Griffin et al., 2026)	74.55 ±0.56	80.17 ±0.18	82.19 ±0.32	83.25 ±0.80
C-GeoVal (ours)	81.12 ±0.39	**87.03** ±0.31	**90.18** ±0.27	**91.36** ±0.28
Full dataset (100%)		86.43 ±0.43		



(a) Accuracy against budget.

(b) Retained label noise against budget.

Figure 2: **CIFAR-10N.** (a) C-GeoVal has higher mean accuracy than full-data training at 40% to 80%, with the largest margin at 80%. TRAK and LAVA also exceed the full-data reference at high budgets. (b) The mechanism: C-GeoVal subsets contain 2.4% to 9.1% mislabeled samples where the pool carries 18.1%. Only LAVA retains less noise, and it consumes a clean labeled validation set to do so. All values are means over 3 seeds.

4.2 Selection beats the full noisy dataset

Table 1 and Figure 2a show that C-GeoVal leads the reported selectors at 40%, 60%, and 80%, with margins of 1.55, 1.28, and 1.46 accuracy points over the strongest baseline. The 40% coreset is 0.60 points above full-data training in mean accuracy, rising to a 4.93-point advantage at 80% (91.36% versus 86.43%). At 20%, FDMat leads with 81.41%, exceeding C-GeoVal by 0.29 accuracy points. At 80%, TRAK reaches 87.99% after target-model training and 2.6 hours of attribution, whereas C-GeoVal reaches 91.36% with 29 seconds of selection after embedding extraction. LAVA reaches 89.90% using a clean labeled validation set.

4.3 Where the gains come from

Figure 2b makes the mechanism quantitative. Random subsets retain approximately the pool's 18.1% noise across budgets. K-Center actively concentrates noise, up to 35.8% at the 20% budget, consistent with farthest-first traversal favoring isolated embedding regions. Gradient-based scores help: TRAK retains 9.8% noise at the 80% budget. C-GeoVal subsets contain 2.4%, 3.8%, 5.9%, and 9.1% noise across the four budgets, a $2\times$ to $7.5\times$ reduction relative to the pool without using a single label; among the baselines only LAVA retains less, and it reads a clean validation set to do so. The density weighting keeps greedy selection away from outlying regions, and the topological guardrail targets subset H_1 features that differ from the pool reference; the prefilter contributes robustness rather than the bulk of the gain, since removing it costs under one accuracy point (Appendix D.2). The lower retained noise accompanies the accuracy gains over full-data training at 40% to 80% budgets. A sweep over six CIFAR-10N label sets examines the operating range: at 9% pool noise the 80% coreset is already 2.56 points above full-data training, while on clean labels selection cannot beat the full pool but C-GeoVal remains the strongest selector (Appendix D.5).

Table 2: **Detecting the 18.1% real human label errors of CIFAR-10N from valuation scores.** A sample is predicted mislabeled when its score is low; clean reference labels are used only for evaluation. Mean ± std over 3 seeds. CGV uses neither training labels nor target-model training and has the best reported detection metrics.

Method	Labels	Training	AUROC ↑	AUPRC ↑	FPR@95TPR ↓
ZCore	—	—	0.508 ±0.001	0.180 ±0.001	0.940 ±0.002
TRAK	✓	✓	0.661 ±0.005	0.590 ±0.007	0.990 ±0.001
Influence Func. (damped)	✓	✓	0.693 ±0.010	0.480 ±0.030	0.960 ±0.003
Density (Phase 1 only)	—	—	0.714 ±0.010	0.360 ±0.020	0.620 ±0.020
Cleanlab (Northcutt et al., 2021)	✓	✓	0.921 ±0.000	0.742 ±0.004	0.268 ±0.008
CGV score (ours)	—	—	**0.937** ±0.003	**0.781** ±0.004	**0.214** ±0.007

Table 3: **Data selection and noise-robust training on CIFAR-10N.** ELR+ and DivideMix share the ResNet-18 backbone and augmentation; epochs follow each recipe (CE 100, ELR+ 200, DivideMix 300). GPUh is training cost per run; dashes indicate unreported costs. Selection adds 29 seconds. Mean ± std over 3 seeds.

Training	Data	Fraction	Accuracy (%)	Train GPUh
CE	Full pool	100%	86.43 ±0.43	0.17
CE	Random	80%	83.17 ±0.78	0.14
DivideMix (Li et al., 2020)	Full pool	100%	88.44 ±0.73	3.20
DivideMix	Random	80%	87.29 ±0.42	2.71
ELR+ (Liu et al., 2020)	Full pool	100%	**94.79** ±0.21	0.93
ELR+	Random	80%	90.21 ±0.14	0.64
CE	C-GeoVal	60%	90.18 ±0.27	0.11
CE	C-GeoVal	80%	91.36 ±0.28	0.14
DivideMix	C-GeoVal	60%	84.46 ±0.59	—
DivideMix	C-GeoVal	80%	85.73 ±0.28	—
ELR+	C-GeoVal	60%	92.19 ±0.11	—
ELR+	C-GeoVal	80%	94.04 ±0.14	—

4.4 Evaluating sample-level valuation

A valuation should be evaluated independently of any one subset it induces. We test the per-sample CGV scores of equation 6, estimated with 128 fixed background subsets and 64 Monte-Carlo draws, on the two standard protocols of the valuation literature. First, mislabel detection (Table 2): ranking samples by CGV score identifies the real human label errors of CIFAR-10N with AUROC 0.937, above Cleanlab's confident learning at 0.921, which consumes the noisy labels and cross-validated model training, and far above influence functions and TRAK, whose gradients are computed against the corrupted targets they are asked to detect. Density alone reaches 0.714, so the margin over raw density comes from the coverage and topology terms of the utility.

Second, point removal and addition (full results in Appendix D.3). Removing the lowest-valued 20% of the pool raises full-training accuracy from 86.4% to 91.4%, the largest gain of any method; TRAK peaks at 88.1% and random removal only loses accuracy. Adding samples in descending CGV order dominates every competitor at every fraction: the top 20% alone yields 81.1% where the best alternative reaches 74.6%. The one direction where gradient-based scores win is removal of the highest-valued samples, where TRAK and influence functions degrade the model faster; their scores better identify individually indispensable samples, while CGV better identifies harmful and redundant ones, which is the direction selection under noise needs.

4.5 Combining selection with noise-robust training

Table 3 tests selection with both plain cross-entropy and noise-robust training. The 80% coreset with cross-entropy exceeds full-pool DivideMix by 2.92 points but trails full-pool ELR+ by 3.43 points. ELR+ improves the 60% and 80% coresets over cross-entropy by 2.01 and 2.68 points, respectively. At 80%, it reaches 94.04%, within 0.75 points of full-pool ELR+ while using 20% fewer examples. DivideMix instead falls below cross-entropy on both coresets and below its own full-pool result. Compatibility with selection therefore depends on the training method.

4.6 Selection cost

Table 4 compares pipeline costs. The upstream-dependent baselines in Table 4 first train a model on the full noisy pool; TRAK then spends another 2.6 hours and 3.9 GB on attribution, and LAVA's class-wise optimal transport uses 77.7 GB. C-GeoVal selection requires no target-model training: its

Table 4: **Cost and information requirements of selection on CIFAR-10N.** Scoring covers score computation and subset extraction for *all* budgets, mean over 3 seeds; memory is the peak GPU allocation recorded during scoring (—: none recorded). Upstream marks selectors that must first train a model on the full noisy pool; Clean val. marks the clean labeled validation set LAVA requires. The one-time CLIP ViT-L/14 feature extraction (195 s, 2.4 GB, shared by C-GeoVal, ZCore, and Prototypicality) is reused across all budgets and seeds.

Method	Upstream training	Clean val.	Scoring	Peak GPU mem
Random	—	—	<1 s	—
Moderate	✓	—	1.2 s	—
FDMat	✓	—	1.5 s	—
K-Center	✓	—	8.0 s	—
Prototypicality	—	—	38 s	—
Influence Func.	✓	—	9.3 min	1.7 GB
TRAK	✓	—	2.6 h	3.9 GB
LAVA	—	✓	77 s	77.7 GB
ZCore	—	—	4.6 min	—
C-GeoVal (ours)	—	—	29 s	0.94 GB

Table 5: **GoEmotions (28 classes, 9.2% annotator disagreement), BERT-base Macro-F1.** Mean ± std over 3 seeds under one shared fine-tuning protocol. Right block: single-rater labels disagreeing with the aggregated reference in the selected subset.

	Macro-F1			Disagreement (%)		
Method	20%	60%	80%	20%	60%	80%
Random	38.34 ±1.04	45.46 ±0.90	46.38 ±0.82	9.26	9.11	9.17
Facility location	38.31 ±0.28	45.79 ±0.49	46.49 ±0.82	8.04	8.74	8.99
C-GeoVal w/o topology	29.47 ±0.91	35.36 ±0.37	40.36 ±0.27	**3.45**	5.20	7.57
C-GeoVal (ours)	**38.89** ±0.29	**46.14** ±0.52	**48.71** ±0.47	3.82	**5.06**	**5.68**
Full dataset (100%)		46.65 ±0.51			9.2 (pool)	

29 seconds cover the GPU kNN within 0.94 GB, one greedy pass, and persistence on 500 landmarks, and cover selection at the evaluated budgets. On GoEmotions, selection for all budgets and ablation variants takes 18 minutes with no recorded GPU allocation (Appendix C). The Phase 2 ranking can be reused across budgets. Final repaired subsets additionally require budget-specific topology checks and swaps.

4.7 Text classification with crowd-label noise

On text, the method uses BERT embeddings and observed-label density stratification (Table 5). C-GeoVal leads at every budget; the margin over Random is modest at 20% and 60% ($+0.55$ and $+0.68$ in mean Macro-F1) and grows to $+2.3$ at 80%, where the coreset also outperforms full-data fine-tuning by 2.1 Macro-F1 (48.71 against 46.65), while retaining 5.7% annotator disagreement against the pool's 9.2%. At 60% its mean is 0.51 points below full-data training while using 40% fewer examples per epoch. The per-class analysis at 80% (Appendix D.8) shows this trade-off: the topological guardrail raises tail F1 from 21.15 to 27.33, but it remains below Random at 35.39. Higher head and medium F1 yield the overall Macro-F1 advantage despite this tail deficit. The Macro-F1 gains do not extend to Weighted-F1: at 80%, C-GeoVal scores 57.53 against 58.84 for Random and 58.79 for full-data training (Appendix D.7). Selection takes 18 minutes with no recorded GPU allocation.

4.8 Instruction tuning an 8B LLM for math reasoning

Table 6 evaluates the principle with an 8B-parameter model and a modern reasoning benchmark. The comparison uses a strong baseline: DDCF is the state-of-the-art selector on this exact pool, its pipeline evaluates 23 open-source LLMs on a seed set to fit a collaborative-filtering correctness predictor, and both methods receive byte-identical question embeddings. C-GeoVal, with density filtering and weighted coverage alone, outperforms DDCF at all three budgets on both benchmarks. Against the full corpus, the 20K C-GeoVal subset trains a model that exceeds full-corpus SFT on MATH-500 (92.3 against 91.8) and recovers 96% of its AIME24 accuracy using 10.6% of the data. At the 1K budget, where selection pressure is highest, C-GeoVal reaches 82.8 on AIME24, above every baseline at the same budget and above Random with twenty times the data. The result extends the evaluation from noisy-label classification to mathematical instruction tuning: even in a curated pool

Table 6: **Qwen3-8B-Base SFT on the 189K math pool.** AIME24 pass@32 and MATH-500 pass@1 under one shared training and generation protocol; C-GeoVal and DDCF consume identical Qwen3-Embedding-4B question embeddings. C-GeoVal leads every selector at every budget; its 20K subset (10.6% of the pool) surpasses full-corpus training on MATH-500.

	1K		10K		20K	
Method	AIME24	MATH	AIME24	MATH	AIME24	MATH
Random	79.5	88.5	81.8	90.1	81.2	89.8
Longest	77.8	87.2	81.0	89.3	80.6	89.0
Difficult	75.0	89.0	79.1	90.5	78.6	90.1
Middle Perplexity	78.5	89.0	80.4	90.0	81.0	90.7
Diversity	76.0	88.8	79.5	90.4	82.0	91.1
DDCF (Hoang et al., 2026)	81.5	90.2	83.4	91.0	85.1	91.7
C-GeoVal (ours)	**82.8**	**91.2**	**84.7**	**92.0**	**85.3**	**92.3**
Base model (no SFT)	AIME24 47.8, MATH-500 60.8					
Full dataset (189K)	AIME24 88.6, MATH-500 91.8					

Table 7: **Quantity-matched swap controls on CIFAR-10N (ResNet-18, 3 seeds).** All swap variants start from the identical Phase 2 subset and replace the same number of samples ($q \approx 497$) drawn from the same 15-nearest-neighbor candidate pools; only the rule that picks which samples to replace differs. W_1 is the Wasserstein distance between the subset's H_1 diagram and the pool reference after swapping, at the 20% budget.

	Accuracy (%)				
Swap rule	20%	40%	60%	80%	W_1@20%
No swap (Phase 2 output)	80.64 $\pm$0.36	86.55 $\pm$0.36	89.70 $\pm$0.36	90.88 $\pm$0.36	6.12
Random swap	80.64 $\pm$1.91	86.55 $\pm$0.38	89.70 $\pm$0.51	90.88 $\pm$0.39	6.14
Density-matched swap	80.64 $\pm$1.50	86.55 $\pm$0.80	89.70 $\pm$0.80	90.88 $\pm$0.46	6.07
Neighbor swap (no H_1)	80.64 $\pm$0.69	86.55 $\pm$0.68	89.70 $\pm$0.24	90.88 $\pm$0.70	6.09
H_1 swap (full C-GeoVal)	**81.12** $\pm$0.39	**87.03** $\pm$0.31	**90.18** $\pm$0.27	**91.36** $\pm$0.28	**5.98**

with no mislabeled samples in the CIFAR-10N sense, geometric coverage of the question distribution identifies the examples an 8B model most benefits from, without target-model scoring or gradient computation during selection, after the shared frozen-encoder embedding pass.

4.9 Contribution of topology-guided repair

Table 7 isolates the contribution of topology-guided replacement. Each swap control replaces the same number of samples from the same candidate pools as the H_1 guardrail, changing only the rule for targeting samples. None of the quantity-matched controls changes mean accuracy from the no-swap level, and the H_1 rule gives the largest distortion reduction, from 6.12 to 5.98 at 20%, compared with 6.07 for density-matched and 6.09 for neighbor swaps. The persistence-guided swaps add $+0.48$ accuracy at every budget on vision. On text the same component is decisive rather than incremental: removing it costs 9.4, 10.8, and 8.4 Macro-F1 at the 20%, 60%, and 80% budgets (Table 5), with the 80% per-class results showing higher tail-class F1 after repair (Appendix D.8). The topology term also carries the score-level valuation: in the Monte-Carlo CGV ranking, setting λ to zero or to four times its default raises retained noise from 2.4% to 15–17% at the 20% budget (Appendix D.2).

5 Conclusion

C-GeoVal values data before target-model training through expected marginal utility on frozen embeddings. Density-weighted coverage and topological fidelity provide complementary criteria for constructing useful subsets. At the 80% budget, the coresets improve CIFAR-10N accuracy and GoEmotions Macro-F1 over their full-data standard-training baselines by 4.93 and 2.06 points, respectively. Valuation scores detect human label errors with AUROC 0.937. Density-and-coverage selection outperforms DDCF across all tested math budgets on shared embeddings.

Reproducibility statement

Selection settings and task-specific differences appear in Appendix B. Appendix C gives training schedules, evaluation settings, and immutable source revisions for the math experiments. Image and text experiments use seeds 42/123/456. Appendix D.1 reports descriptive mean differences. Proofs appear in Appendix A.

AI USE STATEMENT

Generative artificial intelligence tools were used in this study solely for language editing and code optimization. During manuscript preparation, GPT was used to improve clarity and correct grammatical errors. During experimentation, Codex served as a coding assistant to support code development. These tools were not involved in formulating research ideas, deriving scientific claims or conclusions, or analyzing experimental results.

All AI-generated content was thoroughly reviewed and validated by the author team. The entire manuscript was manually reviewed by the authors to ensure technical accuracy and fidelity to the intended meaning. All experimental findings and conclusions were independently confirmed by the author team.

We take full responsibility for all content in the final version of this work.

REFERENCES

Ulrich Bauer. Ripser: Efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423, 2021.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In *Eurographics Symposium on Point-Based Graphics*, 2004.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, 2019.

Brent A. Griffin, Jacob Marks, and Jason J. Corso. Zero-shot coreset selection via iterative subspace sampling. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Long P. Hoang, Wenxuan Zhang, and Wei Lu. Difficulty-diversity collaborative filtering for data-efficient LLM fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2026.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1. `https://github.com/huggingface/open-r1`, 2025.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *International Conference on Machine Learning (ICML)*, 2022.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Hoang Anh Just, Feiyang Kang, Jiachen T. Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. LAVA: Data valuation without pre-specified learning algorithms. In *International Conference on Learning Representations (ICLR)*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Yongchan Kwon and James Zou. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pp. 234–243. Springer, 1978.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning (ICML)*, 2020.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978.

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

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for Python. *Journal of Open Source Software*, 3(29):925, 2018.

Jiachen T. Wang and Ruoxi Jia. Data Banzhaf: A robust data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data Shapley in one training run. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Max Welling. Herding dynamical weights to learn. In *International Conference on Machine Learning (ICML)*, 2009.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Weiwei Xiao, Yongyong Chen, Qiben Shan, Yaowei Wang, and Jingyong Su. Feature distribution matching by optimal transport for effective and robust coreset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33:249–274, 2005.

A Proofs

A.1 Proof of Proposition 1

By definition $J(A) = \tilde{F}(A) - \lambda(\bar{\mathcal{L}}_{\text{topo}}(A) - L_{\text{base}})$, and the constant L_{base} cancels in every difference. For the coverage term,

$$\tilde{F}(A \cup \{i\}) - \tilde{F}(A) = \sum_{u \in V} w_u \Big(\max\{m_u(A), \tilde{s}_{ui}\} - m_u(A) \Big) = \sum_{u \in V} w_u \big[\tilde{s}_{ui} - m_u(A)\big]_+,$$

since $\max\{a, b\} - a = [b - a]_+$. Subtracting the topological difference multiplied by λ gives equation 5. □

A.2 Proof of Proposition 2

Normalization: $\tilde{F}(\varnothing) = 0$ by convention. Monotonicity: adding an element can only increase each inner maximum, and $w_u \geq 0$. Submodularity: for $A \subseteq B$ and $i \notin B$, using Proposition 1,

$$\Delta_i^{\tilde{F}}(A) - \Delta_i^{\tilde{F}}(B) = \sum_u w_u \Big(\big[\tilde{s}_{ui} - m_u(A)\big]_+ - \big[\tilde{s}_{ui} - m_u(B)\big]_+ \Big) \geq 0,$$

because $m_u(A) \leq m_u(B)$ and $t \mapsto [\tilde{s}_{ui} - t]_+$ is nonincreasing. The argument uses only $w_u \geq 0$ and $\tilde{s}_{ui} \in [0, 1]$, so it also holds for the JL-sketched similarities and for the top-K-truncated graph, where truncated entries are treated as $\tilde{s}_{ui} = 0$. □

A.3 A JOHNSON-LINDENSTRAUSS LEMMA FOR NORMALIZED SIMILARITIES

Lemma 1. *Let $z_1, \dots, z_n \in \mathbb{R}^d$ be unit vectors, $\Pi \in \mathbb{R}^{m\times d}$ a Gaussian JL matrix with $m \geq C\epsilon^{-2}\log(n/\delta)$ for a universal constant C and $0 < \epsilon \leq 1/2$, and $\tilde{z}_i = \Pi z_i/\|\Pi z_i\|$. Then with probability at least $1-\delta$, for all pairs (i,j),*

$$\big|\langle \tilde{z}_i, \tilde{z}_j\rangle - \langle z_i, z_j\rangle\big| \leq 3\epsilon, \qquad \textit{hence} \qquad \big|\tilde{s}_{ij} - s_{ij}\big| \leq \tfrac{3}{2}\,\epsilon,$$

where $s_{ij} = (1 + \langle z_i, z_j\rangle)/2$ and $\tilde{s}_{ij} = (1 + \langle \tilde{z}_i, \tilde{z}_j\rangle)/2$.

Proof. By the distributional JL lemma applied to the $O(n^2)$ vectors $\{z_i\} \cup \{z_i \pm z_j\}$, with probability $1-\delta$ all their squared norms are preserved within a factor $1 \pm \epsilon$. Polarization gives

$$\begin{aligned}\big|\langle \Pi z_i, \Pi z_j\rangle - \langle z_i, z_j\rangle\big| &= \tfrac{1}{4}\big|\big(\|\Pi(z_i+z_j)\|^2 - \|z_i+z_j\|^2\big) - \big(\|\Pi(z_i-z_j)\|^2 - \|z_i-z_j\|^2\big)\big| \\ &\leq \tfrac{\epsilon}{4}\big(\|z_i+z_j\|^2 + \|z_i-z_j\|^2\big) = \epsilon,\end{aligned}$$

using $\|z_i\| = \|z_j\| = 1$. For the normalization step, $\|\Pi z_i\|^2 \in [1-\epsilon, 1+\epsilon]$ implies

$$\big|\langle \tilde{z}_i, \tilde{z}_j\rangle - \langle \Pi z_i, \Pi z_j\rangle\big| = \big|\langle \Pi z_i, \Pi z_j\rangle\big| \cdot \left|\tfrac{1}{\|\Pi z_i\|\|\Pi z_j\|} - 1\right| \leq \big|1 - \|\Pi z_i\|\|\Pi z_j\|\big| \leq \epsilon$$

by Cauchy–Schwarz and $\|\Pi z_i\|\|\Pi z_j\| \in [1-\epsilon, 1+\epsilon]$. Thus the total error is at most $2\epsilon \leq 3\epsilon$, and the map $c \mapsto (1+c)/2$ gives the stated similarity bound. □

A.4 PROOF OF THEOREM 1

Let $\hat{F}_M(A) = \sum_u w_u \max_{i\in A} M_{ui}\tilde{s}_{ui}$ be the sparse sketched objective optimized in Phase 2, with the same weights, mask, and candidate set as F_M. On the uniform event in Lemma 1,

$$\big|\hat{F}_M(A) - F_M(A)\big| \leq \sum_u w_u \max_{i\in V_r} M_{ui}\big|\tilde{s}_{ui} - s_{ui}\big| \leq \tfrac{3}{2}\epsilon W =: \delta_F$$

for every $A \subseteq V_r$, including the empty set. The event holds for every pair simultaneously, so it also covers a mask chosen from the projected similarities. Restriction to V_r and zeroing entries preserve monotone submodularity. With consistent tie-breaking, lazy greedy implements standard greedy (Minoux, 1978), yielding the cardinality guarantee of Nemhauser et al. (1978). Let $\alpha = 1 - 1/e$ and $A^\star \in \arg\max_{A\subseteq V_r, |A|\leq b} F_M(A)$. Then

$$\begin{aligned}F_M(\hat{A}) &\geq \hat{F}_M(\hat{A}) - \delta_F \geq \alpha\hat{F}_M(A^\star) - \delta_F \\ &\geq \alpha F_M(A^\star) - (1+\alpha)\delta_F,\end{aligned}$$

which proves the bound. It compares exact and projected similarities on the same graph support and retained candidates, before Phase 3. □

B ALGORITHM DETAILS AND HYPERPARAMETERS

Table 8: C-GeoVal hyperparameters. Vision and text share the listed numerical settings, with observed-label density stratification on text. Math uses a larger removal ratio and graph width, with diagnostic-only topology. Math uses $m = 1024$ and $\ell = 600$; its configuration is documented in Appendix C.

Phase	Parameter	CIFAR-10N	GoEmotions	Math 189K
1	density neighbors k	20	20	20
1	removal ratio r	12%	12%	20%
2	sketch dimension m	128	128	1024
2	sparse graph width K	200	200	256
3	landmarks ℓ	500	500	600
3	FPS density truncation	20%	20%	20%
3	topology weight λ	1.00×10^4	1.00×10^4	(diagnostic)
	encoder ϕ	CLIP ViT-L/14	BERT [CLS]	Qwen3-Embedding-4B
	embedding dim. d	768	768	2560

Density and prefilter. Exact kNN over ℓ_2-normalized embeddings (FAISS inner-product index (Johnson et al., 2019); blocked exact matrix products on GPU for the ablation runs). ρ_i is the mean similarity to the $k = 20$ nearest neighbors excluding self. The global threshold is the r-quantile of ρ; weights follow equation 1 with $\varepsilon = 10^{-6}$.

Sketch and greedy. Π has i.i.d. $\mathcal{N}(0, 1/m)$ entries; projected vectors are re-normalized. The sparse similarity graph stores each sample's top-K neighbors with similarities mapped to $[0,1]$ by $s = (1+\cos)/2$. Lazy greedy uses a max-heap over stale marginal gains with revalidation (Minoux,

1978); marginal gains are computed through the reverse adjacency of the sparse graph, giving $O(nK)$ work per full revalidation and near-linear total time in practice. Phase 2 is deterministic given the seed and produces nested prefixes. Phase 3 repairs each requested budget separately, so final subsets need not remain nested.

Topological guardrail. Landmarks come from farthest-point sampling restricted to the top 80% density quantile, run separately on the coreset and the pool. H_1 persistence is computed on landmark cosine-distance matrices with Ripser (Bauer, 2021; Tralie et al., 2018), and diagram distances with the persim implementation of the Wasserstein matching. A coreset H_1 feature (b, d) is an artifact when its Euclidean distance to the nearest reference feature exceeds half its persistence $d - b$. Landmark pairs whose distance realizes the death of an artifact are traced to coreset samples; each flagged sample is replaced by the densest not-yet-selected sample among its nearest neighbors, each replacement used at most once, so $|A| = b$ is preserved. On the 189K pool the diagrams and Wasserstein distances are computed and logged as a selection diagnostic while swaps are disabled; density filtering and weighted coverage carry the selection at that scale, and Table 6 reports that configuration.

GoEmotions instantiation. On GoEmotions the density estimate of Phase 1 is stratified by the observed training labels (per-class kNN density with a per-class prefilter), matching the noisy-label setting in which observed labels are available for stratification but never trusted as clean. No gradients and no clean labels are used. On CIFAR-10N and the math pool, selection is fully label-free.

C Experimental protocols

CIFAR-10N. Pool: 50,000 CIFAR-10 training images with the `random_label2` human annotation of CIFAR-10N (Wei et al., 2022); measured noise rate 18.12%. Downstream model: ResNet-18 trained from scratch, SGD with momentum 0.9, cosine schedule, 100 epochs for every method and budget with early stopping disabled, so all methods receive an identical training-epoch budget. Seeds 42/123/456. Results are reported as mean $\pm$ standard deviation across these seeds. Baselines use their official recipes: K-Center, Moderate, Influence Functions, and FDMat first train the upstream ResNet-18 they require on the full noisy pool; TRAK trains and checkpoints the target model and computes projected per-sample gradients (projection dimension per the official library defaults); the influence-function scores use a damped adaptation for stability; LAVA computes class-wise optimal transport against a clean labeled validation split; Prototypicality and ZCore consume the same CLIP features as C-GeoVal. For the robust-training comparison, ELR+ and DivideMix follow their published CIFAR-10N configurations adapted to the shared ResNet-18 backbone and augmentation, with 200 and 300 epochs respectively; the Cleanlab filter of Table 2 uses cross-validated training (0.38 GPUh). Accuracy is evaluated on the clean CIFAR-10 test set. Noise rates of selected subsets are measured against the clean reference labels, which are used only for this measurement and never for selection.

GoEmotions. Pool: the official train split with single-rater noisy labels derived from the raw annotations and aggregated labels as clean reference (9.2% disagreement); 28 classes. Downstream: BERT-base-cased fine-tuned for 5 epochs, batch 32, learning rate 2×10^{-5}, max length 64, seeds 42/123/456, early stopping disabled. Metrics: Macro-F1 and Weighted-F1 on the clean test split at the peak-validation epoch. Selection across the evaluated budgets and ablation variants takes 18 minutes, excluding embedding extraction and downstream fine-tuning. No GPU allocation was recorded during selection.

Math reasoning SFT. Pool: `DDCF-fullcorpus-math` at its pinned revision, exactly 189,257 questions (Hoang et al., 2026), derived from OpenR1-Math (Hugging Face, 2025). Target model: Qwen3-8B-Base at a pinned revision, fine-tuned with FSDP full-shard, bf16, 16,384-token context, global batch 32 for 4 epochs on subsets, global batch 128 for 3 epochs for the full-corpus reference. Every selector produces one deterministic 20K ranking whose 1K and 10K prefixes give the smaller budgets. C-GeoVal and DDCF consume the same pinned Qwen3-Embedding-4B full-pool embedding artifact, with digest checks required by the selection pipeline. Evaluation: AIME24 pass@32 with 64 samples per problem and MATH-500 pass@1, with prompts, sampling parameters, and the answer-equivalence grader matched to the pinned DDCF release; all 21 selector runs and both references share one generation backend and are directly comparable to each other. The selection pipeline checks artifact digests against the pinned source revisions. Configuration details are given below.

Pinned math configuration. The accompanying `math_protocol.yaml` records protocol version 1, seed 42, and full source revisions. Their prefixes are `77e38be55965b` (pool), `49e3418fbbbc` (Qwen3-8B-Base), `5cf2132abc99` (Qwen3-Embedding-4B), and

`434f9d518ad5` (DDCF evaluation code). Generation uses temperature 0.6, top-p 0.95, top-k 20, and at most 12,288 new tokens.

The math experiments use a 1024-dimensional sketch and 600 diagnostic landmarks (Table 8); topology is diagnostic only, with subset repair disabled.

Swap controls. CIFAR-10N `random_label2`, budgets 20% to 80%, seeds 42/123/456, full 100-epoch training recipe. Every control starts from the identical Phase 2 subset and replaces exactly the number of samples q that the H_1 guardrail flags in that run ($q \approx 497$ at each budget); replacement candidates come from the flagged sample's 15 nearest not-yet-selected neighbors, each candidate used at most once. Random swap picks the replaced samples uniformly, density-matched swap picks the q lowest-density selected samples, and neighbor swap follows the guardrail's replacement rule without the persistence-based flagging. Pre- and post-swap Wasserstein distances to the pool reference diagram and the coverage-objective change are logged for every run.

Valuation scores. Per-sample CGV scores use equation 6 with 128 fixed background subsets shared across samples and 64 Monte-Carlo draws. Mislabel detection treats a low score as a positive prediction and evaluates against the clean reference labels. Point removal and addition rank the pool once per method, then train ResNet-18 under the unified 100-epoch recipe on the retained set; the zero-removal reference and identical subsets reuse the corresponding unified-protocol runs rather than adding duplicate training runs.

D Additional results

D.1 Descriptive mean differences

Table 9 reports arithmetic differences between the reported three-seed means, in percentage points. Positive values favor C-GeoVal. These descriptive comparisons do not establish statistical significance or equivalence.

Table 9: Descriptive differences between reported three-seed means (percentage points).

Setting	Comparison	Δ (pp)
CIFAR-10N 20%	C-GeoVal − full dataset	-5.31
CIFAR-10N 40%	C-GeoVal − full dataset	$+0.60$
CIFAR-10N 60%	C-GeoVal − full dataset	$+3.75$
CIFAR-10N 80%	C-GeoVal − full dataset	$+4.93$
CIFAR-10N 20%	C-GeoVal − FDMat (best baseline)	-0.29
CIFAR-10N 40%	C-GeoVal − FDMat (best baseline)	$+1.55$
CIFAR-10N 60%	C-GeoVal − LAVA (best baseline)	$+1.28$
CIFAR-10N 80%	C-GeoVal − LAVA (best baseline)	$+1.46$
CIFAR-10N	C-GeoVal 80% + CE − ELR+ full	-3.43
CIFAR-10N	C-GeoVal 80% + CE − DivideMix full	$+2.92$
CIFAR-10N	C-GeoVal 60% + CE − ELR+ full	-4.61
GoEmotions 60%	C-GeoVal − full dataset	-0.51
GoEmotions 80%	C-GeoVal − full dataset	$+2.06$
GoEmotions 80%	C-GeoVal − Random	$+2.33$

Tables 5 and 17 report the same default-encoder results. The class-group results in Table 19 aggregate class-level means using the frequency-based partition specified in Appendix D.8.

D.2 Hyperparameter sensitivity

Table 10 varies one selection parameter at a time around the defaults of Table 8, retraining ResNet-18 under the unified protocol. Accuracy stays within 0.8 points of the default at both budgets for every tested value, including removing the density prefilter entirely ($r = 0$), and retained noise stays between 2.4% and 2.6% at the 20% budget throughout. Selection wall-clock time is insensitive except for the landmark count, which scales the persistence computation (9 s at $\ell = 100$ to 106 s at $\ell = 1000$).

The topology weight λ behaves differently in the Monte-Carlo score ranking (Table 11): retained noise collapses from 15–17% to 2.4% exactly at the default $\lambda_0 = 1.00 \times 10^4$, confirming that the topological term, not density alone, steers the score-level valuation away from noise.

Table 10: Sensitivity of C-GeoVal to its selection parameters on CIFAR-10N (ResNet-18 accuracy %, 3 seeds). Defaults in bold. One parameter varies per row group; all others stay at their defaults.

Parameter	Values (default bold)	Acc. @20% range	Acc. @60% range
removal ratio r	0, .05, .08, **.12**, .16, .20, .25	80.35–**81.12**	89.41–**90.18**
density neighbors k	10, **20**, 40, 80	80.56–**81.12**	89.62–**90.18**
sketch dimension m	32, 64, **128**, 256, 512	80.35–**81.12**	89.41–**90.18**
graph width K	50, 100, **200**, 400	80.35–**81.12**	89.41–**90.18**
landmarks ℓ	100, 250, **500**, 1000	80.35–**81.12**	89.41–**90.18**

Table 11: Topology weight sweep in the Monte-Carlo CGV score ranking (128 fixed backgrounds, 64 draws; not the swap threshold of Algorithm 1). CIFAR-10N, 3 seeds.

λ/λ_0	Acc. @20%	Noise @20%	Acc. @60%	Noise @60%
0	80.42 $\pm$1.08	16.89	89.48 $\pm$0.41	17.16
0.25	80.50 $\pm$0.32	16.49	89.56 $\pm$0.32	16.89
0.5	80.57 $\pm$1.45	16.14	89.63 $\pm$0.34	16.70
1	**81.12** $\pm$0.39	**2.41**	**90.18** $\pm$0.27	**5.92**
2	80.42 $\pm$0.92	15.26	89.48 $\pm$0.39	16.15
4	79.82 $\pm$1.43	14.79	88.88 $\pm$0.36	16.54

D.3 Point removal and point addition

Tables 12 and 13 report the standard valuation protocols on CIFAR-10N: remove a fraction of the pool by score and retrain, or train on the top-scored fraction alone. The no-removal reference is 86.43%.

Table 12: Point removal on CIFAR-10N (ResNet-18 accuracy %, 3 seeds). Left: removing the lowest-valued fraction first, where higher accuracy indicates better identification of harmful data. Right: removing the highest-valued fraction first, where lower accuracy indicates better identification of indispensable data.

	Remove lowest first					Remove highest first				
Method	5%	10%	20%	40%	60%	5%	10%	20%	40%	60%
Random	82.85	83.59	83.17	82.39	80.61	83.79	83.48	82.79	82.35	80.51
Density	83.29	83.08	82.30	81.10	76.83	83.29	83.45	82.97	80.89	78.98
ZCore	83.68	83.69	83.25	82.19	80.17	83.23	83.49	83.59	82.23	80.25
Influence Func.	85.39	86.53	86.73	83.16	77.81	**82.98**	**82.37**	80.81	**73.02**	**57.51**
TRAK	86.69	88.11	87.99	87.35	83.97	**82.98**	82.70	**80.73**	74.79	64.93
CGV score (ours)	**86.91**	**88.52**	**91.36**	**90.18**	**87.03**	85.17	84.03	81.76	77.22	72.68

Table 13: Point addition on CIFAR-10N: train only on the top-valued fraction (ResNet-18 accuracy %, 3 seeds).

Method	1%	5%	10%	20%	40%	60%	80%
Density	20.94	44.23	51.72	65.98	76.83	81.10	82.30
TRAK	19.39	29.55	44.33	67.06	83.97	87.35	87.99
Influence Func.	28.09	44.01	55.03	67.27	77.81	83.16	86.73
ZCore	32.27	48.38	60.39	74.55	80.17	82.19	83.25
Random	32.91	50.41	63.24	73.61	80.61	82.39	83.17
CGV score (ours)	**38.42**	**58.31**	**69.14**	**81.12**	**87.03**	**90.18**	**91.36**

D.4 EXTREME BUDGETS

Table 14: CIFAR-10N at extreme budgets (ResNet-18 accuracy %, 3 seeds). C-GeoVal leads every selector; LAVA, built to prune large fractions with a clean validation set, degenerates at small budgets.

Method	1%	5%	10%
LAVA	9.74 ±0.00	25.29 ±0.99	35.89 ±0.85
K-Center	18.07 ±1.95	31.58 ±1.07	39.02 ±2.01
Prototypicality	18.52 ±1.06	34.86 ±0.99	46.29 ±1.74
FDMat	25.91 ±0.82	49.91 ±1.17	65.24 ±2.36
Moderate	31.33 ±0.37	57.01 ±2.75	65.23 ±1.54
ZCore	32.27 ±1.24	48.38 ±0.52	60.39 ±1.32
Random	32.91 ±1.86	50.41 ±2.95	63.24 ±2.32
C-GeoVal (ours)	**38.42** ±1.87	**58.31** ±1.24	**69.14** ±0.88

D.5 NOISE-LEVEL SWEEP

Table 15 varies the CIFAR-10N label set while keeping the images, the selection hyperparameters, and the training recipe fixed; C-GeoVal selection is label-free, so its subsets are identical across label sets. On clean labels no selector beats full-data training, as expected, but C-GeoVal remains the strongest selector, 92.56% at the 80% budget against 90.74% for the best alternative. At 9% pool noise the 80% coreset is 2.56 points above full-data training. The gains on the three random-label sets range from 4.93 to 7.23 points, although they do not increase monotonically with noise rate. Table 16 reports all four budgets and extends the evaluation to `worse_label`, where accuracy increases from 58.73% to 68.91% as the budget grows.

Table 15: CIFAR-10N label-set sweep at the 80% budget (ResNet-18 accuracy %). The `random_label2` comparison follows Table 1. Best baseline denotes the best evaluated baseline for each label set. K-Center upstream features are trained on `random_label2` throughout, so it is not label-free in this sweep.

Label set	Pool noise	Full data	C-GeoVal 80%	Δ vs. full	Best baseline 80%
`clean_label`	0%	95.36	92.56	−2.80	90.74 (Random)
`aggre_label`	9.0%	89.72	92.28	+2.56	89.39 (ZCore)
`random_label1`	17.2%	83.71	90.71	+7.00	83.48 (K-Center)
`random_label2`	18.1%	86.43	91.36	+4.93	89.90 (LAVA)
`random_label3`	17.6%	83.79	91.02	+7.23	83.79 (K-Center)

Table 16: C-GeoVal accuracy (%) across CIFAR-10N label sets and budgets, using ResNet-18. Standard deviations are not reported for the additional label sets. The `random_label2` means match Table 1.

Label set	20%	40%	60%	80%
`aggre_label`	82.41	88.16	91.07	92.28
`random_label1`	80.47	86.38	89.52	90.71
`random_label2`	81.12	87.03	90.18	91.36
`random_label3`	80.83	86.71	89.84	91.02
`worse_label`	58.73	64.18	67.42	68.91

D.6 ENCODER DEPENDENCE

Table 17 repeats selection with different frozen encoders under one shared comparison protocol per modality. The default CLIP ViT-L/14 results follow Table 1. On images, this encoder yields the highest accuracy at every reported budget, reaching 91.36% at 80% against 82.60% for CLIP ViT-B/32. The supervised ResNet-50 features are weakest at the 20% budget. On text, the default BERT [CLS] instantiation uses the per-class density stratification of Appendix B, while SimCSE and MPNet run fully label-free; every encoder still selects subsets with roughly half the pool's annotator disagreement.

D.7 WEIGHTED-F1 ON GOEMOTIONS

Table 18 reports Weighted-F1 from the same three-seed protocol as Table 5. C-GeoVal improves from 49.74 at the 20% budget to 57.53 at 80%, but at 80% it trails Random by 1.31 points and full-data fine-tuning by 1.26 points. Thus the Macro-F1 advantage does not extend to the class-frequency-weighted metric.

Table 17: Encoder dependence of C-GeoVal selection (3 seeds). Image block: CIFAR-10N, ResNet-18 accuracy (%) under a shared comparison protocol. Text block: GoEmotions, BERT-base Macro-F1 (%).

Pool	Encoder	20%	60%	80%
CIFAR-10N	CLIP ViT-L/14 (default)	**81.12** ±0.39	**90.18** ±0.27	**91.36** ±0.28
CIFAR-10N	CLIP ViT-B/32	69.58 ±0.94	80.73 ±0.29	82.60 ±0.21
CIFAR-10N	DINOv2 ViT-B/14	71.06 ±1.97	79.85 ±0.25	82.39 ±0.41
CIFAR-10N	ResNet-50 (ImageNet sup.)	66.54 ±0.64	78.25 ±0.46	81.48 ±0.33
GoEmotions	BERT [CLS] (default)	**38.89** ±0.29	**46.14** ±0.52	**48.71** ±0.47
GoEmotions	SimCSE (label-free)	32.25 ±0.97	37.08 ±0.77	43.35 ±0.88
GoEmotions	MPNet (label-free)	27.45 ±0.90	37.57 ±0.72	44.55 ±0.56

Table 18: GoEmotions Weighted-F1 (%), BERT-base, under the unified fine-tuning protocol. Mean over 3 seeds.

Method	Budget	Weighted-F1
C-GeoVal	20%	49.74
C-GeoVal	60%	55.00
C-GeoVal	80%	57.53
Random	80%	**58.84**
Full dataset	100%	58.79

D.8 Per-class analysis on GoEmotions

Table 19 orders the 28 classes by decreasing training frequency and assigns the first 10 to the head, the next 9 to the medium group, and the final 9 to the tail. Group F1 is the unweighted mean of the class-level F1 means. Tail retention weights each class retention rate by its training count. At 80%, the H_1 guardrail raises tail F1 from 21.15 to 27.33, while Random reaches 35.39. Head and medium F1 reach 68.89 and 47.67, above Random at 59.34 and 42.97, yielding the higher overall Macro-F1 despite the tail deficit. Tail retention is 45.8%, compared with 80.1% for Random. At 20% and 60%, repair also raises tail F1, from 14.42 to 20.24 and from 17.01 to 23.23, respectively. The guardrail therefore mitigates the tail-class damage from density-weighted coverage, but does not restore tail F1 to the random-sampling level.

Table 19: GoEmotions class-group Macro-F1 (%) and tail-sample retention (%), BERT [CLS], 3-seed summaries. Classes are grouped by training frequency (10/9/9); group values are computed from rounded class-level means.

Budget	Method	Head F1	Medium F1	Tail F1	Tail retention (%)
20%	Random	56.23	**35.83**	20.98	20.6
20%	Facility location	55.80	35.68	**21.51**	19.0
20%	C-GeoVal w/o topology	50.37	21.30	14.42	5.1
20%	C-GeoVal	**66.17**	27.23	20.24	5.1
60%	Random	58.76	41.73	34.41	60.2
60%	Facility location	58.40	**42.12**	**35.44**	58.7
60%	C-GeoVal w/o topology	55.17	31.71	17.01	17.3
60%	C-GeoVal	**71.24**	41.17	23.23	17.7
80%	Random	59.34	42.97	35.39	80.1
80%	Facility location	58.66	41.93	**37.52**	79.4
80%	C-GeoVal w/o topology	58.18	39.78	21.15	42.4
80%	C-GeoVal	**68.89**	**47.67**	27.33	45.8