UNSUPPORTEDNESS IS NOT USEFULNESS: TASK-RELEVANT INFORMATION ACQUISITION FOR OPEN-WORLD ACTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world active learning requires selecting annotations from pools that contain known classes, task-relevant novel classes, and irrelevant out-of-distribution (OOD) samples. Existing acquisition strategies often rely on representation novelty, uncertainty, or coverage as proxies for annotation value. We show that this assumption can fail: samples that are highly unsupported by the labeled set may correspond either to useful novel concepts or to irrelevant distribution shifts. We formalize this mismatch through unsupported directions and introduce Task-Relevance-conditioned Information Gain (TRIG), an acquisition framework that separates task relevance from information value. TRIG learns relevance from queried feedback and incorporates it as a continuous weight inside a batch log-determinant information objective, rather than using relevance as a hard filtering criterion. On a controlled ImageNet-200 open-world benchmark, TRIG achieves the strongest mean final in-task top-1 accuracy among comparable non-oracle selectors, discovers all 80 novel classes, and achieves a favorable Pareto trade-off among task performance, discovery, and irrelevant OOD annotation cost. The same behavior persists across DINOv2 and CLIP (Contrastive Language–Image Pre-training) representations and against recent open-set active-learning strategies. These results show that open-world acquisition should explicitly model task relevance instead of treating representation novelty as a surrogate for annotation value.

1 INTRODUCTION

Active learning increasingly operates over large pools encoded by pretrained visual representations. Such pools may contain known task classes, unseen categories that should expand the task, and irrelevant visual domains. The acquisition problem is therefore not only which example is uncertain or diverse, but which new information is useful for an evolving task.

Many closed-set criteria rely on uncertainty, coverage, gradient diversity, or weak support under the labelled representation (Sener & Savarese, 2018; Ash et al., 2020). Their geometric intuition is useful: an example far from labelled data may add information. The hidden assumption is that the new direction belongs to the task. In an open-world pool, task-relevant novelty and irrelevant OOD variation may both lie far from labelled support, even though only the former should expand the task. Consequently,

$$\boxed{\text{unsupportedness} \neq \text{usefulness}}. \tag{1}$$

Figure 1 illustrates the ambiguity. A deficiency score can find missing representation support yet prefer an irrelevant direction over useful novelty. Annotation value must therefore distinguish task relevance from conditional information and batch diversity.

We instantiate this principle as TRIG. TRIG learns a relevance proxy only from annotations already returned by the acquisition interface, combines it with predictive utility, and optimizes a labelled-set-relative log-determinant batch objective. The active ontology expands when a novel class accumulates enough queried evidence; hidden candidate states remain evaluator-only.

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.

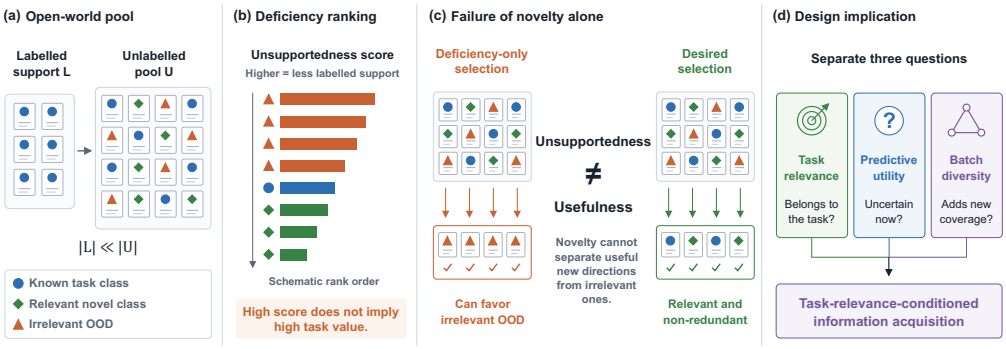

Figure 1: The unsupportedness failure and its design implication. (a) The open-world pool mixes known task classes, relevant novel classes, and irrelevant OOD. (b) Representation deficiency ranks candidates by weak labelled support. (c) Because both relevant novelty and irrelevant variation can be unsupported, deficiency-only selection can favor OOD rather than useful novelty. (d) Open-world acquisition must separate task relevance, predictive utility, and batch diversity. The three-seed ImageNet-200 diagnostic confirms this ambiguity empirically.

We evaluate on ImageNet-200 with 120 initially known classes, 80 discoverable classes, and five near- and far-OOD sources. The soft-deficiency failure criterion is satisfied in all three seeds.

Among comparable non-oracle selectors, TRIG obtains the strongest mean final Global accuracy and promotes all 80 novel classes. BADGE favors earlier learning and discovery at higher OOD cost, while relevance-heavy selectors spend fewer labels on OOD; TRIG provides a favorable trade-off. CLIP ViT-B/16 and matched contemporary open-set active learning (OSAL) adaptations preserve the central conclusion.

The contributions of this work are:

1. We identify and formalize a failure mode of representation-deficiency acquisition in open-world pools: unsupported directions need not be useful directions.
2. We formulate task-relevance-conditioned information acquisition and instantiate it as TRIG, combining queried-feedback relevance with a monotone submodular batch objective under an evolving ontology.
3. We provide an ImageNet-200 study with near- and far-OOD, contemporary baselines, Pareto analysis, and cross-representation validation, showing that the failure persists and TRIG balances task performance, discovery, and OOD annotation cost.

2 Related Work

Closed-set acquisition. Classical active learning assumes that the pool belongs to the target label space. Uncertainty sampling queries decision-boundary points, core-set methods seek representative coverage, BADGE combines uncertainty with gradient-space diversity, and TypiClust favors diverse locally typical samples (Sener & Savarese, 2018; Ash et al., 2020; Hacohen et al., 2022). Bayesian uncertainty, learned loss prediction, adversarial sampling, batch mutual information, and scalable or budget-robust selectors extend this toolbox (Gal et al., 2017; Yoo & Kweon, 2019; Sinha et al., 2019; Kirsch et al., 2019; Citovsky et al., 2021; Yehuda et al., 2022; Bae et al., 2025a;b; Li et al., 2025). These approaches differ in whether diversity is measured in input, representation, prediction, or gradient space. However, their acquisition scores usually presume that every candidate belongs to the task. Our K/N/O evaluation isolates this assumption without exposing candidate identity to the selector.

Open-set and open-world acquisition. Open-set recognition rejects unseen categories, while generalized category discovery models labelled and unlabelled data spanning seen and unseen classes (Bendale & Boult, 2016; Vaze et al., 2022a). Open-set active learners include LfOSA's

activation statistics, EOAL's known–unknown entropy signals, EAOA's energy-based uncertainty, and E^{2}OAL's use of queried unknowns (Ning et al., 2022; Safaei et al., 2024; Zong & Huang, 2025; Zong et al., 2026). Open-world discovery adds a second unknown: a sample may be absent from the classifier yet remain task-relevant; Joda therefore combines OOD filtering and discovery (Schmidt et al., 2025). ALOE prioritizes likely discovery targets, whereas EB-OSAL filters likely unknowns before informativeness scoring (Xie et al., 2025; Lyu & Beksi, 2026). SECOS uses cross-modal semantic knowledge and Stylist removes environment-biased dimensions (Liu et al., 2026; Smeu et al., 2025). Many of these methods treat unknownness mainly as a rejection signal or a gate before conventional acquisition. In our K/N/O setting, this is insufficient because N should expand the task while O should be downweighted. TRIG learns that distinction from paid feedback and uses continuous relevance inside, rather than before, the information objective. Matched adaptations and native author protocols are reported separately.

Representation and batch information. Null-space residuals and spectral leverage identify directions weakly supported by the labelled representation, but do not reveal whether the missing direction is useful. This distinction becomes important with foundation representations: a strong encoder can expose many coherent directions outside current labelled support, including both useful semantic structure and irrelevant domains. DINOv2 provides the principal representation, CLIP supplies a cross-family check, and OpenOOD separates near- from far-OOD (Oquab et al., 2024; Radford et al., 2021; Zhang et al., 2024). Log-determinant, D-optimal, determinantal, and submodular objectives reward complementary batches and support standard greedy guarantees (Nemhauser et al., 1978). TRIG connects these ideas by measuring information relative to labelled coverage while conditioning each contribution on learned task relevance and predictive utility.

3 The unsupportedness failure in open-world acquisition

3.1 Pool, information access, and evolving ontology

Let the opaque unlabelled pool be

$$\mathcal{U} = \mathcal{K} \cup \mathcal{N} \cup \mathcal{O}, \tag{2}$$

where $\mathcal{K}$ contains initially known classes, $\mathcal{N}$ contains relevant but undiscovered classes, and $\mathcal{O}$ contains irrelevant OOD samples. This partition is evaluator-only. The acquisition policy receives candidate features, model outputs, opaque pool indices, and previous annotations, but not the hidden state, source dataset, or future class label of an unqueried candidate.

The initial labelled set contains samples from 120 known ImageNet-200 classes. A queried novel sample becomes positive feedback for relevance, but enters the supervised classifier only after its class is promoted. Promotion occurs after ten observed labels and expands the active ontology.

Signal	Acquisition	Evaluator
Frozen feature and probe output	yes	yes
Previously queried annotation	yes	yes
Candidate K/N/O state or class	no	yes
OOD source or shift identity	no	yes

3.2 Objectives

Open-world acquisition has several competing objectives. We report global open-world accuracy, area under the global learning curve (AULC), novel-class discovery area under the curve (AUC), promoted classes, useful-query precision, and OOD query rate, with near-OOD and far-OOD reported separately. The multi-objective view distinguishes policies that learn the target task from policies that mainly avoid OOD or maximize early discovery.

For a query sequence with cumulative annotation costs c_t and values y_t, we compute the normalized area

$$\mathrm{AUC}(c, y) = \frac{1}{c_T - c_0} \int_{c_0}^{c_T} y(c)\, dc, \tag{3}$$

using the same piecewise-linear interpolation for all methods. All queries, including irrelevant OOD queries, consume the same annotation budget.

3.3 Unsupportedness is not usefulness

The formal motivation uses normalized representations and the local decomposition

$$\mathbb{R}^d = \mathcal{S}_L \oplus \mathcal{T} \oplus \mathcal{O}, \tag{4}$$

where $\mathcal{S}_L$ is the labelled-supported subspace, $\mathcal{T}$ is unsupported but task-relevant, and $\mathcal{O}$ is unsupported and task-irrelevant. Thus the unsupported subspace is $\mathcal{N} = \mathcal{T} \oplus \mathcal{O}$.

For a normalized representation z, the hard null-space score is

$$s_{\mathrm{NS}}(z) = \|P_{\mathcal{N}} z\|_2^2. \tag{5}$$

If $z_O \in \mathcal{O}$ is normalized, then $s_{\mathrm{NS}}(z_O) = 1$. Let $G \succeq 0$ be a task-information operator with range contained in $\mathcal{T}$. Then $z_O^\top G z_O = 0$. Conversely, for

$$z_N = \sqrt{1-\delta}\, z_S + \sqrt{\delta}\, z_T, \qquad 0 < \delta < 1, \tag{6}$$

with orthogonal unit vectors $z_S \in \mathcal{S}_L$ and $z_T \in \mathcal{T}$, we have $s_{\mathrm{NS}}(z_N) = \delta < 1$ while $z_N^\top G z_N > 0$ whenever $z_T^\top G z_T > 0$. Hence geometric unsupportedness cannot determine task value.

The same issue remains for the soft spectral score. With

$$A_L = \lambda I + \sum_{i \in L} z_i z_i^\top, \qquad d_{\mathrm{spec}}(z) = z^\top A_L^{-1} z, \tag{7}$$

low-eigenvalue directions in both $\mathcal{T}$ and $\mathcal{O}$ can have high leverage. This motivates task relevance as a separate factor rather than as a post-hoc interpretation of novelty.

4 Task-Relevant Information Acquisition

The failure in Section 3 motivates a design principle: relevance decides whether a candidate belongs to the evolving task, predictive utility measures current learning value, and labelled-set-relative geometry controls batch redundancy. TRIG is the frozen instantiation shown in Figure 2.

4.1 Relevance from observed feedback

At round zero, no external OOD negatives are available, so $q_{\mathrm{rel}}(x) = 1$ for every candidate. Later rounds refit a deterministic linear binary head on frozen 768-dimensional DINOv2 features. Positives are all initially labelled known samples and all previously queried task-relevant samples, including unpromoted novel samples; negatives are previously queried irrelevant-OOD samples. The head uses full-batch Adam, class-balanced weights, no probe logits, and no unqueried pool labels. The resulting score estimates whether a candidate contributes to the evolving task ontology rather than being novel only in feature space:

$$q_{\mathrm{rel}}(x) = \sigma\big((a^\top z_x + b)/2\big) \in [0, 1]. \tag{8}$$

The temperature 2 is fixed. If no queried OOD negative exists, the estimator returns $q_{\mathrm{rel}}(x) = 1$. Zero initialization and fixed full-batch order make the CPU fit deterministic. The rule keeps evaluator truth outside the acquisition API and makes relevance an evolving task signal.

4.2 Information utility and labelled geometry

Let $H_{\mathrm{norm}}(x)$ be the probe's entropy normalized to the logarithm of the active-class count. The reported TRIG implementation uses

$$u(x) = \beta + (1-\beta) H_{\mathrm{norm}}(x), \qquad \beta = 0.2, \tag{9}$$

and the candidate weight

$$w(x) = q_{\mathrm{rel}}(x) u(x). \tag{10}$$

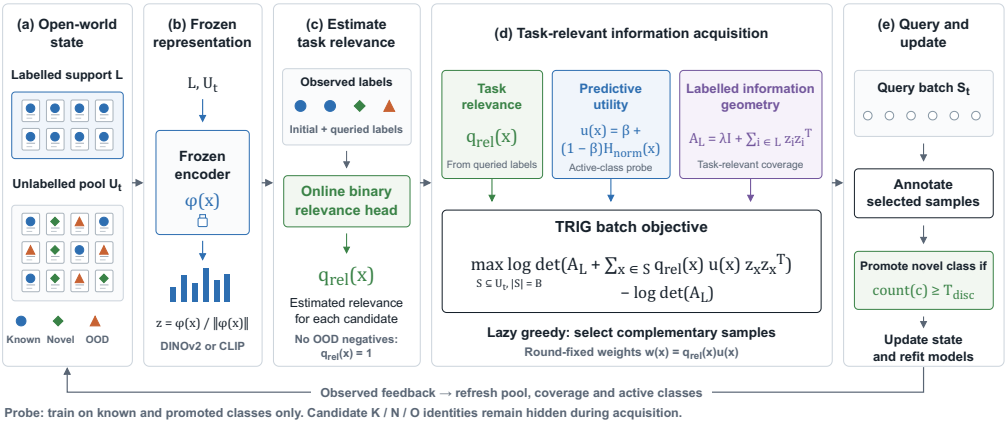

Figure 2: Overview of TRIG. A frozen encoder maps the labelled support and current pool to normalized features. Previously observed annotations train an online relevance head, while the active-class probe supplies predictive utility and the labelled set defines information geometry. TRIG combines these signals inside a relevance-conditioned log-determinant batch objective, queries the selected samples, promotes sufficiently observed novel classes, and updates the next-round state. Candidate K/N/O identities remain hidden during acquisition.

The constant term prevents the selector from suppressing all candidates that are confident under the current known-class probe.

With normalized features, the labelled information matrix is

$$A_L = \lambda I + \sum_{i \in L} z_i z_i^\top, \tag{11}$$

where the frozen protocol uses $\lambda = 10^{-2}$ and unit labelled weights. Here $L = L_{\text{cov}}$ contains all initial known samples and all previously queried task-relevant samples, including unpromoted novel samples; queried irrelevant OOD and unqueried candidates are excluded. This coverage set is broader than the linear probe's supervised set, which includes a novel class only after promotion. For a batch S, TRIG maximizes

$$f(S) = \log\det\left(A_L + \sum_{x \in S} w(x) z_x z_x^\top\right) - \log\det A_L. \tag{12}$$

The round-fixed weights $w(x)$ define a continuous relevance relaxation. The batch is constructed by deterministic lazy greedy with marginal gain $\log(1 + w(x) z_x^\top A_S^{-1} z_x)$. The implementation solves for A_L^{-1}, applies Sherman–Morrison rank-one updates, breaks ties by stable opaque candidate order, keeps heap control flow on CPU, and executes matrix operations on CUDA.

4.3 Batch objective and optimization

For fixed nonnegative candidate weights, Equation 12 is normalized, monotone, and submodular. Greedy selection therefore has the standard cardinality-constrained approximation guarantee. We use deterministic lazy greedy with stable opaque-index tie breaking; full proofs, numerical qualifications, and the distinction from exact Bernoulli expected information appear in Appendix A.4.

5 Main Results

5.1 Benchmark and evaluation protocol

The primary benchmark uses OpenOOD's ImageNet-200 subset (Deng et al., 2009; Yang et al., 2022; Zhang et al., 2024) with 120 initial known classes and 80 discoverable novel classes. Each

Table 1: Compact ImageNet-200 comparison under the unified protocol (three-seed mean ± sample standard deviation). Representative rows expose the principal task, discovery, deficiency, relevance-filtering, and contemporary-OSAL regimes; the complete 18-method comparison is in Appendix B.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Discovered ↑	OOD ↓
Random	0.9450 ± 0.0026	0.8901 ± 0.0030	0.8633 ± 0.0038	$\mathbf{80.0000 \pm 0.0000}$	0.2505 ± 0.0008
BADGE	0.9517 ± 0.0032	$\mathbf{0.9047 \pm 0.0059}$	$\mathbf{0.8973 \pm 0.0101}$	79.6667 ± 0.5774	0.8005 ± 0.0020
TypiClust	0.9353 ± 0.0040	0.8719 ± 0.0053	0.8462 ± 0.0050	$\mathbf{80.0000 \pm 0.0000}$	0.3033 ± 0.0016
Null-space baseline	0.7533 ± 0.0124	0.6905 ± 0.0145	0.3463 ± 0.0297	40.6667 ± 2.5166	0.9003 ± 0.0077
Pure spectral	0.6690 ± 0.0226	0.6370 ± 0.0101	0.2048 ± 0.0266	23.0000 ± 4.5826	0.8932 ± 0.0082
Relevance-weighted spectral	0.9483 ± 0.0029	0.8153 ± 0.0178	0.6681 ± 0.0426	79.6667 ± 0.5774	0.1578 ± 0.0048
EOAL-Adapted	0.9290 ± 0.0078	0.8308 ± 0.0094	0.7127 ± 0.0262	78.3333 ± 0.5774	0.2087 ± 0.0040
LfOSA-Adapted	0.5917 ± 0.0075	0.5787 ± 0.0060	0.0469 ± 0.0236	7.0000 ± 1.7321	$\mathbf{0.0477 \pm 0.0092}$
E^2OAL-Adapted	0.6653 ± 0.0320	0.6264 ± 0.0177	0.1644 ± 0.0585	21.6667 ± 8.0829	0.0595 ± 0.0039
TRIG	$\mathbf{0.9637 \pm 0.0040}$	0.8872 ± 0.0025	0.8371 ± 0.0072	$\mathbf{80.0000 \pm 0.0000}$	0.6065 ± 0.0009

method starts from 1,200 labelled examples, uses ten rounds of 2,000 queries, and consumes exactly 20,000 annotations. A novel class is promoted after ten labels. The encoder is frozen DINOv2 ViT-B/14 (Oquab et al., 2024) and a linear probe is trained from scratch for the current active ontology at each round.

The primary ImageNet-200 pool retains all 257,751 ID candidates and adds the complete available OOD sources without replacement: NINCO (Bitterwolf et al., 2023) and SSB-hard (Vaze et al., 2022b) as near-OOD, and iNaturalist (Van Horn et al., 2018), OpenImage-O (Wang et al., 2022; Kuznetsova et al., 2020), and Describable Textures (DTD) (Cimpoi et al., 2014) as far-OOD. The resulting pool has 85,908 OOD candidates and the same shuffled opaque index order is used by all methods for a given seed. Hidden state and source labels are accessed only after querying.

We compare uncertainty sampling (Settles, 2009), Core-set (Sener & Savarese, 2018), BADGE (Ash et al., 2020), TypiClust (Hacohen et al., 2022), deficiency-only controls, and contemporary OSAL selectors: EOAL (Safaei et al., 2024), LfOSA (Ning et al., 2022), EAOA (Zong & Huang, 2025), E^2OAL (Zong et al., 2026), and Joda (Schmidt et al., 2025). The compact main table shows representative operating regimes; the complete 18-method matrix appears in Appendix B. Relevance-weighted rows are controlled variants derived from our relevance signal. The suffix "-Adapted" denotes EOAL, LfOSA, and E^2OAL selectors whose published selection structures are evaluated through the shared frozen-feature, no-leakage interface. Native author-protocol reproductions remain separate in Appendix I; Joda-Reimpl is an equation-level reimplementation. This protocol separation follows calls for realistic and transparent evaluation of deep active learning (Mittal et al., 2025).

The pre-specified failure test computes tie-aware novel-versus-OOD area under the receiver operating characteristic curve (AUROC) from evaluator-side states; a seed passes when soft spectral AUROC is below 0.5, and the aggregate gate requires at least two of three seeds. Main outcomes are final Global accuracy, global AULC, novel-class discovery AUC, promoted classes, and irrelevant-OOD query rate. *Global* is evaluator-side top-1 accuracy on a fixed in-task test set spanning K∪N. Not-yet-promoted novel classes remain in the denominator and receive zero credit because their labels are absent from the probe output space. Semantic OOD is excluded from classification and measured through irrelevant-query statistics. Three-seed means use sample standard deviations; bold cells mark the best displayed mean in the column direction.

5.2 Deficiency confounds novelty with irrelevant OOD

The formal ImageNet-200 failure test confirms the ambiguity illustrated in Figure 1. Soft spectral novel-versus-OOD AUROC is below 0.5 in every seed, confirming the failure across seeds. The result is consistent with Proposition 1: weak labelled support does not determine whether a direction is useful to the task. We report the hard null-space score as an auxiliary diagnostic, while the soft score defines the failure test.

5.3 Main ImageNet-200 comparison

Table 1 reports a compact view of the unified comparison.

Table 2: Mechanistic comparison on ImageNet-200. Standard BADGE is relevance-free; rel-BADGE is a controlled variant that uses the same learned q_{rel} as TRIG only for a top-$\min(|\mathcal{U}|, 10B)$ relevance prescreen before projected BADGE. TRIG instead uses q_{rel} continuously inside its batch information objective. Values are three-seed mean $\pm$ sample standard deviation.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Discovered ↑	OOD ↓
BADGE	0.9517 ± 0.0032	$\mathbf{0.9047 \pm 0.0059}$	$\mathbf{0.8973 \pm 0.0101}$	79.6667 ± 0.5774	0.8005 ± 0.0020
rel-BADGE	0.9340 ± 0.0079	0.9017 ± 0.0045	0.8915 ± 0.0013	78.3333 ± 0.5774	$\mathbf{0.0547 \pm 0.0007}$
TRIG	$\mathbf{0.9637 \pm 0.0040}$	0.8872 ± 0.0025	0.8371 ± 0.0072	$\mathbf{80.0000 \pm 0.0000}$	0.6065 ± 0.0009

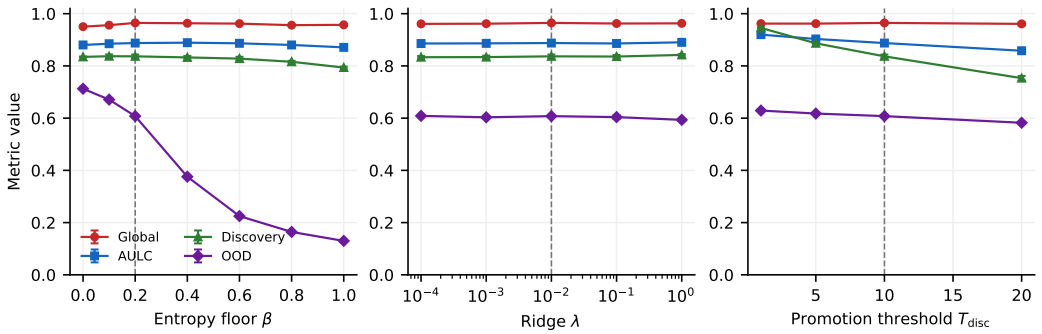

Figure 3: Three-seed one-factor-at-a-time sensitivity of the frozen TRIG configuration on ImageNet-200. Dashed lines mark the main setting. Global performance is stable over broad λ and promotion-threshold ranges; varying β exposes a smooth exploration–OOD trade-off. Curves show means; per-seed values and sample standard deviations are reported in the supplementary files.

TRIG achieves the highest mean final Global accuracy among the comparable non-oracle selectors and promotes all 80 novel classes. The null-space baseline and pure spectral selection degrade sharply and query large amounts of OOD in the full results, empirically connecting the benchmark to the unsupportedness failure. LfOSA-Adapted and E^{2}OAL-Adapted occupy the opposite extreme: they reduce the OOD query rate to 0.0477 and 0.0595, respectively, but reach only 0.5917 and 0.6653 Global accuracy and promote 7.0 and 21.7 novel classes on average. These results make the central trade-off explicit: aggressively avoiding unknown samples can conserve labels while also suppressing task-relevant novel-class discovery.

Different alternatives optimize different coordinates. Standard BADGE, which never uses our relevance estimator, achieves stronger AULC and discovery AUC through aggressive gradient-space exploration, while spending about four fifths of its queries on irrelevant OOD. Relevance-weighted BADGE (rel-BADGE) is a controlled variant: it prescreens candidates with the same learned q_{rel} used by TRIG and then runs standard BADGE. Table 2 shows that rel-BADGE reduces OOD from 0.8005 to 0.0547, but does not improve final Global or complete promotion. TRIG instead uses relevance continuously inside the information objective, yielding the strongest mean final Global and complete promotion. The shared GPU runner also recorded cumulative acquisition-stage wall times lower by factors of 2.3–3.1 for TRIG than for projected BADGE across matched CLIP and covariate-shift repetitions, with moderately higher peak CUDA memory.

5.4 Sensitivity to acquisition settings

Figure 3 tests one factor at a time around the frozen setting $(\beta, \lambda, T_{\text{disc}}) = (0.2, 10^{-2}, 10)$ using the same three seeds, pools, budgets, and ViT-B/14 features as the main benchmark. Global accuracy remains stable across $\lambda \in [10^{-4}, 1]$ and the tested promotion thresholds. Increasing β reduces OOD acquisition but eventually lowers AULC and discovery AUC, exposing the expected relevance–exploration trade-off rather than a narrow optimum.

The harder ImageNet-200 setting exposes a genuine trade-off among early discovery, final task performance, and irrelevant-query cost. Matched three-seed benchmarks on CIFAR-100 (Krizhevsky et al., 2009) and TinyImageNet (Le et al., 2015) provide complementary direct-performance tests.

Table 3: Cross-dataset summary of the matched secondary benchmarks in Tables 11 and 12 (three-seed mean ± sample standard deviation). The same standard Coreset selector is used as the reference method for every metric and both datasets; this fixed reference is not a metric-wise composite baseline. Higher is better except OOD.

Dataset	Metric	Coreset reference	**TRIG**
CIFAR-100	Global ↑	0.5127 ± 0.0081	$\mathbf{0.6943 \pm 0.0136}$
CIFAR-100	AULC ↑	0.5103 ± 0.0106	$\mathbf{0.5616 \pm 0.0201}$
CIFAR-100	Discovery AUC ↑	0.0000 ± 0.0000	$\mathbf{0.1388 \pm 0.0250}$
CIFAR-100	Useful ↑	0.4323 ± 0.0130	$\mathbf{0.8740 \pm 0.0089}$
CIFAR-100	OOD ↓	0.5677 ± 0.0130	$\mathbf{0.1260 \pm 0.0089}$
CIFAR-100	Discovered ↑	0.0 ± 0.0	$\mathbf{21.0 \pm 1.0}$
TinyImageNet	Global ↑	0.5730 ± 0.0095	$\mathbf{0.6540 \pm 0.0072}$
TinyImageNet	AULC ↑	0.4725 ± 0.0098	$\mathbf{0.5210 \pm 0.0015}$
TinyImageNet	Discovery AUC ↑	0.1083 ± 0.0131	$\mathbf{0.2783 \pm 0.0059}$
TinyImageNet	Useful ↑	0.6651 ± 0.0039	$\mathbf{0.8517 \pm 0.0017}$
TinyImageNet	OOD ↓	0.3349 ± 0.0039	$\mathbf{0.1483 \pm 0.0017}$
TinyImageNet	Discovered ↑	38.7 ± 2.5	$\mathbf{65.3 \pm 1.5}$

Table 4: Exact TRIG component ablation on ImageNet-200 (three-seed mean ± sample standard deviation). Each row removes exactly one frozen component.

Variant	Global ↑	AULC ↑	Discovery AUC ↑	OOD ↓
Full TRIG	$\mathbf{0.9637 \pm 0.0040}$	$\mathbf{0.8872 \pm 0.0025}$	$\mathbf{0.8371 \pm 0.0072}$	0.6065 ± 0.0009
Without relevance	0.9067 ± 0.0090	0.7772 ± 0.0041	0.5677 ± 0.0095	0.9058 ± 0.0009
Without entropy utility	0.9573 ± 0.0038	0.8710 ± 0.0031	0.7938 ± 0.0037	$\mathbf{0.1298 \pm 0.0006}$
Without batch log-det	0.9583 ± 0.0046	0.8631 ± 0.0041	0.7850 ± 0.0066	0.3174 ± 0.0222

Table 3 summarizes Tables 11 and 12 in the appendix. To keep the cross-dataset comparison matched, the summary uses the same standard Coreset selector as the reference method for every metric and dataset; the complete appendix tables retain all independent selectors.

Across both datasets, TRIG exceeds the same Coreset reference on all six displayed metrics, including final task accuracy, early learning, discovery, useful-query fraction, OOD avoidance, and promoted-class count. The complete per-method tables and the shared mini-protocol are retained in the appendix.

5.5 Component ablation

The exact one-component removals in Table 4 support the intended decomposition. Without relevance, OOD acquisition rises to 0.906 and both task learning and discovery deteriorate. Removing entropy utility changes OOD much more than final accuracy: OOD falls from 0.6065 to 0.1298, while Global decreases by 0.0064. Removing batch log-det retains all classes but reduces AULC and Discovery AUC, consistent with weaker complementary batch coverage. Thus relevance, predictive utility, and non-redundant batch information govern distinct parts of the open-world trade-off.

The ablation also explains why TRIG is not an OOD filter followed by an ordinary active learner. Constant relevance recovers the unsupportedness failure; stronger filtering makes the query stream purer but can suppress downstream task learning. The entropy ablation identifies the complementary axis: predictive exploration buys earlier learning and discovery at higher irrelevant-query cost.

5.6 Cross-setting validation

Four checks support the same claim without being pooled into one leaderboard. CLIP ViT-B/16 (Radford et al., 2021) tests representation transfer: TRIG obtains Global 0.9240 ± 0.0036, promotes all 80 classes, and reduces OOD from BADGE's 0.5582 to 0.1425. The two secondary datasets give three-seed replications in which TRIG leads independent selectors across all displayed metrics. The modern-OSAL track shows stronger task-performance means for TRIG than for EOAL-Adapted, LfOSA-Adapted, and E^{2}OAL-Adapted, while the latter two expose the cost of

Table 5: Evaluator-only relevance diagnostics on task-relevant ImageNet-R and ImageNet-V2 shifts (three-seed mean ± sample standard deviation). Mean AUC averages rounds 1–9; final recall is measured at a 10% irrelevant-OOD false-positive rate. Shift samples remain outside acquisition. The complete four-method comparison is in Appendix Table 18.

Shift	Method	Mean AUC ↑	Final AUC ↑	Final recall@10 ↑
ImageNet-R	BADGE	0.7714 ± 0.0096	0.7757 ± 0.0014	0.3753 ± 0.0052
ImageNet-R	**TRIG**	$\mathbf{0.7937 \pm 0.0069}$	$\mathbf{0.8262 \pm 0.0019}$	$\mathbf{0.5116 \pm 0.0032}$
ImageNet-V2	BADGE	0.8898 ± 0.0018	0.8890 ± 0.0023	0.6318 ± 0.0099
ImageNet-V2	**TRIG**	$\mathbf{0.9041 \pm 0.0008}$	$\mathbf{0.9190 \pm 0.0005}$	$\mathbf{0.7482 \pm 0.0030}$

optimizing OOD rejection too strongly. Full matrices and protocols are in the appendix. Appendix Table 17 separately varies query granularity at a fixed total budget. Smaller batches mainly improve feedback timing, raising AULC and discovery AUC while keeping final Global stable.

6 Discussion

The relevance estimator models membership in the evolving task, not generic proximity to the training distribution. This matters because task-related covariate shifts may look unusual without being irrelevant. ImageNet-R (Hendrycks et al., 2021) and ImageNet-V2 (Recht et al., 2019) remain outside acquisition and are used only for evaluator-side diagnostics. Table 5 shows that TRIG ranks both shifts above semantic OOD more reliably than the same BADGE reference throughout acquisition and at the final round. The single-seed ImageNet-C diagnostic in Appendix H.1 shows the same qualitative behavior.

The score acquires its semantics from feedback rather than from a fixed definition of the training distribution. Queried known and novel samples provide positive evidence, including before novel-class promotion, whereas queried irrelevant samples provide negative evidence. Relevance therefore changes with the active ontology as annotation feedback accumulates.

Open-world acquisition couples final task learning, ontology discovery, and irrelevant-query cost. BADGE emphasizes early learning and discovery at high OOD cost, while conservative relevance-heavy selectors can reduce OOD but delay task expansion. TRIG instead favors final task coverage: it combines the strongest mean final Global with complete ontology promotion at an intermediate OOD cost. The matched-cost views in Appendix G expose these different operating preferences rather than collapsing them into one score.

Limitations. Our controlled study freezes DINOv2 or CLIP features and retrains a linear probe to isolate acquisition behavior; end-to-end representation adaptation remains to be tested. The relevance head also has a cold start until queried feedback supplies irrelevant examples. The evaluation uses simulated labels, fixed promotion thresholds, and finite OOD sources. Extending the protocol to ambiguous labels, annotator disagreement, open-vocabulary relations, and variable annotation costs is an important next step.

7 Conclusion

Our study shows that representation novelty alone is insufficient for open-world active learning: task-relevant novel classes and irrelevant OOD samples can both appear as unsupported relative to the labeled set. TRIG learns task relevance from queried annotation feedback and integrates it with predictive utility through a labelled-set-relative log-determinant objective. On ImageNet-200, TRIG achieves the strongest mean final Global accuracy among comparable non-oracle selectors, promotes all 80 novel classes, and provides a favorable trade-off among task learning, discovery, and OOD annotation cost. More broadly, open-world acquisition should identify which unsupported samples represent meaningful task expansion instead of optimizing novelty alone. The resulting principle is simple: first determine whether missing information belongs to the evolving task, then evaluate how informative and complementary it is.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading, improving the clarity and readability of the manuscript, and assisting with the writing, debugging, and refinement of code. We did not use generative AI tools for generating research ideas, designing algorithms, conducting experiments, analyzing results, or producing scientific claims. All AI-assisted text modifications and code were carefully reviewed, verified, and tested by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Wonho Bae, Junhyug Noh, and Danica J. Sutherland. Generalized coverage for more robust low-budget active learning. In *European Conference on Computer Vision*, pp. 318–334, 2025a.

Wonho Bae, Danica J. Sutherland, and Gabriel Oliveira. Uncertainty herding: One active learning method for all label budgets. In *International Conference on Learning Representations*, 2025b.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506. PMLR, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192. PMLR, 2017.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8175–8195, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

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

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899, 2025. doi: 10.1109/TNNLS.2024.3396463.

Hezhao Liu, Jiacheng Yang, Junlong Gao, Mengke Li, Yiqun Zhang, Shreyank N. Gowda, and Yang Lu. SECOS: Semantic capture for rigorous classification in open-world semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39627–39636, 2026.

Zongyao Lyu and William J. Beksi. Energy-based open-set active learning for object classification. In *Pattern Recognition – ICPR 2026*, volume 16817 of *Lecture Notes in Computer Science*, pp. 570–586. Springer, 2026. First online 3 August 2026.

Sudhanshu Mittal, Joshua Niemeijer, Oezguen Cicek, Maxim Tatarchenko, Jan Ehrhardt, Jörg P. Schäfer, Heinz Handels, and Thomas Brox. Realistic evaluation of deep active learning for image classification and semantic segmentation. *International Journal of Computer Vision*, 133(7): 4294–4316, 2025.

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.

Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–49, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2024. URL `https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/`.

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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019.

Bardia Safaei, Vibashan VS, Celso M. de Melo, and Vishal M. Patel. Entropic open-set active learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4686–4694, 2024.

Sebastian Schmidt, Leonard Schenk, Leo Schwinn, and Stephan Günnemann. Joint out-of-distribution filtering and data discovery active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25677–25687, 2025.

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

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Stefan Smeu, Elena Burceanu, Emanuela Haller, and Andrei Liviu Nicolicioiu. Robust novelty detection through style-conscious feature ranking. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 7409–7418, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7492–7501, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022b.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Tian Xie, Jifan Zhang, Haoyue Bai, and Robert D. Nowak. Deep active learning in the open world. *Transactions on Machine Learning Research*, 2025.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 35, 2022.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. *Advances in Neural Information Processing Systems*, 35:22354–22367, 2022.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD V1.5: Enhanced Benchmark for Out-of-Distribution Detection. *Data-centric Machine Learning Research*, 2:1–32, 2024. URL `https://mlanthology.org/dmlr/2024/zhang2024dmlr-openood/`.

Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10153–10162, 2025.

Chen-Chen Zong, Yu-Qi Chi, Xie-Yang Wang, Yan Cui, and Sheng-Jun Huang. Revisiting unknowns: Towards effective and efficient open-set active learning. *arXiv preprint arXiv:2603.07898*, 2026.

A Theory and experimental protocol

A.1 Exact ImageNet-200 protocol

The primary ImageNet-200 pool contains 257,751 ID candidates and 85,908 OOD candidates. The OOD sources are NINCO (Bitterwolf et al., 2023) and SSB-hard (Vaze et al., 2022b) (near-OOD),

and iNaturalist (Van Horn et al., 2018), OpenImage-O (Wang et al., 2022; Kuznetsova et al., 2020), and DTD Textures (Cimpoi et al., 2014) (far-OOD). The initial labelled set has 1,200 samples, the query schedule is ten rounds of 2,000, and the promotion threshold is ten labels per novel class. All methods share the same per-seed shuffled pool order and exact 20,000 query budget.

A.2 No-leakage boundary

The acquisition API is given frozen candidate features, current probe outputs, opaque candidate indices, and annotations returned for previously queried samples. Candidate-side K/N/O membership, source names, hidden class labels, and covariate-shift labels are evaluator-owned. The evaluator reconstructs K/N/O, near- and far-OOD membership, useful precision, promotion timing, and discovery curves only after queries have been selected.

A.3 Formal method definitions

The exact frozen definitions are:

- spectral: deterministic top-B by d_{spec};
- relevance-weighted entropy: deterministic top-B by $q_{\mathrm{rel}} H_{\mathrm{norm}}$;
- relevance-weighted spectral: deterministic top-B by $q_{\mathrm{rel}} d_{\mathrm{spec}}$;
- relevance-weighted BADGE: deterministic top-$\min(|\mathcal{U}|, 10B)$ relevance prescreen followed by projected BADGE;
- relevance-weighted Coreset: the same relevance prescreen followed by cosine k-center;
- TRIG: greedy log-determinant selection with candidate weight $q_{\mathrm{rel}}[0.2 + 0.8 H_{\mathrm{norm}}]$, unweighted labelled matrix, and the configured positive ridge.

The selector always returns the exact available budget and records fallback behavior when a method reaches a degenerate case.

A.4 Deficiency and log-determinant results

Let the normalized labelled feature matrix have eigendecomposition

$$\sum_{i \in L} z_i z_i^\top = V \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2) V^\top, \tag{13}$$

and write v_j for the columns of V.

Theorem 1 (Hard-to-soft deficiency limit)**.** *Partition the eigendirections into a supported set $\mathcal{S}$ and an unsupported set $\mathcal{U}$. For any normalized z,*

$$\lambda d_{\mathrm{spec}}(z) = \sum_{j=1}^{d} \frac{\lambda}{\lambda + \sigma_j^2} (v_j^\top z)^2. \tag{14}$$

Moreover,

$$\begin{aligned} \left| \lambda d_{\mathrm{spec}}(z) - \|P_{\mathcal{U}} z\|_2^2 \right| \le & \max_{j \in \mathcal{S}} \frac{\lambda}{\lambda + \sigma_j^2} \|P_{\mathcal{S}} z\|_2^2 \\ & + \max_{j \in \mathcal{U}} \frac{\sigma_j^2}{\lambda + \sigma_j^2} \|P_{\mathcal{U}} z\|_2^2. \end{aligned} \tag{15}$$

Thus the ridge-scaled soft score approaches hard unsupported-subspace residual energy when supported eigenvalues dominate λ and unsupported eigenvalues are negligible relative to λ.

Proof. Diagonalizing $A_L = V \operatorname{diag}(\lambda + \sigma_j^2) V^\top$ gives the first identity. Subtracting the indicator-weighted expansion of $\|P_{\mathcal{U}} z\|_2^2$ and bounding the supported and unsupported sums separately gives the inequality. □

Table 6: Representation-deficiency failure on ImageNet-200. Soft spectral novel-versus-OOD AUROC below 0.5 is the pre-specified seed-level failure criterion.

Seed	Soft AUROC ↓	Hard AUROC	Top-250 N	Top-250 O	Top-250 OOD
0	0.3830	0.4353	250	0	0.0000
1	0.3352	0.3834	216	22	0.0880
2	0.3280	0.3783	196	47	0.1880

Proposition 1 (Unsupportedness is not sufficient)**.** *Suppose $\mathbb{R}^d = \mathcal{S}_L \oplus \mathcal{T} \oplus \mathcal{O}$, where $\mathcal{T}$ is task-relevant and $\mathcal{O}$ is task-irrelevant. Let $G \succeq 0$ have range contained in $\mathcal{T}$. There exist normalized z_O, z_N such that*

$$s_{\mathrm{NS}}(z_O) > s_{\mathrm{NS}}(z_N) \quad \textit{but} \quad z_O^\top G z_O < z_N^\top G z_N. \tag{16}$$

Proof. Choose a unit $z_O \in \mathcal{O}$ and $z_N = \sqrt{1-\delta} z_S + \sqrt{\delta} z_T$ for orthogonal unit vectors $z_S \in \mathcal{S}_L$, $z_T \in \mathcal{T}$, and $0 < \delta < 1$ with $z_T^\top G z_T > 0$. Orthogonality gives $s_{\mathrm{NS}}(z_O) = 1 > \delta = s_{\mathrm{NS}}(z_N)$, whereas the range condition gives $z_O^\top G z_O = 0 < \delta z_T^\top G z_T = z_N^\top G z_N$. □

The preceding theorem transfers this counterexample to the soft score whenever both $\mathcal{T}$ and $\mathcal{O}$ lie in weakly supported directions: soft spectral deficiency converges to the same unsupported energy that cannot distinguish their task value. This is an existence result and does not claim that every empirical pool realizes the construction.

Theorem 2 (Round-fixed TRIG objective)**.** *Fix one acquisition round, $A_L \succ 0$, and nonnegative candidate weights w_x. Define*

$$f(S) = \log\det\left(A_L + \sum_{x \in S} w_x z_x z_x^\top\right) - \log\det A_L. \tag{17}$$

Then f is normalized, monotone, and submodular. Under a cardinality constraint $|S| \le B$, exact greedy selection achieves the standard $(1 - 1/e)$ approximation to the optimal batch.

Proof. Normalization follows from $f(\varnothing) = 0$. With $A_S = A_L + \sum_{x \in S} w_x z_x z_x^\top$, the matrix determinant lemma gives

$$\Delta(x \mid S) = \log\bigl(1 + w_x z_x^\top A_S^{-1} z_x\bigr) \ge 0, \tag{18}$$

which proves monotonicity. If $S \subseteq T$, then $A_S \preceq A_T$ and hence $A_S^{-1} \succeq A_T^{-1}$, so $\Delta(x \mid S) \ge \Delta(x \mid T)$. This is diminishing returns and therefore submodularity. The cardinality-constrained greedy guarantee follows from the classical bound for normalized monotone submodular maximization. The implementation's lazy heap computes the same greedy choices in exact arithmetic; float32 execution is an implementation approximation, not a stronger theorem. □

Finally, q_{rel} weighting is a continuous relaxation rather than exact marginalization. If $R_x \sim \mathrm{Bernoulli}(q_x)$, concavity of log determinant implies

$$\mathbb{E}\left[\log\det\left(A_L + \sum_x R_x B_x\right)\right] \le \log\det\left(A_L + \sum_x q_x B_x\right), \tag{19}$$

so substituting $q_x B_x$ gives a deterministic soft surrogate (and an upper bound on this expected log determinant), not equality in general.

B Complete ImageNet-200 results

Table 7 retains every independent selector and internal mechanism control from the unified comparison; the BADGE, rel-BADGE, and TRIG mechanism slice is reported in the main paper in Table 2. The total OOD query rate is retained in the full matrix, while the representative near- and far-OOD annotation composition is visualized in Figure 4. Oracle diagnostics are excluded from this leaderboard and remain upper-bound analyses.

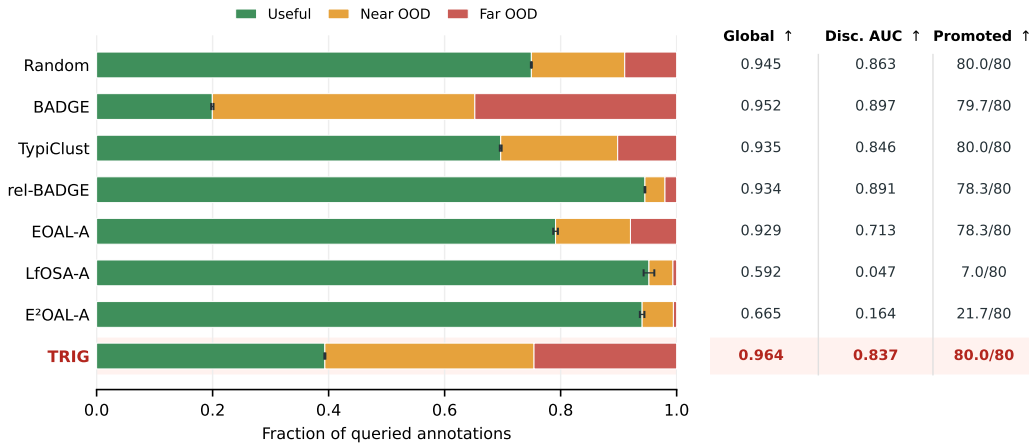

Figure 4: Annotation composition and task outcomes for representative ImageNet-200 operating regimes (three-seed means). Bars partition the query budget into useful task samples, near OOD, and far OOD; thin whiskers mark the sample standard deviation of total OOD at the useful/OOD boundary. The aligned columns report final Global accuracy, novel Discovery AUC, and promoted novel classes. Random and TypiClust provide non-relevance references; BADGE and rel-BADGE isolate aggressive exploration and relevance prescreening; "-A" denotes a protocol-matched adapted competitor. The complete metrics and all 18 methods are reported in Table 7.

Table 7: Full ImageNet-200 unified comparison (three-seed mean ± sample standard deviation). Qualified "-Adapted" rows preserve contemporary selector structures under the shared feature protocol; relevance-weighted rows are controlled variants.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Discovered ↑	OOD ↓
Random	0.9450 ± 0.0026	0.8901 ± 0.0030	0.8633 ± 0.0038	**80.0000 ± 0.0000**	0.2505 ± 0.0008
Entropy	0.8893 ± 0.0086	0.7992 ± 0.0051	0.6285 ± 0.0117	70.3333 ± 2.0817	0.8705 ± 0.0057
Margin	0.9413 ± 0.0015	0.8888 ± 0.0061	0.8585 ± 0.0194	79.0000 ± 1.0000	0.8548 ± 0.0018
Coreset	0.8963 ± 0.0229	0.7688 ± 0.0071	0.5431 ± 0.0212	69.6667 ± 3.5119	0.7495 ± 0.0027
BADGE	0.9517 ± 0.0032	**0.9047 ± 0.0059**	**0.8973 ± 0.0101**	79.6667 ± 0.5774	0.8005 ± 0.0020
TypiClust	0.9353 ± 0.0040	0.8719 ± 0.0053	0.8462 ± 0.0050	**80.0000 ± 0.0000**	0.3033 ± 0.0016
Null-space baseline	0.7533 ± 0.0124	0.6905 ± 0.0145	0.3463 ± 0.0297	40.6667 ± 2.5166	0.9003 ± 0.0077
Pure spectral	0.6690 ± 0.0226	0.6370 ± 0.0101	0.2048 ± 0.0266	23.0000 ± 4.5826	0.8932 ± 0.0082
Relevance-weighted entropy	0.9523 ± 0.0055	0.8601 ± 0.0050	0.7833 ± 0.0146	**80.0000 ± 0.0000**	0.4854 ± 0.0276
Relevance-weighted spectral	0.9483 ± 0.0029	0.8153 ± 0.0178	0.6681 ± 0.0426	79.6667 ± 0.5774	0.1578 ± 0.0048
Relevance-weighted BADGE	0.9340 ± 0.0079	0.9017 ± 0.0045	0.8915 ± 0.0013	78.3333 ± 0.5774	0.0547 ± 0.0007
Relevance-weighted Coreset	0.8450 ± 0.0060	0.7814 ± 0.0116	0.5750 ± 0.0315	58.6667 ± 1.1547	0.0659 ± 0.0008
Relevance-weighted null-space	0.9040 ± 0.0135	0.8117 ± 0.0181	0.6623 ± 0.0434	72.3333 ± 2.0817	0.0748 ± 0.0059
Joda-Reimpl†	0.7503 ± 0.0025	0.7018 ± 0.0126	0.3658 ± 0.0313	40.0000 ± 1.7321	0.1359 ± 0.0129
EOAL-Adapted	0.9290 ± 0.0078	0.8308 ± 0.0094	0.7127 ± 0.0262	78.3333 ± 0.5774	0.2087 ± 0.0040
LfOSA-Adapted	0.5917 ± 0.0075	0.5787 ± 0.0060	0.0469 ± 0.0236	7.0000 ± 1.7321	**0.0477 ± 0.0092**
E^2OAL-Adapted	0.6653 ± 0.0320	0.6264 ± 0.0177	0.1644 ± 0.0585	21.6667 ± 8.0829	0.0595 ± 0.0039
TRIG	**0.9637 ± 0.0040**	0.8872 ± 0.0025	0.8371 ± 0.0072	**80.0000 ± 0.0000**	0.6065 ± 0.0009

C Secondary Benchmarks and Backbone Controls

Auxiliary failure maps reproduce the novel-versus-OOD inversion on CIFAR-100 and TinyImageNet. These evaluator-only diagnostics support the failure principle but are not pooled with the ImageNet-200 benchmark.

C.1 Backbone scale controls

The ImageNet-200 ViT-S/14 single-seed control preserves the qualitative ordering between TRIG and deficiency-only selectors. A three-seed TinyImageNet comparison further shows similar TRIG discovery and OOD behavior across ViT-S/14 and ViT-B/14; these controls do not replace the stronger cross-family CLIP result.

Table 8: Auxiliary evaluator-only failure diagnostics. A soft spectral AUROC below 0.5 means that the deficiency score ranks irrelevant OOD above task-relevant novel samples on average. The K/N/O labels are used only after acquisition; they are not exposed to the selector.

Dataset	Backbone	Soft novel-versus-OOD AUROC ↓	Top-250 OOD ↓	Top-250 K/N/O
CIFAR-100	ViT-S/14	0.2116	0.9840	0/4/246
TinyImageNet	ViT-S/14	0.2057	0.8400	4/36/210
TinyImageNet	ViT-B/14	0.3272	0.5600	1/109/140

Table 9: ViT-S/14 scale-control experiment on the same ImageNet-200 pool. This is a single-seed robustness diagnostic, not a three-seed main result.

Method	Global	AULC	Discovery AUC	Discovered	OOD
BADGE	0.9300	**0.8753**	**0.8800**	**80**	0.7123
Null-space baseline	0.6690	0.6179	0.2087	30	0.9299
Pure spectral	0.6280	0.5927	0.1300	18	0.9422
TRIG	**0.9460**	0.8632	0.8413	**80**	**0.3587**

C.2 Matched secondary benchmark

The secondary benchmark uses the same frozen DINOv2 ViT-S/14 feature track and a symmetric three-seed protocol on CIFAR-100 (Krizhevsky et al., 2009) and TinyImageNet (Le et al., 2015), which is derived from ImageNet (Deng et al., 2009). We report the independent baseline selectors, including TypiClust (Hacohen et al., 2022), and TRIG under the same initial-label, query, promotion, and evaluator-only no-leakage rules within each dataset. Relevance-weighted variants are controlled variants and are not presented as secondary competitors. These tables are auxiliary cross-dataset evidence; they are not pooled with the ImageNet-200 primary benchmark.

D Ablations and Sensitivity

D.1 Exact TRIG component ablations

The ablations remove one component at a time under the unchanged ImageNet-200 protocol. They are reported as diagnostic evidence for the decomposition and do not define a new method variant.

D.2 Relevance-estimator ablation

Table 14 changes only q_{rel} while retaining the frozen TRIG utility and batch objective. Constant relevance (R0) spends most labels on OOD. The positive-only R1 model nearly eliminates OOD queries, but its weak novel/OOD ranking collapses class discovery, illustrating that fitting the known task support too tightly is an unsafe open-world relevance rule. Signed graph propagation (R3) improves final-round known-versus-OOD (K/O) and novel-versus-OOD (N/O) relevance AUROC over the binary head and reduces OOD annotations by 24.6 percentage points while still discovering all novel classes. Its final Global accuracy is 0.006 below the binary head and its AULC differs by only 0.0018. We therefore retain the pre-specified binary-head TRIG configuration as the main method. The signed-PPR result shows that relevance calibration and annotation efficiency can improve without monotonically improving task accuracy. The Oracle gap further shows that relevance quality remains consequential, especially for label efficiency and early discovery. Signed PPR constructs its deterministic Hierarchical Navigable Small World (HNSW) graph on CPU while probe training and TRIG log-det selection remain on CUDA; it is an appendix robustness ablation and is excluded from the primary efficiency claim. Bold cells mark the best non-oracle mean; the Oracle row is an evaluator-only upper bound.

D.3 Novel diversity and promotion timing

Across rounds 1–9, the mean number of unique novel classes per round is 53.81 for BADGE, 45.26 for relevance-weighted BADGE, and 68.22 for TRIG. Corresponding normalized novel-class entropies are 0.9403, 0.7349, and 0.9307, while Gini–Simpson diversity is 0.9707, 0.9041, and 0.9750.

Table 10: Three-seed TinyImageNet scale-control diagnostic for TRIG under the same five-round mini-protocol. The table tests qualitative stability across backbone scale; it is not a complete baseline comparison.

Backbone	Global	AULC	Discovery AUC	Discovered	OOD
ViT-S/14	0.6540 ± 0.0072	0.5210 ± 0.0015	0.2783 ± 0.0059	65.3 ± 1.5	$\mathbf{0.1483 \pm 0.0017}$
ViT-B/14	$\mathbf{0.7737 \pm 0.0070}$	$\mathbf{0.5998 \pm 0.0077}$	$\mathbf{0.3004 \pm 0.0166}$	$\mathbf{70.3 \pm 1.2}$	0.1557 ± 0.0076

Table 11: Three-seed CIFAR-100 secondary benchmark under the matched ViT-S/14 mini-protocol (2,000 initial labels, ten rounds of 250 queries). Values are mean $\pm$ sample standard deviation over three seeds. The table contains independent baseline selectors and TRIG; relevance-weighted variants are controlled variants and are not treated as competitors. All selectors use the same opaque pool order per seed; K/N/O and OOD source labels are evaluator-only.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Useful ↑	OOD ↓	Discovered ↑
Random	0.5157 ± 0.0060	0.5118 ± 0.0100	0.0000 ± 0.0000	0.6752 ± 0.0073	0.3248 ± 0.0073	0.0 ± 0.0
Entropy	0.5597 ± 0.0337	0.5224 ± 0.0231	0.0392 ± 0.0361	0.3847 ± 0.0434	0.6153 ± 0.0434	5.3 ± 2.5
Margin	0.5340 ± 0.0229	0.5161 ± 0.0112	0.0042 ± 0.0031	0.5045 ± 0.0293	0.4955 ± 0.0293	2.0 ± 1.7
Coreset	0.5127 ± 0.0081	0.5103 ± 0.0106	0.0000 ± 0.0000	0.4323 ± 0.0130	0.5677 ± 0.0130	0.0 ± 0.0
BADGE	0.5227 ± 0.0108	0.5150 ± 0.0098	0.0017 ± 0.0019	0.5793 ± 0.0164	0.4207 ± 0.0164	0.7 ± 0.6
subsample-BADGE	0.5167 ± 0.0060	0.5126 ± 0.0090	0.0000 ± 0.0000	0.5845 ± 0.0118	0.4155 ± 0.0118	0.0 ± 0.0
TypiClust	0.5137 ± 0.0107	0.5104 ± 0.0110	0.0000 ± 0.0000	0.6532 ± 0.0042	0.3468 ± 0.0042	0.0 ± 0.0
Null-space baseline	0.5123 ± 0.0144	0.5089 ± 0.0121	0.0000 ± 0.0000	0.2191 ± 0.0180	0.7809 ± 0.0180	0.0 ± 0.0
Pure spectral	0.5123 ± 0.0145	0.5089 ± 0.0124	0.0000 ± 0.0000	0.2668 ± 0.0314	0.7332 ± 0.0314	0.0 ± 0.0
Joda-Reimpl†	0.5323 ± 0.0150	0.5143 ± 0.0098	0.0013 ± 0.0013	0.7631 ± 0.0567	0.2369 ± 0.0567	1.0 ± 1.0
TRIG	$\mathbf{0.6943 \pm 0.0136}$	$\mathbf{0.5616 \pm 0.0201}$	$\mathbf{0.1388 \pm 0.0250}$	$\mathbf{0.8740 \pm 0.0089}$	$\mathbf{0.1260 \pm 0.0089}$	$\mathbf{21.0 \pm 1.0}$

The evidence therefore points to delayed early novel mass and promotion rather than a persistent lack of within-novel diversity.

D.4 Online feedback diagnostic

In a three-round CIFAR-100 diagnostic, OOD queries fall from 82.8% at cold start to 8.8% and 4.8% after feedback, while useful-query precision rises from 17.2% to 91.2% and 95.2%. This short diagnostic illustrates adaptation and is not a causal no-feedback ablation.

E Cross-Representation Validation

E.1 Second-foundation-encoder robustness

Table 16 repeats a frozen six-method key matrix with the OpenAI CLIP ViT-B/16 image tower. The experiment uses only 512-dimensional normalized image embeddings: it does not use the text tower, prompts, class names, or any hidden pool state. Pool construction, budgets, promotion, probe training, and acquisition hyperparameters remain matched to the primary ImageNet-200 protocol.

The cross-encoder result shows that TRIG has the highest mean final Global accuracy (0.9240), narrowly above BADGE (0.9210), while reducing the mean OOD query rate from 0.5582 to 0.1425. BADGE and Random have higher AULC and discovery AUC than TRIG. Relevance-weighted spectral selection filters OOD most aggressively but loses final accuracy and class coverage, whereas the null-space baseline and Pure spectral attract substantially more OOD and degrade sharply. Thus the distinction between representation unsupportedness and task usefulness persists under a second foundation encoder, and TRIG retains a useful accuracy–annotation-cost trade-off. The small final-accuracy differences are interpreted conservatively.

F Sensitivity Analysis

F.1 Sensitivity to Acquisition Settings

Figure 3 summarizes one-factor-at-a-time sensitivity around the frozen setting $(\beta, \lambda, T_{\mathrm{disc}}) = (0.2, 10^{-2}, 10)$, using the same three seeds, pools, budgets, and ViT-B/14 features as the main

Table 12: Three-seed TinyImageNet secondary benchmark under the matched ViT-S/14 mini-protocol (2,000 initial labels, five rounds of 250 queries). Values are mean ± sample standard deviation over three seeds. The table contains independent baseline selectors and TRIG; relevance-weighted variants are controlled variants and are not treated as competitors. All selectors use the same opaque pool order per seed; K/N/O and OOD source labels are evaluator-only.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Useful ↑	OOD ↓	Discovered ↑
Random	0.5020 ± 0.0096	0.4560 ± 0.0077	0.0371 ± 0.0036	0.6344 ± 0.0106	0.3656 ± 0.0106	17.0 ± 1.7
Entropy	0.5027 ± 0.0031	0.4582 ± 0.0075	0.0413 ± 0.0109	0.3717 ± 0.0186	0.6283 ± 0.0186	15.0 ± 2.0
Margin	0.5077 ± 0.0050	0.4575 ± 0.0059	0.0358 ± 0.0145	0.4539 ± 0.0122	0.5461 ± 0.0122	16.7 ± 3.8
Coreset	0.5730 ± 0.0095	0.4725 ± 0.0098	0.1083 ± 0.0131	0.6651 ± 0.0039	0.3349 ± 0.0039	38.7 ± 2.5
BADGE	0.5540 ± 0.0082	0.4724 ± 0.0064	0.0879 ± 0.0119	0.5840 ± 0.0090	0.4160 ± 0.0090	30.3 ± 3.2
subsample-BADGE	0.5633 ± 0.0147	0.4723 ± 0.0014	0.0900 ± 0.0119	0.5904 ± 0.0097	0.4096 ± 0.0097	33.3 ± 3.8
TypiClust	0.5603 ± 0.0124	0.4663 ± 0.0076	0.0737 ± 0.0050	0.7451 ± 0.0040	0.2549 ± 0.0040	35.7 ± 4.2
Null-space baseline	0.4660 ± 0.0017	0.4512 ± 0.0076	0.0146 ± 0.0073	0.3216 ± 0.0165	0.6784 ± 0.0165	5.0 ± 2.6
Pure spectral	0.4603 ± 0.0091	0.4504 ± 0.0084	0.0104 ± 0.0059	0.2659 ± 0.0149	0.7341 ± 0.0149	3.0 ± 1.0
Joda-Reimpl[†]	0.4940 ± 0.0321	0.4615 ± 0.0107	0.0658 ± 0.0148	0.5965 ± 0.0517	0.4035 ± 0.0517	16.0 ± 7.0
TRIG	$\mathbf{0.6540 \pm 0.0072}$	$\mathbf{0.5210 \pm 0.0015}$	$\mathbf{0.2783 \pm 0.0059}$	$\mathbf{0.8517 \pm 0.0017}$	$\mathbf{0.1483 \pm 0.0017}$	$\mathbf{65.3 \pm 1.5}$

Table 13: Exact three-seed component ablations of TRIG on the unified ImageNet-200 protocol. Each row removes exactly one frozen component; values are mean ± sample standard deviation.

Variant	Global ↑	AULC ↑	Discovery AUC ↑	Useful ↑	OOD ↓	Discovered ↑
TRIG without relevance	0.9067 ± 0.0090	0.7772 ± 0.0041	0.5677 ± 0.0095	0.0942 ± 0.0009	0.9058 ± 0.0009	74.3 ± 1.2
TRIG without entropy utility	0.9573 ± 0.0038	$\mathbf{0.8710 \pm 0.0031}$	$\mathbf{0.7938 \pm 0.0037}$	$\mathbf{0.8702 \pm 0.0006}$	$\mathbf{0.1298 \pm 0.0006}$	$\mathbf{80.0 \pm 0.0}$
TRIG without batch log-det	$\mathbf{0.9583 \pm 0.0046}$	0.8631 ± 0.0041	0.7850 ± 0.0066	0.6827 ± 0.0222	0.3174 ± 0.0222	$\mathbf{80.0 \pm 0.0}$

benchmark. Performance is stable across $\lambda \in [10^{-4}, 1]$ (Global accuracy 0.961–0.965). Increasing β suppresses OOD queries from 0.713 at $\beta = 0$ to 0.130 at $\beta = 1$, but eventually lowers discovery AUC and AULC. Final Global accuracy is also stable across promotion thresholds (0.961–0.965). Discovery AUC decreases as T_{disc} increases because the threshold directly delays the metric's promotion event.

Table 17 separately fixes the total annotation budget at 20,000 while varying query granularity. Reducing the batch from 2,000 to 500 raises AULC from 0.8875 to 0.9075 and discovery AUC from 0.8367 to 0.8833, consistent with more frequent annotation feedback and probe updates. Final Global remains within 0.0034 for batches 500–2,000, whereas the coarse 5,000-query schedule falls to 0.9530 and has weaker early learning and discovery. Total OOD changes by only 1.4 percentage points over the full range. Thus the main result is not specific to a 2,000-query batch, although coarse feedback measurably delays learning; the main protocol is retained rather than replaced by the post-hoc smaller-batch setting.

G Matched-Cost Analysis

Figure 5 re-parameterizes frozen trajectories by cumulative useful and irrelevant-OOD annotations. Each seed is interpolated only on support shared by every compared method and seed. Over the shared useful-label interval, normalized Global areas are 0.862 for BADGE, 0.833 for relevance-weighted BADGE, and 0.817 for TRIG; discovery areas are 62.82, 57.58, and 53.85. On the common 0–1,079 OOD-annotation interval, TRIG has Global and discovery areas of 0.577 and 2.78, compared with 0.719 and 33.09 for relevance-weighted BADGE and 0.707 and 30.47 for Random. These descriptive evaluator-only views reject universal cost dominance and support the multi-objective interpretation; they are not new acquisition runs.

H Covariate-Shift Diagnostics

ImageNet-R (Hendrycks et al., 2021), ImageNet-V2 (Recht et al., 2019), and ImageNet-C (Hendrycks & Dietterich, 2019) remain outside acquisition and are treated as task-relevant shifts rather than irrelevant OOD. Table 18 measures relevance separation induced by four annotation trajectories; it is not a shifted-image accuracy benchmark.

Table 14: Three-seed relevance-estimator ablation under the unchanged ImageNet-200 TRIG objective. K/O and N/O denote known-versus-OOD and novel-versus-OOD evaluator-only relevance AUROCs in the final acquisition round. R1 uses positive-only cross-fitted Ledoit–Wolf affinity; R3 uses two-channel signed personalized PageRank (PPR). Oracle is a diagnostic upper bound. Values are mean ± sample standard deviation.

Estimator	Global ↑	AULC ↑	Discovery AUC ↑	Useful ↑	OOD ↓	Discovered ↑	K/O AUC ↑	N/O AUC ↑
R0: $q_{\mathrm{rel}} = 1$	0.9067 ± 0.0090	0.7772 ± 0.0041	0.5677 ± 0.0095	0.0942 ± 0.0009	0.9058 ± 0.0009	74.3 ± 1.2	0.5000 ± 0.0000	0.5000 ± 0.0000
R1: one-class affinity	0.6430 ± 0.0147	0.6307 ± 0.0135	0.1806 ± 0.0331	$\mathbf{0.9847 \pm 0.0045}$	$\mathbf{0.0153 \pm 0.0045}$	17.0 ± 2.6	0.9179 ± 0.0042	0.6276 ± 0.0112
Binary head (main)	$\mathbf{0.9647 \pm 0.0031}$	$\mathbf{0.8875 \pm 0.0030}$	0.8367 ± 0.0069	0.3923 ± 0.0017	0.6077 ± 0.0017	$\mathbf{80.0 \pm 0.0}$	0.9320 ± 0.0022	0.9259 ± 0.0030
R3: signed PPR	0.9587 ± 0.0035	0.8858 ± 0.0020	$\mathbf{0.8375 \pm 0.0050}$	0.6379 ± 0.0112	0.3621 ± 0.0112	$\mathbf{80.0 \pm 0.0}$	$\mathbf{0.9517 \pm 0.0031}$	$\mathbf{0.9545 \pm 0.0048}$
Oracle relevance	0.9667 ± 0.0012	0.9191 ± 0.0010	0.9058 ± 0.0036	1.0000 ± 0.0000	0.0000 ± 0.0000	80.0 ± 0.0	1.0000 ± 0.0000	1.0000 ± 0.0000

Table 15: Online relevance feedback in a three-round CIFAR-100 diagnostic. The first round is cold-start; subsequent rounds use only observed annotation feedback.

Round	Queries	Useful precision ↑	OOD rate ↓
0	250	0.1720	0.8280
1	250	0.9120	0.0880
2	250	0.9520	0.0480

H.1 ImageNet-C corruption stress diagnostic

The corruption diagnostic follows the frozen ImageNet-200 ViT-B/14 TRIG trajectory for ten rounds of 2,000 queries and uses ImageNet-C (Hendrycks & Dietterich, 2019). We extract features for a fixed 2,000-image, 15-corruption ImageNet-C evaluation set, but keep those samples outside the acquisition pool. The evaluator computes q_{rel} after each round and compares the ImageNet-C scores with the remaining irrelevant-OOD scores. No corruption identity, K/N/O membership, or hidden shift label is available to the acquisition API. Results are single-seed evidence about relevance calibration under corruption shift and should not be interpreted as a corruption active-learning leaderboard.

I Contemporary OSAL comparisons and reproduction details

The unified results use matched frozen features and retain qualified "-Adapted" labels. Native author-protocol executions are reported separately because they train raw-image ResNet systems with open-set components on source-specific datasets and schedules. Completed author-code runs include E^{2}OAL on TinyImageNet (final target accuracy 65.1%, 3:50:39), EAOA on TinyImageNet (65.7%, 2:19:14), and LfOSA on CIFAR-100 (53.45%, 45:00:22). The LfOSA rerun uses only compatibility and output-isolation patches; its ten-round accuracy trajectory correlates at 0.991 with the reference trajectory, with a 2.25-point final difference. Its conservative behavior is directionally consistent with LfOSA-Adapted, although their absolute accuracies are not directly comparable across datasets, representations, and training protocols. These heterogeneous native results establish executability and external comparability, not a shared leaderboard. A native EOAL execution did not finish within the fixed compute budget: as its queried training set expanded and the native pipeline retrained its models each round, observed completed-round training time rose from 7:27 initially to 9:34:09 in round seven and 38:13:47 in round eight. This is an empirical run-time observation, not an asymptotic quadratic-complexity claim. The incomplete trajectory is not reported, and the controlled comparison retains only the explicitly qualified EOAL-Adapted result.

I.1 Runtime and device policy

Formal GPU-first runs record acquisition, probe, and projection devices, peak CUDA memory, and acquisition wall time. Candidate bookkeeping, annotation history, lazy-greedy control flow, evaluator statistics, and hash generation remain CPU-side. Projected gradient methods stream candidate chunks and do not materialize the full class-by-feature gradient embedding. TypiClust is intentionally retained as the CPU reference-style implementation and is not used to support a primary efficiency claim.

Table 16: Three-seed second-foundation-encoder key matrix on ImageNet-200 using the frozen OpenAI CLIP ViT-B/16 image tower. No text prompts or class names are used. All methods share the main protocol; values are mean ± sample standard deviation.

Method	Global ↑	AULC ↑	Discovery AUC ↑	Useful ↑	OOD ↓	Discovered ↑
Random	0.9133 ± 0.0067	0.8599 ± 0.0075	0.8633 ± 0.0038	0.7495 ± 0.0008	0.2505 ± 0.0008	$\mathbf{80.0 \pm 0.0}$
BADGE	0.9210 ± 0.0020	$\mathbf{0.8718 \pm 0.0047}$	$\mathbf{0.8946 \pm 0.0149}$	0.4418 ± 0.0031	0.5582 ± 0.0031	$\mathbf{80.0 \pm 0.0}$
Null-space baseline	0.8287 ± 0.0172	0.7098 ± 0.0144	0.5008 ± 0.0319	0.4120 ± 0.0118	0.5879 ± 0.0118	68.7 ± 1.5
Pure spectral	0.7577 ± 0.0064	0.6672 ± 0.0094	0.3740 ± 0.0300	0.2941 ± 0.0046	0.7059 ± 0.0046	53.0 ± 2.6
rel-spectral	0.8990 ± 0.0155	0.7770 ± 0.0123	0.6608 ± 0.0300	$\mathbf{0.9352 \pm 0.0010}$	$\mathbf{0.0649 \pm 0.0009}$	76.7 ± 2.5
TRIG	$\mathbf{0.9240 \pm 0.0036}$	0.8414 ± 0.0061	0.8225 ± 0.0175	0.8575 ± 0.0019	0.1425 ± 0.0019	$\mathbf{80.0 \pm 0.0}$

Table 17: Three-seed query-batch granularity sensitivity for the frozen TRIG configuration on ImageNet-200. Every row uses exactly 20,000 queries; smaller batches therefore receive more feedback and classifier updates. The bold row is the frozen main protocol. Values are mean ± sample standard deviation, and no alternative setting replaces the main configuration.

Batch	Rounds	Global ↑	AULC ↑	Discovery AUC ↑	OOD ↓	Near OOD ↓	Far OOD ↓
500	40	0.9617 ± 0.0021	0.9075 ± 0.0025	0.8833 ± 0.0036	0.6025 ± 0.0032	0.3583 ± 0.0034	0.2442 ± 0.0002
1000	20	0.9613 ± 0.0006	0.8992 ± 0.0022	0.8650 ± 0.0038	0.6048 ± 0.0019	0.3594 ± 0.0034	0.2455 ± 0.0041
2000	10	0.9647 ± 0.0031	0.8875 ± 0.0030	0.8367 ± 0.0069	0.6077 ± 0.0017	0.3605 ± 0.0006	0.2472 ± 0.0022
5000	4	0.9530 ± 0.0010	0.8572 ± 0.0019	0.7656 ± 0.0000	0.6163 ± 0.0023	0.3598 ± 0.0014	0.2565 ± 0.0012

I.2 Implementation qualifications

Joda-Reimpl uses explicit candidate-database construction, SISOM and SISOMe score computation, energy-based OOD handling, Joda class-score adjustment, and deterministic top-q ranking. It is an equation-level reimplementation and is not the official executable Joda implementation. EOAL-Adapted, LfOSA-Adapted, and E^{2}OAL-Adapted preserve the published selection structures on the frozen DINOv2 feature track. EOAL uses a deterministic cosine K-means fallback because the published FINCH dependency is unavailable; LfOSA uses per-class two-component maximum-logit Gaussian mixture models (GMMs); E^{2}OAL uses the published Jensen–Shannon (JS) uncertainty product and class-balanced prefix rule with a fixed hidden-label-free precision target. A separate projected-gradient diagnostic failed the predefined minimum-bridge criterion and is not used to support the reported TRIG method.

J Additional Diagnostics

Projected gradient-information candidate. A projected BADGE pseudo-gradient variant improved early novel-query mass but increased OOD acquisition and failed the pre-specified minimum-bridge criterion. It was evaluated only at seed 0, was not promoted to the method, and triggered no additional parameter sweep or modification of the frozen TRIG configuration.

Stochastic-relevance diagnostic. The Jensen gap between the deterministic soft singleton score and a Bernoulli expected singleton score is approximately $7.30\times$ larger on OOD than on novel samples. Replacing the singleton ranking preserves novel retrieval but reduces the top-batch OOD fraction by only 7.02 percentage points on average, below the frozen 10-point requirement in every seed. The upgrade gate therefore failed, and the reported method remained the frozen TRIG configuration.

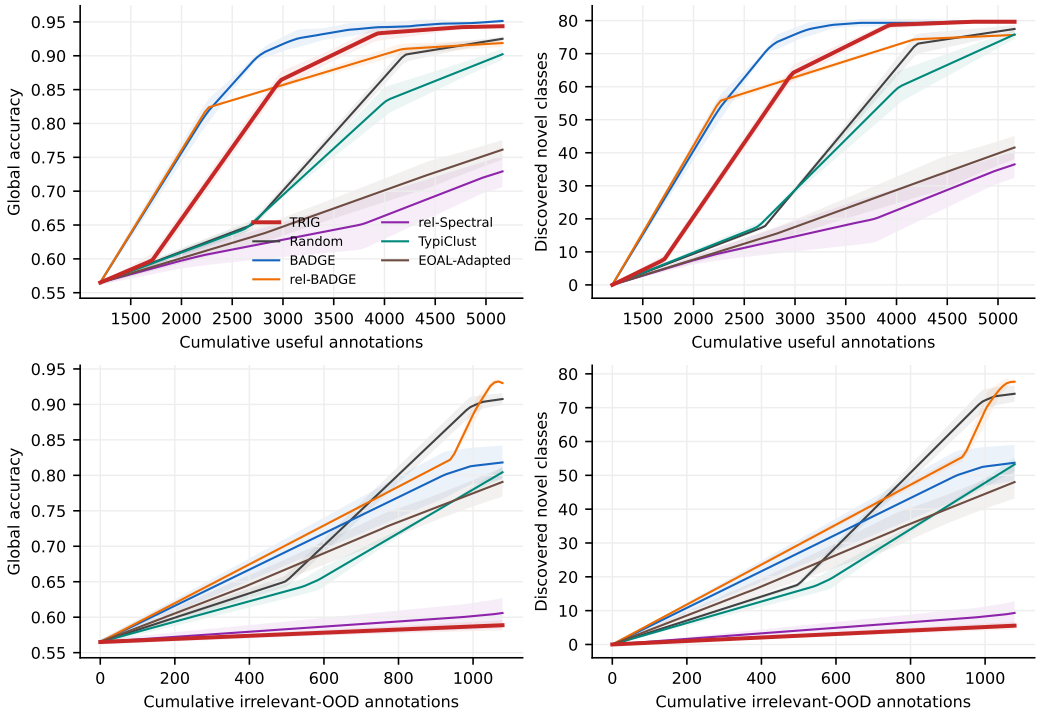

Figure 5: Performance re-parameterized by cumulative useful (top) or irrelevant-OOD (bottom) annotations. Lines and bands are three-seed means and sample standard deviations over shared support; raw integer checkpoints are provided in the supplementary files.

Table 18: Three-seed evaluator-only relevance diagnostics on task-relevant ImageNet-R and ImageNet-V2 shifts. Mean AUC averages rounds 1–9 within each seed before aggregation. Recall is measured at 10% irrelevant-OOD false-positive rate in round 9. Shift samples remain outside the acquisition pool; values are mean ± sample standard deviation.

Shift	Method	Mean AUC ↑	Final AUC ↑	Final recall@10 ↑
ImageNet-R	**TRIG**	0.7937 ± 0.0069	0.8262 ± 0.0019	0.5116 ± 0.0032
ImageNet-R	BADGE	0.7714 ± 0.0096	0.7757 ± 0.0014	0.3753 ± 0.0052
ImageNet-R	rel-BADGE	0.6944 ± 0.0174	0.6574 ± 0.0156	0.2996 ± 0.0277
ImageNet-R	rel-Coreset	0.7044 ± 0.0415	0.6720 ± 0.0389	0.3200 ± 0.0334
ImageNet-V2	**TRIG**	0.9041 ± 0.0008	0.9190 ± 0.0005	0.7482 ± 0.0030
ImageNet-V2	BADGE	0.8898 ± 0.0018	0.8890 ± 0.0023	0.6318 ± 0.0099
ImageNet-V2	rel-BADGE	0.8697 ± 0.0043	0.8576 ± 0.0051	0.6073 ± 0.0188
ImageNet-V2	rel-Coreset	0.8552 ± 0.0018	0.8414 ± 0.0012	0.5867 ± 0.0078

Table 19: Evaluator-only ImageNet-C corruption stress diagnostic for TRIG. The frozen ViT-B/14 selector follows the ten-round ImageNet-200 trajectory; ImageNet-C features remain outside the acquisition pool. q_{rel} is summarized for ImageNet-C and the remaining irrelevant-OOD group, while AUC and recall measure separation at fixed irrelevant-OOD false-positive rates. This is a single-seed diagnostic, not a corruption active-learning benchmark.

Round	$q_{\mathrm{rel}}(C)$	$q_{\mathrm{rel}}(O)$	C-vs-O AUC ↑	Recall@5%	Recall@10%	Recall@20%
0	1.000	1.000	0.500	–	–	–
1	0.511	0.395	0.661	0.088	0.198	0.394
2	0.428	0.309	0.698	0.155	0.268	0.476
3	0.398	0.274	0.726	0.168	0.306	0.512
4	0.378	0.253	0.735	0.181	0.317	0.538
5	0.367	0.241	0.740	0.185	0.328	0.540
6	0.368	0.234	0.752	0.199	0.354	0.571
7	0.366	0.228	0.757	0.204	0.375	0.574
8	0.368	0.224	0.764	0.212	0.387	0.586
9	0.366	0.220	0.768	0.216	0.394	0.590