QUALIFICATION BEFORE EXPANSION: AUDITABLE SYNTHETIC-ONLY CLASS ADDITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data can make a new class trainable before any real example exists, but trainability does not show that the expanded classifier will recognize the class in deployment. We formulate *pre-deployment qualification*: deciding, from historical classes and candidate synthetic data alone, whether the evidence supports admitting a class. Qualification-Guided Synthetic Class Expansion (Q-SCE) replays each historical class as a pseudo-missing candidate under the same information budget, tests whether candidate-observable descriptors predict revealed real recall better than a historical-mean reference, and abstains when they do not. We froze one instantiation, a four-descriptor Ridge qualifier with task, support, and retention gates and an empirical lower-score tier, before opening any candidate target data on DomainNet Clipart→Real (45 candidates) and Stable-Diffusion→Caltech-256 (101). Candidate-recall MAE falls from .259 to .224 and from .138 to .111. Admission gains are small. On Caltech the policy admits 78 classes with one failure, a detector-p10$\geq .50$ screen does as well, and CLIP zero-shot already succeeds on 99 classes; on DomainNet the policy admits one class and abstains on the rest. The results support replay-based qualification with frozen abstention. They do not show a better ranker than simple scores or a guarantee on the failure rate among admitted classes.

1 INTRODUCTION

A deployed classifier may need to recognize a class that was absent from its training set, is dangerous or expensive to photograph, or can be rendered or generated before real examples exist. Synthetic data makes such a class trainable immediately, and a large literature studies how to train with it (He et al., 2023; Sariyildiz et al., 2023; Azizi et al., 2023). We ask an earlier operational question: *before any real example of the candidate class is observed, what evidence justifies trusting the expansion?*

A benchmark may contain many classes for which synthetic-only expansion eventually works while offering too little candidate-specific evidence to know in advance which ones will. We separate three questions. Transfer is the outcome revealed after real evaluation. Qualification asks whether observable pre-reveal evidence supports a specific inference about that outcome. Admission asks whether supported evidence clears an application requirement. A predictor can improve absolute error while ranking classes poorly, or rank well while being biased at a fixed threshold, so we check each property separately.

Transferability estimators such as LEEP, H-score, LogME, and GBC use an observed target dataset, often with labels (Nguyen et al., 2020; Bao et al., 2019; You et al., 2021; Pándy et al., 2022). Label-free accuracy estimators remove target labels but still use unlabeled target inputs (Deng & Zheng, 2021; Guillory et al., 2021; Garg et al., 2022; Baek et al., 2022; Yu et al., 2022). In our setting the candidate's real distribution is absent altogether.

Q-SCE uses historical classes as controlled pseudo-missing trials. It hides a historical class's real data, reintroduces that class synthetically, computes only the evidence a future candidate would have, and reveals the real outcome afterward. Replay then asks whether a candidate-specific rule predicts across classes better than a reference that knows historical transfer but nothing about the candidate. If it does not, Q-SCE abstains. We use standard predictors and tests and contribute the decision problem, a validation protocol that mirrors the future information budget, and a prospective audit of one frozen decision policy.

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.

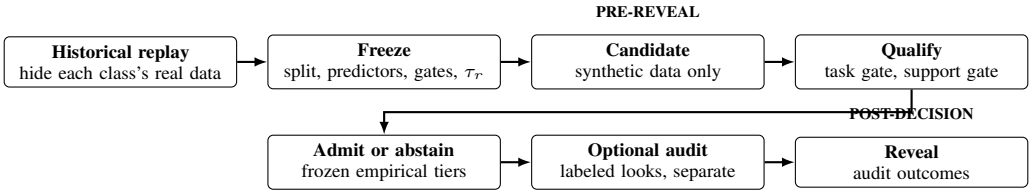


Figure 1: Chronology. Candidate real data stay sealed through the freeze, evidence computation, qualification, and the zero-real decision. The optional labeled audit changes the information budget and is never pooled with zero-real admission.

Our contributions are:

1. A formulation of synthetic-only class addition as pre-deployment qualification, with pseudo-missing trials and a task gate that suppresses candidate-specific claims unless held-class evidence supports them. We estimate the gate's false-pass rate and power in a stylized simulation.
2. A concrete abstaining policy that combines a frozen Ridge-4 recall qualifier, empirical lower-score tiers, and task, support, and retention gates, with zero-real admission kept separate from any later labeled audit.
3. Two prospectively frozen visual evaluations on 146 candidate classes whose target data were sealed until reveal (the pretrained encoders may still know the concepts), protocol-isolated CLINC150 regeneration runs (decisions frozen per run, on intents whose outcomes an earlier holdout had revealed), and labeled post-reveal diagnostics showing where the policy helps, where simpler scores tie or win, and where it abstains too often.

The prospectively evaluated policy. The headline policy is fixed Ridge-4 plus the task, support, and retention gates, followed by the empirical L90 tier; the same freeze records L80 as a lower tier. DomainNet and Caltech-256 test this fixed instantiation. The nested five-expert selector, conformal Rule B, single-descriptor rankings, and labeled audit are retrospective or downstream analyses. The prospectively supported claim is that candidate-observable evidence improves recall prediction over a historical-mean reference on both prospective splits. We call the protocol *auditable* in a narrow sense: splits, operating points, and every zero-real decision are recorded in hashed manifests before reveal, so a reader can check that no candidate target data informed a decision.

Deployment scope. Q-SCE targets a fixed operational classifier or detector that must be expanded while real inputs from the candidate class are unavailable. If a validated zero-shot model already covers the deployment taxonomy, synthetic expansion may be unnecessary, as on Caltech.

2 Related Work

Synthetic data for new classes. Generated images can train or augment classifiers (He et al., 2023; Sariyildiz et al., 2023; Azizi et al., 2023), feature-generating zero-shot learning synthesizes features for classes with no real examples (Xian et al., 2018), and synthetic test sets can support evaluation of small subgroups (van Breugel et al., 2023). These works ask how to train or evaluate with synthetic data; we ask when a synthetic-only class can be trusted before any real example exists. Domain adaptation uses target observations (Sun et al., 2016; Ganin et al., 2016), and covariate-shift theory explains why source training need not transfer (Shimodaira, 2000; Sugiyama et al., 2007; Ben-David et al., 2010). Class-incremental methods receive real examples of new classes (Rebuffi et al., 2017; Li & Hoiem, 2016; Tao et al., 2020).

Performance estimation. Transferability scores use an observed target task (Nguyen et al., 2020; Bao et al., 2019; You et al., 2021; Pándy et al., 2022). AutoEval regresses accuracy on statistics of synthetic meta-datasets (Deng & Zheng, 2021), which is close in spirit to pseudo-missing trials, and difference of confidences, agreement-on-the-line, ATC, and the projection norm estimate accuracy from unlabeled target inputs (Guillory et al., 2021; Baek et al., 2022; Garg et al., 2022; Yu et al., 2022). LOVM predicts vision-language model performance from language alone (Zohar et al., 2023). Table 21 in the appendix compares the information each setting uses.

Abstention and conformal inference. Selective prediction rejects individual inputs (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). Conformal methods give marginal coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023; Barber et al., 2021) and can be adapted to drift (Tibshirani et al., 2019; Barber et al., 2023); risk-control methods bound expected losses (Bates et al., 2021; Angelopoulos et al., 2021; 2024). Conformal selection controls the false-discovery rate among selected units (Jin & Candès, 2023). When observations are nested in groups, as images are in classes, hierarchical exchangeability is the relevant assumption (Lee et al., 2023). Cross-validated errors from overlapping training sets are dependent, which complicates inference on them (Bengio & Grandvalet, 2004; Bates et al., 2024). The suitability filter uses post-deployment unlabeled data (Pouget et al., 2025).

3 Problem Formulation

Let K be the historical classes (*anchors*), each with real data R_k and synthetic data S_k. Candidate u has synthetic data S_u only, so the observable set is $\mathcal{O}_u = \{R_k, S_k : k \in K\} \cup \{S_u\}$. An expansion procedure trains an expanded classifier f_e from $\mathcal{O}_u$; after reveal it attains real candidate recall $r_u = P_{R_u}(f_e(x) = u)$. Let $\Delta_K(u)$ be the resulting change in known-class macro-recall.

Why cross-class structure is required. Recall on u is not identified by $\mathcal{O}_u$: two worlds that agree on $\mathcal{O}_u$ can place R_u entirely inside or entirely outside the region f_e assigns to u, giving $r_u = 1$ and $r_u = 0$ with identical observables, so every predictor has worst-case error at least $1/2$ (App. A). Any method must adopt an assumption linking candidates to historical classes. Q-SCE assumes that (descriptor, recall) pairs are exchangeable across semantic classes. The task gate tests a different property: whether the descriptors carry transferable signal about recall on held classes. Exchangeability can hold when they carry none (the gate should then fail), and the gate can pass when a particular future candidate is not exchangeable with the anchors.

For a predictor q of recall from a candidate-observable descriptor z_u, and a reference q_0 that uses historical transfer only, we summarize relative absolute error by $Q_{\text{task}} = 1 - \text{MAE}(q)/\,\text{MAE}(q_0)$. Given required recall τ_r and tolerated retention loss τ_d, a candidate is a *failure* if admitted with $r_u < \tau_r$. We use $\tau_r = .50$ and $\tau_d = .02$. These operating points were fixed during development and were not externally preregistered: τ_r and the task-gate threshold in the cross-benchmark development analysis, the 1% detector false-positive target in the visual protocols, and τ_d and the support rule in the final Adaptiope runner. DomainNet, Caltech-256, and all CLINC150 runs inherited them unchanged.

4 Qualification-Guided Synthetic Class Expansion

4.1 Pseudo-missing trials, qualifier variants, and the task gate

For historical class k, hide R_k from every operation concerning the missing class, reintroduce k using S_k, compute z_k, freeze a prediction, and reveal R_k to measure r_k. This mirrors the information budget a future candidate faces, with one fewer known class.

The class is the unit of replay. Each replay yields one (descriptor, recall) pair; images within a class are not independent qualification examples. Holding out images while keeping the class's real prototype would expose exactly the information that is missing at deployment. The held class is therefore also hidden from alignment, the realness classifier, and detector negatives, and its real outcome becomes training data for the qualifier only after its simulated decision is recorded.

We evaluate two qualifier implementations. The *retrospective selector*, used in the 152-candidate audit and the full-pipeline simulations, is fully nested. For each outer held anchor h, the expansion option (RAW, MeanShift, or CORAL alignment of frozen features) is chosen on $K \setminus \{h\}$, and an inner leave-one-anchor-out loop evaluates five pre-specified experts: historical mean, four-descriptor Ridge, 5-NN, margin-only Ridge, and detector-p10 Ridge. The one-standard-error rule (Hastie et al., 2009) selects the simplest expert statistically indistinguishable from the best. The staged Adaptiope and CLINC150 evaluations and both prospective visual evaluations instead use a *fixed*

Ridge-4 qualifier: four-descriptor Ridge was locked before candidate reveal, and the other experts serve only as retrospective comparisons.

The *task gate* requires the active qualifier's outer predictions to beat the historical mean: a one-sided paired sign-flip test at $p < .05$ together with a paired-bootstrap (Efron, 1979) 95% interval for $|q_0 - r| - |q - r|$ whose lower end is above zero. Early development used a weaker rule ($Q_{\text{task}} > 0$); every reported admission uses the two-part gate, which we call the *strict* gate. The gate certifies improvement in absolute error over the reference, not ranking quality or accuracy near τ_r. Outer errors share training anchors, so these tests lack exact independent-unit calibration (Bengio & Grandvalet, 2004; Bates et al., 2024); Sec. 6.1 measures their behavior by simulation. If the gate fails, every candidate-specific admission is suppressed.

4.2 Evidence, feature support, and retention

The descriptor has four entries: (i) *realness*, the mean candidate-synthetic probability under a real-versus-synthetic classifier trained on known classes; (ii) cosine distance from the candidate synthetic prototype to the nearest real known prototype (its negative is *prototype proximity*); (iii) *detector p10*, the tenth percentile of the candidate detector's probability on held candidate synthetic samples; and (iv) *margin*, that percentile minus the 95th percentile on synthetic known negatives. *Expansion.* Features from a frozen encoder (DINOv2-small for images (Oquab et al., 2024); MiniLM or E5 for text) are optionally aligned from the synthetic to the real domain using known classes only, by a mean shift (MeanShift) or by mean and covariance matching (CORAL, Sun et al., 2016), and ℓ_2-normalized. Known classes are predicted by nearest real prototype. The candidate is added singly by a class-balanced logistic-regression detector trained on candidate synthetic positives and real and synthetic known negatives. An input is routed to the candidate when the detector's probability exceeds a threshold γ, set on a known-class validation split to route about 1% of known inputs; each candidate has its own detector, so γ differs across candidates. Candidate recall r_u is the fraction of its real inputs routed to it. Joint addition of several candidates is not studied.

What the descriptors measure. Realness asks whether generated examples resemble the historical real domain in the frozen representation; prototype distance places the candidate relative to the known taxonomy; detector p10 asks whether even weakly recognized synthetic positives score highly; margin compares that score with synthetic negatives. None observes real recall, since a detector can confidently recognize synthetic artifacts that do not transfer, so replay must decide whether they are informative in the current task. Ridge-4 combines them into an absolute recall estimate, and Sec. 6.3 tests whether the combination beats single scores.

A support gate standardizes descriptors on anchors, computes the mean distance to the five nearest anchors, and abstains when the candidate's empirical tail probability is below .05 (App. D). A retention gate requires measured known-class macro-recall change of at least $-.02$. In the staged runner, γ is set with the known-class `real_val` partition, while retention is computed on the separate known-class partition named `real_test`; the latter serves as gate-calibration data, and neither split contains candidate-real examples. The minimum measured retention change was $-.019$ on Adaptiope, $-.0104$ on DomainNet, and $-.0132$ on Caltech-256, so the retention gate changed no decision, although on Adaptiope it came within .001 of binding.

4.3 Admission rule A: empirical lower-score tiers

For each outer held anchor h, a Ridge scale model trained on inner cross-fitted residuals gives $s(z) = \text{clip}(\exp\{g(z)\}, .02, 1)$. With normalized residuals $a_h = (\hat{r}_h - r_h)/s_{-h}(z_h)$ and $q_{1-\alpha}$ their $\lceil (n+1)(1-\alpha) \rceil$-th order statistic over n anchors,

$$L_u^{1-\alpha} = \max\{0, \hat{r}_u - \max(0, q_{1-\alpha})\, s(z_u)\}. \tag{1}$$

A candidate enters the *L90 tier* if $L_u^{90} \geq \tau_r$ and the *L80 tier* if $L_u^{80} \geq \tau_r > L_u^{90}$; otherwise it is abstained. All admissions also require the task, support, and retention gates. Eq. 1 pairs the full-data prediction with leave-one-out residuals, as in the classical jackknife, and carries no finite-sample guarantee; a jackknife+ construction (Barber et al., 2021) would be needed for one. Even a valid marginal bound would not bound failures among admitted candidates. The names L90 and L80 identify the residual quantile used, not an achieved coverage: realized L90 coverage was .844 on DomainNet and .950 on Caltech.

Table 1: Evidence hierarchy. DomainNet and Caltech-256 were frozen before any candidate target data were opened; the frozen protocols did not name a single primary endpoint.

Evaluation	Cand. + anchors	Qualifier	Role and caveat
Prospective			
DomainNet Clipart→Real	45 + 254	Fixed Ridge-4	No candidate Real image opened before freeze. Clipart is a hand-drawn proxy source rather than rendered or generated data.
SD1.5→Caltech-256	101 + 155	Fixed Ridge-4	Generated source, sealed authentic target, plus a frozen CLIP baseline.
Protocol-isolated replication (outcomes of these classes seen earlier)			
CLINC150 regeneration	$6 \times (50 + 100)$	Fixed Ridge-4	Three seeds each for Qwen/MiniLM and Mistral/E5, all source text regenerated and decisions frozen per run. The 50 candidate intents were revealed in an earlier staged holdout (App. B.6).
Adaptiope seed-42 resplit	$2 \times (41 + 82)$	Fixed Ridge-4	Rendered Synthetic→Product and →Real; outcome-blind split of classes seen in development.
Development and sensitivity			
GTSRB, SIP-17, VisDA, Adaptiope	43, 15, 12, 2×41	Nested selector	Retrospective audit on the original candidate splits; candidates share anchors.
Other (appendix)		As labeled	Earlier Adaptiope holdouts, ResNet-50 backbone, CLINC published split, Mistral/E5, procedural intents, multi-split replay.

From prediction to a decision. A high point prediction does not by itself admit a candidate. The residual adjustment measures how far comparable historical predictions overstated recall, with a scale that varies with the descriptor; the task gate checks usefulness across historical classes, the support gate checks extrapolation in descriptor space, and retention checks the existing taxonomy. Abstention therefore means the frozen policy lacks enough evidence at τ_r, not that the candidate will fail. On DomainNet, 26 of the 42 abstained candidates succeed.

Post-reveal selected-set diagnostic (Rule B). Because Rule A does not control failures among admitted classes, we also analyze conformal selection with Benjamini–Hochberg (Benjamini & Hochberg, 1995; Jin & Candès, 2023) on disjoint fit and calibration anchors (App. B.13; Sec. 6.4). Rule B was analyzed only after reveal and is not part of the evaluated policy.

Limited-real audit (separate stage). If labels are permitted after abstention, the frozen runners inspect $n \in \{5, 10, 15, 20, 25\}$ labeled real examples and admit when a 98% one-sided Clopper–Pearson (Clopper & Pearson, 1934) lower bound reaches .50; Bonferroni spending gives 90% family confidence per candidate. The audit stops at its first admission or at 25 labels. The $n = 5$ look cannot admit (its maximum bound is .457). A post-reveal correction removes it and spends the same family error over $n \in \{10, 15, 20, 25\}$ with 97.5% bounds; frozen decisions are unchanged, and Sec. 6.5 reports both schedules.

5 Experimental Design

GTSRB/Synset (Stallkamp et al., 2011; Sielemann et al., 2024), SIP-17 (Zhu et al., 2023), VisDA-2017 (Peng et al., 2017), and Adaptiope (Ringwald & Stiefelhagen, 2021) supply rendered or simulated sources. DomainNet (Peng et al., 2019) freezes 254 anchors and 45 candidates from 299 eligible classes, and Caltech-256 freezes 155 anchors and 101 candidates. Rule A qualification and bound calibration use all anchors. Each freeze also records a fit/calibration split (190/64 and 116/39) for later Rule-B analysis, but no Rule-B decision was frozen. On Caltech, Stable Diffusion 1.5 (Rombach et al., 2022) generates 30 source images per class from frozen category names and prompt templates, and the 60 authentic images per candidate stay sealed. Both pre-reveal manifests record zero candidate-target openings. The frozen Ridge-4 expansion selected CORAL alignment on DomainNet and MeanShift

Table 2: Left: retrospective audit on the original candidate splits (152 candidates). AURC is the area under the risk–coverage curve when candidates are admitted in order of the selector's prediction (lower is better; random equals the base failure rate). For SIP-17 and VisDA, AURC scores the selector's historical-mean fallback as a constant. Right: fraction of simulated runs in which the gate passes for the full five-expert selector, by number of historical classes n and population R^2 of the descriptor signal (120–200 runs per cell). Row b.e. is the break-even signal at which ridge and the historical mean have equal expected error (ridge-only gate, 800 runs per cell).

Direction	**Gate**	**MAE / null**	**AURC**	**Rand.**	R^2	n=12	17	43	82	123
GTSRB/Synset	pass	**.131** / .190	**.558**	.837	0	.045	.015	.000	.000	.008
SIP-17	fail	.226 / **.224**	.236[†]	.267	.10	.075	.035	.050	.125	.075
VisDA-2017	fail	.229 / **.225**	.633[†]	.583	.25	.090	.110	.260	.515	.767
Adaptiope S→Real	pass	**.200** / .235	**.263**	.463	.50	.260	.355	.815	.985	1.00
Adaptiope S→Product	pass	**.212** / .246	**.091**	.244	b.e.	.061	.076	.023	.019	.014

[†] With the leave-one-out historical-mean fallback left in place, AURC is .339 and .855 (see text). Adaptiope rows use the original candidate split, not the seed-42 resplit of Table 8.

on Caltech, using historical anchors only. For CLINC150 (Larson et al., 2019), each regeneration produces 30 source utterances per intent and freezes decisions before reusing the previously revealed 100 real utterances per candidate.

Baselines. We report two detector screens that use no qualifier. The γ-*screen* admits a candidate when its detector p10 is at least the detector's own threshold γ and the retention gate passes; this is the qualifier-free baseline used throughout. The *.50-screen* instead admits when detector p10 is at least .50. Because γ is set per candidate and can lie on either side of .50, neither screen's admitted set contains the other's. Both are scored on the same revealed-recall outcome as Q-SCE. For the frozen expansion choices, γ ranges from .100 to .987 across DomainNet's 45 candidates and from .053 to .940 across Caltech's 101 candidates. The .50-screen was added as a post-reveal diagnostic; it was not specified as a frozen prospective comparator.

Four evaluation questions. We report prediction (candidate MAE against the historical-mean reference), ordering (rank correlation and matched-coverage failures against simple scores), calibration (empirical coverage of the lower scores), and decisions (admissions and failures). A method can succeed on one and fail on another, so we report each separately and never quote a failure count without its admission count.

6 Results

6.1 What the task gate can and cannot detect

Where the gate passes (Table 2, left), ranking by the selector beats random by .15 to .28 AURC. In the two directions where it fails, the selector falls back to the historical mean for 13/15 and 11/12 candidates, and ranking is close to random (.236 vs .267; .633 vs .583) once the fallback is scored as a constant. Scored naively, the fallback looks much worse than random (.339, .855), because a leave-one-out mean for held class h equals $(\sum_k r_k - r_h)/(n-1)$ and so decreases in r_h.

The simulation (right; App. B.18) runs the complete nested selector and gate on synthetic tasks. Without signal the gate passes in 0–4.5% of runs; at break-even signal a ridge-only gate passes slightly above .05 with 12–17 classes and at most .023 from 43 up. Power grows with the number of classes: at $R^2 = .25$ the gate passes in 9–11% of runs with 12–17 classes and 77% with 123. The SIP-17 and VisDA failures are thus compatible with weak but real signal; had the gate been ignored, the tiers would have admitted two SIP-17 candidates, both successful, and none on VisDA. The fixed Ridge-4 gate of the prospective runs, simulated with the runners' $\lambda = 10$ (400 runs per cell), passes in at most 3.5% of no-signal runs and never at $n \geq 43$; at break-even signal it passes in .011 of runs at $n = 155$ and .010 at $n = 254$; and at $R^2 = .25$ it passes in 90% of runs at $n = 155$ and 99% at $n = 254$. At $R^2 = .10$ these fall to 23% and 50%.

Table 3: Prospective visual holdouts and CLINC150 regeneration runs. Seed rows report mean±SD over three runs and per-run admission ranges. Coverage is empirical 80/90%, averaged over seeds for the seed rows. The γ-screen and .50-screen are the qualifier-free baselines of Sec. 5; Audit counts are labeled-audit admissions among candidates the zero-real tiers abstained on.

Run	Task gate	Cand./null MAE	ρ	80/90 cov.	L90	L80 or L90	γ-screen	.50-screen	Audit
DomainNet Clipart→Real	pass	.224/.259	.462	.778/.844	1 (0)	3 (0)	13 (1)	11 (2)	16 (0)
SD1.5→Caltech-256	pass	.111/.138	.376	.851/.950	78 (1)	85 (2)	88 (2)	83 (1)	7 (0)
Qwen/MiniLM, 3 seeds	3/3	.182±.006/.261±.008	.672±.038	.860/.913	9–18 (1–2)	22–25 (2–3)	–	–	7–8 (0)
Mistral/E5, 3 seeds	2/3	.226±.024/.268±.024	.582±.138	.820/.920	0–3 (0)	0–12 (0–2)	–	–	14–23 (0)

Entries are admissions (failures). Admitting every candidate would fail at 16/45 (.356) on DomainNet and 8/101 (.079) on Caltech.

What the gate has cost. The gate failed in four settings with retained ungated decisions: SIP-17, VisDA, Mistral/E5 seed 1001, and the ResNet-50 Real seed-53 partition. In two of them, switching it off would have added successful admissions (two on SIP-17 under the nested-selector gate and eleven on seed 1001 under the fixed Ridge-4 gate); in VisDA and the ResNet-50 partition it would have added none. We have no instance in which the gate prevented a failure, so its justification rests on the simulation and the non-identifiability argument.

6.2 Prospective visual holdouts and protocol-isolated replications

On DomainNet the gate passes (held-anchor MAE .195 vs .230), candidate MAE improves from .259 to .224 (paired interval $[.005, .064]$), and all three zero-real admissions succeed. Empirical 80/90% coverage is .778/.844, below nominal. Both qualifier-free screens admit more DomainNet classes and include failures (13 with 1; 11 with 2). On Caltech-256 the gate passed before any of the 6,060 candidate target images were opened (anchor MAE .132 vs .158), and candidate MAE improved by .0267 (paired interval $[.0099, .0434]$). Eight of the 101 candidates fail under the detector outcome. L90 admits 78 with one failure (one-sided 95% upper bound .059) and the L80-or-L90 policy admits 85 with two. The closest comparator is the .50-screen, which admits 83 with one failure; random admission of 78 classes would give 6.18 expected failures. The one L90 failure has recall .417, so at $\tau_r = .40$ the L90 tier (then 84 Caltech admissions) has no failure; the .059 bound is therefore sensitive to the choice of τ_r.

Calibrated CLIP (Radford et al., 2021) predicts that all 101 Caltech classes clear .50 zero-shot recall, and 99 do. Here CLIP zero-shot recall is the fraction of a class's 60 authentic images assigned to it among all 256 text labels, a different classifier and outcome from the Q-SCE detector. A practitioner willing to deploy CLIP would not need synthetic expansion on Caltech; qualifying a CLIP-based expansion with the same protocol remains future work. As a pre-reveal ranker on the detector outcome, CLIP synthetic recall (the share of 30 generated images assigned to the candidate among the 155 anchor labels plus the candidate) has AURC .0284 against .0127 for Ridge-4, and the paired-bootstrap interval for the difference includes zero.

In the regeneration runs, point-estimate gains appear in every seed, but two Mistral intervals include zero (App. B.8). Qwen/MiniLM improves candidate MAE by $.080 \pm .004$ and passes the task gate in all three runs; Mistral/E5 passes in two; seed 1001 fails despite good hindsight ranking ($\rho = .701$), and seed 1003 admits 9 with 2 failures. Across all six runs L90 makes 48 repeated decisions with four failures, all in Qwen runs. At the intent level, 28 distinct intents are admitted and three fail at least once, a one-sided 95% upper bound of .254 when each intent counts once. Within-run L90 upper bounds range from .238 to .632.

6.3 Do four descriptors add value?

Ridge-4 is the only single frozen policy evaluated throughout, but it does not uniformly beat the best simple score chosen after reveal (Table 4). It ties on Adaptiope Product and on Caltech, where margin and detector p10 also make one failure at both coverages and realness, prototype proximity, and CLIP synthetic recall make three to seven (Figure 2). It avoids one failure on Adaptiope Real and trails prototype proximity by one on DomainNet, where proximity's rank correlation with recall is only .010. Across full curves Ridge-4 has the lowest AURC on the Adaptiope seed-42 resplit, but most pairwise differences are within bootstrap uncertainty, margin and detector p10 rank better on the original Adaptiope Product split, and margin has lower mean AURC on the Qwen regeneration

Table 4: Matched-coverage summary on the common detector-recall outcome. Entries are failures among the top-k candidates ranked by each score, not the admissions of the tiered policy (on Caltech the L80/L90 policy admits 85 with two failures). The best simple score is chosen after reveal. Full curves, AURC intervals, and all scores are in App. B.4.

Evaluation and coverage	Ridge-4	Best simple score	Simple score
Adaptiope Product, top 11	1	1	Realness
Adaptiope Real, top 11	2	3	Margin
DomainNet, top 10	1	0	Prototype proximity
Caltech, top 78 / 85	1 / 1	1 / 1	Margin; detector p10

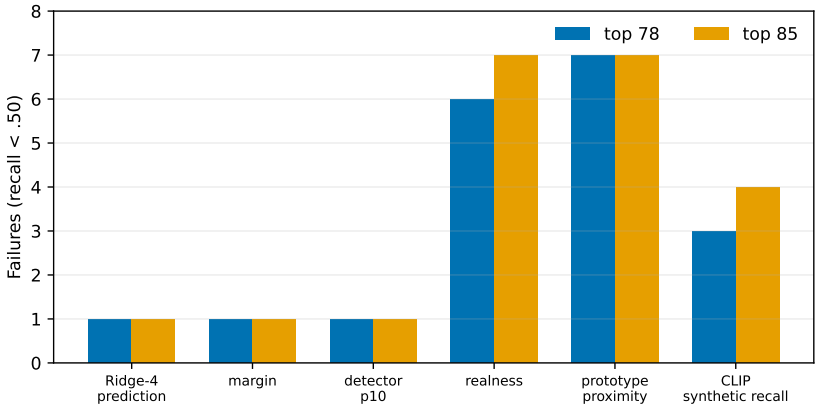


Figure 2: Caltech-256 failures among the top 78 and top 85 candidates ranked by each pre-reveal score, on the common detector-recall outcome (post-reveal diagnostic).

runs (.148 vs .159). The case for four descriptors is that one fixed score transfers across benchmarks without per-benchmark selection. The decision layer did not help on Caltech: the top 85 by Ridge-4 contain one failure, while the 85 admitted by the tiers and gates contain two.

Qualifier-free screens. The γ-screen fails often on Adaptiope (11 of 38 admissions on Product, 20 of 37 on Real) and on the six CLINC150 regeneration runs (219 admissions with 45 failures, .205, against 92 with 9, .098, for Q-SCE). Over six heterogeneous single-run evaluations, including the procedural-intent run, it admits 178 with 43 failures (.242) against 52 with 5 (.096); without that run the counts are 158/43 and 44/5. Of the 43 failures, 31 come from Adaptiope. On the prospective benchmarks it admits 13 with 1 failure on DomainNet and 88 with 2 on Caltech. These comparisons use unmatched coverage. On Adaptiope, where synthetic evidence is most misleading, the gates avoid most screen failures; elsewhere the screens are competitive.

6.4 Selected-set risk: tiers versus conformal selection

Post-reveal Rule B on Caltech retains 37 calibration and 97 candidate units (p-value floor $1/38$, so BH at $q = .10$ needs at least 26 rejections) and selects 92 with 5 failures at .10 and 97 with 7 at .20. Its candidate descriptors use all 155 anchors, so the $|K|$ versus $|K| - 1$ asymmetry remains and no false-discovery guarantee is validated. The symmetric version selects nothing on DomainNet at either level and nothing on Mistral at $q = .10$; at .20 it selects 14 (1 failure) and 11 (2) on two Mistral seeds. Qwen selects 22–25 at .10 and 34–41 at .20.

6.5 Does zero-real screening save labels?

Zero-real admission trades labels for unverified admissions. On Caltech, L80-or-L90 screening followed by the frozen five-look audit uses 310 labels for 92 admissions with two failures, against 1,520 labels for 76 admissions and no failure with the audit alone. On DomainNet screening

Table 5: Admissions (failures) and one-sided 95% Clopper–Pearson upper bounds, which are descriptive and treat admitted classes as independent. Rule-B rows are post-reveal diagnostics that were never frozen. [a]Over the five runs with admissions.

Evidence	Rule	Admit (false)	95% upper bound
DomainNet (prospective)	L80 or L90 tier	3 (0)	.632
Caltech-256 (prospective)	L90 / L80 or L90 tier	78 (1) / 85 (2)	.059 / .072
Adaptiope seed-42 Product	L90 / L80 or L90 tier	2 (0) / 11 (1)	.776 / .364
CLINC150, per regeneration run	L80 or L90 tier	0–25 (0–3)	.221–.550[a]
DomainNet Rule B (symmetric)	$q = .10$ / .20	0 / 0	–
Caltech Rule B (asymmetric)	$q = .10$ / .20	92 (5) / 97 (7)	–
Qwen, Rule B (symm.)	$q = .10$ / .20	22–25 (2–3) / 34–41 (5–10)	–
Mistral, Rule B (symm.)	$q = .10$ / .20	0 / 0–14 (0–2)	–

saves little: 855 rather than 890 labels for the same 19 admissions, about 45 labels per admission. Adaptiope, CLINC, the corrected four-look schedule, and the decision-layer ablation are in App. B.14 and App. B.15.

7 Discussion and Limitations

What is prospectively supported. Under one frozen Ridge-4 policy, candidate-observable evidence improves recall prediction over the historical mean on both prospective visual splits. Admission evidence is weaker: Caltech supplies 78 of the 79 L90 admissions but is a benchmark CLIP already solves and the .50-screen matches, and DomainNet supplies one. These results support information-budget replay and auditable abstention. They do not establish that Ridge-4 is the best ranker, that the tier controls selected-set risk, or that broad admission will hold on harder ontologies.

Operational interpretation. The intended use is expanding a fixed deployed system when candidate real inputs are absent and replacing that system with a zero-shot model is not an option. There, a gate failure or an unsupported candidate tells the operator to keep the old system or collect labels. When the gate passes but few candidates clear L90, as on DomainNet, the labeled audit supplies most admissions (16 of 19) and uses 855 labels in total, most of them on candidates it never admits. On Caltech the balance reverses: the audit adds 7 admissions to the 85 made without labels.

Exchangeability and atypical classes. Random or hash splits make classes exchangeable by construction and cannot test exchangeability. A post-reveal official-domain CLINC holdout improves candidate MAE for Qwen and Mistral (App. B.6), but it is not an unseen visual ontology, and the Adaptiope atypical groups are small (App. B.20). The support gate passed all 45 DomainNet candidates, so descriptor-space extrapolation was not what limited admissions there.

What remains open. Only two benchmarks are prospective, and their pretrained encoders and generator may know the concepts. We did not run a harder prospective benchmark, qualify a CLIP-head expansion, freeze the post-hoc rankers or Rule B, or build a semantic visual holdout. A decisive test would freeze, before reveal, a benchmark on which CLIP zero-shot fails often, with the .50-screen, margin, 5-NN, and a label-free estimator as co-primary comparators and one pre-declared primary endpoint. We kept the 1.08% of Caltech source images that the safety checker blacked out, to avoid outcome-informed regeneration.

8 Conclusion

Q-SCE treats synthetic-only class expansion as a qualification decision made before candidate real data exist, using replayed historical classes and a policy frozen before reveal. In both prospective visual evaluations the fixed Ridge-4 instantiation predicts candidate recall better than a historical-mean reference. Its admissions add little on Caltech, where simple screens and CLIP do as well, and it admits one class at L90 on DomainNet, where simple screens admit failures. Harder prospective ontologies with frozen strong comparators are the next test of its practical value.

AI USE STATEMENT

Generative AI assisted literature discovery, feedback, drafting, code development, visualization, and critique. The authors ran the experiments, verified the reported claims and citations, and take responsibility for the manuscript and artifacts. The corresponding disclosure is also provided in the submission form.

REPRODUCIBILITY STATEMENT

The supplement supports integrity verification, retained-output checks, and inspection of the staged runners; it is not a turnkey raw-data reproduction package. Full reruns require the external datasets, model checkpoints, compute, and input preparation listed in App. F and the submitted runbook.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *Transactions on Machine Learning Research*, 2023.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, 2022.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing*, pp. 2309–2313, 2019. doi: 10.1109/ICIP.2019.8803726.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, volume 9908 of *Lecture Notes in Computer Science*, pp. 614–629. Springer, 2016. DOI: `https://doi.org/10.1007/978-3-319-46493-0_37`.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using Bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9172–9182, 2022.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. VisDA: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019. doi: 10.1109/ICCV.2019.00149.

Angéline Pouget, Mohammad Yaghini, Stephan Rabanser, and Nicolas Papernot. Suitability filter: A statistical framework for classifier evaluation in real-world deployment settings. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49603–49627. PMLR, 2025.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

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*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017. doi: 10.1109/CVPR.2017.587.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Tobias Ringwald and Rainer Stiefelhagen. Adaptiope: A modern benchmark for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 101–110, 2021. doi: 10.1109/WACV48630.2021.00015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Anne Sielemann, Lena Loercher, Max-Lion Schumacher, Stefan Wolf, Masoud Roschani, Jens R. Ziehn, and Juergen Beyerer. Synset Signset Germany: A synthetic dataset for German traffic sign recognition. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems*. IEEE, 2024. doi: 10.1109/ITSC58415.2024.10920175.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German traffic sign recognition benchmark: A multi-class classification competition. In *2011 International Joint Conference on Neural Networks*, pp. 1453–1460. IEEE, 2011. doi: 10.1109/IJCNN.2011.6033395.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2058–2065, 2016. doi: 10.1609/aaai.v30i1.10306.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2526–2536, 2019.

TorchVision maintainers. ResNet-50 model and IMAGENET1K_V2 weights. TorchVision model documentation, 2026. URL `https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html`. Accessed 2026-09-23.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. In *Advances in Neural Information Processing Systems*, 2023.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. doi: 10.48550/arXiv.2212.03533.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Xiaomeng Zhu, Talha Bilal, Pär Mårtensson, Lars Hanson, Mårten Björkman, and Atsuto Maki. Towards sim-to-real industrial parts classification with synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4454–4463, 2023.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. LOVM: Language-only vision model selection. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

A NON-IDENTIFIABILITY ARGUMENT

Fix $\mathcal{O}_u$ and the expansion procedure, so f_e and the candidate region $A_u = \{x : f_e(x) = u\}$ are fixed; assume A_u and its complement are both nonempty. Let P and P' be joint distributions over all historical real and synthetic data and S_u that agree on the distribution of $\mathcal{O}_u$, and differ only in the conditional law of R_u given $\mathcal{O}_u$: under P, R_u is supported on a subset of A_u; under P', on a subset of A_u^c. Both are valid because R_u is not a component of $\mathcal{O}_u$ and both regions are nonempty. Then $r_u = 1$ under P and $r'_u = 0$ under P', while $q(\mathcal{O}_u)$ takes the same value under both. Hence $\max\{|q-1|, |q-0|\} \geq 1/2$. If the admission rule admits at this $\mathcal{O}_u$, the admission is false under P' for any $\tau_r > 0$. □

The argument only shows that some assumption linking the candidate to other classes is unavoidable; it says nothing about which assumption is appropriate.

B ADDITIONAL VALIDATION DETAILS

B.1 OPERATING-POINT PROVENANCE AND γ-SCREEN BASELINE

Operating points as set in the retained runners: the development analysis fixed $\tau_r = .50$ and the task-level sign-flip threshold at .05. The visual runners used a detector false-positive target of .01; the final Adaptiope runner fixed $\tau_d = .02$ and support $p = .05$. The empirical quantile in Eq. 1 is the $\lceil (n+1)(1-\alpha) \rceil$-th order statistic, and the bound is infinite (no admission) when that rank exceeds n.

Table 6: γ-screen on the seed-42 Adaptiope candidate split, using the known-class threshold γ and retention gate. A failure has revealed candidate recall below .50.

Target	Policy	Admit	False	False rate	Min. recall
Product	γ-screen	38/41	11	.289	.04
Product	Q-SCE L90 tier	2/41	0	.000	.54
Product	Q-SCE L80 or L90 tier	11/41	1	.091	.15
Real	γ-screen	37/41	20	.541	.02
Real	Q-SCE L90 tier	0/41	0	–	–

B.2 EARLIER STAGED ADAPTIOPE HOLDOUTS

In the first Adaptiope S→Real execution, predictions were frozen before its 4,100 candidate target images were processed. Candidate MAE was .218 versus .275 for the historical mean (paired-improvement interval $[.019, .098]$; sign-flip $p = .0034$; Spearman (Spearman, 1904) $\rho = .563$). The L90 tier admitted none. The later S→Product holdout froze an adaptive-tier policy before Product candidate images were processed. Candidate MAE was .214 versus .261, and five L90-tier and five L80-tier candidates were admitted; all ten exceeded .50 recall (mean .853, minimum .570). Because Product shares semantic classes with Real and the method evolved after the Real run, we describe both as staged holdouts; neither was preregistered.

B.3 ADAPTIOPE SEED-42 RESPLIT

Table 7: Seed-42 resplit prediction and bound results. Improvement is paired absolute-error reduction relative to the historical mean.

Target	Q MAE	Null MAE	ΔMAE 95% CI	Sign p	ρ	80/90 cov.
Product	**.187**	.220	[.0002,.0685]	.0307	.380	.780/.951
Real	**.197**	.234	[.0084,.0678]	.0072	.497	.805/.976

The frozen qualifier improves over the historical mean in both directions. Product's secondary one-sided Wilcoxon (Wilcoxon, 1945) test gives $p = .0558$ while the pre-specified sign-flip gate and bootstrap interval pass; Real's Wilcoxon gives $p = .0061$. We report all tests. Product admits two

L90-tier candidates (recall .95 and .54) and nine L80-tier candidates, of which eight succeed; Axe is predicted at .765 with $L^{80} = .555$ and realizes .15. Real makes no zero-real admission. Among admitted Product candidates, the mean fraction of real known validation inputs routed to the new class is 1.1%; the maximum over all candidates is 2.4% (Product) and 2.0% (Real).

B.4 Matched-coverage and single-descriptor diagnostics

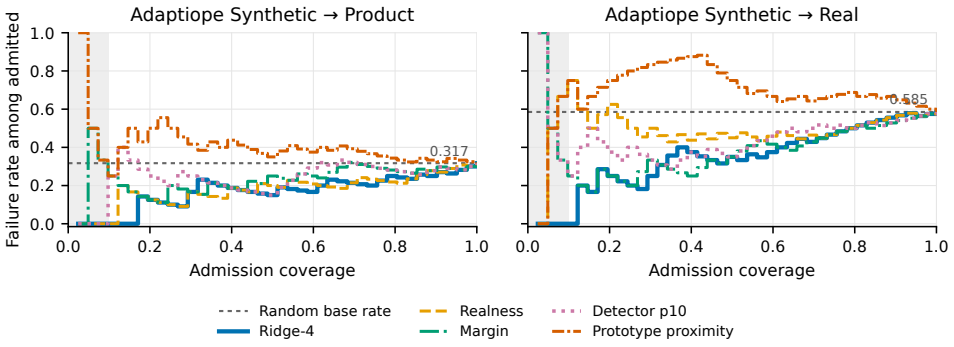


Figure 3: Risk–coverage curves on the Adaptiope seed-42 resplit (41 candidates per direction). Candidates are admitted in order of each score. The shaded region (coverage below .10) is determined by one to four classes.

Table 8: Full matched-coverage comparison. Entries are failures among the top-k candidates, AURC, and Spearman ρ with realized recall. ΔAURC is the feature's AURC minus the Q-SCE prediction's, with an unadjusted 95% bootstrap interval over candidates; after Holm adjustment across the eight comparisons, only the Real prototype-proximity difference remains significant. Adaptiope values are post hoc from retained seed-42 outputs; DomainNet top-k values are post-reveal diagnostics.

	Adaptiope Product (13/41 fail)					**Adaptiope Real** (24/41 fail)					**DomainNet**	
Score	@5	@11	AURC	ρ	ΔAURC [95% CI]	@5	@11	AURC	ρ	ΔAURC [95% CI]	@5	@10
Q-SCE prediction	**0**	**1**	**.173**	.38	–	**1**	**2**	**.356**	.50	–	1	1
Realness	1	**1**	.183	.13	+.010 [−.066, .083]	3	5	.499	.41	+.143 [−.016, .306]	1	2
Margin	1	2	.241	.37	+.067 [−.036, .178]	**1**	3	.401	.47	+.046 [−.079, .143]	**0**	2
Detector p10	2	3	.242	.32	+.069 [−.046, .192]	2	4	.453	.34	+.097 [−.036, .204]	**0**	2
Prototype proximity	2	5	.404	−.40	+.231 [.020, .413]	3	9	.692	−.34	+.337 [.102, .539]	**0**	**0**
Random (expected)	1.59	3.49	.317	0		2.93	6.44	.585	0			

The detailed comparison is mixed. Prototype proximity is safest at DomainNet top-10, yet anti-correlates with Adaptiope recall. Ridge-4 has the lowest AURC on the Adaptiope seed-42 resplit, but its advantage over several features is within bootstrap uncertainty; on the original Adaptiope Product split margin and detector p10 rank better (Table 22), and margin also has lower mean AURC on the Qwen regeneration runs (.148 vs .159; Table 13). On DomainNet, Ridge-4's AURC (.193) is statistically indistinguishable from realness (.235), margin (.237), detector p10 (.259), and prototype proximity (.196); a post-hoc 5-NN has lower MAE (.205 vs .224). Nested isotonic calibration of detector p10 passes the DomainNet gate but is worse than Ridge-4 in candidate MAE/AURC (.239/.270 vs .224/.193). On Caltech at top-78/top-85 coverage, Ridge-4, margin, and detector p10 each make 1/1 failures, CLIP synthetic recall 3/4, realness 6/7, and prototype proximity 7/7 (App. B.11). No descriptor is uniformly necessary or sufficient across benchmarks.

B.5 Frozen-backbone sensitivity

Replacing DINOv2-small with frozen `torchvision/resnet50:IMAGENET1K_V2` features (TorchVision maintainers, 2026) on the seed-42 split, the held-anchor qualifier still beats its null in both directions, but candidate-MAE improvement intervals include zero and no candidate reaches a zero-real tier.

Table 9: Frozen ResNet-50 sensitivity. Q/null are nested held-anchor MAEs; the interval is revealed paired candidate-MAE improvement.

Target	Q/null	Q p	Cand./null MAE	ΔCI	ρ	80/90 cov.
Product	.193/.235	.0025	.227/.257	$[-.016, .075]$	.384	.829/.902
Real	.147/.175	.0090	.164/.197	$[-.007, .073]$	.519	.854/.878

B.6 Staged CLINC150 holdout on the published split

Before opening any real utterance from a candidate intent, a deterministic domain-stratified split selected 100 anchor and 50 candidate intents (10/5 per official domain). Qwen2.5-3B-Instruct (Qwen et al., 2024) at revision `aa8e72537993ba99e69dfaafa59ed015b17504d1` generated 30 utterances per intent from the intent and domain names only. Frozen `sentence-transformers/all-MiniLM-L6-v2` embeddings (Wang et al., 2020; Reimers & Gurevych, 2019) at revision `1110a243fdf4706b3f48f1d95db1a4f5529b4d41` replaced image embeddings. The held-anchor qualifier has MAE .234 versus .289 ($Q_{\text{task}} = .190$; bootstrap improvement $[.025, .084]$; sign-flip $p = .0002$). After reveal, candidate MAE is .180 versus .266 (improvement $[.043, .127]$), Spearman $\rho = .614$, and adaptive 80/90% coverage is .94/.98. All five L90-tier admissions succeed; one of ten L80-tier admissions fails (Table 10). The audit separately admits 19 candidates with no failure.

Table 10: All CLINC150 zero-real admissions on the published split. L80/L90 were frozen before candidate reveal.

Intent	Tier	Pred.	L80	L90	Actual
distance	L90	.839	.599	.511	.63
goodbye	L90	.864	.610	.517	.59
order checks	L90	.873	.644	.561	1.00
rollover 401k	L90	.870	.648	.566	1.00
vaccines	L90	.909	.648	.553	1.00
book hotel	L80	.839	.570	.471	.77
card declined	L80	.825	.565	.470	.95
confirm reservation	L80	.796	.503	.396	.82
exchange rate	L80	.780	.502	.401	.89
insurance change	L80	.880	.597	.493	.98
pin change	L80	.835	.514	.397	.96
report lost card	L80	.800	.547	.455	.82
share location	L80	.815	.557	.463	.64
translate	L80	.815	.554	.459	.41
where are you from	L80	.803	.536	.438	.91

This holdout revealed the outcomes of all 50 candidate intents, which is why the later regeneration runs on the same split are classed as protocol-isolated rather than prospective. A post-reveal leakage audit compares all 5,000 candidate real utterances (100 per intent) with the generated source text: 10 are exact normalized duplicates (.2%), none has nearest character-ngram similarity at least .90, and the median nearest similarity is .388. This cannot rule out exposure of the public dataset during pretraining.

B.7 Second model pair and procedural novel intents

The model-pair sensitivity retains the published split but regenerates the source side with Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and uses E5-base-v2 (Wang et al., 2022). Because primary CLINC150 outcomes had been exposed, this is post-reveal sensitivity. The procedural novelty experiment reuses 100 public anchors but replaces all candidates with 20 newly minted identifiers; Mistral receives only each alias and definition, and target utterances come from a separately sealed procedural grammar. Its large MAE gain (.082 vs .310) reflects a level shift, since procedural targets are uniformly easy, while its rank correlation is negative ($-.339$) in the published partition and

in every alternative partition. We treat it only as a leakage stress test and exclude it from pooled comparisons in the main text.

Table 11: Text sensitivity results. Admission columns give admissions (failures) for the L90 tier and for L80 or L90.

Run	Cand./null MAE	ΔCI	ρ (p)	90 cov.	L90	L80 or L90
Mistral/E5 sensitivity	.224/.296	[.022,.122]	.596 (5.05×10^{-6})	.940	1 (0)	15 (3)
Procedural novel intents	.082/.310	[.161,.270]	$-.339$ (.144)	.950	6 (0)	8 (0)

The procedural audit finds no exact source–target duplicates and no character-ngram similarity at least .90. The labeled audit admits 12 (Mistral/E5) and 11 (procedural) candidates with no failures.

B.8 End-to-end regeneration replications

Each replication starts from the same 100-anchor/50-candidate split, regenerates all source utterances, recomputes embeddings, refits the fixed Ridge-4 qualifier and scale model, freezes decisions and hashes, and only then evaluates outcomes. Seeds are 20271001–20271003 for both pairs. Within each run no candidate real utterance was read before the freeze, but candidate outcomes had been revealed by the earlier published-split holdout, so these runs isolate generation and fitting noise rather than testing on unseen classes.

Table 12: All six regeneration runs on the same 50 target intents. L90 and All give admissions (failures); Audit uses labeled target evidence.

Pair	Seed	Task?	Cand./null MAE	ρ	L90	All	Audit
Qwen/MiniLM	1001	yes	.178/.254	.710	18 (2)	25 (2)	8 (0)
Qwen/MiniLM	1002	yes	.178/.261	.634	9 (1)	24 (3)	8 (0)
Qwen/MiniLM	1003	yes	.188/.269	.671	18 (1)	22 (2)	7 (0)
Mistral/E5	1001	no	.230/.286	.701	0 (0)	0 (0)	23 (0)
Mistral/E5	1002	yes	.200/.241	.614	3 (0)	12 (0)	14 (0)
Mistral/E5	1003	yes	.247/.278	.431	0 (0)	9 (2)	17 (0)

Across Qwen seeds, held-anchor nested MAE is $.229 \pm .002$ versus $.279 \pm .012$, and $Q_{\text{task}} = .179 \pm .043$; across Mistral seeds, $.243 \pm .011$, $.272 \pm .003$, and $.105 \pm .049$. Qwen abstains on 25, 26, and 28 of 50 candidates; Mistral on 50, 38, and 41. The seed whose gate failed is retained in every summary. The per-seed paired 95% bootstrap intervals for candidate-MAE improvement are [.037,.112], [.051,.114], [.037,.124] for Qwen and [.025,.087], [$-.001$,.083], [$-.007$,.070] for Mistral. Within-run one-sided 95% Clopper–Pearson upper bounds on the failure rate are, for the L90 tier, .310, .429, .238 (Qwen) and .632 (Mistral 1002), and for the L80-or-L90 tiers .231, .292, .259 (Qwen) and .221, .550 (Mistral 1002, 1003). Pooling across runs would treat repeated admissions of the same intent as independent and is not reported.

Table 13: Post-reveal ranker sensitivity over three regeneration seeds per model pair. Entries are mean AURC with the observed seed range; lower is better.

Score	Qwen/MiniLM	Mistral/E5
Frozen Ridge-4	.159 [.137,.187]	.168 [.116,.248]
Realness	.278 [.197,.389]	.344 [.302,.396]
Margin	.148 [.136,.155]	.170 [.117,.233]
Detector p10	.166 [.127,.186]	.188 [.118,.284]
Prototype proximity	.467 [.450,.501]	.471 [.424,.557]

B.9 Prospectively frozen DomainNet evaluation

The protocol uses the official Clipart and Real lists and requires at least 50 Clipart and 100 Real images per class, leaving 299 classes. Each class contributes 50 Clipart and 100 Real images. The pre-reveal task gate passes with nested MAE .195 versus .230, $Q_{\text{task}} = .152$, Monte Carlo sign-flip

$p = 5.0 \times 10^{-6}$ (the smallest value attainable with 200,000 draws), and bootstrap interval $[.022, .048]$. The runner used all 254 anchors for held-anchor qualification and bound calibration; the recorded 190/64 split was reserved for Rule B. The frozen protocol did not designate a single primary endpoint, so we do not retrofit one. After reveal, candidate MAE is .224 versus .259, Spearman $\rho = .462$ ($p = .0014$), and adaptive 80/90% coverage is .778/.844. One L90-tier and two L80-tier candidates are admitted, all successful; the audit admits 16 with no failure, using 12.8 labels on average per audit-admitted candidate.

B.10 DomainNet post-reveal diagnostics and threshold sensitivity

These diagnostics were computed after outcomes were available and are not frozen competitors.

Table 14: Post-reveal DomainNet diagnostics. False@5/10 is the number of failures among the top 5/10; ΔAURC is relative to Ridge-4 with a 95% candidate-bootstrap interval (20,000 resamples). The random AURC is the base failure rate, $16/45 = .356$.

Score	AURC	ΔAURC [95% CI]	ρ	@5	@10
Frozen Ridge-4	.193	–	.462	1	1
Realness	.235	+.042 [−.040,.139]	.438	1	2
Margin	.237	+.044 [−.020,.119]	.281	0	2
Detector p10	.259	+.066 [−.007,.155]	.220	0	2
Prototype proximity	.196	+.003 [−.118,.115]	.010	0	0

Table 15: DomainNet threshold sensitivity; entries are admitted/false.

Tier	$\tau_r = .40$	$\tau_r = .50$	$\tau_r = .60$	$\tau_r = .70$
L80 or L90 tier (L^{80})	14/2	3/0	1/0	0/0
L90 tier only (L^{90})	3/0	1/0	0/0	0/0

B.11 Prospectively frozen Stable Diffusion to Caltech-256

We use the 256 object categories of Caltech-256 (Griffin et al., 2007), excluding clutter. A SHA-derived split assigns 155 anchor and 101 candidate classes. Each class has 30 Stable Diffusion 1.5 source images and 60 authentic target images; the candidate target paths were sealed before source generation. The generator saw only the category name and ten fixed prompt templates. The safety checker replaced 83/7,680 outputs with black images (49 anchor, 34 candidate); all were retained under the frozen rule. Candidate target images opened before the manifest freeze: zero.

The DINOv2 Ridge-4 qualifier has held-anchor MAE .132 versus .158 ($Q_{\text{task}} = .1616$ from unrounded values; paired-improvement interval $[.0109, .0405]$). After reveal, candidate MAE is .111 versus .138, Spearman $\rho = .376$, and adaptive 80/90% coverage is .851/.950. The L90 tier admits 78 with one failure and L80-or-L90 admits 85 with two. Both failures have recall .417. A calibrated CLIP source-only predictor has MAE .083 for CLIP zero-shot recall, but admits all 101 classes; 99 exceed .50. On the shared Q-SCE detector outcome, CLIP synthetic-recall AURC is .0284 versus .0127 for Ridge-4 ($\Delta = .0157$, bootstrap interval $[-.0001, .0600]$); prototype proximity is worse ($\Delta = .0940$, $[.0305, .1729]$).

Failures at matched coverage for every score appear in Figure 2.

At $\tau_r = .30, .40, .50, .60, .70, .80$, the prospective-pool L90 policy admits respectively 100/1, 87/0, 79/1, 64/1, 4/0, and 0/0 (admitted/failed); the L80-or-L90 policy admits 123/6, 105/3, 88/2, 80/3, 63/5, and 0/0. The γ-screen admits 101 pooled candidates (13 DomainNet, 88 Caltech) at every threshold, with respectively 1, 1, 3, 5, 14, and 31 failures under the corresponding success criterion. The operating point .50 was fixed before both reveals. The L90 drop between .60 and .70 arises because 60 gate-passing Caltech candidates have frozen L90 scores in $[.60, .70)$, so they cross the threshold together; only four remain at .70.

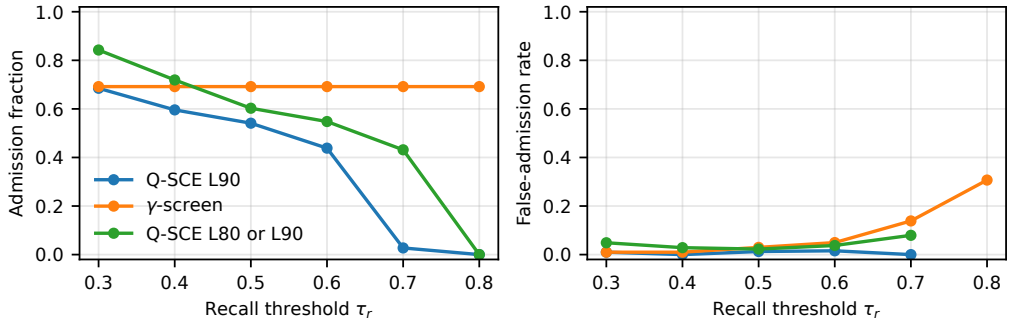


Figure 4: Threshold sensitivity pooled descriptively over 45 DomainNet and 101 Caltech candidates. The γ-screen curve uses detector p10 at the known-class threshold γ (plus retention); its selected set stays fixed as the recall success threshold τ_r changes. Counts are not independent benchmark replications.

B.12 Calibrated, drop-one, rank-gate, and semantic diagnostics

All results in this subsection are post-reveal. Nested isotonic regression of target recall on detector-p10 improves on the historical mean in 9 of 12 evaluated settings under the strict task gate, but does not uniformly beat Ridge-4: on DomainNet its candidate MAE/AURC is .239/.270 versus .224/.193, while on the three Qwen seeds its MAE ranges .174–.192 versus .178–.188 for Ridge-4. Dropping one descriptor at a time changes the strict gate from pass to fail only when margin is dropped on Adaptiope Product and prototype proximity is dropped on Adaptiope Real; no feature is uniformly indispensable across 12 settings.

A Spearman permutation gate passes all 12 retained settings (DomainNet, Caltech, Adaptiope Product and Real, three Qwen and three Mistral CLINC regeneration seeds, and the published-split Qwen and Mistral CLINC runs), including the Mistral seed whose error-based gate fails, showing that ranking and absolute calibration test different claims. A finite-sample jackknife+ diagnostic is conservative but low-powered: at nominal .90 it makes no admission in 8/12 settings and at most two in any seed run; its empirical coverage ranges .86–1.00. Finally, leave-one-official-domain CLINC prediction gives positive candidate-MAE improvement for published Qwen (.0867, interval $[.0362, .1386]$), published Mistral (.0737, $[.0142, .1445]$), and all Qwen seeds; the Mistral seed-1003 interval crosses zero ($[-.0195, .0776]$).

B.13 Symmetric Rule-B diagnostic

To target selected-set risk directly, let $\mathcal{C}$ be calibration anchors disjoint from those used to fit $\hat{\mu}$. For candidate u, test $H_u : r_u < \tau_r$ with

$$p_u = \frac{1 + \sum_{c \in \mathcal{C}} \mathbb{I}[r_c < \tau_r,\ \hat{\mu}(z_c) \geq \hat{\mu}(z_u)]}{|\mathcal{C}| + 1}, \tag{2}$$

then admit the Benjamini–Hochberg rejections at level q (Benjamini & Hochberg, 1995; Jin & Candès, 2023). Under class exchangeability, conformal selection controls the unconditional expected false-discovery proportion. Emptying the selected set with the task gate cannot increase that unconditional rate, but the rate conditional on the gate passing is not controlled because the gate depends on calibration outcomes. Every fitted component excludes $\mathcal{C}$, and the support score is permutation-equivariant. The p-value floor $1/(|\mathcal{C}|+1)$ also limits power: at $q = .10$, BH needs at least 7 selections on DomainNet and 26 on Caltech. Across the 24 calibration-unit simulations (20, 40, or 80 anchors; two image counts; four correlations), class-level calibration yields .898–.907 coverage at nominal .90, whereas treating correlated images as independent yields .970–.999 (App. B.16).

The original post-hoc implementation fitted candidate descriptors with all known classes but calibration-anchor descriptors with one fewer. We corrected this by deterministically dropping one hashed fit anchor for every candidate, so candidate and calibration descriptors both use $|K| - 1$ known classes. The change is small in MAE but material for selection.

Table 16: Symmetric Rule B at $q = .10$ (post-reveal diagnostic). Entries are selected/failed after the task gate.

Evaluation	Selected/failed
DomainNet	0/0
Qwen/MiniLM seeds 1001, 1002, 1003	25/2, 22/3, 23/2
Mistral/E5 seeds 1001, 1002, 1003	0/0, 0/0, 0/0

At $q = .20$ the symmetric construction still selects none on DomainNet but 14/1 and 11/2 on two Mistral seeds and 34/5, 35/6, and 41/10 on Qwen.

For Caltech, a separate post-reveal reconstruction uses the frozen 116/39 fit/calibration split and the same pooled support count, but retains the original asymmetry: candidate descriptors use all 155 anchors. Thirty-seven calibration anchors and 97 candidates pass support. At $q = .10$, the minimum p-value is $1/38$ and BH selects 92 of the 97 supported candidates with five failures; at .20 it selects 97 with seven. It does not test the symmetric Rule-B guarantee.

B.14 Expanded label-cost comparison

Table 17: Label-cost comparisons. Failures in screening-plus-audit rows arise from zero-real admissions; the labeled audit uses exact one-sided binomial bounds.

Policy	Labels	Admit	False
DomainNet audit only	890	19	0
DomainNet L80 then audit	855	19	0
Caltech frozen five-look only	1520	76	0
Caltech corrected four-look only	1500	83	0
Caltech L80 + frozen five-look	310	92	2
Caltech L80 + corrected four-look	310	93	2

The frozen five-look schedule uses $n = 5, 10, 15, 20, 25$ with per-look confidence .98. At $n = 5$, even five successes give lower bound $.02^{1/5} = .457 < .50$, so that look cannot admit. Removing it and spending $\alpha = .10$ over four looks (confidence .975 each) adds one screened Caltech admission without extra labels; the corrected policy was not frozen and is reported only as a design correction.

For the Qwen seed runs, audit-only uses 875/860/910 labels and has 27/28/24 admissions without failure; L80-screen-then-audit uses 530/550/610 labels and has 33/32/29 admissions with 2/3/2 failures. For Mistral, the corresponding label counts are 965/930/950 versus 965/765/805; final failures are 0/0/0 versus 0/0/2. Label savings are substantial in several runs but arise by converting some candidates from verified to zero-real admission. On Adaptiope Product, L80-or-L90 screening saves 25.4% of labels (630 vs 845) but admits the failed Axe class, while L90 screening saves 4.1% with no failure.

B.15 Pooled comparison and full ablation

For completeness we pool six heterogeneous evaluations (Adaptiope Product and Real seed-42, CLINC Qwen and Mistral/E5 on the published split, the procedural novel intents, and DomainNet; 247 candidate instances). The γ-screen admitted 178 with 43 failures (.242) and Q-SCE 52 with 5 (.096); excluding Adaptiope, 103/12 (.117) versus 41/4 (.098); DomainNet plus CLINC-Qwen alone, 45/2 (.044) versus 18/1 (.056). Excluding the procedural run, the γ-screen admits 158 with 43 failures (.272) and Q-SCE 44 with 5 (.114). These pooled contrasts are descriptive, dependent, and dominated by Adaptiope; they are not population rates. Sec. 6.3 discusses them.

Table 18 removes one decision component at a time. On the six regeneration runs, removing the task gate adds 11 successful admissions, removing the support gate adds four admissions of which three fail, and global rather than adaptive scaling yields 73 admissions with 7 failures instead of 92 with 9. In the six-evaluation pool, removing support instead adds 13 admissions and no failures. Drop-one Ridge retraining changes the gate outcome only for Product without margin and Real

without prototype proximity, and candidate rankings can improve or worsen after any drop. The retention gate changed no decision in any evaluation.

Table 18: Component ablation (admitted/false). The six-evaluation column includes the procedural novel-intent run.

Policy	Six evaluations	Six regeneration runs
Full Q-SCE	52/5	92/9
No task gate	52/5	103/9
No support gate	65/5	96/12
No adaptive scaling	43/4	73/7
No retention gate	52/5	92/9
γ-screen	178/43	219/45

B.16 Calibration-unit simulation and Axe diagnosis

The simulation crosses anchor counts 20, 40, and 80, images per anchor 20 and 100, and within-class correlations .01, .10, .30, .60, with 10,000 repetitions for each of $3 \times 2 \times 4 = 24$ settings. At nominal .90, one residual per held class gives .8984–.9074 coverage; treating correlated images as independent gives .9701–.9989, which is valid but needlessly conservative because the effective sample size is overstated.

Axe is not a support-gate extrapolation case. On Product its support p-value is .627, prediction .765, adaptive L80/L90 .555/.400, and recall .15; on Real, support $p = .843$, prediction .535, L80/L90 .277/.162, recall .02. The L90 tier rejects it in both directions. A local safeguard designed after observing Axe also rejects it, but because it is post hoc we do not add it to Q-SCE.

B.17 Retrospective multi-split robustness

For five settings we replay the published class partition and four SHA-derived partitions (seeds 17, 29, 53, 97), reusing representations and revealed outcomes. Because the partitions overlap, these are sensitivity analyses and should not be counted as independent trials.

Table 19: Class-partition sensitivity: minimum–maximum over five partitions. "Gate pass" counts partitions where the task gate passed. L90 gives the range of L90-tier admissions and of failures.

Setting	Gate pass	Q_{task}	Cand. MAE	ρ	90 cov.	L90 (false)
ResNet-50 Product	5/5	.146–.178	.166–.227	.384–.657	.902–1.000	0 (0)
ResNet-50 Real	4/5	.094–.171	.138–.164	.352–.576	.878–.976	0 (0)
Qwen/MiniLM	5/5	.190–.268	.180–.224	.562–.699	.860–.980	2–17 (0–4)
Mistral/E5	5/5	.113–.212	.181–.241	.206–.596	.880–.980	0–1 (0–1)
Procedural novel	5/5	.110–.170	.054–.186	$-.378$–$-.020$	.950–1.000	0–6 (0)

Point candidate-MAE improvement is positive in all 25 replays, and 18/25 paired-bootstrap intervals exclude zero. The ResNet-50 Real seed-53 partition fails the task gate; removing only that gate still produces 0/41 L80 and 0/41 L90 admissions after support and retention checks.

B.18 Task-gate simulation

Each simulated task has n historical classes with descriptors $z \sim \mathcal{N}(0, I_4)$ and recall $r = \mathrm{clip}(0.55 + 0.22\,[\sqrt{R^2}\, w^\top z/\|w\| + \sqrt{1-R^2}\,\varepsilon], 0, 1)$ with $w = (1, .5, .5, 0)$ and $\varepsilon \sim \mathcal{N}(0,1)$. The zero weight makes one descriptor pure noise, so the selector must also learn to ignore an uninformative input. The gate is the two-part rule of Sec. 4.1 (one-sided sign-flip test with 2,000 flips at $p < .05$ and a 2,000-resample paired-bootstrap interval whose lower end exceeds zero), applied against the leave-one-out mean.

Full-pipeline rows ($R^2 \in \{0, .10, .25, .50\}$) run the complete nested selector: for each outer held class, an inner leave-one-out loop scores the historical mean, ridge on all four descriptors, 5-NN, and two single-descriptor ridges (margin alone and detector p10 alone), five experts in all; the

one-standard-error rule picks the simplest eligible expert in the order historical mean, margin ridge, detector ridge, four-descriptor ridge, 5-NN; and the chosen expert predicts the outer class. There are 200 runs per cell for $n \leq 82$ and 120 for $n = 123$ (Monte Carlo standard error at most .046). Expansion-option selection is not simulated.

Fixed Ridge-4 gate. Because the staged and prospective runs use fixed Ridge-4, we also simulate exact leave-one-out ridge ($\lambda = 10$, matching the runners) on all four standardized descriptors, 400 runs per cell. Pass rates for $n = 12, 17, 43, 82, 123, 155, 254$ are .035, .018, 0, 0, 0, 0, 0 at $R^2 = 0$; .080, .053, .040, .080, .175, .225, .495 at $R^2 = .10$; .093, .095, .273, .528, .780, .900, .993 at $R^2 = .25$; and .320, .443, .838, .998, .998, 1.00, 1.00 at $R^2 = .50$. At grid-estimated break-even signal it passes in .011 of 800 runs at $n = 155$ ($R^2 = .023$) and .010 at $n = 254$ ($R^2 = .014$).

Break-even row. For each n we locate by grid search (300 runs per grid point) the R^2 at which exact leave-one-out ridge on all four descriptors and the leave-one-out mean have approximately equal expected absolute error ($R^2 = .126, .136, .058, .045, .027$ for $n = 12, 17, 43, 82, 123$), then run the ridge-only gate 800 times at that value (standard error at most .009).

B.19 Leave-one-out mean artifact

In the SIP-17 and VisDA audits, the selector's nested Spearman correlation on held anchors is $-.996$ and -1.000, which is what a leave-one-out mean produces by construction. Replacing the fallback predictions with a class-independent constant (the mean over all classes, ties broken at random, averaged over 2,000 draws) changes selector AURC from .339 to .236 (random .267) on SIP-17 and from .855 to .633 (random .583) on VisDA. On the three gate-passing directions the selector never falls back, and the correction has no effect.

B.20 Atypical-candidate check

We split the 41 seed-42 candidates in each Adaptiope direction at the upper tercile of support distance (the mean descriptor-space distance to the five nearest anchors), giving 14 atypical candidates. The base failure rate is higher among them (.43 vs .26 on Product; .71 vs .52 on Real).

Table 20: Seed-42 Adaptiope candidates split by support distance. MAE is candidate recall error; coverage is empirical adaptive-bound coverage.

Target	Subset	Q-SCE / null MAE	80/90 cov.	Admit (false)
Product	Atypical (14)	.180 / .258	.786 / 1.000	2 (0)
	Typical (27)	.191 / .200	.778 / .926	9 (1)
Real	Atypical (14)	.171 / .238	.929 / 1.000	0 (0)
	Typical (27)	.210 / .233	.741 / .963	0 (0)

C Information Setting and Complete Procedure

Table 21: Information used before prediction or decision.

Setting	Hist. real	Cand. synthetic	Cand. real inputs	Cand. real labels
Unsupervised domain adaptation	✓	optional	✓	–
Transferability estimation	optional	optional	✓	often ✓
Label-free accuracy estimation	✓	–	✓	–
Q-SCE zero-real qualification	✓	✓	–	–
Q-SCE limited-real audit	✓	✓	✓	✓

1. Freeze the task, expansion options, τ_r, τ_d, predictor library, support rule, tier levels (L80, L90), and Rule-B level q.
2. For every historical anchor, hide its real examples and select the expansion option using only the remaining anchors.

3. Compute the four descriptors; reveal the held anchor only after its prediction is frozen.
4. Use the predeclared qualifier. For the retrospective selector, evaluate the predictor library in an inner loop and apply the one-standard-error rule; for the staged and prospective runs, fit the pre-fixed Ridge-4 rule. Predict each outer anchor without its outcome.
5. Compare outer errors with historical-mean errors. If the task gate fails, suppress all candidate-specific admissions.
6. Evaluate the feature-support and retention gates.
7. Rule A: cross-fit the scale model, form L^{80} and L^{90}, and assign tiers. Rule B: apply the support rule to calibration anchors and candidates alike, compute Eq. 2 on retained calibration anchors, and apply BH at level q.
8. If labels are permitted, run the frozen sequential audit on abstentions and report it separately.

D Nested Selector and Evidence

If expert e^* has the lowest inner mean error, expert e is eligible when $\overline{\ell}_e \leq \overline{\ell}_{e^*} + \mathrm{SE}(\ell_e - \ell_{e^*})$. The exact simplicity order implemented by the retrospective selector is historical mean, margin-only Ridge, detector-p10 Ridge, four-feature Ridge, then 5-NN. An exact eligibility tie between the two single-feature experts resolves to margin-only Ridge.

Table 22: Expert choices by the outer selector and AURC of each expert's ranking in the 152-candidate audit. Random AURC equals the base failure rate.

	Times selected					AURC					
Direction	Hist.	Ridge-4	5-NN	Margin	Det. p10	Selector	Ridge-4	5-NN	Margin	Det. p10	Random
GTSRB/Synset	0	7	2	34	0	.558	.558	.564	.558	.591	.837
SIP-17	13	0	0	2	0	.339	.443	.450	.430	.425	.267
VisDA-2017	11	0	0	1	0	.855	.881	.780	.845	.881	.583
Adaptiope S→Real	0	41	0	0	0	.263	.263	.300	.319	.319	.463
Adaptiope S→Product	0	41	0	0	0	.091	.091	.100	.071	.063	.244

The prototype distance is $d_u = \min_{k \in K} \big(1 - \bar{\phi}(S_u)^\top \bar{\phi}(R_k) / (\|\bar{\phi}(S_u)\| \, \|\bar{\phi}(R_k)\|)\big)$ for frozen embedding ϕ. For the support gate, with D_i the mean distance of anchor i to its five nearest other anchors in standardized descriptor space and D_u the candidate's, $p_{\mathrm{sup}}(u) = \big(1 + \sum_i \mathbb{I}[D_i \geq D_u]\big)/(|K| + 1)$, and candidates with $p_{\mathrm{sup}} < .05$ abstain. The support gate rejects Flat iron, Vacuum cleaner, and Wristwatch for Product (recalls .36, .09, .63) and Flat iron, Knife, and Vacuum cleaner for Real (all below .50); it passes all 45 DomainNet candidates.

E Seed-42 Admissions and Chronology

Table 23: All Product zero-real admissions in the seed-42 resplit (tier: L90 or L80). Bold marks the only failure.

Candidate	**Tier**	$\hat{r}$	L^{80}	L^{90}	**Actual**
Bicycle helmet	L90	.849	.666	.532	.95
Baseball bat	L90	.846	.661	.524	.54
Tyrannosaurus	L80	.793	.595	.449	.61
Brachiosaurus	L80	.796	.592	.441	.98
Binoculars	L80	.779	.569	.414	.99
Electric guitar	L80	.763	.556	.402	.60
Skeleton	L80	.772	.559	.401	.85
Computer mouse	L80	.763	.551	.394	.82
RC car	L80	.757	.546	.391	.78
Compass	L80	.733	.532	.384	.52
Axe	L80	.765	.555	.400	**.15**

The official Adaptiope archive matched SHA-256 `93378c8d5f89ebd7f2015d38ba8d3e48c00bae959ebeeaf822e79e6fcfc7fd67`. Both seed-42 directions used the same runner and

deterministic class split. Both pre-reveal artifacts were completed and hashed before either reveal, and every file in each pre-reveal checksum list verified. All 123 Adaptiope class outcomes had appeared in earlier development runs (12 seed-42 candidates were earlier candidates and 29 earlier anchors), so this run is protocol-isolated and its classes had been seen before.

Table 24: Label cost on the Adaptiope seed-42 resplit (post hoc, from retained audit outputs). *Audit only* applies the sequential audit to every candidate; the Q-SCE rows admit zero-real candidates first and audit the rest. Labels count real examples consumed (at most 25 per candidate).

Target	**Policy**	**Labels**	**Admit (false)**	**Zero-real admit (false)**
Product (28 successes)	Audit only	845	15 (0)	–
	L90 tier, then audit	810	16 (0)	2 (0)
	L80 or L90 tier, then audit	630	21 (1)	11 (1)
Real (17 successes)	Audit only	930	8 (0)	–
	Either Q-SCE policy	930	8 (0)	0 (0)

F REPRODUCIBILITY INVENTORY

The artifact is supplied as `QSCE_Supplementary.zip` in the submission's supplementary-material field. Its experiment archives retain runners, frozen protocols, class and custody manifests, pre-reveal decisions and checksums, reveal outputs, logs, environments, and source tables. Each result is labeled as prospective, protocol-isolated, development, or sensitivity; these roles should not be pooled as exchangeable trials. The supplementary README identifies the γ-screen used in Fig. 4 and distinguishes it from the separate variable-threshold p10 diagnostic retained in another archive. It also points to component digests and verification, external inputs, two excluded Caltech CLIP caches, and upstream Adaptiope documentation.

Table 25: Top-level submitted artifact and six key component archives. Full component digests are listed in `ARCHIVE_SHA256SUMS.txt`; the submitted README points to that manifest and its verification procedure.

Artifact and SHA-256
Master artifact archive `d0781355f084f155aba13f07bdb5a32783616e4475fb4c8a69cbfdbb2648e2ab`
Prospective Caltech-256 archive `01223fc42c0dec452483e4988bf2c75e7bcef52bbc2b03cec8091f5a30850b17`
Retained-output diagnostic archive `5d36c98bf0fab3920f109dc4d3ae0d0f40d0e6445cbcb3efe24ce5235f2beaee`
Symmetric Rule-B / semantic-holdout archive `05329df7276359e75d871c728f9a691dee1d483674951490017aa2b35b80b3fe`
Post-reveal analysis archive `5a8c4808b4625017bb56c0a31e41110ee5de4069ba8296e7d386bc43cb2286b5`
Task-gate simulation archive `efd797165d86e162fb25df50daff628e940551743993df06efc73d795d927484`
Retained-output checks archive `28df0efff6367b158ff0988f867e8fbda26c64e37f8e4eb177d9e12fcd469d8c`