ROUTER-UNIFORM SELECTIVE PREDICTION UNDER CANDIDATE-SET SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language systems often choose candidate labels after observing an image. Even when each candidate mechanism is individually reliable, input-dependent routing can concentrate its errors and invalidate selective-risk guarantees. We characterize this failure using labeled replay: one image is evaluated under every registered mechanism. Two joint events, a selectable accepted error and an unavoidable correct acceptance, determine a router-uniform risk envelope that is exact for deterministic labels. We prove that synchronizing acceptance gates preserves worst-router coverage and cannot increase worst-router risk. The resulting objective is to predict and certify whether any mechanism errs. For set-independent scores, two nested endpoint sets additionally certify all intermediate sets. These results yield SETGUARD-R, which combines a learned shared gate with exact image-level binomial or remaining-pool hypergeometric bounds. On Flowers-102, a label-free router raises the risk of per-mechanism MSP gates from 0.054 to 0.230. Joint replay instead certifies 44.9% coverage for a DeepSets gate at risk target 0.10, while a marginal union-bound certificate for the same score rejects all inputs. A logistic gate attains 48.8% mean coverage across five training seeds and improves on ViLU with the same replay certificate by 4.1–4.8 percentage points across six dataset–model systems.

1 INTRODUCTION

Vision-language models enable recognition over labels specified in natural language (Radford et al., 2021; Zhai et al., 2023). Classifying an image of a husky is a different problem when its candidate labels are unrelated objects than when they are similar dog breeds. In deployed vision-language systems, a retriever proposes the labels, a service sets the shortlist size, and a catalog filter may remove the correct answer. The image is unchanged, yet both the prediction and its confidence can change. This flexibility has motivated work on VLM calibration (LeVine et al., 2023; Tu et al., 2024) and learned failure prediction (Lafon et al., 2025).

Selective prediction trades coverage for accuracy by abstaining on uncertain inputs (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Candidate-set shift adds a routing problem: a router may select the shortlist after seeing the image, concentrating each mechanism on its weak cases and changing both errors and acceptances. With accept-all gates, suppose two mechanisms err on disjoint, observable regions, each of probability 0.08. Each mechanism and every input-independent mixture has risk 0.08, yet choosing the failing mechanism in each region raises risk to 0.16 (Proposition 1), with the image–label distribution and scoring model fixed.

Replaying every registered mechanism on the same image lets one label evaluate all their decisions. The joint pattern of errors and abstentions yields a risk envelope that holds uniformly over image-dependent routers, including routers never seen during calibration (Theorem 2). Its denominator identifies the correct acceptances that no router can avoid. This leads to a shared-gate principle: intersecting mechanism-specific gates preserves worst-router coverage and cannot increase worst-router risk (Theorem 3). The robust target then becomes the image-level event that any mechanism in the portfolio errs.

SETGUARD-R learns to predict this event with a logistic, MLP, or DeepSets (Zaheer et al., 2017) scorer and certifies a shared acceptance threshold on held-out labeled replay, following established

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.

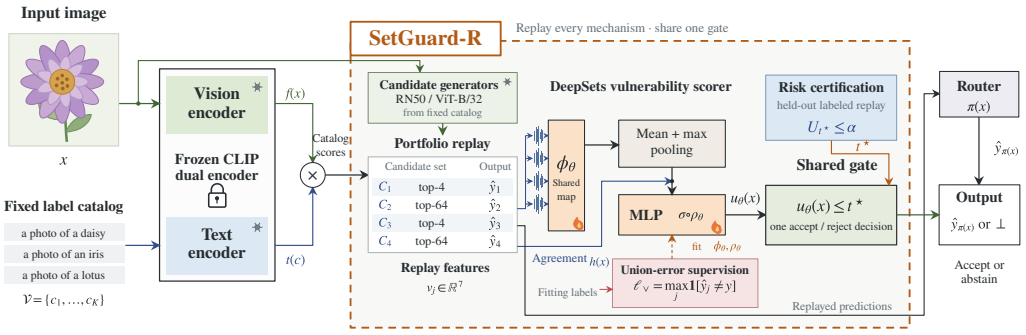


Figure 1: **SETGUARD-R with a DeepSets scorer.** Frozen CLIP scores are replayed over a registered portfolio of candidate mechanisms. The scorer is trained on the image-level union-error target $\ell_\vee$ (whether any mechanism errs), and independent labeled replay certifies a shared threshold. The router chooses among the replayed predictions, but all mechanisms share one acceptance decision. The finite-pool union-risk bound controls every in-portfolio router on the remaining records with confidence $1 - \delta_R$. Logistic and MLP scorers use the same certification procedure.

risk-control principles (Bates et al., 2021; Angelopoulos et al., 2025). Exact binomial and hypergeometric bounds give population and remaining-pool guarantees. For set-independent scores, replaying the endpoints of a nested family also certifies every intermediate candidate set (Proposition 6); top-4/top-64 endpoints cover 2^{60} sets in one interval. Our experiments isolate these effects through matched-score certificate comparisons, routing tests, and matched-feature gate ablations across six dataset–target systems.

2 CANDIDATE MECHANISMS AND DEPLOYMENT ROUTING

Let $(X, Y) \sim P$, and let $\mathcal{V}$ be a fixed label catalog. A portfolio $\mathcal{G} = \{g_1, \ldots, g_J\}$ consists of deterministic, label-free mechanisms, each returning a nonempty candidate set $C_j(x) = g_j(x) \subseteq \mathcal{V}$. A mechanism comprises the retriever, shortlist size, filters, preprocessing, and tie-breaking rule, and its version is part of the certified contract. The vision-language model (VLM) is frozen and predicts

$$\hat{y}_j(x) = \arg\max_{c \in C_j(x)} s(x, c; C_j(x)), \qquad L_j(x, y) = \mathbf{1}\{\hat{y}_j(x) \neq y\}. \tag{1}$$

Omitting the true label counts as an error: $Y \notin C_j(X)$ implies $L_j = 1$. The theory allows set-dependent scores; our implementation uses dual encoders, for which $s(x, c; C) = s(x, c)$.

A selector equips each mechanism with a gate $A_j(x) \in \{0, 1\}$. A router $\pi(j \mid x)$ chooses a mechanism after observing the image and any deterministic replay features; $\Pi_\mathcal{G}$ denotes the class of all such measurable stochastic routers. The deployment law is $Q_\pi(dx, dy, j) = P(dx, dy)\, \pi(j \mid x)$, with coverage and selective risk

$$\mathcal{C}_\pi(A) = \mathbb{E}_P \sum_j \pi(j \mid X) A_j(X), \qquad \mathcal{R}_\pi(A) = \frac{\mathbb{E}_P \sum_j \pi(j \mid X) A_j(X) L_j(X, Y)}{\mathcal{C}_\pi(A)}. \tag{2}$$

Risk is defined only for routers with positive coverage; a reject-all system has zero coverage and does not count as a success. We assign worst-router risk zero when no router has positive coverage, solely to state boundary cases. The minimax objective is to maximize $\inf_{\pi \in \Pi_\mathcal{G}} \mathcal{C}_\pi(A)$ subject to $\sup_{\pi \in \Pi_\mathcal{G} : \mathcal{C}_\pi(A) > 0} \mathcal{R}_\pi(A) \leq \alpha$.

We call this setting *pure* candidate-set shift: changing π changes the distribution of the candidate set while preserving P_{XY}. The guarantee covers new routers, not arbitrary new sets (Table 9). Every portfolio mechanism must be replayable on the certification images; logs that record only one mechanism per image are insufficient. Learned retrievers are frozen before certification, and randomized mechanisms are handled through an explicit seed portfolio (Appendix D).

Proposition 1 (Marginal certification does not compose). *For any $0 < \beta < 1/2$, there exist two fixed candidate mechanisms, a fixed VLM, and a distribution with deterministic labels such that, with*

accept-all gates, each mechanism has risk β, every input-independent mixture has risk β, and an image-dependent router has risk 2β.

The construction (Appendix A) partitions the image space into two disjoint error regions and a common correct region; only the router changes. More generally, separate per-mechanism bounds control constant mixtures (an acceptance-weighted average) but not input-dependent mixtures.

3 THE REPLAY ENVELOPE AND A SHARED-GATE PRINCIPLE

For arbitrary gates, define three image-level indicators:

$$E_A(x,y) = \mathbf{1}\{\exists j : A_j(x) = 1,\, L_j(x,y) = 1\}, \quad S_A(x,y) = \mathbf{1}\{\forall j : A_j(x) = 1,\, L_j(x,y) = 0\},$$
$$D_A(x) = \min_j A_j(x). \tag{3}$$

An image with $E_A = 1$ *offers* a selectable accepted error; one with $S_A = 1$ *forces* a correct acceptance under every router. Every remaining image has no accepted error and at least one rejecting mechanism, so it cannot supply a reliable denominator: a router can avoid its correct predictions by choosing rejection. Let $e_A = P(E_A = 1)$, $s_A = P(S_A = 1)$, and $d_A = P(D_A = 1)$.

Theorem 2 (Two-event routing envelope)**.** *For fixed mechanisms, predictions, and gates,*

$$\sup_{\pi \in \Pi_{\mathcal{G}} : \mathcal{C}_\pi(A) > 0} \mathcal{R}_\pi(A) \;\le\; B(A) := \frac{e_A}{e_A + s_A}, \qquad \inf_{\pi \in \Pi_{\mathcal{G}}} \mathcal{C}_\pi(A) = d_A, \tag{4}$$

with $B(A) = 0$ when $e_A + s_A = 0$. When $e_A + s_A > 0$, the bound is attained by a router that observes (X, Y) (outside $\Pi_{\mathcal{G}}$), and by an image-only router in $\Pi_{\mathcal{G}}$ whenever $Y = f_*(X)$ almost surely for some measurable f_*. With stochastic labels, the bound remains valid but may be conservative.

Why the denominator matters. For any router, the accepted error mass is at most e_A and the accepted correct mass is at least s_A. A label-aware router attains both by choosing an accepted error on E_A, any mechanism on S_A, and a rejecting mechanism elsewhere. This is a population characterization, not a proposal to use labels at deployment; Appendix A treats the zero-mass cases.

Theorem 3 (Synchronizing gates is minimax-dominant)**.** *Given mechanism-specific gates $A_1, \ldots, A_J$, define the synchronized gates $\bar{A}_j(x) = D_A(x)$ for every j. Then*

$$\inf_\pi \mathcal{C}_\pi(\bar{A}) = \inf_\pi \mathcal{C}_\pi(A), \qquad B(\bar{A}) \le B(A), \tag{5}$$

and the worst-case risk over image-only routers cannot increase, even with stochastic labels. Shared gates therefore lose nothing under the minimax coverage objective, whether the risk constraint is the envelope B or the exact worst-case image-only risk; for deterministic labels the two coincide.

Synchronization removes only predictions that some router could already avoid: it leaves the forced-correct event unchanged and can only remove exposed errors. It does *not* imply dominance in average coverage under a known router, nor that shared gates are cheaper; those are different objectives.

For a shared gate $A_j(x) = a(x)$, define the portfolio loss $\ell_\vee(x,y) = \max_{j \le J} L_j(x,y)$. Then $E_A = a\,\ell_\vee$, $S_A = a(1 - \ell_\vee)$, and

$$B(a) = P(\ell_\vee = 1 \mid a = 1), \qquad \mathcal{C}_\pi(a) = P(a = 1) \quad \text{for all } \pi, \tag{6}$$

with $B(a) = 0$ when $P(a = 1) = 0$. Minimizing the replay envelope therefore targets *any portfolio error*. The following threshold principle characterizes optimal coverage under the envelope constraint (Appendix A).

Proposition 4 (The coverage-optimal shared gate)**.** *Let $\eta(x) = P(\ell_\vee = 1 \mid X = x)$, and consider randomized shared gates $a(x) \in [0,1]$, where $a(x)$ is the probability of acceptance. If $\mathbb{E}[\eta(X)] \le \alpha$, accept-all is optimal. Otherwise, suppose that some gate with $\mathbb{E}[a(X)] > 0$ satisfies $B(a) = \mathbb{E}[a\ell_\vee]/\mathbb{E}[a] \le \alpha$. Among all such gates, an optimum is a lower level set of η, randomized on its boundary: $a^*(x) = \mathbf{1}\{\eta(x) < c^*\} + r\mathbf{1}\{\eta(x) = c^*\}$, where $c^* \in [0,1]$ is the largest feasible level and $r \in [0,1]$ is the largest boundary acceptance probability that keeps $\mathbb{E}[a^*(X)(\eta(X) - \alpha)] \le 0$ (take $r = 0$ if the boundary has zero mass). This includes the case $c^* = \alpha$, in which all points with $\eta = \alpha$ may be accepted. If no positive-coverage feasible gate exists, only rejection remains. Any strictly increasing transform of η induces the same gates.*

What the union target learns. Let $G(x) = \mathbf{1}\{\hat{y}_1(x) = \cdots = \hat{y}_J(x)\}$. For single-label classification, the following identity holds pointwise:

$$\ell_\vee(x, y) = 1 - G(x) + G(x)\mathbf{1}\{\hat{y}_1(x) \neq y\}. \tag{7}$$

Disagreement is an observable failure of portfolio-wide correctness; agreement still requires learning whether the common answer is wrong. Thus any feasible shared gate has $P(G = 0 \mid a = 1) \leq \alpha$. The union target incorporates both sources of error, whereas averaging mechanism losses can underweight cases in which only one mechanism fails. This identity also explains why an agreement filter alone is insufficient.

Corollary 5 (The price of portfolio expansion)**.** *With the gates of existing mechanisms fixed, adding mechanisms can only increase e_A and decrease s_A and d_A, so the routing envelope cannot improve. For a fixed shared gate, coverage is unchanged and the portfolio loss cannot decrease. Expansion thus needs a new certificate unless the added losses are pointwise dominated by the certified portfolio loss.*

The certificate carries no multiplicity penalty in J, but this is not free robustness: more mechanisms create more opportunities for error and can reduce certifiable coverage. A portfolio containing an always-wrong mechanism has $\ell_\vee = 1$ almost surely and admits no positive-coverage certificate at any $\alpha < 1$. This limitation is intrinsic to the routing guarantee we seek.

3.1 Endpoint replay covers a candidate-set interval

A finite portfolio need not enumerate every candidate set it certifies. Suppose $\varnothing \neq C^-(x) \subseteq C^+(x)$ and the target uses a fixed score $s(x, c)$ with one global, set-independent tie order. Define the interval $[C^-, C^+](x) = \{C : C^-(x) \subseteq C \subseteq C^+(x)\}$, and let $L(x, y, C)$ be the loss when predicting from C.

Proposition 6 (Nested candidate closure)**.** *For every (x, y) and every $C \in [C^-, C^+](x)$, $L(x, y, C) \leq \max\{L(x, y, C^-(x)), L(x, y, C^+(x))\}$. Consequently, a shared gate certified on the endpoints of several intervals remains certified, unchanged, under arbitrary input-dependent choice of any set in their union, without replaying interior sets or assuming anything about their frequencies.*

If both endpoints predict y, the inner set contains y and no label in the outer set outranks it, so every sandwiched set also predicts y. The two endpoints thus capture complementary failures: omission at the inner boundary and competition at the outer boundary. For top-4 and top-64 retrieval, a single interval contains 2^{60} sets, of which only the two endpoints are replayed. The closure is specific to set-independent scoring and generally fails for option-conditioned generative VLMs; a set outside the declared intervals does not inherit the certificate.

4 SetGuard-R: learning and certifying replay reliability

4.1 A portfolio-error predictor

The target is a frozen dual encoder with normalized image and text embeddings $f(x)$ and $t(c)$ and cosine score $s(x, c) = f(x)^\top t(c)$. The catalog $\mathcal{V}$ is side information available to every method; no ground-truth label is used at inference. The full-catalog score vector is cached once per image, which is practical for the recognition benchmarks below but does not address unbounded catalogs.

For each replayed set $C_j(x)$ with two largest scores $s_{j,(1)} \geq s_{j,(2)}$, the feature vector of mechanism j is

$$v_j = \big[s_{j,(1)},\ s_{j,(1)} - s_{j,(2)},\ H(p_j)/\log|C_j|,\ b_j,\ \log|C_j|,\ r_j/|\mathcal{V}|,\ \Delta_j\big], \tag{8}$$

For a singleton set, the margin and normalized entropy are defined as zero. Here p_j is the within-set softmax at a temperature τ fixed on fitting/tuning data, $r_j \in \{0, \ldots, |\mathcal{V}| - 1\}$ is the zero-based catalog rank of $\hat{y}_j$, Δ_j is a standardized pooled-Gaussian distance difference combined with the margin (Appendix C), and the anchor $b_j = s(x, \hat{y}_j) - \tau \log \frac{1}{|\mathcal{V}|} \sum_{c \in \mathcal{V}} e^{s(x,c)/\tau}$ is an ordinary fixed-catalog confidence, candidate-independent given the winning label; it is not a new out-of-distribution (OOD) score. We learn portfolio vulnerability from these features. The shared score may be logistic, an MLP,

or a permutation-invariant DeepSets network. The DeepSets instance (Zaheer et al., 2017) takes the form

$$u_\theta(x) = \sigma\Big(\rho_\theta\Big[J^{-1}\textstyle\sum_j \phi_\theta(v_j),\ \max_j \phi_\theta(v_j),\ h(x)\Big]\Big), \tag{9}$$

where $h(x)$ contains the fraction of mechanisms that agree with the modal prediction and the entropy of the predicted labels. We fit u_θ to $\ell_\vee$ with a binary cross-entropy loss on the fitting images and accept when $a_t(x) = \mathbf{1}\{u_\theta(x) \le t\}$. Each image contributes a single training target; its J replays are not J independent examples. All mechanisms share one acceptance decision without imposing a hard agreement constraint. Proposition 4 justifies the target and the threshold form; whether u_θ estimates η well enough to beat simple rules is an empirical question, examined in Section 5. The logistic and MLP variants instead concatenate the registered mechanism features with $h(x)$ and use the same target and certificate. Their ordered inputs fix the portfolio size, whereas DeepSets accommodates a variable number of mechanisms; a changed portfolio still requires a new certificate. Optimization details are in Appendix C.

4.2 An image-level finite-sample certificate

Let $\mathcal{T}$ contain T thresholds chosen on independent tuning data, and condition on all fitting and tuning decisions, including the portfolio. On n i.i.d. certification images, define $n_t = \sum_i a_t(X_i)$ and $k_t = \sum_i a_t(X_i)\ell_\vee(X_i, Y_i)$, and the one-sided exact (Clopper–Pearson) binomial bounds

$$\begin{aligned} U_t &= \begin{cases} 1, & n_t \in \{0, k_t\}, \\ \mathrm{Beta}^{-1}(1-\frac{\delta_R}{T}; k_t{+}1, n_t{-}k_t), & \text{otherwise}, \end{cases} \\ V_t &= \begin{cases} 0, & n_t = 0, \\ \mathrm{Beta}^{-1}(\frac{\delta_C}{T}; n_t, n{-}n_t{+}1), & \text{otherwise}. \end{cases} \end{aligned} \tag{10}$$

We select the largest threshold with $U_t \le \alpha$, or reject all inputs if none passes. A prespecified minimum $V_t \ge c_0$ may be required for deployability; missing it is a failed coverage target, not a risk violation.

Theorem 7 (Router-uniform certification)**.** *Under the conditions above, with probability at least $1 - \delta_R - \delta_C$ over the certification images, every passing threshold simultaneously satisfies*

$$\mathcal{R}_\pi(a_t) \le \alpha \quad \textit{and} \quad \mathcal{C}_\pi(a_t) \ge V_t \qquad \textit{for all } \pi \in \Pi_\mathcal{G} \textit{ with positive coverage}. \tag{11}$$

The statement holds for a threshold selected from the passing set after seeing the certification counts; it needs no independence across mechanisms on the same image and no union bound over routers.

The proof applies binomial inversion and a union bound over the frozen thresholds after the routing reduction. With no observed errors, $U_t = 1 - (\delta_R/T)^{1/n_t}$, so $T = 40$ and $\delta_R = 0.025$ require 71 accepted calibration images at $\alpha = 0.10$ and 144 at $\alpha = 0.05$. Appendix A.8 gives a fixed-sequence alternative for the i.i.d. population protocol. Both guarantees condition on the fitted scorer and the frozen portfolio; selecting among models or seeds requires an additional error allocation.

Directly auditing local gates. Under the i.i.d. population protocol, arbitrary fixed gates can also be audited without synchronization. Let $K = \sum_i E_A(X_i, Y_i)$ and $M = \sum_i (E_A + S_A)(X_i, Y_i)$. The two events are disjoint and, conditional on M, $K \sim \mathrm{Binomial}(M, B(A))$ whenever $e_A + s_A > 0$; the same one-sided bound with $n_t = M$ and $k_t = K$ therefore controls the routing envelope, and $M = 0$ yields the trivial bound. Worst-router coverage is bounded separately from $\sum_i D_A(X_i)$ with its own budget. This audit shows why local gates that succeed individually may fail jointly: their correct but avoidable predictions do not contribute to S_A.

4.3 Certification on a finite evaluation pool

An official benchmark is a finite collection that may have class quotas, duplicates, or a distribution different from its training split. The following certificate conditions on the entire evaluation pool instead of assuming its records are i.i.d. draws; it is deliberately a different guarantee from Theorem 7.

Freeze all predictors, gates, and threshold grids using external fitting and tuning data. From a pool of N records, draw n certification indices by simple random sampling without replacement,

independently of the records' contents, and score all records without labels. At threshold t, let N_t be the number of accepted pool records, n_t the number of these drawn for certification, k_t their portfolio errors, and K_t the unknown number of accepted portfolio errors in the pool. Conditional on n_t, the accepted certification indices form a uniformly random subset of size n_t of the N_t accepted records, so $k_t \mid n_t \sim \mathrm{Hypergeom}(N_t, K_t, n_t)$. With $F_{N,K,n}$ the hypergeometric CDF and $\bar{\delta} = \delta_R/T$, define

$$K_t^U = \max\{K \in \{k_t, \ldots, N_t - n_t + k_t\} : F_{N_t,K,n_t}(k_t) \geq \bar{\delta}\}. \tag{12}$$

When $N_t > n_t$, the union risk $B_{\mathrm{rem},t} = (K_t - k_t)/(N_t - n_t)$ of the accepted records in the remaining blind pool satisfies, simultaneously for all registered thresholds with probability at least $1 - \delta_R$,

$$B_{\mathrm{rem},t} \leq U_t^{\mathrm{HG}} := \frac{K_t^U - k_t}{N_t - n_t}, \qquad \mathcal{C}_{\mathrm{rem},t} = \frac{N_t - n_t}{N - n}. \tag{13}$$

The remaining-pool coverage in (13) is known exactly from unlabeled scores. A threshold passes only if $U_t^{\mathrm{HG}} \leq \alpha$ and $N_t > n_t$. If $n_t = 0$, then $K_t^U = N_t$ and the bound equals one, so no favorable zero-error inference is made without labels.

Via the pointwise replay envelope, the bound controls every router on the remaining pool. Its only randomness is the certification draw, and its coverage is an exact remaining-pool fraction (Appendix B).

Algorithm 1: SETGUARD-R on a finite evaluation pool

1. Fit a union-error scorer on fitting replay; freeze it, the portfolio, and the T thresholds obtained from independent tuning images.
2. Score the N pool records without labels. Uniformly draw n indices without replacement and obtain their labels.
3. At every threshold count (N_t, n_t, k_t) and compute K_t^U and U_t^{HG} by (12)–(13).
4. Select the largest threshold with $N_t > n_t$ and $U_t^{\mathrm{HG}} \leq \alpha$; if none passes, reject all remaining records.
5. Use this shared gate for every registered router on the remaining pool.

5 EXPERIMENTS

Our experiments separate the contribution of joint certification from that of the score used to rank images. Logistic, MLP, and DeepSets gates instantiate the same SETGUARD-R framework; the comparison does not presume that permutation invariance improves coverage for a fixed portfolio.

5.1 PROTOCOL AND COMPARISONS

Systems and data. The primary system is Flowers-102 (Nilsback & Zisserman, 2008) with a frozen CLIP ViT-B/16 target (Radford et al., 2021). Its four candidate mechanisms retrieve the top-4 and top-64 labels under each of two proposal encoders, CLIP RN50 and ViT-B/32; candidate generation never uses the image label. The fitting, tuning, certification, and remaining evaluation splits contain 1,632, 408, 1,844, and 4,305 images, and each labeled image evaluates all mechanisms. Cross-system comparisons add CUB-200-2011 (Wah et al., 2011) and ImageNet-1K (Deng et al., 2009) and use both ViT-B/16 and ViT-L/14 targets; gates are fitted separately for each system. Training records (training and validation combined for Flowers-102) are split $80/20$ into fitting and tuning sets; ImageNet uses a label-independent $50{,}000$-image training subset. Certification samples 30% of each evaluation pool uniformly without replacement and evaluates the selected gate on the complement.

Certification and controls. The primary certificate is the exact remaining-pool hypergeometric bound, with $\delta_R = 0.05$ and $\alpha \in \{0.05, 0.10\}$. Each of at most $T = 40$ tuning-derived thresholds receives budget δ_R/T. We choose the largest passing threshold; if none passes, coverage is zero and the conditional risk is undefined. Budgets apply separately to each frozen scorer and protocol. Scoring baselines, such as maximum softmax probability (MSP), are pooled by their worst mechanism and use the same joint certificate. For likelihood-ratio selectors, we report our pooled LR-inspired adaptation (Appendix C) and the class-conditional Δ-MDS–RLog selector (Heng & Soh, 2026).

Table 1: Flowers-102, CLIP ViT-B/16, seed 0, $\alpha = 0.10$. All rows use joint replay. Coverage is a percentage; UCB and realized union risk are probabilities. $T = 1$ for synchronized local gates and $T \leq 40$ otherwise. Reject-all has zero coverage and undefined conditional risk. Strict-risk results are in Table 11.

Shared gate	Risk UCB	Coverage (%)	Union risk
MSP (Hendrycks & Gimpel, 2017)	.098	43.3	.068
Temp.-scaled MSP (Guo et al., 2017)	.098	43.3	.068
MCM (Ming et al., 2022), $\tau = 1$	.094	22.6	.057
Energy (Liu et al., 2020)	–	0.0	–
LR-inspired (pooled) (Heng & Soh, 2026)	.090	41.5	.059
Δ-MDS–RLog (Heng & Soh, 2026)	.091	42.1	.058
Worst margin	.099	43.1	.068
Agreement + worst margin	.098	42.2	.057
Catalog agreement + margin	.098	42.2	.057
ViLU (Lafon et al., 2025) + replay	.095	44.6	.054
DeepSets (Zaheer et al., 2017), mean target	.100	43.9	.051
Synchronized local MLP ($T = 1$)	.089	47.8	.058
SETGUARD-R: DeepSets, union	.094	44.9	.063
SETGUARD-R: MLP, union	.096	45.5	.052
SETGUARD-R: logistic, union	.096	**47.9**	.061

ViLU (Lafon et al., 2025) is trained on matched mechanism–image examples, pools predicted failure by the maximum, and is reported as "ViLU + replay." The logistic, MLP, and DeepSets union gates share seven mechanism features, agreement statistics, and labeled fitting images. The mean-target control changes only the DeepSets supervision. Synchronized local MLP gates fix their thresholds on tuning data, intersect their accepted sets, and certify the resulting single shared gate ($T = 1$).

5.2 CERTIFIED COVERAGE

Table 1 reports the primary split. At $\alpha = 0.10$, the logistic union gate retains 47.9% of the remaining records, compared with 44.6% for ViLU + replay and 43.3% for MSP. Synchronized local gates reach 47.8% with a smaller threshold family, and DeepSets retains 44.9%. The same certificate thus supports competitive coverage across architectures. With the DeepSets score held fixed, the marginal union-bound certificate returns zero coverage, whereas joint replay certifies 44.9% with upper confidence bound (UCB) 0.094 and realized union risk 0.063. This comparison isolates the information retained by joint replay. For that gate, 52 errors among 865 accepted certification images yield $K_t^U = 234$ accepted errors in the full pool. The remaining accepted count is $1{,}932$, giving $(234 - 52)/1{,}932 = 0.0942$; the observed remaining error count is 122, giving $122/1{,}932 = 0.0631$. Certification and evaluation therefore use different numerators.

At $\alpha = 0.05$, MLP certifies 19.4% coverage and the mean-target DeepSets control certifies 20.0%, whereas the primary union-target DeepSets gate rejects all inputs. The union target is thus not uniformly better at the stricter level. Across five training seeds on the fixed split, logistic, MLP, and DeepSets attain 48.8 ± 1.1, 47.7 ± 1.6, and $46.5 \pm 1.0\%$ coverage at $\alpha = 0.10$. Logistic and MLP certify nonzero coverage at $\alpha = 0.05$ for all five seeds, compared with one seed for DeepSets. These runs vary the fitted scorer while keeping the data split and certification draw fixed.

Table 2 reports all six dataset–target systems. In five-seed mean coverage, logistic exceeds ViLU by 4.1–4.8 percentage points. All four scoring methods improve with ViT-L/14. The gain over ViLU does not grow uniformly with the target backbone, and because each system is fitted separately, these results do not test the transfer of a gate across datasets.

5.3 SEPARATING SCORING FROM CERTIFICATION

We compare three finite-pool certification protocols with each score held fixed (Table 3) under the same total risk-failure budget $\delta_R = 0.05$. Joint-grid bounds the union error directly, allocating δ_R/T to each frozen threshold. Marginal-grid bounds the error of each mechanism on the *same shared accepted set*, allocating $\delta_R/(JT)$ to each mechanism–threshold pair, and sums these bounds, capped

Table 2: Certified coverage (%) at $\alpha = 0.10$, mean $\pm$ standard deviation over five training seeds. All methods use joint replay. Logistic, MLP, and DeepSets (DS) are union-target SETGUARD-R variants.

Dataset	Target	Logistic	MLP	DS	ViLU (Lafon et al., 2025)
Flowers-102	ViT-B/16	48.8 ± 1.1	47.7 ± 1.6	46.5 ± 1.0	44.1 ± 1.2
Flowers-102	ViT-L/14	54.2 ± 0.9	53.5 ± 1.1	52.1 ± 0.8	49.8 ± 1.0
CUB-200-2011	ViT-B/16	45.3 ± 1.2	44.6 ± 1.4	43.8 ± 1.1	41.2 ± 1.3
CUB-200-2011	ViT-L/14	51.8 ± 0.8	50.9 ± 1.0	49.7 ± 0.9	47.3 ± 1.1
ImageNet-1K	ViT-B/16	58.4 ± 0.6	57.9 ± 0.7	56.8 ± 0.5	53.6 ± 0.8
ImageNet-1K	ViT-L/14	65.2 ± 0.5	64.7 ± 0.6	63.9 ± 0.4	60.5 ± 0.7

Table 3: Score–certificate comparison on Flowers-102, CLIP ViT-B/16, seed 0, $\alpha = 0.10$, $\delta_R = 0.05$. Entries are certified remaining-pool coverage (%). Marginal-grid certifies zero coverage for all scores. Both grid protocols use the same score-specific threshold family ($T \leq 40$); tuning-one tests a single tuning-selected threshold.

Score	Joint-grid	Tuning-one
MSP (Hendrycks & Gimpel, 2017)	43.3	44.1
LR-inspired (pooled) (Heng & Soh, 2026)	41.5	42.4
Δ-MDS–RLog (Heng & Soh, 2026)	42.1	43.0
ViLU (Lafon et al., 2025) + replay	44.6	45.2
SETGUARD-R: logistic, union	47.9	48.6
SETGUARD-R: MLP, union	45.5	46.2
SETGUARD-R: DeepSets, union	44.9	45.3
DeepSets (Zaheer et al., 2017), mean target	43.9	44.5

at one. This union bound requires no independence between mechanism errors. Tuning-one fixes a single threshold on tuning data and tests it at level δ_R; if it fails, no other threshold is tested on the certification labels.

Marginal-grid returns zero coverage for every score in this comparison, whereas joint replay retains the correlation of errors across mechanisms instead of summing marginal upper bounds. On the same remaining records, the ungated mean mechanism error is 33.8% and the union error is 37.5%, so much of the error mass is shared across mechanisms (Table 10); summing marginal bounds pays for this overlap repeatedly. The joint certificate benefits the external scores as well as the learned union gates. Tuning-one increases coverage by 0.4–0.9 percentage points on this split: it trades the grid's choice after certification for a single tuning-selected operating point. This outcome does not imply that a single-threshold protocol always dominates a grid.

5.4 ROUTING, ABLATIONS, AND CLOSURE

Per-mechanism MSP calibration at $\alpha = 0.10$ yields risk 0.054 under uniform routing but 0.230 under the strongest tested label-free router (Table 4). Local MLP gates remain below the target, even in their oracle envelopes, so this is a concrete failure of local calibration rather than an empirical failure shared by every local score. For the primary shared DeepSets gate, all residual accepted errors happen to be shared by the four mechanisms, so every routed risk equals 0.063 on this split. Shared acceptance guarantees common coverage and a router-uniform bound on union risk; it does not in general make routed risks identical. The certified bound in this run is 0.094.

Ablations and endpoint closure. Removing the catalog anchor and rank features reduces the seed-average DeepSets coverage from 46.5 to 43.8%, and removing the likelihood-ratio feature reduces it to 41.2% (Table 6). Removing the agreement features changes it to 46.4% and leaves strict-level feasibility at one of five seeds, so agreement is not essential in this system. Interior top-8, top-16, top-32, and thinned candidate sets produce no pointwise closure violations, and their extended union error equals the endpoint union error (Table 7). For separately refitted portfolios with $J = 2, 4, 8$, coverage is 47.9, 44.9, and 44.9% (Table 5); the additional degraded proposal views provide no coverage gain here. Deployment nevertheless requires both proposal-encoder passes as well as the target pass; a small gate cost does not remove this replay overhead.

Table 4: Routing on Flowers-102 at $\alpha = 0.10$. Worst LF is the largest risk among the tested label-free routers, with its coverage (%) in parentheses. Oracle is the label-aware risk envelope. MinCov is the minimum coverage over in-portfolio routers, and MeanCov is the uniform-router coverage (both %).

Gate / calibration	Uniform	Worst LF (cov.)	Oracle	MinCov	MeanCov
MSP (Hendrycks & Gimpel, 2017) / pooled	.045	.072 (15.9)	.194	6.6	14.6
MSP / per mechanism	.054	.230 (6.3)	1.000	0.0	26.2
Local MLP / pooled	.048	.060 (47.3)	.074	37.7	42.4
Local MLP / per mechanism	.046	.058 (44.2)	.070	37.5	41.4
SETGUARD-R: shared DeepSets	.063	.063 (44.9)	.063	44.9	44.9

6 RELATED WORK

Selective prediction and risk control. Selective classification trades coverage against accepted error (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019); our optimal shared gate is a portfolio version of thresholding conditional error. Distribution-shift methods improve scoring and characterize selection (Xia & Bouganis, 2022; Liang et al., 2024; Heng & Soh, 2026); risk-controlling prediction and learn-then-test separate fitting from finite-sample validation (Bates et al., 2021; Angelopoulos et al., 2025; 2024). We adopt this separation after reducing routing risk to joint replay events.

Vision-language uncertainty. The calibration of vision-language models can change with the label vocabulary and image–text geometry (LeVine et al., 2023; Tu et al., 2024; Wang et al., 2025). ViLU predicts failures from image and candidate-text embeddings (Lafon et al., 2025); MCM uses concept matching for OOD detection (Ming et al., 2022); and likelihood-ratio selectors contrast the feature distributions of correct and incorrect predictions (Heng & Soh, 2026). These methods supply scores, whereas our certificate controls how errors compose under input-dependent candidate routing. We therefore evaluate their scores under a common replay certificate rather than treating uncertainty estimation itself as a routing guarantee.

Guarantees under changing policies. Weighted conformal prediction handles covariate shift through density ratios (Tibshirani et al., 2019), and conditional guarantees can cover a specified class of reweightings (Gibbs et al., 2025). Conformal Policy Control uses policy-dependent likelihood ratios to control losses as policies change (Prinster et al., 2026). Replay instead supplies full information about a finite candidate portfolio: one labeled image evaluates every mechanism. This permits router-uniform control without estimating propensities, provided the image–label law (or a declared finite pool) stays fixed.

7 LIMITATIONS AND CONCLUSION

The guarantee covers the frozen portfolio and its proven endpoint closure. Changing the retriever, catalog, image distribution, or endpoint preprocessing requires a new analysis or certificate. The finite-pool certificate concerns the remaining records, not future images. Adding mechanisms can expose more errors and reduce coverage, and full replay adds proposal-encoder passes. Benchmark overlap with CLIP pretraining data remains possible, and aggregate risk control does not establish groupwise reliability.

Candidate mechanisms that are reliable in isolation need not remain reliable when an input-dependent router chooses among them. Replay exposes their joint error structure, and shared acceptance removes the router's ability to manipulate the acceptance denominator. The resulting union-error target admits exact certification and can be learned with several scoring architectures. Our experiments show that joint replay retains useful coverage, with the logistic variant achieving the highest average coverage. At the stricter risk level, the comparison with mean-target supervision shows that score fitting and certification power must be evaluated together.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) solely for language polishing, including improving grammar, clarity, and readability. These tools were not used to generate substantive content or to perform any research tasks, including developing methods or theoretical frameworks, formulating claims or hypotheses, constructing proofs, designing experiments, generating or processing data, analyzing or interpreting results, writing code, searching the literature, translating text, or preparing figures. All AI-assisted revisions were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including all claims, results, and references.

REFERENCES

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

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021.

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

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

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, 2019.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Alvin Heng and Harold Soh. Know when to abstain: Optimal selective classification with likelihood ratios. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RsfaFXBFzM.

Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphaël Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. ViLU: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17807–17817, 2025.

Will LeVine, Benjamin Pikus, Pranav Raja, and Fernando Amat Gil. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=dmxMGW6J7N.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35087–35102, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2603.02196.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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, 2019.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48791–48808, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63467–63489, 2025.

Guoxuan Xia and Christos-Savvas Bouganis. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 1995–2012, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

A Proofs and illustrative constructions

A.1 Proof of Proposition 1

Partition the image space into disjoint measurable regions H_1, H_2, H_0 with probabilities $\beta, \beta, 1-2\beta$. Let $Y = 0$ everywhere, and fix a dual-encoder decision rule that predicts label 0 from $\{0\}$ and

label 1 from $\{1\}$. Mechanism j returns $\{1\}$ on H_j and $\{0\}$ elsewhere; both mechanisms depend on x alone, and both gates always accept. Each mechanism therefore has risk β, and so does every constant mixture. The image-dependent router that selects mechanism 1 on H_1, mechanism 2 on H_2, and either mechanism elsewhere has risk 2β. The distribution of (X, Y) and the scoring rule are unchanged. Singleton sets keep the construction short; adding lower-scoring labels yields the same example at any set size. The example establishes non-composition, not the impossibility of every calibration procedure. □

A.2 Proof of Theorem 2

For a router, write $b = \mathbb{E}\sum_j \pi_j A_j L_j$ and $c = \mathbb{E}\sum_j \pi_j A_j (1 - L_j)$. Pointwise, $\sum_j \pi_j A_j L_j \le E_A$ and $\sum_j \pi_j A_j (1 - L_j) \ge S_A$, so $b \le e_A$, $c \ge s_A$, and $b/(b + c) \le e_A/(e_A + s_A)$ whenever the denominators are positive, because $b/(b + c)$ is nondecreasing in b and nonincreasing in c. If $e_A = s_A = 0$, there are almost surely no accepted errors, so every positive-coverage router has risk zero. For attainment, choose the first index that accepts an error on E_A, index 1 on S_A, and the first rejecting index elsewhere; the latter exists because there is no accepted error there and not every mechanism accepts correctly. This measurable finite-index oracle has error mass e_A and correct mass s_A, and it attains the bound whenever $e_A + s_A > 0$. With deterministic labels, $L_j(x, f_*(x))$ is a function of x, so the oracle is an image-only router, whether or not it is learnable. For coverage, $\sum_j \pi_j A_j \ge \min_j A_j = D_A$ pointwise, and choosing a rejecting index whenever one exists attains equality using only x. □

A.3 Proof of Theorem 3

For the synchronized gates $\bar{A}_j = D_A$, the all-accepted indicator is still D_A, so the minimum coverage is unchanged. The forced-correct event is $S_{\bar{A}} = \{D_A = 1, \forall j : L_j = 0\} = S_A$, and the exposed-error event is $E_{\bar{A}} = \{D_A = 1, \exists j : L_j = 1\} \subseteq E_A$. Holding s_A fixed, $e/(e + s_A)$ is nondecreasing in e (with the stated zero conventions), so $B(\bar{A}) \le B(A)$. For the exact image-only risk, fix any router under $\bar{A}$; in the original system, copy it where $D_A = 1$ and choose the first rejecting mechanism where $D_A = 0$. This router uses only x and reproduces the same accepted predictions and losses, so every positive-coverage risk attainable after synchronization is also attainable before it, including with stochastic labels. The comparison is between policy classes that can evaluate all gates; an information-constrained class of local gates poses a different problem. □

A.4 Proof of Proposition 4

If accept-all is feasible, it maximizes coverage. If no positive-coverage feasible gate exists, rejection is the fallback. In the remaining case, for a randomized shared gate $a(x) \in [0, 1]$ with $\mathbb{E}[a(X)] > 0$, the constraint $B(a) \le \alpha$ is equivalent to $\mathbb{E}[a(X)(\eta(X) - \alpha)] \le 0$. Fix any feasible gate with coverage $m = \mathbb{E}[a(X)]$. Among all gates with the same coverage, the bathtub principle places acceptance mass on the smallest values of η; hence there is a lower level set with boundary randomization, $\tilde{a}(x) = \mathbf{1}\{\eta(x) < c\} + \rho \mathbf{1}\{\eta(x) = c\}$, satisfying $\mathbb{E}[\tilde{a}] = m$ and $\mathbb{E}[\tilde{a}\eta] \le \mathbb{E}[a\eta]$. Thus $\tilde{a}$ is feasible and has the same coverage. For existence, let Q be a quantile function of $\eta(X)$ and set $H(m) = \int_0^m (Q(u) - \alpha)\, du$ for $m \in [0, 1]$. The smallest achievable constraint value at coverage m is $H(m)$, realized by that threshold gate. Since H is continuous, the feasible set $\{m : H(m) \le 0\}$ is compact and has a largest element m^*. A quantile threshold with boundary randomization realizes m^* and is optimal. When $c^* = \alpha$, its boundary contributes zero to the constraint and may be accepted in full. Strictly increasing transforms preserve all sublevel sets. □

A.5 Proof of Corollary 5

Adding a mechanism expands the existential event E_A and contracts both universal events S_A and D_A; the ratio is nondecreasing in its first argument and nonincreasing in its second. With a fixed shared gate, only the union-error event changes. Refitting the gate can move coverage in either direction; the monotonicity statement does not compare independently refitted systems. □

A.6 Proof of Proposition 6

If either endpoint is wrong, the right-hand side is one. Otherwise, both endpoint predictions equal y: then $y \in C^-$, and y wins among all labels in C^+ under the fixed score and tie order. Every set that contains C^- and is contained in C^+ includes y and no label ranked above it, so its prediction is y. Taking maxima across intervals gives pointwise domination by the endpoint union loss; multiplying by the unchanged shared gate proves the routing claim. Because the endpoints belong to the family, its worst label-aware error equals the endpoint union error. □

A.7 Proof of Theorem 7

Condition on the portfolio, VLMs, scorer, hyperparameters, and threshold grid; each certification pair is then an i.i.d. draw from P. A zero-acceptance gate has $n_t = 0$ and cannot pass a nontrivial certificate. Otherwise, conditional on the accepted index set, the n_t retained image-level union losses are i.i.d. Bernoulli with parameter $B(a_t) = P(\ell_\vee = 1 \mid a_t = 1)$, and (10) is an exact one-sided upper confidence bound; a union bound over the T thresholds gives $B(a_t) \le U_t$ simultaneously with probability at least $1 - \delta_R$. Moreover, $n_t \sim \mathrm{Binomial}(n, P(a_t = 1))$, so binomial inversion and another union bound give $P(a_t = 1) \ge V_t$ for all thresholds with probability at least $1 - \delta_C$. On the intersection of these events, (6) and Theorem 2 transfer the bounds to every router, and selecting a threshold within that event does not invalidate them. No independence between mechanisms is used. □

A.8 Two-chain fixed-sequence certification

Order the frozen thresholds as $t_{(1)} < \cdots < t_{(T)}$. An *anchor* t° is chosen from tuning data alone: for each threshold, the tuning acceptance rate is scaled to the certification sample size, the accepted tuning risk is replaced by its exact binomial upper bound at level $1/2$ plus one binomial standard deviation, and t° is the largest threshold whose predicted exact bound at level δ_R is at most α (or the threshold with the smallest predicted bound if none qualifies). Two chains are then tested on the certification data, each stopped at its first failure: an upward chain $t^\circ, t^\circ_{+1}, \ldots$ and a coarse downward chain $t^\circ_{-4}, t^\circ_{-8}, \ldots$. Each hypothesis $H_t : B(a_t) > \alpha$ is tested at level $\delta_R/2$ with the binomial bound (10) at $T = 1$ under the i.i.d. population protocol. When the anchor is the smallest threshold, only the upward chain exists, and it runs at level δ_R. The largest threshold certified by either chain is selected. Because the chains' orders and levels depend only on tuning data and each chain stops at its first non-rejection, a false certification within a chain requires rejecting the first true null in that chain's order. This has probability at most $\delta_R/2$ in the two-chain case, giving a total error of at most δ_R by a union bound; in the single-chain case, it is at most δ_R directly (Angelopoulos et al., 2025). The certified set is a contiguous block above the anchor plus a sparse set below it, not a prefix of the family. The coverage bound V_t retains its union bound over the grid. With $\delta_R = 0.025$, the zero-error requirements are 42 ($\alpha = 0.10$) and 86 ($\alpha = 0.05$) accepted images per test, compared with 71 and 144 under the union bound; the downward chain rescues cases in which the anchor is too aggressive, at the cost of a more conservative selected threshold. This argument concerns the fixed population risks $B(a_t)$; it does not by itself establish a fixed-sequence guarantee for the random remaining-pool risks. The primary experiments use the hypergeometric grid certificate with a Bonferroni correction.

A.9 A finite-sample power calculation

For zero observed errors, $U = 1 - \bar{\delta}^{1/m}$ with $\bar{\delta} = \delta_R/T$, so the smallest accepted count that can certify level α is $m_{\min} = \lceil \log(\delta_R/T)/\log(1-\alpha) \rceil$: 71 ($\alpha = 0.10$) and 144 ($\alpha = 0.05$) for $T = 40$ and $\delta_R = 0.025$, and 42 and 86 per test under the two-chain fixed-sequence scheme (36 and 72 for a single chain at level δ_R). If the true accepted union risk is q, the conditional power at accepted sample size m is exactly $\sum_{k=0}^{m} \binom{m}{k} q^k (1-q)^{m-k} \mathbf{1}\{U_{\mathrm{CP}}(k, m; \bar{\delta}) \le \alpha\}$. Power, decisive counts for local gates, and coverage lower bounds distinguish a poor ranking score from insufficient labeled data. Reducing the number of preregistered thresholds improves power; tuning the grid after seeing certification errors is not allowed.

B PROOF AND SCOPE OF THE FINITE-POOL CERTIFICATE

This appendix proves the remaining-pool guarantee of Section 4.3. The pool and the acceptance set of every threshold are fixed before the certification indices are sampled.

Validity. The hypergeometric CDF $F_{N,K,n}(k)$ is nonincreasing in K, so $\{K_t^U < K_t\} = \{F_{N_t,K_t,n_t}(k_t) < \bar{\delta}\}$ for any fixed K_t, and the latter event has probability at most $\bar{\delta}$ by discrete CDF inversion, conditionally on n_t and therefore unconditionally. A union bound over the frozen family gives simultaneous validity, and subtracting the observed k_t from the bound on total errors gives the remaining-pool statement. The replay envelope and the nested closure are pointwise and therefore apply unchanged to the uniform distribution over the remaining records. Neither the gate nor the threshold family may be adapted to the revealed labels. Per-mechanism baselines allocate their budget over every mechanism–threshold pair; a pooled-mixture baseline first assigns one router draw to every pool record, freezes it, and applies the same inversion to the routed predictions. These baselines certify their respective fixed decisions, not arbitrary switching.

C IMPLEMENTATION AND EVALUATION PROTOCOL

C.1 MODEL AND MECHANISM CONTRACT

The evaluated targets are OpenAI CLIP `ViT-B/16` and `ViT-L/14` (Radford et al., 2021); the proposal encoders are OpenAI CLIP `RN50` and `ViT-B/32`, which differ from every target. The reference pipeline loads the models through `open_clip` with their released preprocessing in evaluation mode and records model identifiers, library versions, preprocessing, and manifest hashes alongside the feature metadata. Each model encodes one dataset-specific template from the CLIP zero-shot prompt set (e.g., "a photo of a {}, a type of flower."); image and text features are ℓ_2-normalized, and the score is their dot product. Class names follow the official class lists. Candidate labels are sorted by their integer catalog identifier to break exact score ties, and the candidate manifest is identical for all target models.

The four endpoint mechanisms are, in registered order, RN50–4, RN50–64, B/32–4, and B/32–64: the top-K catalog labels under the proposal encoder's cosine score. $J = 2$ is the RN50 interval alone, and $J = 4$ is the primary portfolio. For $J = 8$, we append both intervals computed after first resizing the proposal image to 112 pixels on the shorter edge and then applying the released preprocessing; the released protocol also defines $J = 12$ with the same intervals at 56 pixels, which we do not report. The target encoder always sees the original image. Within each interval, we also replay $K = 8, 16, 32$ and a thinned subset that keeps the four inner labels and deletes every second outer-only label in catalog order; these interior sets are used only to test Proposition 6. The gate always uses endpoint features, and candidate generation never uses Y.

C.2 SPLITS AND SAMPLE UNITS

We use the official training/evaluation partitions, keep their class lists, and record image identifiers. Official training images (train and validation, where both exist) are split $80/20$ into fitting/tuning by a fixed permutation (NumPy PCG64, seed 2027); for ImageNet, 50,000 training identifiers are first selected without labels (seed 2026). The scorer, portfolio, and tuning grid are then frozen. The official evaluation pool is permuted without access to labels (seed 2028); the first $\lfloor 0.30N \rfloor$ records reveal their labels for certification, and the remainder forms the blind test set. This departs from the standard benchmark split and is documented with a manifest. No class quotas are imposed on the certification draw, resampling until a favorable certificate appears is prohibited, and remaining-pool claims are transductive.

C.3 SCORER AND OPTIMIZATION

For the Flowers-102 implementation, the temperature τ shared by the within-set softmax and the anchor is selected from $\{0.01, 0.03, 0.07, 0.10\}$ by the tuning log-loss of the seed-0 union gate and then frozen for every learned gate. Feature means and standard deviations are computed on fitting images only and shared across mechanisms. The mechanism encoder has widths $d \to 64 \to 64$ with

GELU activations, where $d = 7$ for the six replay features plus the likelihood-ratio score ($d = 6$ in the no-LR ablation); mean and maximum pooling give 128 features, and together with the two agreement features they feed a $130 \to 128 \to 1$ head (21,569 parameters in total). We train with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256 images) for at most 100 epochs, stopping early after ten epochs without improvement in tuning log-loss. No class reweighting is used. Fitted seeds are 0–4; the selected Flowers-102 temperature is $\tau = 0.07$. Seed 0 is the primary frozen scorer and is never replaced by the best test seed.

The mean-target control replaces $\ell_\vee$ by $J^{-1} \sum_j L_j$ and leaves the architecture, optimizer, and budget unchanged. The MLP variant of SETGUARD-R concatenates the J endpoint vectors in registered order with the agreement features and uses two hidden layers whose width is chosen so that its parameter count is closest to that of DeepSets. For $J = 4$, its input dimension is 30, hidden width is 131, and parameter count is 21,485. The logistic variant uses the same 30 standardized inputs and has 31 parameters; its learning rate is 10^{-2}. Both retain AdamW weight decay 10^{-4}, batch size 256, the 100-epoch limit, and ten-epoch tuning patience. The full-catalog agreement gate rejects a portfolio if any prediction differs from the full-catalog winner and otherwise ranks by the full-catalog margin; it is implemented as a hard mask multiplied by the gate, so that $t = +\infty$ never re-accepts a disagreement. Plain agreement rejects any inter-mechanism disagreement and ranks by the worst margin. Pooled confidence controls take the minimum over mechanisms of MSP at $\tau = 0.01$, of MSP at the temperature minimizing the fitting negative log-likelihood of the full-catalog softmax (temperature scaling), of MSP at $\tau = 1$ (MCM), and of the negative energy. The LR-inspired control is a pooled-Gaussian adaptation of Heng & Soh (2026). For each mechanism, correct and incorrect fitting predictions define means $\mu_{j,c}, \mu_{j,w}$. The within-group residual covariance $\widehat{\Sigma}_j$ is regularized as $\Sigma_j = 0.9\widehat{\Sigma}_j + 0.1\,\mathrm{tr}(\widehat{\Sigma}_j)I/d_f$, where d_f is the target embedding dimension. Using its pseudoinverse, define $D_j(x) = \|f(x) - \mu_{j,w}\|^2_{\Sigma_j^{-1}} - \|f(x) - \mu_{j,c}\|^2_{\Sigma_j^{-1}}$. The seventh feature in (8) is $\Delta_j = z_j(D_j) + \lambda_j z_j(s_{j,(1)} - s_{j,(2)})$, where each z_j uses fitting means and standard deviations and $\lambda_j \in \{0, 0.25, 0.5, 1, 2, 4\}$ minimizes tuning AURC for mechanism j. The pooled baseline uses $\min_j \Delta_j$ as confidence. Tables 1 and 3 report this pooled adaptation and the original class-conditional Δ-MDS–RLog selector as separate baselines. The no-catalog-anchor/rank ablation removes only b_j and $r_j/|\mathcal{V}|$; the LR feature and other candidate-dependent features remain.

C.4 FROZEN THRESHOLDS AND CERTIFICATION ROUTINE

The grid contains 39 tuning-score quantiles at probabilities $r/40$, $r = 1, \ldots, 39$, plus $+\infty$; duplicates are removed, and the resulting count $T \le 40$ is used in every allocation. Reject-all is a separate fallback, not a threshold. Accepted sets are nested in the threshold because smaller predicted vulnerability means greater confidence, so the largest passing threshold maximizes acceptance among passing thresholds, although risk bounds need not be monotone. The population routine is as follows:

1. Freeze the catalog, target model, portfolio, fitted scorer, and grid.
2. Replay all J mechanisms for each certification image and compute $\ell_{\vee,i} = \max_j \mathbf{1}\{\hat{y}_{ij} \neq y_i\}$ once per image.
3. For each threshold, count n_t and k_t and compute U_t and V_t (Bonferroni, $\delta_R = \delta_C = 0.025$), or run the fixed-sequence test of Appendix A.8.
4. Select the largest passing threshold, or reject all.
5. At deployment, replay the same portfolio, evaluate the shared gate, and return the routed mechanism's prediction only if the gate accepts.

The finite-pool routine replaces step 3 by the hypergeometric bounds (13), with Bonferroni allocation δ_R/T, $\delta_R = 0.05$, and exact remaining-pool coverage; it is the routine behind the primary tables. The marginal competitor uses the same shared gate, per-mechanism hypergeometric bounds with budget $\delta_R/(JT)$, and $U_{\mathrm{marginal}}(t) = \min\{1, \sum_j U_{j,t}\}$. Pooled calibration of a local score predraws one mechanism per pool record from the uniform mixture (seed 0) and certifies the routed decisions; per-mechanism calibration certifies each mechanism's own score on its own grid with budget $\delta_R/(JT)$. The synchronized-local-gates ablation chooses per-mechanism thresholds on tuning data at the target risk, intersects the gates, and certifies the resulting single shared gate ($T = 1$). A new scorer seed, catalog, endpoint retriever, endpoint preprocessing rule, or gate changes the certified object. The 0.05

budget applies separately to each fixed scorer and dataset–model system, not simultaneously to all methods, seeds, or systems; selecting among fitted models requires an additional allocation.

C.5 Routing rules and out-of-contract tests

Constant routers include each mechanism alone, the uniform distribution, and a distribution placing 0.8 on the first registered mechanism and 0.2 uniformly on the rest. The entropy router selects the mechanism with the largest normalized target entropy; its gate-aware variant does so among the mechanisms that the gate accepts. The disagreement router selects the least frequent predicted label, breaking ties by entropy and then by mechanism order. The learned adversarial router is a 64-unit MLP on the concatenated replay features, trained on fitting images to predict all J errors and early-stopped on tuning images; it routes to the mechanism with the largest predicted error (gate-aware variant: among accepted mechanisms). The denominator-seeking variant first chooses the accepted mechanism with largest predicted error. If that error is below θ and a rejecting mechanism exists, it instead chooses rejection. We select θ from 20 tuning-score quantiles between probabilities 0 and 0.95, maximizing tuning risk subject to at least 20 accepted tuning images; ties use the first threshold and an infeasible search defaults to zero. This is the router attaining 0.230 in Table 4. Its scorer and threshold search use no certification or remaining-pool labels; its routing decision may observe the calibrated gates. Worst LF reports the maximum over this stated finite family, not an estimate of the supremum over all label-free routers. The label-aware oracle chooses an accepted error when one exists, otherwise a rejecting mechanism when available, and otherwise any mechanism; its risk is computed from the E_A, S_A counts and equals the accepted union error for a shared gate. Out-of-contract tests (a third proposal encoder, $K = 2$ or $K = \min(128, |\mathcal{V}|)$, true-label deletion, and image corruption) are part of the released protocol, are labeled *uncertified*, and are not reported here; a guarded service checks membership in a registered interval without using the true label and rejects requests outside it.

C.6 Reporting and computational cost

The reference replay format contains image identifiers, labels, candidate memberships, predictions, feature vectors, scores, and mechanism hashes. Certification requires one frozen score per pool record and an $n \times J$ calibration-loss matrix. The Flowers-102 results report E_A, S_A, D_A counts, union omissions, ranking-only errors, disagreement, and wrong agreement. Remaining-pool rates are exact finite-population quantities once all labels are revealed; the bootstrap intervals over test images stored in the result files are descriptive stability analyses, not additional guarantees. Contrastive replay reuses the cached text and image embeddings, so certification and deployment cost one forward pass per proposal encoder plus one target pass per image; full-catalog scoring is counted for every method that uses it.

C.7 Additional results

Table 5: Endpoint-portfolio study on Flowers-102, CLIP ViT-B/16 at $\alpha = 0.10$ (finite-pool Bonferroni certificate, seed 0). The DeepSets gate is refitted and recertified for each portfolio. "Accepted errors / count" are the certification-sample counts k_t/n_t; coverage is the exact remaining-pool coverage (%).

J	Union error	Accepted errors / count	Risk UCB	Remaining cov.	Test risk
2	35.4	58 / 923	0.096	47.9	0.054
4	37.5	52 / 865	0.094	44.9	0.063
8	40.2	54 / 872	0.096	44.9	0.062

D Guarantee boundaries and numerical checks

Randomized mechanisms. The theorems condition on fixed functions g_j. If a mechanism draws a fresh seed U, either include a fixed, finite set of seeded mechanisms in the portfolio, or define the example as (X, U) and replay all policies under the same recorded seed distribution; certification and deployment must share that augmented distribution. This does not certify adaptive seed search.

Table 6: Ablations on Flowers-102, CLIP ViT-B/16 at $\alpha = 0.10$ with matched label budgets (seed 0; seed mean±sd over five seeds in parentheses). Omission and ranking-only masses partition the accepted-error rate. The local-gate row (per-mechanism calibrated local MLP) uses images accepted by at least one mechanism as its denominator and reports worst-router / uniform-router coverage; its envelope is in Table 4.

Variant	Union risk	Omission	Ranking-only	Coverage (seeds)
Union target, shared gate	0.063	0.027	0.036	44.9 (46.5±1.0)
Mean target, same network	0.051	0.034	0.017	43.9 (44.2±0.4)
No catalog anchor/rank	0.058	0.038	0.020	44.7 (43.8±1.0)
No agreement features	0.062	0.026	0.036	46.3 (46.4±0.9)
Feature MLP, union target	0.052	0.033	0.019	45.5 (47.7±1.6)
Logistic, union target	0.061	0.030	0.031	47.9 (48.8±1.1)
No likelihood-ratio feature	0.058	0.037	0.021	41.3 (41.2±0.4)
Worst margin, no learning	0.068	0.049	0.019	43.1
Synchronized local gates (T=1)	0.058	0.033	0.025	47.8
Local gates (test-only; accepted by any)	0.058	0.039	0.020	37.5 / 41.4

Table 7: Nested-closure test on Flowers-102, CLIP ViT-B/16: interior sets inside the registered intervals, their test error, their accepted risk under the unchanged SETGUARD-R DeepSets gate ($\alpha = 0.10$), and pointwise violations of Proposition 6 over all images.

Interior set	Error rate	Accepted risk	Pointwise violations
RN50-8	33.1	0.063	0
RN50-16	33.5	0.063	0
RN50-32	33.8	0.063	0
RN50-thinned	33.8	0.063	0
ViT-B-32-8	33.6	0.063	0
ViT-B-32-16	33.7	0.063	0
ViT-B-32-32	33.9	0.063	0
ViT-B-32-thinned	34.3	0.063	0
Endpoint union (certified)		0.063	

Table 8: Per-seed certified coverage (%) of the matched-feature learned gates on Flowers-102, CLIP ViT-B/16 at $\alpha = 0.10$ (finite-pool joint replay with Bonferroni correction), the seed mean±sd, and at $\alpha = 0.05$ the number of seeds that certify (mean coverage of those in parentheses). Seed 0 is the primary frozen scorer.

Gate	s0	s1	s2	s3	s4	mean±sd	passes at α=0.05
SETGUARD-R (DeepSets, union)	44.9	47.3	45.8	47.2	47.2	46.5±1.0	1/5 (16.5)
DeepSets (Zaheer et al., 2017), mean target	43.9	43.9	43.7	44.5	44.8	44.2±0.4	4/5 (20.4)
No catalog anchor/rank	44.7	44.4	43.2	42.0	44.4	43.8±1.0	5/5 (22.9)
No agreement features	46.3	47.3	45.6	45.4	47.5	46.4±0.9	1/5 (12.0)
Feature MLP, union	45.5	47.0	49.5	47.1	49.4	47.7±1.6	5/5 (20.2)
Logistic, union	47.9	50.0	49.4	47.1	49.7	48.8±1.1	5/5 (15.0)

Table 9: What the guarantee covers. The image–label law, or the declared finite evaluation pool, remains fixed. "Outside contract" is not a certified success even if empirical risk is small.

Change	Example	Guarantee
Constant mixing	New frequencies over the four endpoints	Yes
Image-dependent routing	Route by entropy, agreement, or learned image features	Yes
Interior candidate sets	Any set between registered inner/outer endpoints	Yes
Loss-aware evaluation oracle	Pick an erroneous mechanism when one exists	Upper envelope
Outside registered intervals	Outputs failing every registered set inclusion	No
Image/label shift	Corruptions, new domain, relabeling	No

Router changes versus sample changes. The guarantee is uniform over measurable routers for the same P_{XY}. It permits choosing a router after certification, but not selecting images based on certification errors, reusing certification images as test cases, or changing P_{XY}. Context variables such as a region filter can be part of X only if they are recorded, replayable, and drawn from the same joint distribution.

Finite-state verification. `verify_theory.py` enumerates finite deterministic problems and routers (all 729 two-mechanism/three-image patterns with exact rational arithmetic, 1,000 random weighted three-mechanism patterns, 224,000 nested-set inequalities, and exact binomial and hypergeometric coverage on parameter grids) to check the envelope, minimum coverage, synchronized-gate dominance, closure, and boundary cases of the confidence bounds. These checks complement the analytical proofs.

E SUPPLEMENTARY EMPIRICAL COMPARISONS

E.1 POPULATION-CERTIFICATE COMPARISON

The population Clopper–Pearson calculation for the primary DeepSets gate gives UCB 0.091, a coverage lower bound of 43.2%, and a realized test union risk of 0.063, with $\delta_R = \delta_C = 0.025$. Its interpretation requires the i.i.d. sampling assumption of the population theorem. This coverage lower bound is not an exact remaining-pool coverage and is not combined with the finite-pool results above.

E.2 ERROR STRUCTURE AND SEED SENSITIVITY

Table 10: Blind-pool portfolio statistics on Flowers-102, CLIP ViT-B/16 (%). Omission means that the true label is absent from the candidate set. The mean and union errors summarize the four mechanisms on the same records.

Mechanism	Error	Omission
RN50-4	32.6	18.9
RN50-64	34.1	0.3
ViT-B-32-4	34.4	19.0
ViT-B-32-64	34.2	0.2
Mean error	33.8	
Union error $\ell_\vee$	37.5	
All agree / agree and wrong	78.1	15.7
Full-catalog accuracy	65.8	

The mean error is 33.8% and the union error is 37.5%, while 15.7% of records receive the same wrong prediction from all four mechanisms. The mechanisms thus share many errors, and agreement alone does not establish correctness. This structure also explains why adding marginal error bounds can be substantially more conservative than directly certifying their union.

Table 8 reports the individual training-seed results. Training seeds vary initialization and minibatch order while the fitting/tuning split and certification draw remain fixed. These comparisons measure sensitivity to score fitting. For Flowers-102, the seed standard deviation is $[5^{-1}\sum_{s=0}^{4}(c_s - \bar{c})^2]^{1/2}$, where c_s is certified coverage for seed s. The confidence statements in Sections 4.2–4.3 concern randomness in the certification sample. Seed 0 is the primary scorer throughout.

E.3 COMPUTATIONAL COST

For the four-mechanism portfolio, replay requires one forward pass from each of two proposal encoders and one target-encoder pass per image. The top-4 and top-64 sets reuse each proposal ranking, and text embeddings are cached. The shared gate has 31 parameters (logistic), $21{,}485$ (MLP), or $21{,}569$ (DeepSets), in addition to the frozen encoders. Full-catalog scoring and proposal-encoder passes are part of the replay computation.

Supplementary implementation. The supplementary material provides the Flowers-102 training and evaluation pipeline, dataset manifest, feature-extraction metadata, per-run result summaries, and image-level certification routines. The routines implement binomial population bounds and hypergeometric remaining-pool bounds. Table 11 gives the corresponding strict-risk comparison.

Table 11: Strict-risk comparison on Flowers-102, CLIP ViT-B/16, seed 0, $\alpha = 0.05$, $\delta_R = 0.05$. All listed gates use joint replay. No passing threshold means zero coverage and undefined conditional risk ("–").

Shared gate	Risk UCB	Coverage (%)	Union risk
MSP / temperature-scaled MSP	–	0.0	–
MCM / energy / pooled LR-inspired	–	0.0	–
Worst margin / agreement controls	–	0.0	–
Synchronized local MLP ($T = 1$)	–	0.0	–
DeepSets, mean target	.050	20.0	.028
SETGUARD-R: DeepSets, union	–	0.0	–
SETGUARD-R: MLP, union	.047	19.4	.016
SETGUARD-R: logistic, union	.047	13.4	.007