ALLOCATING ADDITIONAL LABELS BETWEEN CASCADE PREDICTORS AND SELECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

A useful learned selector does not establish that the next labels should train the selector rather than the predictors. Controlled image-cascade experiments separate these questions. In an adapted CIFAR-100 long-tailed condition, class-conditioned routing improves on a strong scalar rule by 2.206 percentage points (pp), but expanding gate support contributes only 0.200pp. Allocating the identical additional labels to the predictors instead improves the source-selected cascade by 0.881pp; two additional paired training repetitions give 1.612pp. Tiny ImageNet preserves the investment advantage, including after SOURCE-selected answer correction. A complete endpoint-support $\times$ sampling-prior factorial identifies a 5.901pp gain from distinct examples at the expanded prior, rising to 7.088pp under a separately fixed fourfold training horizon. Prior-only and routing-interaction contrasts remain unresolved. Component controls connect the support effect to an executable alternative: source-selected upgraded cheap-only prediction exceeds the old Tiny cascade by 2.965pp. Corrected Tiny and stronger CIFAR predictors retain fixed-budget investment gains without crossing the corresponding old-pair hindsight ceilings. When corrected budgets are also SOURCE-selected, cheap-only prediction has lower resident request time, but its same-label accuracy advantage remains unresolved. Thus useful routing, the return to further gate labels, and the answers made available by predictor training are distinct quantities. Reported intervals condition on fitted models and exposed images; the comparisons do not establish a universal allocation rule or a lifecycle speedup.

1 INTRODUCTION

An image cascade first runs an inexpensive predictor and acquires an expensive prediction only on selected examples. Better routing can improve accuracy at the same acquisition count. A separate decision arises when additional labeled examples become available: should the selector receive those examples, or should the predictors improve first? A comparison against literal confidence cannot answer that investment question. Even a substantial class-conditioned gain can coexist with a small return to further selector supervision.

The distinction is concrete in the present positive control. In a long-tailed CIFAR-100 condition, a learned class policy exceeds a strong scalar rule by 2.206pp. Increasing honest selector support from 2k to 5k, however, adds only 0.200pp. Giving those same additional 3k examples to the predictors rather than the selector improves the actual source-selected system by 0.881pp. Thus a routing gain and a marginal investment gain have different baselines, interventions, and meanings.

Three linked comparisons locate the source and boundary of this result. First, matched cheap-error and relative-value learners distinguish predicting a cheap mistake from predicting a useful paid replacement. Second, class-budget replay asks whether the recorded advantage requires new within-class ordering. Third, an endpoint-support $\times$ sampling-prior factorial separates access to distinct labeled examples from altered training class frequencies. Every allocation arm receives the same opportunity to fit and select a policy on source data. Figure 1 maps the allocation and factorial controls; Appendix Table 5 records their conclusions and limits.

The contribution is a controlled comparison of where the same next labels help, not a new optimal-deferral theorem. The primary evidence first establishes useful routing, then compares label recipients, and finally separates distinct examples from sampling prior and training horizon. Component

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.

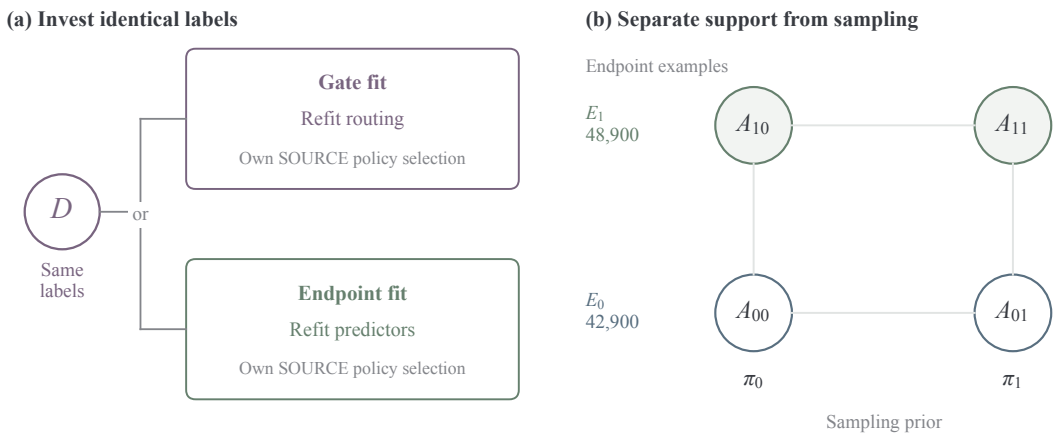


Figure 1: Alternative uses of the same label donation. (a) The whole donation D goes to either the gate or endpoints: 3,000 CIFAR or 6,000 Tiny labels. Both arms refit and select their own SOURCE policy; their final cascade accuracies compare label recipients. Endpoint-trained rows cannot supervise an honest gate for that endpoint. (b) The Tiny factorial crosses endpoint support and sampling prior while holding honest gate support at 4,000. The diagram is schematic. Equal labels do not imply equal compute.

and no-acquisition controls test a concrete consequence of that sequence. The stronger CIFAR comparison preserves the practical gain without reproducing every Tiny oracle witness. Compact-policy superiority over the strong simple control and routing compensation remain unresolved.

2 Related work and scope

Deferral and acquisition. Confidence sufficiency and comparative-risk learning are established questions in cascades (Jitkrittum et al., 2023); consistent learning to defer also predates the present experiments (Mozannar & Sontag, 2020). Selective classification studies abstention rather than a necessarily useful paid replacement (Geifman & El-Yaniv, 2017; 2019). Budgeted model selection and early-exit networks allocate computation across examples (Nan & Saligrama, 2017; Huang et al., 2018). Consistent two-stage objectives address joint predictor and selector learning (Regol et al., 2025). Dynamic feature acquisition studies which unobserved information to purchase, often sequentially (Covert et al., 2023; Gadgil et al., 2024; Norcliffe et al., 2025). The present study holds the candidate paid predictor fixed within a routing comparison and then changes the allocation of supervision between components. No new consistency guarantee or canonical reproduction of all cited algorithms is claimed.

Long-tail learning and calibration. Class-prior adjustment, balanced objectives, and representation/classifier decoupling provide strong alternatives to a learned routing explanation (Menon et al., 2021; Ren et al., 2020; Kang et al., 2020). Effective-sample weighting and margin-based objectives likewise distinguish imbalance from unweighted example count (Cui et al., 2019; Cao et al., 2019). Temperature calibration is a standard control (Guo et al., 2017); label-shift correction requires assumptions beyond observing class-frequency differences (Lipton et al., 2018). These approaches motivate controls rather than a claim that calibration universally explains routing. The factorial below changes actual sampling frequencies and distinct endpoint examples separately, without substituting a post-hoc score transformation for predictor training.

Targets, policies, and label investment. A matched target contrast fixes features, capacity, training recipe, and action constraint while changing E_f to V. A practical policy contrast compares complete source-selected procedures at fixed q. An investment contrast reallocates the same labeled identifiers and permits each resulting system to choose a policy. These contrasts separately test target information, policy performance, and label allocation.

Table 1: The strong comparison is a complete procedure, not a target name. Paid-blind ERROR fitting does not make the final system menu paid-blind. No actor observes true class, paid prediction, or correctness before acquisition. Matched target tests share capacity and row streams; practical menus deliberately compare different source-selected families.

Control	Cheap-time input	FIT supervision	SOURCE selection criterion
Literal confidence	Maximum confidence C	None	Fixed rule
Strong scalar	C; no class ID	Declared error/value targets	TUNE chooses scalar/bin member
ERROR-CHEAP	C, predicted class K	Cheap error E_f	Cheap-error TUNE Brier
VALUE-CLASS	Same C, K	Relative value $E_f - E_h$	Paid-informed TUNE utility
Compact menu	Frozen $\hat{e}$, K	Paid-risk offset or affine value	TUNE selects fixed small menu
Complete system	Declared candidate scores	Candidate-specific	TUNE utility; no TARGET selection

3 PROTOCOLS THAT PRESERVE THE COMPARISON

Let f denote the cheap predictor and h the paid predictor. The selector $g(O_0) \in \{0, 1\}$ observes cheap-time information O_0, not the paid prediction or the label Y. Define $E_f = \mathbf{1}\{f \neq Y\}$, $E_h = \mathbf{1}\{h \neq Y\}$, and relative value $V = E_f - E_h$. At acquisition fraction $q = \mathbb{E}[g]$ and constant acquisition fee c,

$$\mathrm{Acc}(g) = 1 - \mathbb{E}[E_f] + \mathbb{E}[gV], \qquad J(g) = \mathbb{E}[gV] - qc. \tag{1}$$

Same-endpoint, same-q utility differences equal accuracy differences. The paid action uses the raw paid argmax. Acquisition count is not GPU time, and a fixed symbolic fee does not establish a dollar cost.

The first image study uses adapted CIFAR-100 (Krizhevsky, 2009) with ResNet-8/56 predictors; the native 64×64 confirmation uses Tiny ImageNet-200 (Le & Yang, 2015) with ResNet-18/50 (He et al., 2016). The names identify concrete recipes, not state-of-the-art benchmark claims. CIFAR LT50 has 50 head classes with 390 endpoint examples each and 50 tail classes with 39 each, totaling 21,450. BALANCED has 39,000 examples; between-condition differences do not isolate imbalance alone. Nested gate banks contain 2k/5k examples, with separate 3k TUNE and 2k source-audit banks. Endpoint fitting and normalization exclude honest gate rows.

Matched CIFAR ERROR and VALUE heads share a ReLU MLP with 64- and 16-unit hidden layers, initialization, 800 Adam updates (Kingma & Ba, 2015), 100k presentations, learning rate $7 \cdot 10^{-4}$, and weight-only L2 penalty $.001$. Scalar input is maximum softmax confidence C; the class-aware input adds the predicted class K; the richer input also adds entropy and the top ten probabilities. The practical scalar menu includes learned scalar and bin controls, not merely literal confidence. Cheap calibration uses temperature $.5/1/2$ with or without source-prior adjustment and retains the original cheap answer.

Selection roles. Gate FIT trains scores. TUNE chooses declared families before evaluation. The ERROR family is selected using cheap-error Brier loss; the complete system menu uses paid-informed TUNE accuracy. Consequently, an ERROR head can be paid-blind during fitting and family selection without making the complete experiment paid-blind. Paid outcomes informed the study design and practical menu. Evaluation labels do not choose a head, hyperparameter, or investment arm.

Table 1 distinguishes class information from practical performance. A learned class head is compared with a learned scalar control; the paid-informed VALUE fit is compared with the complete source-selected policy menu. Endpoint-trained rows cannot supply honest gate targets for the same endpoint; the label-allocation arms enforce that restriction rather than compensate with a larger gate bank.

Training and repeated models. The Tiny factorial uses three paired training realizations, termed parents below. Within each architecture, cells share initialization, augmentation rules, optimizer, Batch Normalization batch size (Ioffe & Szegedy, 2015), and total presentations. Original training uses 9.78M presentations; the separately specified fourfold horizon uses 39.12M and proportionally scaled learning-rate milestones. Each fourfold corner is trained afresh. The horizon comparison therefore changes both optimization length and the corresponding fixed schedule, not just elapsed wall time. Additional two-parent allocation checks remain separate from the original three-parent factorial.

Uncertainty and exposure. Evaluation retains 9,988 CIFAR images and 9,993 Tiny validation images after the recorded exact-duplicate rules. Parents share those images. Paired per-image differences are first averaged across fitted parents; common image indices are then resampled with frozen actions. Reported intervals condition on these models, source choices, and cohort masks. No interval estimates training-seed-population variability or a freshly reranked deployment cohort. Parent-level points accompany key contrasts. The distinction between models and independent data sets is substantive (Demšar, 2006). Previously exposed benchmark follow-ups also differ from new test populations (Recht et al., 2018; 2019; Liu & He, 2025). No retrospective equivalence margin is introduced.

4 Useful routing: targets, class budgets, and expert specificity

The adapted CIFAR LT50 class policy gains 2.206pp over a strong scalar control, with adjusted interval $[1.876, 2.550]$pp. The 5k $-$ 2k support increment is only 0.200pp, $[0.027, 0.377]$pp. Rich sorted posterior input adds -0.007pp over the class-aware input, $[-0.130, 0.114]$pp. Thus useful class access is established in this condition, but neither a large marginal gate-label return nor a further rich-input gain follows.

Relative value changes the matched learner. At $q = .5$ and a fixed five-percent addition/removal budget, VALUE exceeds ERROR by 0.400pp in LT50 and 0.354pp in BALANCED. Family-adjusted intervals are $[0.214, 0.591]$ and $[0.080, 0.624]$pp. The latter parent points are $-0.130/0.551/0.641$pp: a positive average is not uniform improvement. The LT50 $-$ BALANCED interaction remains unresolved. The practical deployment menu gives a different comparison. LT50 cheap calibration averages 33.040% accuracy, compared with 32.943% for selected VALUE and 32.980% for the complete selected menu. The target effect does not establish general VALUE superiority.

Replay holds class budgets fixed. Starting from the same scalar reference mask, record VALUE's additions and removals within each predicted class. A replay adds the highest frozen ERROR scores and removes the lowest scores within those exact class budgets, with stable identifier ties. The resulting mask preserves total acquisitions, changes, and per-class acquisition counts. Exactly,

$$J(g_V) - J(b) = \{J(g_Q) - J(b)\} + \{J(g_V) - J(g_Q)\}. \tag{2}$$

At 5k support, VALUE and replay masks coincide in all six regime–parent cells. The within-class remainder is exactly zero on these actions on the exposed images, not merely nonsignificant. The replay copies VALUE's quotas, however, and is not a free deployable replacement for training VALUE. The first term is not a causal class-mediated fraction. A 2k BALANCED parent has 68 disagreements, excluding universal class-only equivalence.

A gain decomposition is an accounting check, not an explanation. For $d = g_V - g_E$ and frozen cheap-error predictor $\hat{e}$, the empirical accuracy difference $\widehat{\Delta}_{VE} = \widehat{\mathrm{Acc}}(g_V) - \widehat{\mathrm{Acc}}(g_E)$ satisfies

$$\widehat{\Delta}_{VE} = \frac{1}{n}\sum_i d_i \hat{e}_i + \frac{1}{n}\sum_i d_i (E_{f,i} - \hat{e}_i) - \frac{1}{n}\sum_i d_i E_{h,i}. \tag{3}$$

The terms record score allocation, cheap-error residual, and paid errors. If g_E is the exact top-q rule for $\hat{e}$, the first term cannot be positive; the fixed-change procedure need not satisfy that premise. The residual mixes estimation error and realized-label variation, not a certified amount of new information. This limitation motivates the explicit quota and crossed-expert controls instead of interpreting a positive target contrast as proof of expert-specific information.

Expert specificity remains unresolved. Matched and crossed gates for two compatible paid experts isolate whether the learned actions identify the particular expert's advantages. The paired contrast is $\frac{1}{2}\mathbb{E}[(g_A - g_B)(V_A - V_B)]$. VALUE gives 0.035pp, $[-0.005, 0.075]$pp; the shared ERROR gate gives an exactly zero contrast. Three cyclic pairings reuse experts and do not constitute three independent expert populations. Relative-value target gains therefore coexist with unresolved expert-specific gains.

Primary and secondary acquisition budgets. Specified secondary $q = .25/.75$ contrasts transfer the $.5$ source-selected scores without budget-specific fitting or reselection. The $.5$ contrast remains primary; Tiny's additional-group primary remains adverse. Appendix D gives the budget-specific values and replay conventions.

5 THE SAME ADDITIONAL LABELS FAVOR PREDICTOR INVESTMENT

Let the nested honest gate banks satisfy $G_2 \subset G_5$, and define $D = G_5 \setminus G_2$, $|D| = 3,000$. Gate investment gives D to the selector. Endpoint investment gives those exact identifiers to both predictors while retaining only G_2 as honest selector support. Each resulting system regenerates source payoffs and receives the same policy-selection menu. A stale gate is not imposed on improved endpoints. TUNE selects within each investment arm; no learned rule chooses the investment itself before training.

The CIFAR Base/Gate/Endpoint mean deployed accuracies are $33.040/32.980/33.861\%$. Hence the endpoint $-$ gate difference is 0.881pp, $[0.384, 1.382]$pp. Original parent differences are $-0.090/2.313/0.421$pp. Two additional paired repetitions give $0.681/2.543$pp, averaging 1.612pp; those repetitions do not replace the original primary. The first original parent is adverse to both investments relative to Base. Equal labels and equal predictor presentations do not guarantee a gain for every training realization.

The Tiny allocation uses 42,900 endpoint examples, honest gate support of 4,000, and a donation of 6,000. Base retains the smaller banks; Gate expands honest selector support to 10k; Endpoint expands predictor support to 48,900 and retains gate support at 4k. Source TUNE and audit remain common. At $q = .5$, endpoint investment exceeds gate investment by 6.268pp in the original three parents and 6.199pp in two additional repetitions. Useful CLASS-RISK routing also transports, with class-over-strong-scalar gains of $1.845/1.591$pp. Primary fixed-change VALUE $-$ ERROR does not transport: -0.143pp is unresolved in the original group and -0.260pp is adverse in the additional group.

These allocations reuse benchmark labels. Giving both Tiny arms the same SOURCE-selected correction $p'_k \propto p_k \pi_k^{-\tau}$, $\tau \in \{0, .5, 1\}$, reduces the original same-label gain from 6.268 to 4.270pp, $[3.759, 4.797]$pp. All nine arm–parent choices select $\tau = 1$, the menu boundary. The same-prior gain is 4.476pp, $[3.943, 5.004]$pp (Appendix H). The investment ranking survives this shared repair; the original factorial below separates support and sampling prior.

A scoped consequence: available answers and deployable alternatives. The Tiny old/new comparison fixes the expanded prior: A_{01} versus A_{11}, not original-prior A_{00}. For a fixed pair, let B count cheap-correct images, C cheap errors corrected by the paid model, H paid errors on cheap-correct images, and W joint errors. With n images, integer $0 \le k \le n$, and $[x]_+ = \max(x, 0)$, the label-hindsight accuracy ceilings for exactly k acquisitions are

$$U_{\text{raw}} = \{B + \min(k, C) - [k - (n - H)]_+\}/n, \quad U_{\text{keep}} = \{B + \min(k, C)\}/n \le 1 - W/n. \tag{4}$$

The second bound permits retaining the cheap answer after acquisition; neither bound covers a third-label predictor or retraining. Tiny's improved cascade exceeds the old pair's union ceiling in all three original parents at both horizons (Figure 2). Upgraded cheap-only exceeds that ceiling at $4\times$, but not at $1\times$. After SOURCE-selected correction at $1\times$, both upgraded cheap-only and cascade accuracies lie below every corrected gate-investment and A_{01} ceiling. The original witnesses concern uncorrected pairs; the corrected investment advantage compares learned procedures, not all hindsight routing rules (Appendix H).

Each old/new pair receives the same SOURCE menu $q \in \{0, .25, .5\}$: choose the smallest acquisition fraction reaching the old $.5$ SOURCE accuracy, with no TARGET fallback. Tiny selects the

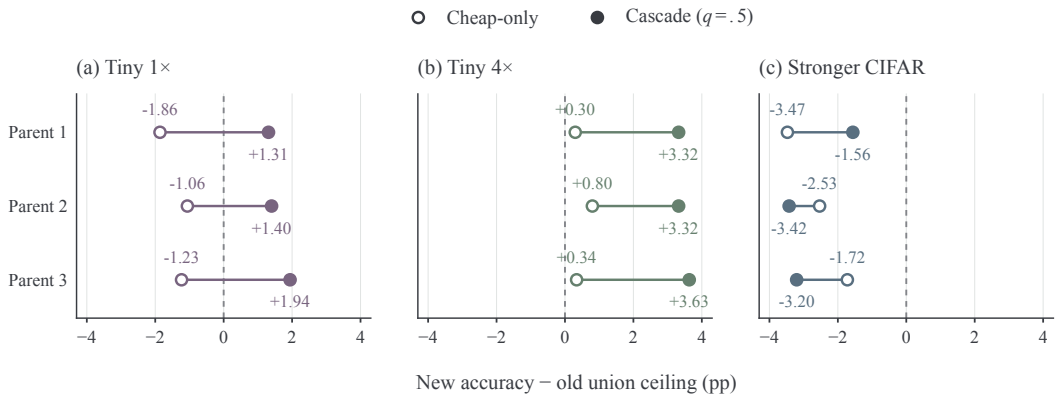


Figure 2: Crossing an old pair's routing ceiling depends on the predictor pair. Each panel has three fitted-parent rows; links join new cheap-only (open) and cascade (filled) accuracies − the original uncorrected pair's union ceiling, in pp. Zero marks that ceiling. Parent indices are condition-local; Tiny horizons use matching parents. Tiny $1\times$ /$4\times$ share 9,993 evaluation images; stronger CIFAR uses 9,988. Images are not training repetitions. No means or confidence intervals are shown. The old union uses label hindsight, not a deployable policy, and does not bound third-label prediction or retraining.

Table 2: Follow-up controls preserve a practical investment gain, not a universal routing limit. Differences are pp with fitted-model-conditional paired-image intervals. Tiny budget selection uses the same SOURCE-selected menu for A_{01}/A_{11}; CIFAR uses a stronger paid predictor with a fixed learning rate in both allocation arms. Component contrasts share gate support; allocation contrasts instead match the total unique-label union. These are separate estimands.

Environment	Contrast	Gain	Conditional interval
Tiny $1\times$	New − old, selected budget	$+2.965$	$[2.478, 3.469]$
Tiny $4\times$	New − old, selected budget	$+4.143$	$[3.672, 4.627]$
CIFAR ResNet-56	Endpoint − gate, same labels	$+3.094$	$[2.503, 3.676]$
CIFAR ResNet-56	Cheap(0) − old gate$(.5)$	$+3.247$	$[2.655, 3.835]$

upgraded cheap-only model at both horizons; actual execution verifies the gains in Table 2. At fixed $q = .5$, updating both components still beats updating only cheap or paid, but procedure interactions are negative, not neural-network synergy. A frozen-body classifier support contrast also gains 7.185pp, $[6.663, 7.722]$pp, at $q = 0$; full representation retraining is not necessary for every observed support benefit.

The CIFAR check fixes the ResNet-56 initial learning rate at .1 for old/new across three paired parents. The practical investment effect survives, but both upgraded cheap-only and the new cascade remain below the stronger old pair's union ceiling. Thus a gain over a deployable policy transports; an impossibility claim about all old-pair routing does not. Appendix G retains component, donation-dose, and execution details without replacing the original primaries.

6 Distinct support survives a sampling-prior control

Denote endpoint support by E_0 or $E_1 = E_0 \cup D$, and training class prior by π_0 or π_1. The original diagonal systems are $A_{00} = (E_0, \pi_0)$ and $A_{11} = (E_1, \pi_1)$. The completed factorial adds $A_{01} = (E_0, \pi_1)$ and $A_{10} = (E_1, \pi_0)$, with each cell training both predictors and selecting a separate policy on source validation data. Gate investment A_G retains E_0, π_0 and expands honest gate data instead (Table 3).

There are 100 head and 100 tail classes. Original endpoint counts are 390 and 39 per class; expanded counts are 420 and 69. Same-prior corners receive byte-identical class-draw streams, matched initialization and augmentation rules, and the same total presentations. Within-class image identities

Table 3: The Tiny factorial separates endpoint identifiers from sampling frequencies. The donation contains 30 examples per class. Identical class streams within a prior make $A_{11} - A_{01}$ and $A_{10} - A_{00}$ distinct-support contrasts, not extra-draw contrasts.

Cell	Endpoint examples	Gate examples	Tail draw mass	Interpretation
A_{00}	42,900	4,000	9.091%	Original support and prior
A_{01}	42,900	4,000	14.110%	Prior change only
A_{10}	48,900	4,000	9.091%	Support change at original prior
A_{11}	48,900	4,000	14.110%	Expanded support and prior
A_G	42,900	10,000	9.091%	Same donation to honest gate

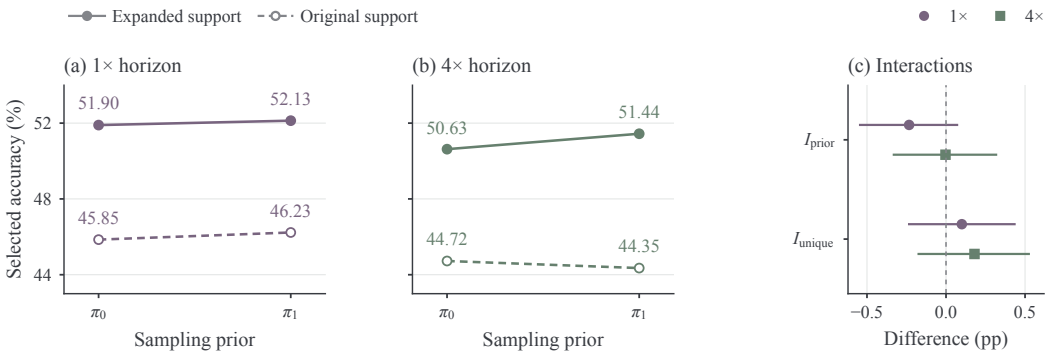


Figure 3: Distinct examples help while routing compensation remains unresolved. Left and center: all eight three-parent finite-bank cascade-accuracy means of the support–prior factorial at $q = .5$, separated by training horizon. Solid/filled and dashed/open series denote expanded and original support. Support/prior contrast intervals are reported in the text. Here S_{ij} is the cheap-error-selected class policy's accuracy advantage over the strong scalar menu in cell ij. Right: $I_{\text{prior}} = S_{01} - S_{00}$ and $I_{\text{unique}} = S_{11} - S_{01}$ retain nominal 97.5% individual paired-image percentile-bootstrap intervals in separate two-contrast families per horizon, conditional on three fitted parents and frozen masks; all four cross zero. Longer training uses an extended schedule.

differ by design. Gate and evaluation rows remain outside endpoint fitting and normalization. Donation rows trained by an endpoint are never counted as honest gate data for that endpoint.

The primary $U_{\pi_1} = A_{11} - A_{01}$ measures distinct-support benefit under the expanded prior. Secondary contrasts are $U_{\pi_0} = A_{10} - A_{00}$ and $R_{E_0} = A_{01} - A_{00}$. At each horizon, paired-image percentile-bootstrap intervals have nominal individual levels of 95% for the primary and 98.75% for these secondaries. At the original horizon, $U_{\pi_1} = 5.901$pp, $[5.410, 6.384]$pp, and $U_{\pi_0} = 6.044$pp, $[5.394, 6.696]$pp. The prior-only estimate is 0.377pp, $[-0.127, 0.887]$pp. The endpoint advantage is therefore not explained by changing class sampling alone.

Longer training does not remove the support effect. Under the fixed fourfold horizon, $U_{\pi_1} = 7.088$pp, $[6.601, 7.575]$pp, and $U_{\pi_0} = 5.901$pp, $[5.287, 6.536]$pp. The former increases by 1.187pp relative to the original horizon, with four-contrast-adjusted interval $[0.484, 1.900]$pp. Prior-only $R_{E_0} = -0.374$pp remains unresolved, $[-0.851, 0.118]$pp. The fourfold endpoint − gate comparison remains positive at 6.881pp, $[6.388, 7.389]$pp. A short optimization horizon alone therefore does not account for the observed investment ranking. This comparison does not prove optimal convergence or independence from all training recipes.

A larger support gap is not a better absolute model. Fourfold training lowers selected-system accuracy in every corner (Figure 3, Table 4). The larger π_1 support contrast is not a general training benefit. Expanded support improves tail accuracy at both horizons; prior changes alone do not reproduce that effect. SOURCE-defined head/tail groups are evaluator summaries, not a class-level causal explanation.

Table 4: Absolute Tiny accuracy (%), averaged across the original three parents. Selected uses the actual source-selected policy at $q = .5$; Head/Tail partition that selected system by the specified true-class groups. Cheap and Paid evaluate each endpoint on the whole cohort. Longer training does not improve these absolute means.

Horizon	Cell	Cheap	Paid	Selected	Head	Tail
1×	A_{00}	42.96	43.52	45.85	70.18	21.53
	A_{01}	43.56	43.70	46.23	69.89	22.57
	A_{10}	48.69	49.96	51.90	69.66	34.13
	A_{11}	49.19	49.59	52.13	70.44	33.82
4×	A_{00}	42.23	42.36	44.72	71.44	18.01
	A_{01}	42.07	42.00	44.35	70.78	17.92
	A_{10}	47.26	48.45	50.63	72.04	29.21
	A_{11}	48.49	48.88	51.44	71.65	31.23

Selected accuracy exceeds either endpoint in these cells: investment improves a useful cascade, with paid corrections and harms retained in paired records.

Useful routing is not established compensation. Define S_{ij} as the cheap-error-selected class policy's advantage over the strong scalar menu in cell ij. The routing interactions are $I_{\text{prior}} = S_{01} - S_{00}$ and $I_{\text{unique}} = S_{11} - S_{01}$. At 1×, the estimates are -0.233pp, $[-0.544, 0.070]$pp, and 0.100pp, $[-0.233, 0.434]$pp. At 4×, the corresponding estimates are -0.003pp, $[-0.330, 0.317]$pp, and 0.180pp, $[-0.173, 0.524]$pp. Neither horizon establishes a reduction in routing value after fixing the prior or expanding support. CLASS-RISK and VALUE alternatives remain separately labeled secondary analyses. These secondary interactions address separate policy targets from the specified cheap-error-class contrast.

For each paired parent, the allocation identity $A_{11} - A_G = U_{\pi_1} + R_{E_0} - (A_G - A_{00})$ holds exactly. Each term has its own source-selected policy, so this accounting is not causal mediation. Distinct-support evidence strengthens the diagonal comparison without supporting a general theory of routing as imbalance compensation.

Two interactions answer different questions. The endpoint–prior interaction is $U_{\pi_1} - U_{\pi_0}$, not the class-over-scalar routing interaction. The original-horizon estimate is -0.143pp, $[-0.861, 0.595]$pp; the fourfold estimate is 1.187pp, $[0.465, 1.895]$pp in the declared secondary family. Expanded support can have a prior-dependent benefit without establishing routing compensation. Horizon changes in I_{prior} and I_{unique} remain unresolved at 0.230pp and 0.080pp, with adjusted intervals $[-0.275, 0.731]$ and $[-0.440, 0.587]$pp.

7 A COMPACT EXPLANATION STILL NEEDS A PRACTICAL COMPARISON

Quota replay does not imply constant within-class paid risk. Positive classwise affine transforms preserve within-class cheap-risk ordering but can change class budgets. Compact scores therefore fit honest gate support without copying evaluation quotas. For frozen cheap-error score $\hat{e}$, the menu contains $M_0 = \hat{e}$, a shrunk paid-risk offset $M_1 = \hat{e} - \hat{r}_K$, and $M_2 = a_K\hat{e} + b_K$ constrained to $a_K \geq 0$. The offset uses honest gate examples to estimate

$$\hat{r}_k = \frac{\sum_{i:K_i=k} E_{h,i} + \alpha\bar{r}}{n_k + \alpha}, \qquad \alpha \in \{10, 50, 200\}, \tag{5}$$

where $\bar{r}$ is pooled paid risk on the same gate bank. The affine menu uses ridge values $.001, .01, .1$. Missing classes back off to pooled parameters. Gate FIT estimates coefficients; TUNE separately selects the compact member and the strong-simple alternative. Evaluation chooses neither coefficients nor menu winners. Fixed shrinkage and the fixed $.01$ affine contrast remain different estimands from selection among compact candidates.

Global positive-affine transforms preserve ranking; classwise transforms can change class budgets. Neither statement automatically applies to multiclass temperature scaling. SOURCE fitting tests the compact explanation without copying evaluation quotas.

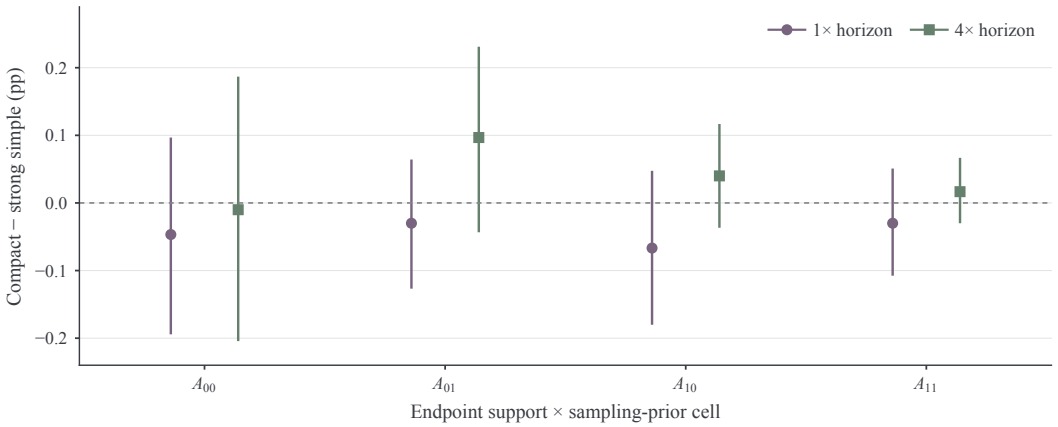


Figure 4: Actual compact-policy execution does not establish superiority over a strong simple alternative. Eight contrasts average three paired parents with separate source fitting/selection. Markers distinguish $1\times$ / $4\times$ horizons; lines show nominal 99.375% paired-image percentile-bootstrap intervals with a Bonferroni 95% family target across both horizons (Appendix E). All intervals contain zero without establishing equivalence. Execution includes cheap inference, source-frozen scoring/ranking, and paid inference on exactly 4,996 of 9,993 images.

The original compact-selected − strong-simple comparisons are -0.053pp in CIFAR LT50, $[-0.350, 0.227]$pp, and -0.020pp in Tiny, $[-0.090, 0.050]$pp. Secondary fixed-affine gains over offsets do not establish superiority over the source-selected simple menu. Gate-FIT cheap-error scores are in-sample, so compact fits inherit that limitation.

The 24 compact fits cross two horizons, three parents, and four corners. Compact, strong-simple, and relative-value policies yield 72 evaluations. Frozen acquisition masks, actual answers, and integer correct counts agree with independent replay. The eight compact − simple mean differences range from -0.067 to $+0.097$pp; all family-adjusted intervals contain zero (Figure 4).

Prediction identity and measured quality remain separate. An earlier Tiny additional-group check found BF16 selected-row predictions differing from cached full-cohort predictions on one image under two policies across four repetitions. All eight discrepancies preserved accuracy and executed endpoint − gate gaps $5.694/6.705$pp with models, choices, and masks fixed. This identity failure is separate from the 72-policy factorial validation.

Corrected policies provide a same-execution quality–time comparison (Appendix J). SOURCE selection chooses new cheap-only prediction against the gate-investment arm at $q = .5$. Same-label accuracy changes by $+.557$pp, with 97.5% interval $[-.043, 1.201]$, while reference/new resident-time ratios are 2.64–2.74. The same-prior accuracy contrast is $+1.328$pp, $[.735, 1.925]$. Same-label superiority, equivalence and noninferiority remain unestablished; these timings exclude training and loading. Earlier uncorrected-policy timings are distinct (Appendix I).

8 Conclusion

Investment rankings depend on data, training, and policy opportunities. Wider-backbone controls (Zagoruyko & Komodakis, 2016) favor endpoints, but image-level VALUE loses to class selectors. The 2k/4k same-label gains $.804/1.815$pp survive SOURCE budget selection; a 4k prior-only parent fails admission, precluding a three-parent mean. Exposed cohorts, three-parent factorials, and incomplete lifecycle costs limit scope.

Same-prior and frozen-body gains separate support from sampling and body updates. Answer correction preserves fixed-budget investment gains but removes the corrected pairs' original oracle witness. Selecting corrected acquisition budgets leaves same-label quality unresolved despite lower resident request time. Routing compensation and compact superiority remain unresolved; useful routing alone cannot determine next-label investment.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, mathematical formulation and proof checking, experimental design and implementation, analysis of routing results, literature search and reference verification, figure preparation, and manuscript drafting and editing. The authors are responsible for the final content.

ETHICS STATEMENT

The experiments reallocate existing labels in public image benchmarks without newly recruited annotation work. Constructed class imbalance is not evidence about fairness across protected human groups. A fixed acquisition fraction does not establish monetary or environmental savings without lifecycle measurement. Reused evaluation data and retained adverse results constrain the scope of deployment claims.

REPRODUCIBILITY STATEMENT

Appendices B and C specify label roles, exposure, and source-policy construction. Appendix D gives paired contrasts and multiplicity conventions; Appendix E records compact-policy execution checks, and Appendix F separates cost categories and retained adverse evidence. Appendices H and I give answer-correction settings, pair ceilings, and request timing. Appendix J links corrected policy quality to resident request time.

REFERENCES

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.07413.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447, 2023. URL https://proceedings.mlr.press/v202/covert23a.html.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1901.05555.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.03301.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://arxiv.org/abs/1512.03385.

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1703.09844.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f09e1ee5035a4c3fe38a5681cae5815-Abstract-Conference.html.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.09217.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Technical report, Stanford University, CS231n, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2403.08632.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.07314.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Feng Nan and Venkatesh Saligrama. Dynamic model selection for prediction under a budget. arXiv:1704.07505, 2017. URL https://arxiv.org/abs/1704.07505.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela van der Schaar, and Pietro Liò. Stochastic encodings for active feature acquisition. In *International Conference on Machine Learning*, pp. 46784–46814, 2025. URL https://proceedings.mlr.press/v267/norcliffe25a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? arXiv:1806.00451, 2018. URL https://arxiv.org/abs/1806.00451.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Florence Regol, Joseph Cotnareanu, Theodore Glavas, and Mark Coates. Is the acquisition worth the cost? surrogate losses for consistent two-stage classifiers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/085ea366002345cab8a1bf0f0ad1b210-Abstract-Conference.html.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/2ba61cc3a8f44143e1f2f13b2b729ab3-Abstract.html.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016. URL https://arxiv.org/abs/1605.07146.

Table 5: Questions, controls, and limits. Accuracy differences are percentage points (pp). The support/prior factorial uses the original three Tiny parents; additional training repetitions remain separate.

Question	Matched comparison	Result	Supported interpretation
Useful routing?	Class versus strong scalar, same acquisition/change budget	CIFAR LT50: $+2.206$pp	A real local routing gain
What changes?	VALUE actions versus own class quotas with cheap-risk order	Exact 5k replay; 2k counterexample	Recorded class allocation matters; no universal equivalence
Where do labels help?	Same donation to selector or predictors; own policy selection	Endpoint: $+0.881$pp CIFAR, $+6.268$pp Tiny	Predictor-favored marginal investment in these settings
Support or prior?	Tiny $A_{11} - A_{01}$; prior held fixed	$+5.901$pp at $1\times$; $+7.088$pp at $4\times$	Distinct-support effect survives longer training
Routing compensation?	Changes in class-over-scalar gain across factorial cells	Prior and support interactions unresolved	Compensation explanation not established
Small deployable substitute?	Source-selected compact versus strong simple; exact serving	Eight adjusted intervals contain zero	No established compact-policy superiority

A Questions, controls, and interpretation

B Experimental roles, exposure, and label accounting

Roles are component-specific. Endpoint FIT trains image predictors and defines normalization statistics. Gate FIT supplies labels and endpoint outputs for routing heads. TUNE selects predefined policy families. Source audit is not used for fitting or policy selection. TARGET denotes the final fixed image cohort. An example becomes ineligible as honest gate data after endpoint training uses that example, even if the example previously belonged to a gate bank. These roles specify which labels may influence each trained component.

Table 6: Unique labeled role counts. The donation is shared across alternative allocation procedures, never counted twice within an honest gate arm. Endpoint use by two predictors does not double the unique-label count.

Dataset	Allocation	Endpoint FIT	Gate FIT	TUNE	Source audit
CIFAR LT50	Base	21,450	2,000	3,000	2,000
	Gate	21,450	5,000	3,000	2,000
	Endpoint	24,450	2,000	3,000	2,000
Tiny	Base / A_{00}	42,900	4,000	6,000	4,000
	Gate / A_G	42,900	10,000	6,000	4,000
	Endpoint / A_{11}	48,900	4,000	6,000	4,000
	Prior-only / A_{01}	42,900	4,000	6,000	4,000
	Support-only / A_{10}	48,900	4,000	6,000	4,000

The CIFAR donation contains 3,000 identifiers; the Tiny donation contains 6,000, exactly 30 per class. The Tiny endpoint base has 100 classes with 390 examples and 100 with 39 examples. Expanded support has 420 and 69 examples per class. Gate support contains 20 examples per class before donation and 50 after donation. Therefore $E_1 \cup G_4$ and $E_0 \cup G_{10}$ have the same unique-label union. Equal union size does not equalize training, paid source-prediction generation, selector search, or model storage.

Data exposure is preserved, not reset. Both image studies reuse public benchmark data. The retained CIFAR evaluation has 9,988 rows; the Tiny evaluation has 9,993 unique rows without exact

training-pixel overlap under the recorded checks. Near-duplicate independence has not been established. Later factorial, longer-horizon, and compact comparisons are exposed-cohort follow-ups. New training initializations do not create new evaluation populations. The experiments establish reproducible differences under the specified cohorts, not an untouched benchmark confirmation.

Predictor and actor information. At deployment, the actor receives the cheap probability vector or its declared summary, predicted class, and frozen score parameters. True class labels and paid predictions are unavailable before acquisition. Class priors used for source correction come from the training design, not from evaluation-label counts. SOURCE-defined true head/tail partitions are permitted for evaluator-only summaries; actors use predicted classes. Every investment arm regenerates source payoffs using its own endpoints.

Reference settings and deviations. ResNet architecture names follow He et al. (2016); the small CIFAR and Tiny recipes at native 64-pixel resolution are concrete adaptations. The wider-backbone boundary uses the family of Zagoruyko & Komodakis (2016). Comparisons with confidence calibration, class-prior adjustment, and cheap-risk heads are explicit controls, not claims of reproducing the full training pipelines of Guo et al. (2017); Menon et al. (2021); Ren et al. (2020); Kang et al. (2020).

C TRAINING AND SOURCE POLICY CONSTRUCTION

Endpoint matching. CIFAR paired endpoint comparisons use 5,491,200 presentations per predictor within the matched recipe. Original Tiny training consumes 9,780,000 image–label presentations over 76,407 optimizer updates: 76,406 batches of 128 and a final batch of 32. The final batch is neither padded nor dropped. Repeated presentations count repeated uses of the prescribed examples; the eight-row preparation pass is excluded. The 256-update continuation is included in these totals, not added to them. The fourfold experiment uses exactly 39,120,000 presentations and 305,625 updates. The model family, initialization rule, augmentation, SGD recipe, and source-only normalization are unchanged between factorial cells. BF16 computation retains FP32 master parameters and the specified Batch Normalization behavior; TF32 is disabled. Learning-rate milestones scale with the fourfold horizon. Each model at the fourfold horizon is trained separately, rather than continued from an original model selected using evaluation results.

For A_{01} and A_{11}, the same class sequence follows π_1; for A_{00} and A_{10}, the sequence follows π_0. Within-class draws use each cell's allowed example bank. Checks of the sampled sequence, class counts, unique examples, and repetition counts verify the realized sampling. Sharing class sequences does not mean sharing image identities after support expansion. The original factorial reuses the completed original diagonal models and trains twelve missing-corner predictors. The fourfold factorial trains all twenty-four predictors fresh; reused original fits are not counted as fourfold evidence.

Head fitting details. Table 1 gives the information, supervision, and selection contract for each main control. CIFAR heads use the same ReLU MLP with 64- and 16-unit hidden layers and one output, identical initialization and row streams, squared-error fitting, and Adam at $7 \cdot 10^{-4}$ for 800 updates/100k presentations. Scalar bin candidates use 8/16/32 bins. Capacity matching is a target comparison within this family, not a claim that every conceivable risk estimator has been exhausted. At $q = .5$, the recorded VALUE policies in both CIFAR conditions and all three parents add 499 acquisitions and remove 499 among 9,988 evaluation images. Both the reference and candidate acquire 4,994 images. These counts specify the realized five-percent change budget relative to the full evaluation cohort.

Tiny source selection retains the same declared family opportunity for each corner: literal confidence, learned scalar, cheap-error class, class-risk bins, relative-value class, and richer value scores. The mechanism primary uses ERROR-CHEAP selected by cheap-risk source loss; CLASS-RISK and VALUE are reported separately. Source selection of a complete system can use paid source outcomes even when a particular ERROR head does not.

Compact-policy fitting scope. A zero affine slope creates ties rather than preserving within-class order; the fixed identifier rule resolves them. Equation 5 and the main text specify the full small

menu, source roles, and ranking controls. In-sample ERROR predictions on gate FIT preclude an out-of-fold or unbiased plug-in interpretation. Frozen SOURCE selection records retain all candidate scores and choices, including adverse members, so the final comparison can be reproduced without evaluation selection.

D PAIRED RESULTS AND MULTIPLICITY

For a realized cohort, expectations in Equation 1 are empirical averages and the fee uses the realized fraction k/n. The nominal Tiny $q = .5$ selects $k = 4,996$ of $n = 9,993$ images, rather than exactly half. Equal-count contrasts cancel the fee even with this rounding. Fixed-budget ranking may use the whole cohort's cheap outputs, but no labels or unacquired paid outputs. All entries below are percentage points. "Parent" means a paired training realization, not an independently sampled test population. Intervals resample common image indices after averaging the paired outcome difference across parents. Top-q masks remain fixed during resampling. These intervals describe fixed-action evaluation uncertainty, not threshold transport, new-cohort ranking, or training-seed uncertainty.

Table 7: Paired observations retained separately. Cyclic expert pairs share paid models. Original and additional groups are not pooled into a new confirmatory result.

Dataset / contrast	Group	Mean	Parent points
CIFAR Endpoint−Gate	Original 3	$+0.881$	$-0.090/2.313/0.421$
	Additional 2	$+1.612$	$0.681/2.543$
Tiny Endpoint−Gate	Original 3	$+6.268$	$6.074/6.354/6.374$
	Additional 2	$+6.199$	$5.694/6.705$
CIFAR VALUE−ERROR, LT50	Original 3	$+0.400$	$0.240/0.461/0.501$
CIFAR VALUE−ERROR, BALANCED	Original 3	$+0.354$	$-0.130/0.551/0.641$
CIFAR matched−crossed expert	3 cyclic pairs	$+0.035$	$0.090/0.075/-0.060$

Table 8: Tiny factorial at $q = .5$. At each horizon, U_{π_1} is the support primary (family one, nominal 95% individual interval). U_{π_0}, R_{E_0}, endpoint–prior interaction, and $A_{01} - A_G$ form a secondary family of four (98.75% individual intervals). I_{prior} and I_{unique} form a separate family of two (97.5%). The separately specified $A_{11} - A_G$ allocation comparison uses 95%. All use 6,000 paired-image percentile-bootstrap resamples; these families apply separately within each horizon.

Horizon	Contrast	Mean	Adjusted interval	Three parent points
1×	U_{π_1}	+5.901	[+5.410, +6.384]	+5.574 / +6.064 / +6.064
1×	U_{π_0}	+6.044	[+5.394, +6.696]	+5.704 / +6.034 / +6.394
1×	R_{E_0}	+0.377	[-0.127, +0.887]	+0.600 / +0.240 / +0.290
1×	Endpoint–prior interaction	-0.143	[-0.861, +0.595]	-0.130 / +0.030 / -0.330
1×	$A_{01} - A_G$	+0.367	[-0.153, +0.879]	+0.500 / +0.290 / +0.310
1×	$A_{11} - A_G$	+6.268	[+5.767, +6.768]	+6.074 / +6.354 / +6.374
1×	I_{prior}	-0.233	[-0.544, +0.070]	+0.170 / -0.140 / -0.731
1×	I_{unique}	+0.100	[-0.233, +0.434]	+0.030 / +0.060 / +0.210
4×	U_{π_1}	+7.088	[+6.601, +7.575]	+6.925 / +6.965 / +7.375
4×	U_{π_0}	+5.901	[+5.287, +6.536]	+5.314 / +6.515 / +5.874
4×	R_{E_0}	-0.374	[-0.851, +0.118]	-0.931 / -0.120 / -0.070
4×	Endpoint–prior interaction	+1.187	[+0.465, +1.895]	+1.611 / +0.450 / +1.501
4×	$A_{01} - A_G$	-0.207	[-0.694, +0.279]	-0.851 / +0.520 / -0.290
4×	$A_{11} - A_G$	+6.881	[+6.388, +7.389]	+6.074 / +7.485 / +7.085
4×	I_{prior}	-0.003	[-0.330, +0.317]	-0.120 / +0.000 / +0.110
4×	I_{unique}	+0.180	[-0.173, +0.524]	+0.440 / +0.070 / +0.030

For these $q = .5$ contrasts, each image's paired difference is averaged across the three fixed parents before resampling image indices with replacement. Equal-tailed percentile intervals use 6,000 resamples (random seed 923231), with lower and upper quantiles $0.025/m$ and $1 - 0.025/m$ for family size m. Quantiles use NumPy's default linear interpolation between adjacent ordered bootstrap means. The $m = 1$ support and allocation contrasts each have an individual 95% interval; the support and allocation contrasts do not form a joint two-contrast family. The $m = 4$ secondary and

$m = 2$ routing families each target nominal 95% family coverage within one horizon. No single family combines both horizons or the whole paper. Parents, selected policies, and cohort masks remain fixed; parents are not resampled. These conditional percentile intervals do not provide exact finite-sample simultaneous coverage.

Additional horizon sensitivity. The main text distinguishes endpoint–prior from routing interactions. The fourfold − original support difference under π_0 is -0.143pp, $[-0.892, 0.607]$pp. The horizon-comparison family contains both support and both routing interactions. Three parent points still limit training-variability claims, and an interval containing zero does not prove invariance.

Budget transfer retains its original primary. Specified $q = .25/.75$ transfer reuses the .5 source-selected scores, without budget-specific fitting or TUNE reselection. In CIFAR, VALUE − ERROR at .25 is 0.093pp in LT50 and -0.200pp in BALANCED, both unresolved; at .75, the estimates rise to 1.061pp and 1.302pp. The own-quota remainder stays exactly zero at 5k. In Tiny, the original/additional groups instead give primary VALUE − ERROR of $-0.143/-0.260$pp but secondary .75 differences of $0.237/0.420$pp. Acquisition budget therefore changes the empirical target comparison. The specified .5 contrast remains primary, including the adverse additional-group result.

Quota convention under budget transfer. Tiny uses literal confidence rather than frozen ERROR ordering inside the copied quotas, so the diagnostic is not numerically identical to CIFAR. Tiny primary VALUE − own-quota is $-0.017/+0.005$pp for the original/additional groups, both unresolved. Main-text secondary target effects do not change this replay convention.

E COMPACT-POLICY RESULTS AND EXECUTION VALIDATION

Table 9: Actual served compact − strong-simple accuracy at $q = .5$. All eight contrasts form one family across both horizons. Nominal 99.375% individual percentile-bootstrap intervals target Bonferroni 95% family coverage. All contain zero; equivalence is not established.

Horizon	Cell	Mean	Adjusted interval	Three parent points
1×	A_{00}	-0.047	[-0.194, +0.097]	+0.100 / -0.140 / -0.100
1×	A_{01}	-0.030	[-0.127, +0.064]	-0.020 / -0.070 / +0.000
1×	A_{10}	-0.067	[-0.180, +0.048]	-0.020 / -0.010 / -0.170
1×	A_{11}	-0.030	[-0.108, +0.051]	-0.020 / -0.050 / -0.020
4×	A_{00}	-0.010	[-0.204, +0.187]	-0.130 / -0.170 / +0.270
4×	A_{01}	+0.097	[-0.043, +0.231]	+0.030 / +0.300 / -0.040
4×	A_{10}	+0.040	[-0.037, +0.117]	-0.010 / +0.070 / +0.060
4×	A_{11}	+0.017	[-0.030, +0.067]	+0.010 / +0.010 / +0.030

The eight contrasts cross the two horizons with all four factorial corners. Each image's compact − strong-simple correctness difference is averaged across the three fixed parents. Equal-tailed percentile intervals use 6,000 paired-image resamples (random seed 2609231124), with quantiles 0.003125 and 0.996875 and default linear interpolation. The resulting 99.375% individual level corresponds to a Bonferroni nominal 95% target for this eight-contrast family. Models, source-selected policies, and cohort masks remain fixed, so the intervals exclude training-seed and policy-selection uncertainty. The family level is nominal, not an exact finite-sample simultaneous guarantee.

The compact source follow-up trains no new CNNs. For each horizon, parent, and corner, a frozen SOURCE record contains fitted parameters, selected compact and simple members, input-identity checks, and the evidence for source-only selection. The actor first performs cheap inference on all 9,993 examples, calculates frozen scores, applies stable whole-cohort ranking, and forwards exactly 4,996 selected examples through the paid predictor. Actual answers, masks, and integer correct counts are stored separately from cached full-cohort predictions. Independent replay reconstructs the score order and selected mask and checks the paid/cheap answer assembly. All 72 served-policy evaluations agree with the precomputed predictions across the 24 horizon–parent–corner settings.

The implementation computes all current menu scores before selecting the declared output, so measured scoring overhead includes unselected heads. Model loading is recorded separately from re-

quest execution. Measurements during concurrent endpoint training are not isolated latency benchmarks. No timing figure is borrowed from a batch-local rule with a different mask. A smaller parameter count is not evidence of a measured speedup.

F ACCOUNTING, RETAINED NEGATIVE EVIDENCE, AND INTERPRETATION

Cost units. The original CIFAR reused predictor pair averages 0.2021 GPUh of measured training-loop time; the matched new endpoint pair averages 0.2036 GPUh. The baseline and gate-investment arms share the reused pair, so displaying both arms does not mean two separate baseline training bills. Tiny original diagonal fitting uses 7.413 training GPUh and the additional two-parent allocation uses 4.853 training GPUh. The original factorial adds 7.347 training GPUh for twelve off-diagonal predictors. The fourfold factorial consumes 59.760 training GPUh for twenty-four fresh predictors. Reserved GPU capacity, idle time across parallel training processes, source inference, CPU head fitting, menu search, and serving are different accounting categories. No sum of incomplete categories is presented as lifecycle cost.

Event-level accounting. Equation 3 gives the exact target-difference identity. Event counts retain corrections, harms, predicted-class allocations, and SOURCE-defined confidence bins rather than interpreting aggregate gain alone.

Retained adverse boundaries. A wider-backbone allocation check favored endpoints by 1.275pp, but an image-level VALUE extension lost to class selectors and all original source choices declined deployment. Earlier recorded-signal checks supported a richer rule over a strong simple control in one setting, not over class conditioning; other reciprocal-expert contrasts remained unresolved. A mean-matching control failed its original criterion in one parent. These controls constrain the tested mechanisms without determining the usefulness of unobserved features under other designs.

Remaining scientific limits. Only two concrete image-cascade families anchor the main investment comparison. The sampling-prior factorial uses the original three Tiny parents, not the additional two repetitions. Longer training changes a specified optimization horizon and schedule; optimization to a universal optimum is not demonstrated. The small compact score class can fail even when a richer class-conditional mechanism exists. Fixed-cohort top-q selection does not establish streaming performance under shifted class mixtures. The study compares completed allocation outcomes rather than training a prospective policy for choosing an investment before fitting. Stronger population, annotation-efficiency, or universal routing claims would require additional designs and independent environments.

Scope of supporting experiments. Component, classifier, stronger-CIFAR, and nested-donation comparisons supplement the original 6k allocation, factorial, and fourfold primary comparisons. The appendix tables report measurement units, uncertainty intervals, and adverse outcomes for each comparison.

G COMPONENT AND OPERATING-POINT CONTROLS

The same-prior Tiny pair cross uses $B_{00} = (f_0, h_0)$, $B_{10} = (f_1, h_0)$, $B_{01} = (f_0, h_1)$, and $B_{11} = (f_1, h_1)$, where subscripts denote old or expanded predictor support. All four procedures receive the same honest 4k gate bank and own SOURCE selection. Define $a_{ij}(g) = \mathrm{Acc}(B_{ij}, g)$. For a frozen old mask g_{00}, the fixed-mask interaction is exactly zero. The own-policy interaction instead records route responses: with $a_{ij} = a_{ij}(g_{ij})$, $a_{11} - a_{10} - a_{01} + a_{00} = T_{11} - T_{10} - T_{01}$, where $T_{ij} = a_{ij}(g_{ij}) - a_{ij}(g_{00})$. Fixed-mask replay is an evaluation diagnostic, not a free deployment of the old cheap model.

No-paid prediction and classifier repair. Under the SOURCE budget-selection rule, old Tiny chooses $q = .5$ and new Tiny chooses $q = 0$ in all three parents at each horizon. Parent gains are $2.402/3.603/2.892$pp at $1\times$ and $3.903/4.443/4.083$pp at $4\times$. Against the same-donation gate arm at .5, forced new cheap-only gains 3.332pp, $[2.842, 3.826]$, and 3.936pp, $[3.456, 4.416]$. The gate arm does not always meet the SOURCE accuracy target; these forced contrasts are not comparisons

Table 10: Tiny component controls at $q = .5$. Each horizon uses three paired fitted parents; the three contrasts form a family per horizon. Parent points and conditional paired-image intervals are pp.

Horizon	Contrast	Mean	Interval	Parent points
$1\times$	Both − cheap update	2.598	$[2.158, 3.025]$	$2.682/2.552/2.562$
	Both − paid update	2.472	$[2.078, 2.882]$	$2.252/2.632/2.532$
	Procedure interaction	-0.831	$[-1.224, -.407]$	$-.640/-.881/-.971$
$4\times$	Both − cheap update	2.809	$[2.395, 3.199]$	$2.912/2.752/2.762$
	Both − paid update	3.009	$[2.592, 3.436]$	$3.032/2.962/3.032$
	Procedure interaction	-1.271	$[-1.631, -.921]$	$-.981/-1.251/-1.581$

after SOURCE budget admission. Frozen-body classifier support adds 7.185pp at $q = 0$ and 4.240pp at .5. Repair − full-refit at $q = 0$ is .494pp, $[-.023, 1.001]$: neither superiority nor equivalence is established. This is a warm-start classifier-only support refit. The frozen A_{01} ResNet-18 body produces 512-dimensional features from normalized 64-pixel images in evaluation mode. BF16 feature extraction returns FP32 arrays; no random crop, flip, color jitter, body update, or Batch Normalization update occurs. Both heads start from the same A_{01} classifier weights and bias and use the expanded-prior class sequence π_1, with 42,900 versus 48,900 allowed labels. FP32 cross-entropy SGD uses learning rate .1, momentum .9, Nesterov acceleration, and weight decay .0005; rate factors of .2 follow 2,934,000, 5,868,000, and 7,824,000 presentations. Each fit consumes 9,780,000 presentations in 76,407 updates, including the final 32-example batch. Own gate fitting uses 4k honest labels and selection uses TUNE6k; the paid A_{01} predictor stays fixed. This differs from the ImageNet-LT cRT implementation (Kang et al., 2020), whose random head initialization, class-aware sampling, online 224-pixel augmentation, and ten-epoch cosine schedule use batch 512 and initial rate .2.

Stronger CIFAR controls. The fixed .1 learning-rate study trains old/new ResNet-56 predictors for three paired parents without selecting a favorable TARGET result. Same-label endpoint − gate parent gains are $4.065/2.743/2.473$pp. New cheap-only versus the strong old gate at .5 gives $2.153/3.634/3.955$pp. Allowing the old arm the same SOURCE budget-selection menu reduces the latter gains to $2.153/2.943/3.404$pp, still positive for these parents. The new cheap and new cascade remain below the strong old union ceiling. Component results with a shared 2,000-row gate bank and equal-total-label comparisons of old endpoints with 5,000 gate rows versus new endpoints with 2,000 gate rows are not interchangeable; the former improves both components beyond a single update but has interaction $-.380$pp, $[-.704, -.043]$.

Smaller donations and SOURCE admission. Nested 2k/4k donations assign 10/20 already labeled images per class, retain three paired parents and 39.12M presentations, and do not replace the 6k primary. The original-prior expanded-support corner A_{10} is not fitted at these doses. This SOURCE budget-selection follow-up was designed after historical TARGET exposure, not as a new independent confirmation. The SOURCE budget selection reuses the frozen .5 score at $q = 0, .25, .5$, rather than fitting a new head for each budget.

Table 11: Nested-donation SOURCE budget-selection contrasts in pp. The six planned contrasts define the multiplicity family. One parent in the 4k prior-only comparison does not meet the SOURCE accuracy target, so no paired interval is reported using only the two qualifying parents. These donation sizes do not establish a scaling law.

Donation	Contrast	Mean	Conditional interval
2k	Same-prior support	1.378	$[.792, 1.933]$
	Same-label endpoint − gate	$.804$	$[.245, 1.369]$
	Prior only	$-.003$	$[-.485, .499]$
4k	Same-prior support	1.761	$[1.151, 2.365]$
	Same-label endpoint − gate	1.815	$[1.172, 2.432]$
	Prior only	2/3 SOURCE admitted; no paired interval	

Agreement between direct execution and precomputed predictions. The 22 unique selected policies each have four execution checks with agreement in masks, answers, correctness, and correct counts. The two additional fixed-head $q = .25$ conditions are included among these policies and account for eight of those checks. Earlier 63-condition execution retained an answer discrepancy on one image and two-mask discrepancies in two other conditions, all with zero correctness change. Those discrepancies occurred under earlier policies, which differ from the selected budget policies. The eight additional requests use four separate processes per condition in two overlapping four-process jobs. Each process loads a new model pair, warms both predictors on 128 images, then times its first full-cohort request. Separate processes do not constitute independent image samples; complete simultaneity of the timed intervals and host exclusivity are unverified. Earlier dose measurements keep the loaded pair for three resident repeats after a first request, with no OS/GPU cache flush. Dose request time sums cheap inference, selected-head scoring/global sorting, and selected paid inference, including transfers, normalization, probability return, and GPU synchronization. Loading, cohort preparation, warmup, output storage, and final-answer assembly are excluded. These dose times are therefore not compared with the broader measurement in Appendix I. Forced cheap-only prediction omits gate scoring, ranking and paid-predictor inference. Annotation, preparation, SOURCE CPU search, and energy remain outside the measured lifecycle costs.

H SOURCE-SELECTED ANSWER CORRECTION AND PAIR CEILINGS

The sensitivity analysis reuses the original three Tiny $1\times$ parents at $q = .5$ on 9,993 exposed images. Each cheap/paid pair shares $\tau \in \{0, .5, 1\}$ in $p'_k \propto p_k \pi_k^{-\tau}$. Strictly positive cached FP32 probabilities are transformed in FP64 and returned to FP32. This adjusts log probabilities without recovering original BF16 pre-softmax logits. Tau zero preserves the original predictions and policies. Training-design priors use E_0 for gate investment A_G and E_1 for A_{01}, A_{11}; no TARGET histogram defines the prior.

Each nonzero candidate recomputes cheap and paid errors, relative value, cheap summaries, and predicted class, then refits the original gate menu on honest labels. The within-tau head selection is unchanged. Across tau, TUNE6k cascade correct count selects the winner, breaking ties toward smaller tau before TARGET evaluation. All nine choices select one, the menu boundary, rather than an established optimum. The 18 additional gate fits take 62.251 seconds of CPU SOURCE fitting; no new CNN training or serving measurement is included. Selected cheap/paid/cascade mean accuracies are $49.825/50.212/51.399\%$ for A_G, $49.038/49.448/51.193\%$ for A_{01}, and $53.971/54.488/55.669\%$ for A_{11}.

Table 12: Answer correction under matched selection opportunities. Differences are pp across the same three parents. The 95% paired-image intervals condition on models, SOURCE choices, and realized masks. Same-label is this follow-up's primary; same-prior is a descriptive secondary, not a jointly adjusted two-contrast family.

Contrast	Original	Corrected [interval]	Corrected parent points
Same-label $A_{11} - A_G$	6.268	4.270 [3.759, 4.797]	3.833/4.713/4.263
Same-prior $A_{11} - A_{01}$	5.901	4.476 [3.943, 5.004]	4.213/4.643/4.573

Same-label compares honest gate support 10k versus 4k; same-prior fixes 4k. The restricted tau menu does not exhaust balancing repairs or establish a universal investment rule.

Why the fixed-pair bounds are exact. The four disjoint correctness groups have sizes C (only paid correct), H (only cheap correct), $B - H$ (both correct), and W (neither correct), so $B + C + W = n$. Raw substitution first uses up to C beneficial acquisitions, then $n - C - H$ neutral acquisitions; only the final H are harmful. Thus exactly k acquisitions produce the numerator $B + \min(k, C) - [k - (n - H)]_+$ in Equation 4. Allowing the cheap answer to be retained removes the harm term and gives $B + \min(k, C) \le B + C = n - W$. These are attainable label-hindsight bounds; a deployable gate need not attain them.

Corrected pairs have different hindsight ceilings. Equation 4 is recomputed after SOURCE correction, without another policy or tau selection. At $k = 4,996$, raw substitution, KEEP, and

union bounds coincide for all six old pairs. Corrected A_{11} cheap-only and cascade accuracies remain below each corrected ceiling. The positive learned-policy contrasts therefore do not establish excess over those oracle policies. Figure 2 remains an exact witness for its original uncorrected pairs; neither calculation ranges over arbitrary answer repairs or all tau values.

Table 13: Corrected Tiny $1\times$ pair counts with $n = 9,993$, $k = 4,996$, and $\tau = 1$. B, C, H, W follow Equation 4. Bound is the common raw, KEEP, and union correct count. Excesses compare corrected A_{11} with the indicated old bound in pp, without confidence intervals. Parent numbers identify the same paired realizations throughout.

Old pair	Parent	B	C	H	W	Bound	Cheap excess	Cascade excess
A_G	1	4968	966	791	4059	5934	-5.394	-3.522
	2	4972	829	922	4192	5801	-3.983	-2.822
	3	4997	887	853	4109	5884	-5.024	-2.962
A_{01}	1	4903	918	820	4172	5821	-4.263	-2.392
	2	4914	834	899	4245	5748	-3.452	-2.292
	3	4884	878	788	4231	5762	-3.803	-1.741

I A COMMON EXECUTOR FOR RESIDENT-COHORT TIMING

The original uncorrected SOURCE budget choices fix old A_{01} at $q = .5$ and new A_{11} at $q = 0$ in all six horizon–parent settings. Both policies use one executor and the same physical NVIDIA H200 GPU within a setting, with PyTorch 2.11.0+cu128, CUDA build 12.8, and TF32 disabled. A request processes all 9,993 images in microbatches of 128, including CPU uint8 transfer, normalization, BF16 inference, FP32 softmax and return to CPU, old-policy selected-head scoring and global top-4,996 sorting, final-answer assembly, and synchronization. The new actor invokes no head, sorting, or paid predictor. Two full-request warmups per policy precede ten alternating old–new/new–old pairs. All 120 requests reproduce the frozen masks, previously executed answers, and cached correctness; repeat-to-repeat disagreement is zero.

Table 14: Uncorrected same-prior operating points on 9,993 images. Accuracy is old/new percent. Time is seconds per resident cohort: medians [25th,75th percentiles] over ten requests, not online single-query latency. Ratio is the median within-pair old/new duration ratio. Repetition quantiles are descriptive, not confidence intervals.

Horizon	Parent	Accuracy old/new	Old time	New time	Ratio
$1\times$	1	46.903/49.305	.92209 [.92095, .92381]	.34298 [.34075, .36630]	2.688
	2	45.592/49.194	.91891 [.91416, .92325]	.34008 [.33970, .34570]	2.690
	3	46.192/49.084	.91684 [.91459, .91747]	.33858 [.33760, .33904]	2.706
$4\times$	1	44.601/48.504	.91790 [.91637, .92605]	.33864 [.33808, .33930]	2.712
	2	44.041/48.484	.91893 [.91779, .92190]	.33976 [.33930, .34013]	2.703
	3	44.411/48.494	.91854 [.91688, .92242]	.33997 [.33926, .34018]	2.698

All three models remain resident together. Loading, cohort preparation, and warmups are excluded. Four GPU processes share the host; 71 compute-process observations, at most .588 seconds apart, detect no unrelated compute process, but activity between observations cannot be excluded. The measurements are neither host-exclusive nor batch-size-optimized throughput estimates. Allocated GPU time is .051723 GPU-hours (46.551 seconds across four GPUs), distinct from request duration.

These operating points have unequal accuracy. The duration ratios are not equal-quality speedups, new SOURCE selection, or measurements of the corrected $\tau = 1$ policies or same-label contrast. Training, SOURCE fitting, model loading, and lifecycle break-even are excluded. No population or training-repeat interval, or ratio against earlier different-executor timings, follows.

J CORRECTED POLICIES WITH JOINTLY MEASURED QUALITY AND REQUEST TIME

This analysis reuses the three original Tiny parents and 9,993 exposed eligible images. No new image or predictor fit is introduced. The fixed-$q = .5$ corrected contrasts in Appendix H remain distinct from the budget-selected policies here. Each arm receives the nine candidates $\tau \in \{0, .5, 1\}$ and $q \in \{0, .25, .5\}$. For each parent and comparison, the corrected reference's best SOURCE accuracy at $q = .5$ defines the target; the selection rule chooses the smallest acquisition fraction among candidates meeting that target, with the fixed cost tie rule. SOURCE admission does not guarantee TARGET noninferiority.

A_{11} selects $q = 0, \tau = .5$ throughout. Gate-investment A_G selects $q = .5, \tau = 1$; same-prior A_{01} selects $q = .25, .5, .25$ across parents with $\tau = 1$. Quality and timing below describe these same executed policies. For accuracy, the three fixed parent differences are averaged per image before 3,000 paired-image bootstrap resamples. Equal-tailed 97.5% individual intervals use the 1.25th/98.75th percentiles for a nominal 95% two-contrast family target. Fixed models, SOURCE choices and realized masks remain outside the resampling uncertainty.

Table 15: Corrected SOURCE-selected operating points. Accuracy differences are percentage points on the same 9,993 images; intervals condition on the three fitted parents.

Contrast	Mean	97.5% interval	Parent differences
Same-label $A_{11} - A_G$	.5571	$[-.0434, 1.2009]$	$.1801/1.4510/.0400$
Same-prior $A_{11} - A_{01}$	1.3276	$[.7355, 1.9247]$	$1.3910/1.3810/1.2108$

The same-label interval contains zero. The positive same-prior contrast does not replace that unresolved primary comparison, and lower request time does not establish an equal-quality speedup. The earlier positive fixed-$q = .5$ same-label result cannot supply a quality guarantee for these different operating points.

Table 16: Resident request durations for the corrected policies in Table 15. Seconds are medians of ten paired requests after two warmups per system. Ratios divide reference by A_{11} medians; repetitions do not constitute independent trained models.

Contrast	Parent	Reference (s)	A_{11} (s)	Ratio
Same-label	1	1.002263	.366730	2.733
	2	.999858	.365474	2.736
	3	.955825	.362223	2.639
Same-prior	1	.666317	.365765	1.822
	2	.994195	.369547	2.690
	3	.692764	.360276	1.923

Two-system execution alternates in ABBA order, yielding 120 timed requests across the six comparisons. The timed path includes cheap inference, answer correction, selected-head scoring, whole-cohort stable top-q selection, selected-row paid inference, answer assembly and synchronization. At $q = 0$, scoring, sorting and paid inference are omitted. Models remain jointly resident on a shared host; loading and cohort preparation are excluded. Annotation, predictor training, lifecycle cost and VRAM savings are not measured.

The cheap forward pass processes the original 10,000-row cohort before retaining 9,993 eligible images, charging the seven excluded rows' cheap computation. Processing only eligible rows changes tail batches and produced two mask/answer discrepancies in each of two first-parent reference checks. Restoring the original batching yields exact agreement of all masks and answers in twelve preliminary executions and all 120 timed requests. Selected-row paid posteriors still differ from full-cohort cached posteriors by up to $.001482$–$.008061$; answer agreement does not imply bitwise posterior or batch invariance.