# GEAR: RESPONSE-AWARE TEST-TIME COMPUTE ALLOCATION FOR FROZEN QUANTIZED SPIKING TRANSFORMERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Frozen quantized spiking classifiers cannot change their weights at deployment, but they can spend extra computation on selected inputs. We study when an alternative view helps rather than merely changes a prediction. GEAR routes low-margin images from a frozen 4-bit QSD-Transformer through nested high-resolution and flipped views, then fuses their predictions. Its empirical response curve models a fixed action's net repair utility—repairs minus newly introduced errors—along the cheap-view margin rank. On ImageNet validation and five ImageNet-derived shifts, a fixed 26:8:4 policy gains 0.79–2.46 percentage points without target-set retuning. A labeled target-like probe calibrates the curve and operating width, evaluated on disjoint held-out rows; geometry interventions show why a source-optimal crop need not transfer. Tests on a second spiking family, synthetic corruptions, detection, and segmentation clarify its scope: margin routing helps when escalation corrects errors concentrated among uncertain inputs, but not when refinement mainly benefits confident ones. The response profile is conditioned on the chosen action and deployment distribution.

## 1 INTRODUCTION

Compressed classifiers are commonly evaluated with one fixed test image, although a serving system may have enough compute to inspect a few inputs again. Spending that budget uniformly can waste views. We ask which images merit another view under a frozen model. In serving, the budget is spent per request batch: a system that already groups 64 images can recompute a fixed fraction at a second view, so the question is which rows to select. The key quantity is not uncertainty alone, but the change in correctness produced by a specific view action.

The quantized spike-driven Transformer (QSD-Transformer) of Qiu et al. (2025) provides a concrete setting. We use a released 4-bit checkpoint with frozen weights, quantization grids, and normalization statistics. GEAR first evaluates a cheap view, then recomputes a fixed number of low-margin images at larger resolutions and fuses their predictions. The same action may repair one error and create another. We therefore measure its *net* response along the margin ranking rather than assume hard examples always benefit.

Our primary contribution is response-aware action valuation for frozen-model inference: an action-conditioned response curve that prices a candidate escalation view by its net repair utility along the cheap-view margin ranking, not by confidence alone. It links margin rank, repairs, breaks, and budget, and governs where to allocate compute and when to stop. GEAR's 26:8:4 nested routing is one concrete instantiation on a released quantized spiking Transformer; the valuation framing is not specific to that family, and the non-spiking controls reproduce the first-stage response. We test the curve's calibration on held-out examples, its transfer across shifts and models, and its failure modes on dense tasks; the rule works well on source-like data, while its profile can shift across target streams.



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.

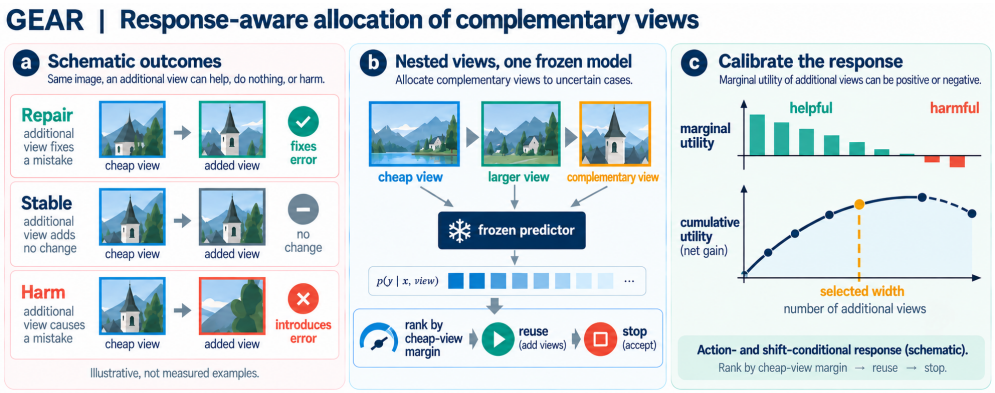


Figure 1: GEAR allocates complementary views to low-margin inputs while keeping one predictor frozen. The repair, stable, and harm cases illustrate possible outcomes, and the marginal-utility and response sketches summarize the mechanism. Quantitative results use the measured curves and tables below.

## 2 Related Work and Positioning

Dynamic networks vary depth, resolution, or spatial focus with the input (Huang et al., 2018; Wu et al., 2018; Wang et al., 2018; Yang et al., 2020; Wang et al., 2022). Test-time augmentation also averages predictions across transformed views (Lyzhov et al., 2020; Mocerino et al., 2021); recent stress-aware policies select them adaptively (Yang, 2026). GEAR is a member of this family; its contribution is not adaptive view selection itself, but a margin-stratified response model that separates sample priority from action value and uses cumulative utility to set budget and stopping point. Confidence-adaptive SNNs instead change internal computation (Li et al., 2023); spiking Transformers establish the model family on which we intervene (Zhou et al., 2023; Qiu et al., 2025).

Distribution shift changes both confidence and view utility. ImageNet-V2 re-samples ImageNet (Recht et al., 2019); Sketch and Renditions alter appearance (Wang et al., 2019; Hendrycks et al., 2021a). Corruption benchmarks expose further sensitivity (Hendrycks & Dietterich, 2019). We use these sets to evaluate transfer of a fixed operating point. Because shift moves both the ranking statistic and the view's value, a deployment rule must be re-priced per target stream rather than fixed once on the source. Unlike test-time adaptation and adaptive test-time defenses that update or optimize during inference (Wang et al., 2021; Croce et al., 2022), our method never updates model parameters.

## 3 Action-Conditioned Response Curves and GEAR

### 3.1 Setting and Decision Rule

Let $f_\theta$ be the frozen quantized classifier, $g_0$ the cheap resize-256/center-crop-224 view, and $z_j(x) = f_\theta(g_j(x))$ the logits for view $j$. All images in a request batch of $B = 64$ first receive $g_0$. Write $P_0 = \mathrm{softmax}(z_0)$ and define the probability margin $m(P) = P_{(1)} - P_{(2)}$, with $P_{(1)}$ and $P_{(2)}$ the two largest class probabilities. The first stage selects the $k_1$ smallest-margin rows, recomputes $g_1$ on those rows, and uses

$$P_1(x) = w_1 P_0(x) + (1 - w_1)\,\mathrm{softmax}(z_1(x)) \quad (x \in S_1), \tag{1}$$

leaving other predictions at $P_0$. The shipped setting is $k_1 = 26$ and $w_1 = 0.3$; $g_1$ is resize-288/crop-288. $k_1$ is a *count per 64-image request*, not a global margin threshold. The policy therefore has deterministic view counts per request, while routed identities vary with the input.

### 3.2 Nested Views Rather Than Uniform Escalation

Among $S_1$, stage 2 re-ranks the current fused margins, selects the lowest $k_2 = 8$, and computes a horizontally flipped $288 \times 288$ crop. Stage 3 re-ranks only those eight rows, selects $k_3 = 4$,

and computes a $320 \times 320$ crop. On selected rows, $P_2 = 0.3P_1 + 0.7(0.2P_0 + 0.8P_{1f})$ and $P_3 = 0.7P_2 + 0.3P_{320}$. Thus $S_3 \subset S_2 \subset S_1$: each stage refines its parent subset and carries forward earlier predictions. Appendix A.1 gives implementation details.

The nominal extra neural-network workload relative to one cheap view is approximately

$$\Delta C \propto k_1\, 288^2 + k_2\, 288^2 + k_3\, 320^2, \tag{2}$$

where constants depend on kernels, batching, and service implementation. Equation 2 captures tensor-shape workload; latency is measured separately on batches from the actual implementation. Changing resize-288/crop-288 to resize-320/crop-288 alters visible content without changing the neural-network tensor shape, so geometry can affect accuracy at the same measured *network* workload. The shipped counts recompute 38 additional views per 64-image request (26+8+4), which the measured service prices at $+44.7$ ms p50 over the cheap pass.

### 3.3 A REPAIR–HARM ACCOUNTING IDENTITY

For any selected set $S$ and final prediction $\hat{y}_i^S$, let $c_i^0$ and $c_i^S$ be correctness indicators before and after escalation. The top-1 accuracy change is

$$\Delta A(S) = \frac{100}{N} \sum_{i \in S} (c_i^S - c_i^0) = \frac{100}{N} (\#\text{repaired} - \#\text{broken}). \tag{3}$$

The identity is accounting; the empirical claim is that a fixed action's *net* utility has a margin-stratified profile that can guide allocation. We use the identity to separate the *selection* effect from the *view* effect. With labels from the same target dataset, the decile plot is a descriptive decomposition of observed gain; source-fitted estimates transfer poorly to ImageNet-Sketch (Appendix C.1).

For a fixed escalation action $a$, let $u_i(a) = c_i^a - c_i^0 \in \{-1, 0, 1\}$ be the net correctness change and let $d_i$ be the cheap-view margin rank. Partition the sorted rows into $B$ blocks and estimate $\bar{u}_b(a) = \mathbb{E}[u_i(a) \mid i \in b]$. The response curve at budget $k$ is the cumulative utility

$$R_a(k) = \frac{100}{N} \sum_{r=1}^{k} u_{(r)}(a), \qquad \widehat{R}_a(k) = \frac{100}{N} \sum_{b=1}^{B} n_b(k) \widehat{\bar{u}}_b(a). \tag{4}$$

Here $u_{(r)}$ is the utility of the $r$th row in margin order and $n_b(k)$ counts the first $k$ rows in block $b$. We call the observed block-level pattern the *response law*: on the classification regimes where margin routing pays, $\bar{u}_b(a)$ is non-increasing with margin rank, so increments of $R_a(k)$ tend to diminish. The law is an empirical regularity of the regimes we test, not an assumption for every action or task; the scope tests below identify where it fails. Positive early blocks justify spending compute on low-margin rows, while flattening or negative later blocks determine where escalation should stop. The law depends on the frozen checkpoint, action, preprocessing, and data distribution. If block means are monotone and block capacities are fixed, an exchange argument makes margin-order filling optimal within this linear block model. At row level the approximation is imperfect: block-optimal minus greedy gaps reach 0.157 pp, while a label-aware oracle remains 1.13–2.81 pp ahead. The declining profile is statistically stable: across the five audited streams the decile-correlation bootstrap intervals exclude zero, the isotonic fit matches the raw curve (MSE ratio $\geq 0.998$), and permutation tests reject the flat null ($p \leq 0.0035$; Appendix E.2).

To estimate the curve, a labeled target-like probe supplies margins and candidate-view utilities; the policy selects the width maximizing $\widehat{R}_a(k)$ on a fixed menu, while disjoint rows evaluate that choice. A safety gate falls back to the minimum width when the probe predicts a decline at larger budgets. Once calibrated, deployment needs only cheap-view margins and the selected width; labels are confined to the probe. Across five model/data conditions, 4,096 probe rows yield a three-draw mean held-out curve error of at most 0.42 pp and a mean selected-width regret of at most 0.11 pp (worst single draw: 0.61 and 0.16 pp). A 512-row probe already stabilises the selected width on source-like streams, and the joint $(k, w)$ selection reaches 0.004–0.012 pp regret on same-distribution settings. A 2,048-row version was frozen after a 36-stream audit and priced six unseen corruption streams within 0.07 units of their held-out optima; Appendix B.1 gives the probe-size sweep and error conventions; Appendix B.4 gives the unified safety-gate audit.

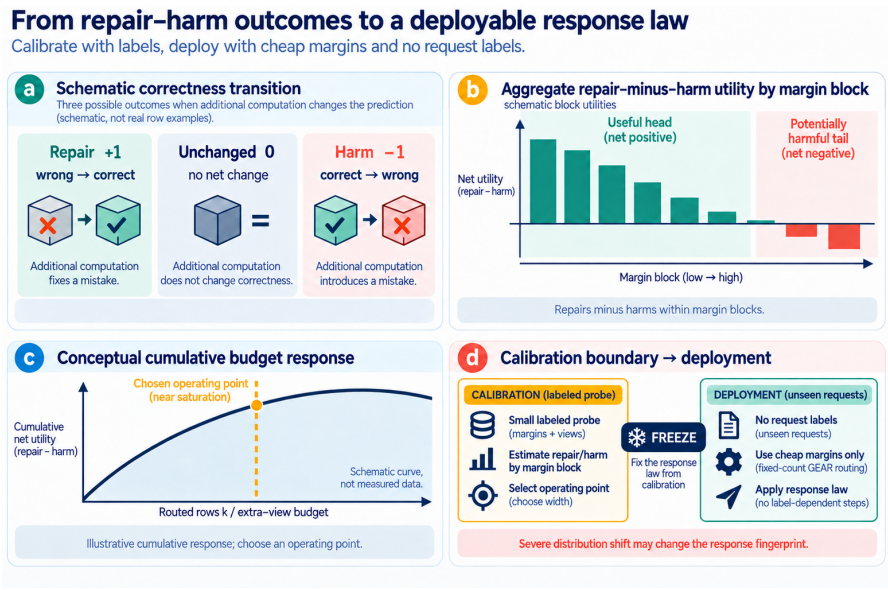


Figure 2: How the response law is measured. Labeled calibration rows reveal repair, stability, or harm after a candidate view; mean net utility by margin block integrates into a budget curve. The plotted bars and curve are conceptual, and labels are used only for calibration, not at deployment.

### 3.4 Mechanism: exact error decomposition, flux accounting, and a four-scalar witness

For any margin stratum, let $A$ and $B$ be the cheap and fused error rates and $J$ their joint error rate. Then $\Pr[\text{repair}] = A - J$, $\Pr[\text{break}] = B - J$, and net utility obeys the exact identity

$$u \;=\; \Pr[\text{repair}] - \Pr[\text{break}] \;=\; A - B \qquad (\text{measured residual} \le 10^{-16}). \tag{5}$$

The law therefore concerns the error advantage $A - B$ across margin strata: repairs and breaks both concentrate at low margin, and their *difference* drives allocation. A Gaussian shared-noise witness parameterized by view advantage, noise scale, correlation, and temperature reproduces the decile shape but not its full amplitude. The identity holds to $9 \times 10^{-17}$; $J/(AB)$ rises from 1.05–1.26 in the lowest-margin decile to over 40 at the top; and the witness matches the shape at correlation 0.979 with a 2.4 pp amplitude gap. Its applicability map places tested families with a mildly stronger escalated view in the payoff regime and assigns the harmful axes to a weaker escalated view at large noise; non-monotone profiles appear only for a weaker view at low noise and grow with the shared-noise correlation (0 of 99 cells at $\rho$=0.3, 46 at $\rho$=0.9). Appendix C.2 provides the full flux and phase-map audits.

## 4 Experimental protocol

**Model and data.** The primary model is a frozen 4-bit QSD-Transformer checkpoint (Qiu et al., 2025); quantization is a complementary efficiency axis to routing (Jacob et al., 2018), and Appendix A.4 checks weight precision and spiking timesteps as alternative test-time compute axes. The primary evaluation uses all 50,000 ImageNet validation images with the same cached preprocessing and class order for every view. The fixed-policy panel uses three 10,000-image ImageNet-V2 variants, ImageNet-Sketch, and ImageNet-R; 7,500 ImageNet-A images (Hendrycks et al., 2021b) test response-profile transfer only (Appendix C.1). R and A use restricted 200-class logits. Three additional QSD exports and a released 6.8M-parameter scale test within-family repeatability; a torchvision ResNet-50 with IMAGENET1K_V2 weights (He et al., 2016) and a torchvision ViT-B/16 (Dosovitskiy et al., 2021) on ImageNet-V2 are non-spiking controls, and a self-trained spiking ResNet trained directly on CIFAR-100 provides a second spiking family. Cache/preprocessing variants have different cheap baselines and are never pooled.

Table 1: Fixed-default top-1 results. Values after the cheap column are *gain over that row's cheap baseline* in percentage points; counts are 26/8/4 per batch of 64. R uses a restricted 200-class protocol.

| Distribution | $N$ | Cheap (%) | Route | + Flip | + Refine |
|---|---|---|---|---|---|
| ImageNet val | 50,000 | 76.902 | 1.668 | 1.938 | 1.966 |
| V2 matched | 10,000 | 65.43 | 1.85 | 2.16 | 2.46 |
| V2 threshold | 10,000 | 74.07 | 1.99 | 2.16 | 2.35 |
| V2 top-images | 10,000 | 79.69 | 1.33 | 1.69 | 2.00 |
| ImageNet-R | 30,000 | 37.39 | 0.63 | 0.80 | 0.79 |
| ImageNet-Sketch | 50.9k | 26.23 | 0.88 | 0.97 | 0.96 |

**Selection and inference.** The default $26{:}8{:}4$ counts and fusion weights are frozen before external-set evaluation. The full 50,000-image ImageNet-validation row summarizes the source operating point and is not treated as independent confirmation. For configurable source-side comparisons, we select hyperparameters on even-indexed validation rows and report on odd-indexed rows; comparisons that scan all 50,000 rows are labelled exploratory. We report accuracy differences in percentage points (pp), paired McNemar tests for specified incremental comparisons, and class-cluster bootstrap intervals where available. Paired tests apply only to the specified incremental comparisons.

**Fixed policy versus probe calibration.** The fixed $26{:}8{:}4$ policy is chosen before external-set evaluation and uses no target labels at inference. The response-curve probe instead uses labeled, target-like examples and candidate-view forwards to calibrate a stream-level width, which is evaluated on disjoint held-out rows.

**Serving scope.** The inline CUDA service runs batch-64 requests on an RTX 4080 SUPER. Its timer starts after all requested image views have been decoded, transformed, and placed on GPU. We therefore report p50/p95 and throughput as *pre-materialized GPU inference* and routing, separately from decoding, CPU preprocessing, transfer, and queueing. The parity audit compares both computations within one service request. Because the shipped rule fixes the view count, its tensor workload is deterministic for full batches, while its gains are measured once per stream against both the cheap pass and uniform escalation at the same fusion weight.

# 5 Results

## 5.1 A fixed route transfers, but its deeper stages do not always pay

Table 1 reports the fixed default on six distributions. ImageNet validation gains 1.966 pp over a 76.902% cheap baseline, and all five external sets gain 0.79–2.46 pp without target-domain retuning. Gains are smaller on ImageNet-R and Sketch: the first routed view remains useful, but flip/refinement nearly saturate. The flip adds $+0.27$ pp on validation and at most $+0.36$ pp on V2, while refinement is flat on R and Sketch ($-0.01$ pp each). The validation contrast is paired: $+1.66$ pp with a bootstrap 95% interval of $[+1.46, +1.86]$ and a Holm-adjusted McNemar $p = 4.6{\times}10^{-58}$ (Appendix B.3).

Paired image-level tests support the first routed view except on the harmful self-trained spiking axis (Appendix B.3). Against uniform escalation, routing differs by at most 0.06 pp on validation and all three V2 variants while recomputing 26 rather than 64 rows; on ImageNet-R and Sketch it trails uniform escalation by 0.41 and 0.30 pp.

## 5.2 Sample priority and action value are distinct

At $k_1 = 26$, margin ranking improves on matched random allocation in validation but narrows on ImageNet-R (Table 2); the gain is 2.4× random on validation and 1.6× on ImageNet-R. Because it uses labels to choose rows, the oracle bounds the view repairability that cheap margins leave unused.

Table 2: Selection at the same 26-of-64 first-view budget. Values are gain over each dataset's cheap prediction (pp); the oracle uses labels. Random is the exact expected gain over random routed subsets, not one sampled run.

| Selection | ImageNet val | ImageNet-R |
|---|---|---|
| Random | 0.684 | 0.389 |
| Margin | 1.668 | 0.630 |
| Label-aware oracle | 3.614 | 3.520 |

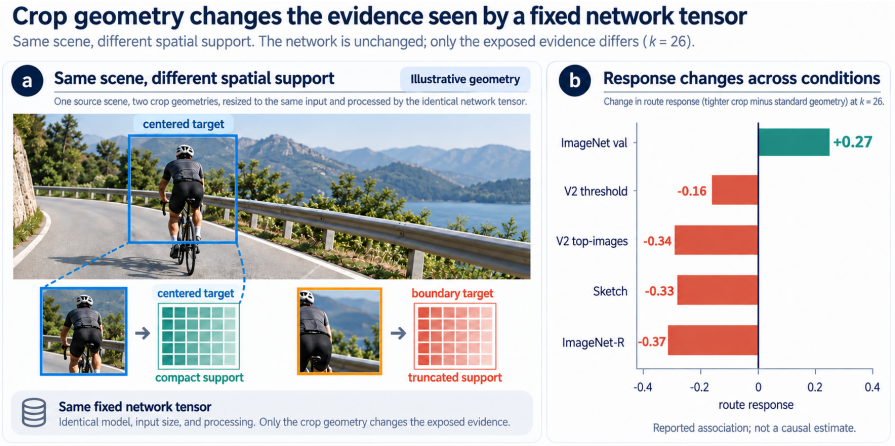


Figure 3: Crop support changes while the network input tensor remains fixed. The scene is schematic; the right panel reports tighter-minus-standard gain for one routed view at $k = 26$.

On held-out odd validation rows, a tightened-geometry three-stage action gains 2.520 pp at 44.1 ms extra pre-materialized GPU inference per batch, versus 1.112 pp at 49.9 ms for a full-batch flip ensemble. Appendix A.2 reports the full action frontier.

### 5.3 Geometry is a performance axis and a transportability risk

Crop geometry changes the evidence at the same input tensor. Tightening the resized short sides from 288/320 to 320/352 raises the full-validation three-stage gain from 1.966 to 2.404 pp at fixed input shape. On held-out odd rows, the corresponding gains are 2.012 and 2.520 pp.

The same intervention reverses on external data. At $k = 26$, tighter versus standard geometry changes single-view route gain by +0.268 pp on validation, but by $-0.160$ to $-0.373$ pp on V2-threshold, V2-top, Sketch, and R (Figure 3). The reversal rules out a universally better high-resolution recipe: action value changes under shift. We ship standard geometry; the cheap-view control is in Appendix A.3.

### 5.4 Budget width, depth, and stopping

The width curve rises quickly and flattens near 26 of 64 rows (Figure 4). At equal 59.4 ms extra inference, 26 two-view rows beat 52 one-view rows by about 0.3 pp; a fourth view gives no consistent held-out gain (Appendix A.2). At a tight validation budget the ladder reduces to plain margin ranking ($+0.04$ pp, $p = 0.65$); separation appears when multiple actions share a budget. At matched cost the ladder exceeds the best adaptive-TTA-style arm in 29 of 30 core stream–budget cells, narrowing on R and Sketch and reversing under severe corruption (Appendix E.1).

### 5.5 Serving and deployment accounting

Table 3 reports an additional, released 6.8M-parameter QSD scale. Its route-only arm improves accuracy, pre-materialized GPU latency, and throughput over the full-batch 288 replacement (281

Table 3: Second-scale QSD: offline gain on 50k images and batch-64 pre-materialized GPU service (three repeats). Energy is GPU-probe J/image. "Full-batch 288 replacement" recomputes every row at 288 without fusion; it is not the fused all-rows escalation arm of Appendix B.3.

| Arm | Gain (pp) | p50 (ms) | img/s | J/img |
|---|---|---|---|---|
| Cheap | 0.000 | 113.5 | 470 | 0.456 |
| Full-batch 288 replacement | 0.550 | 200.3 | 266 | 0.808 |
| Route 26/64 | 1.266 | 189.6 | 281 | 0.881 |
| Cascade 26:8 | 1.608 | 211.7 | 251 | 0.969 |

versus 266 images/s), although the GPU-only energy probe rises from 0.808 to 0.881 J/image. The 26:8 cascade adds 0.342 pp at higher latency. These results use this scale's own 80.816% cheap baseline and are reported separately from Table 1.

The standard $26{:}8{:}4$ service measures 89.31 ms p50 per pre-materialized 64-image request versus 44.64 ms for cheap inference, excluding the data path. The $2{,}048$-row probe keeps mean selected-width regret within $0.10$ pp, and its one-off cost amortizes to $0.54\%$ over ten thousand requests (Appendix F.1). Under a real decode/transform pipeline, routing improves wall-clock throughput in every tested supply regime ($+38\%$ with ten prefetch workers, $+7\%$ with four, $+60\%$ with a predecoded cache) and cuts full-pipeline GPU energy from $0.57$ to $0.37$ J/image; on the second scale it reaches $+55\%$ wall-clock throughput and $38\%$ lower energy than uniform-288. Its end-to-end energy scope differs from the GPU-only probe of Table 3.

### 5.6 What low margin explains across distribution shift

On validation, measured repair rates fall from 0.272 in the lowest-margin decile to 0.005 by the sixth; on ImageNet-R, the first six deciles remain near 0.058–0.070. This flatter profile explains why the narrow route leaves gains on R and Sketch; widening from 26 to 40 rows raises first-stage gains from 0.63 to 0.96 and from 0.88 to 1.09 pp. A source-fitted curve transfers poorly: source-fitted width selection incurs 0.487 pp regret on R and 1.360 pp on ImageNet-A (restricted 200-class protocol; Appendix C.1). Held-out curve reconstruction is nevertheless stable within a distribution: nested even-to-odd maximum errors are 0.062/0.098 pp on the original scale and 0.369/0.347 pp on the second scale.

Changing the cheap crop can inflate the gain by weakening the baseline (Appendix A.3); we retain the training-aligned crop for transfer.

### 5.7 Within-family replication and negative controls

The second released QSD scale confirms the first route (+1.266 pp) and eight-row cascade (+0.342 pp), but not a reliable third-stage increment (Table 3). ResNet-50 and ViT-B/16 controls on V2 gain 0.42 and 1.08 pp. Additional selectors, adaptive fusion, and a fourth view do not close the label-aware oracle gap.

### 5.8 Statistic audits: within-class, class-balanced, and calibration

Within-class and class-balanced decile audits preserve the declining block response, so the trend is not a class-mixture artefact; temperature rescaling changes the gain by at most 0.12 pp. Individual-row margin–utility correlations remain weak (AUC 0.67–0.73 for ranking repair candidates), so the law describes grouped means, not per-image outcomes (Appendix B.2).

### 5.9 Synthetic corruption streams

Utility decreases with margin decile in all nineteen corruption streams; the three weakened or degenerate cases (strong blur, both pixelate settings) are where escalation loses its advantage: strong blur gains $+0.62$ pp and pixelate $-0.13$ pp from an 8.06% cheap baseline, while grayscale and strong noise gain $+2.74$ and $+2.36$ pp (Appendix D.1). Corruption severity alone therefore does not determine view utility. The shipped cascade re-chains under corruption with the same stopping

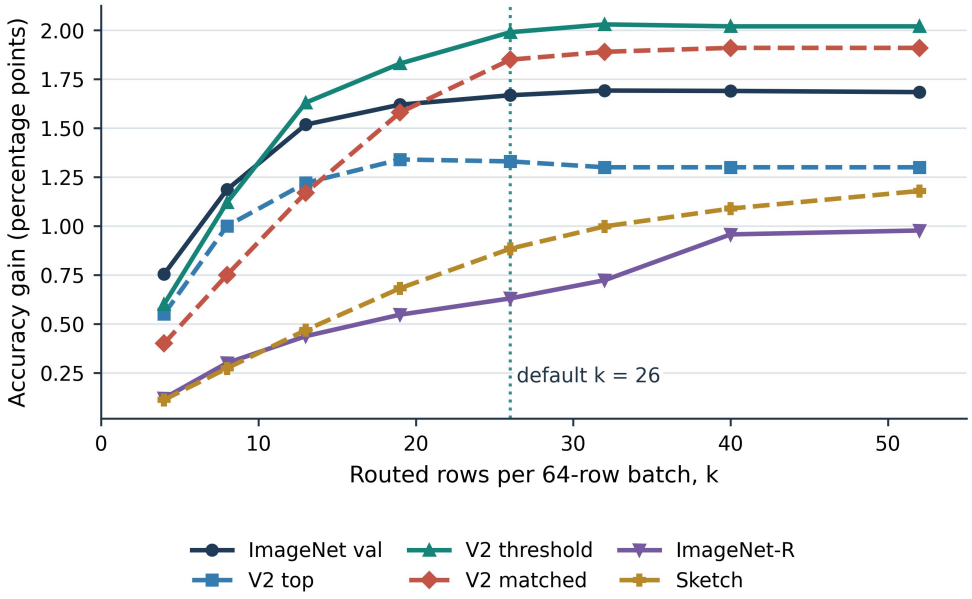


Figure 4: First-stage route-width curves across six distributions. Source-like curves saturate earlier than severe appearance shifts. Target labels measure these curves; the fixed policy uses none, and the calibrated variant uses labels only in its separate probe, not at serving inference.

pattern as the clean stream (stage-2 increments near zero or negative, stage 3 positive wherever repairable structure survives, e.g. grayscale $+0.295$ pp, $p = 10^{-4}$), and the routed gain survives every information-preserving corruption with retention ratios near one.

### 5.10 A SECOND, SELF-TRAINED SPIKING FAMILY

An independently trained, unquantized spiking ResNet on CIFAR-100 supplies a second SNN family with a different architecture and training recipe. We use it as an adverse action condition: its training-mismatched $32 \to 40$ view is harmful, with the default route losing 2.9 pp and uniform escalation 7.6 pp. Repair rate still declines with margin, but breaks are spread more widely, reversing net utility. A held-out probe instead selects a minimal, cheap-weighted action within 0.08 pp of its odd-half optimum. The escalated view is weaker (38.42% versus 60.13% cheap; 0.3/0.7 fusion reaches 52.56%), so the reversal is a property of the action, not of training. Details are in Appendix G.2.

### 5.11 DETECTION AND SEGMENTATION: A SCOPE CONDITION

Detection and segmentation test whether the response extends beyond top-1 correctness. For these scope tests, the probe calibrates on a sign-based per-image improve/degrade surrogate; the dataset-level AP/IoU changes below are evaluated separately, not optimized directly. On VOC detection with SSD-Lite (Sandler et al., 2018), flip merging raises mean per-image AP by 1.19; routing retains 0.87 (73%) at 41% of the escalation budget and beats matched random routing by about 0.4. A harmful scale action ($-0.88$ under uniform escalation) becomes slightly positive under Route13 ($+0.17$). On PASCALContext segmentation, flip and scale augmentation are beneficial uniformly but help high-margin images more; low-margin routing adds little. Across these dense tasks, routing helps when the added view's benefit is concentrated among uncertain inputs.

**Paired evidence.** The detection flip gain has a paired bootstrap 95% interval of $[+0.79, +1.60]$ AP; the harmful scale action $[-1.53, -0.18]$. A disjoint second split preserves the direction of the detection and segmentation findings. The same gate also refuses escalation on the two harmful dense axes ($320 \to 384$ and RetinaNet $600 \to 800$), landing within 0.05 units of their best cell where

Table 4: Previously unseen corruption streams. A frozen 2,048-row probe selects a width; regret is against the label-aware best width on disjoint evaluation rows (pp). All six favor escalation; Appendix F.1 evaluates the refusal path.

| | Posterize | Solarize | Equalize | Emboss | Sharpness | RGB swap |
|---|---|---|---|---|---|---|
| Selected $k$ | 32 | 40 | 26 | 64 | 26 | 32 |
| Regret (pp) | 0.02 | 0.03 | 0.03 | 0.00 | 0.01 | 0.07 |

uniform escalation loses 1.13 and 3.20 units (Appendix B.4). Appendix G.1 reports the full family and multiplicity corrections.

### 5.12 A FROZEN RULE ON PREVIOUSLY UNSEEN STREAMS

After a 36-stream audit, we froze a 2,048-row probe rule and evaluated six new 20,000-image corruption streams once. All were escalation-positive ($+1.8$ to $+2.8$ units); Table 4 reports selected width and held-out regret against the label-aware odd-half optimum, with full pricing in Appendix D.2.

Widths span 26–64 with median/worst regret 0.025/0.07 pp. Across the retrospective 36-stream audit of the unified rule, the probe predicts the sign of the measured slope on 32/36 streams, the gate outcome is insensitive to $\tau$ over $[0.05, 0.5]$, and every one of the six negative streams is held within 0.05 units of its best cell. On twelve later stressful streams (Appendix F.1), the gate refuses the most damaged stream (2.29% cheap accuracy), the other two negative streams lose at most 0.13 pp, and useful streams gain up to $+9.7$ pp.

## 6 DISCUSSION AND LIMITATIONS

The response curve separates sample priority from action value: uncertainty orders inputs, while repair–break determines whether a view is worth its cost. The flat R profile, crop reversal, and harmful spiking action show that margin or resolution alone is insufficient.

Four boundaries delimit the claim. *Budget saturation*: deeper stages add nothing on R and Sketch. *Action transfer*: a source-optimal crop reverses under shift. *Action failure*: a weaker view makes escalation harmful ($-2.9$ pp routed and $-7.6$ pp uniform on the second SNN). *Ordering failure*: segmentation benefits high-margin images more, so low-margin routing adds little. Other limits are shared label semantics, a self-trained second SNN, synthetic corruptions, and serving measurements that exclude queueing. Calibration also assumes a labeled probe per stream; annotation and drift-detection costs are not priced, whereas the fixed policy needs no target labels.

**Deployment recipe.** Form batch-64 requests, collect one labeled, target-like probe per stream, and keep the width that maximizes the probe's estimated cumulative utility, withholding escalation when the safety gate predicts a decline; then freeze that width and run margin-ordered escalation at serving time. The probe stays within 0.23 units across 36 audited streams and 0.07 on six prospective streams; outside calibrated regimes, use the cheap view.

## 7 CONCLUSION

GEAR allocates complementary views in a frozen 4-bit spiking Transformer using an action-conditioned response curve that links cheap-view margin, net repair utility, and compute budget. A fixed policy improves all six ImageNet-derived distributions tested, but useful escalation stays action- and distribution-dependent. The broader lesson is that sample priority and action value are distinct: the curve prices a candidate action before extra compute is committed, and on the tested streams a single labeled target-like probe sets the operating point without updating model weights.

## REPRODUCIBILITY STATEMENT

The paper specifies the checkpoint family, views, fixed routing counts, fusion rule, dataset protocols, selection split, and measured serving scope; Appendix A.1 gives the implementation order and Appendix G.3 records evidence sources and remaining reproduction checks. The submitted artifact contains the evaluation and analysis code behind the paper: the frozen-checkpoint harness and the GEAR policy primitives (margin-rank selection, nested routing, and probability fusion), the official geometry, precision-ladder, and serving benchmarks, the response-law, cascade, width, action-frontier, and paired-statistics analyses, the table-assembly scripts, and the supplement scripts for the probe, corruption, mechanism, prospective, dense-task, and second-family studies, together with unit tests and READMEs that map scripts to the corresponding sections and tables. Base-model training and evaluation build on the released QSD-Transformer code. Datasets, checkpoints, prediction caches, and per-row result files remain external because of size or access constraints, so full numerical reproduction additionally requires matching the dataset and preprocessing identifiers, model and cache hashes, and the latency configuration summarized in Appendix G.3.

## AI USE STATEMENT

Generative AI assisted with literature discovery, experiment-audit ideas and scripts, interpretation checks, manuscript structure, English drafting, LaTeX authoring, and reference formatting. Analyses and text were checked against available experimental artifacts and primary sources. No generative AI system is an author; the human authors retain responsibility for the code, numerical claims, citations, and submitted manuscript.

## REFERENCES

Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4421–4435, 2022. URL https://proceedings.mlr.press/v162/croce22a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021a. doi: 10.1109/ICCV48922.2021.00823. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.pdf.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

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://ai.meta.com/research/publications/multi-scale-dense-networks-for-resource-efficient-image-classification/.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html.

Chen Li, Edward G. Jones, and Steve Furber. Unleashing the potential of spiking neural networks with dynamic confidence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13350–13360, 2023. doi: 10.1109/ICCV51070.2023.01228. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_Unleashing_the_Potential_of_Spiking_Neural_Networks_with_Dynamic_Confidence_ICCV_2023_paper.html.

Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1308–1317, 2020. URL https://proceedings.mlr.press/v124/lyzhov20a.html.

Luca Mocerino, Roberto G. Rizzo, Valentino Peluso, Andrea Calimera, and Enrico Macii. AdapTTA: Adaptive test-time augmentation for reliable embedded ConvNets. In *2021 IFIP/IEEE 29th International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 1–6, 2021. doi: 10.1109/VLSI-SoC53125.2021.9606980. URL https://ieeexplore.ieee.org/document/9606980/.

Xuerui Qiu, Malu Zhang, Jieyuan Zhang, Wenjie Wei, Honglin Cao, Junsheng Guo, Rui-Jie Zhu, Yimeng Shan, Yang Yang, and Haizhou Li. Quantized spike-driven transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5J9B7Sb8rO.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, volume 97, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pp. 409–424, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Xin_Wang_SkipNet_Learning_Dynamic_ECCV_2018_paper.html.

Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20062–20072, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_AdaFocus_V2_End-to-End_Training_of_Spatial_Dynamic_Networks_for_Video_CVPR_2022_paper.html.

Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8817–8826, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_BlockDrop_Dynamic_Inference_CVPR_2018_paper.html.

Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2369–2378, 2020. doi: 10.1109/CVPR42600.2020.00244. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yang_Resolution_Adaptive_Networks_for_Efficient_Inference_CVPR_2020_paper.html.

Youla Yang. SA-TTS: Stress-aware test-time scaling for vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11480–11487, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/Viscale/html/Yang_SA-TTS_Stress-Aware_Test-Time_Scaling_for_Vision_Models_CVPRW_2026_paper.html.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=frE4fUwz_h.

# APPENDIX

The appendix provides implementation details, complete sweeps, per-condition statistics, and robustness analyses that expand on the method and results stated in the main paper.

## A IMPLEMENTATION AND CORE ABLATIONS

### A.1 EXECUTABLE POLICY AND NOTATION

For each batch, compute the cheap logits once and the probability margin of every row. Stable-sort by increasing margin and select 26 of 64. Recompute the selected rows at resize-288/crop-288; apply $P_1 = 0.3P_0 + 0.7P_1^{\text{view}}$ and leave all other rows unchanged. Select the eight lowest post-mixture margins *inside* this set, compute their horizontally flipped 288@288 view, and apply $P_2 = 0.3P_1 + 0.7(0.2P_0 + 0.8P_{1f})$. Select four lowest post-stage-2 margins inside the eight-row set, recompute at 320@320, and apply $P_3 = 0.7P_2 + 0.3P_{320}$. Each stage re-ranks only its parent set; prior predictions are carried forward rather than recomputed. Interpolation and image-normalization settings are part of the implementation configuration; incomplete batches lie outside the evaluated batch-64 protocol.

### A.2 ACTION AND BUDGET FRONTIERS

This section expands the action- and budget-frontier analyses of Sections 5.3–5.4. Figure 5 plots measured batch-64 pre-materialized GPU latency against accuracy gain on odd validation rows. At 59.4 ms per request, a second view on 26 routed rows overtakes a first view spread over 52 rows by about 0.3 pp; the tight-crop point is an in-distribution variant, while the fixed cross-domain policy uses standard geometry.

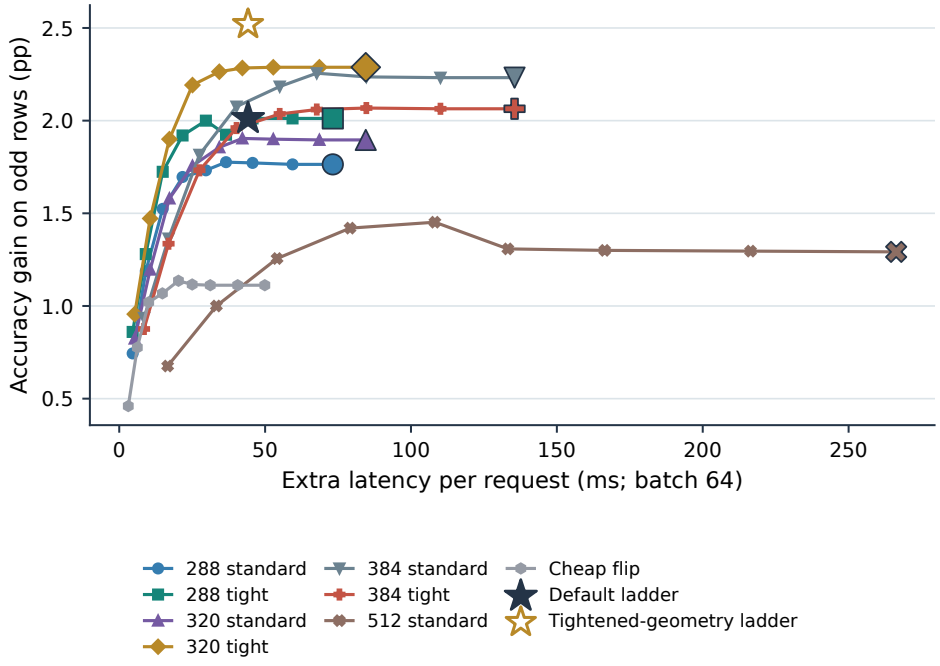


Figure 5: Measured action frontier on ImageNet validation. View actions are priced by batch-64 pre-materialized GPU inference. The filled star is the shipped default ladder; the open star is the tightened-geometry variant, an in-distribution option only. The cross-domain default remains the standard geometry.

### A.3 CHEAP-VIEW BASELINE CONTROL

Replacing 224@256 with 224@320 reduces the cheap ImageNet baseline from 76.902% to 76.008%, while the final standard-geometry system rises from 78.868% to 79.238%. The apparent gain over cheap becomes 3.230 pp, but final accuracy rises only 0.370 pp. Source gains from adjusting cheap and escalated crops are largely non-additive, and the adjusted cheap crop loses on Sketch and R; we therefore retain the training-aligned crop for transfer.

### A.4 ALTERNATIVE COMPUTE AXES: WEIGHT PRECISION AND SPIKING TIMESTEPS

Weight precision is not treated as a deployment-time knob: the released checkpoint was trained jointly at its fixed four-bit grid. Post-hoc re-quantisation moves accuracy below the released level on both scales; the response profile is unchanged, so no bit-width claim enters the method. The timestep axis is saturated in the same way: multiplying $T$ by eight costs $10.2$–$10.9\times$ and moves accuracy by at most $+0.07$ pp, while one geometry step costs $1.79\times$ and buys $+1.37$ pp (Table 5); a re-run at a different batch size flips the sign of the $T{=}2$ increment, so the whole timestep effect is inside the numerical realisation noise. Together with the weight/statistics lock-in of the main text, the practical reading is that the controllable axis is the input, not the arithmetic.

Table 5: Spiking timesteps versus the geometry axis on 10k stratified rows. Absolute accuracy can differ by up to 0.13 pp from the batch-64 protocol because this run uses batch 32; only deltas are compared. Multiplying $T$ by eight costs $10.2$–$10.9\times$ and changes accuracy by at most $+0.07$ pp, while one geometry step costs $1.79\times$ and gains $+1.37$ pp.

| Geometry | $T{=}1$ | $T{=}2$ | $T{=}4$ | $T{=}8$ | $\Delta T{=}8$ | cost ratio |
|---|---|---|---|---|---|---|
| 224@256 | 77.25 | 77.41 | 77.25 | 77.22 | $-0.03$ | $10.9\times$ |
| 288@288 | 78.62 | 78.63 | 78.67 | 78.69 | $+0.07$ | $10.2\times$ |

Cost p50 (ms/batch): 224@256: 45/101/224/487; 288@288: 80/180/400/818.

## B CALIBRATION, STATISTICS, AND ALLOCATION AUDITS

**Appendix shorthand.** Labels such as QSD-T/QSD-L and suffixes such as official/asymmetric/mirror/adaptive identify distinct cached experimental configurations used by individual audits. Their cheap baselines are reported locally and are never pooled or compared across caches.

### B.1 PROBE-SIZE AND CLASS-ROUTER AUDITS

The response curve probe uses $N$ labeled even-indexed rows to define margin-decile boundaries and estimate block utility for candidate views; we score the estimated curve and selected operating point on the odd half, averaging three probe draws. At $N = 4{,}096$, the three-draw mean curve error is 0.29 pp (QSD-T), 0.36 pp (QSD-L), 0.20 pp (ImageNet-V2/QSD), 0.14 pp (ResNet-50-V2) and 0.42 pp (ViT-B/16), with mean selected-width regret of 0.000–0.107 pp; the worst single draw on the hardest condition is 0.61 pp curve error and 0.16 pp regret (Table 6 reports the worst draw per condition). At $N = 1{,}024$, the mean curve error ranges from 0.47 to 1.30 pp across those settings. Joint $(k, w)$ selection on same-distribution validation settings has 0.004–0.012 pp regret for the two QSD scales; cross-domain or cross-architecture settings reach 0.52 pp. The probe-size result therefore estimates an operating point for a sampled target-like stream; cross-domain transfer remains a separate problem.

A class-conditioned router — margin minus a fitted per-predicted-class fixability term with the held-out-optimal coefficient — never beats plain margin at $k{=}26$ or $k{=}40$ on any of the five conditions, so this closed negative is recorded as a remark rather than a table; the complete probe-size sweep is tabulated below.

Table 6: Probe-size sweep on the classification panel: curve error and selected-width regret as the labelled probe grows. All numbers are the worst of three probe draws; the three-draw means are roughly one quarter to one half lower (e.g. val-T at $N$=4,096: mean 0.29 pp curve error, 0.013 pp regret). The $N$=1,024 regret column prices the joint $(k, w)$ selection; the remaining columns score width-only selection.

| Condition | Rows | Best $k$ | 1k err | 1k regret | 2k regret | 4k err | 4k regret |
|---|---|---|---|---|---|---|---|
| imagenet-val-T | 50,000 | 32 | 1.07 | 0.004 | 0.032 | 0.427 | 0.032 |
| imagenet-val-L | 50,000 | 26 | 0.84 | 0.152 | 0.152 | 0.606 | 0.152 |
| imagenetv2-matched-T | 10,000 | 40 | 1.66 | 0.340 | 0.040 | 0.220 | 0.000 |
| resnet50-v2-on-v2 | 10,000 | 19 | 0.76 | 0.020 | 0.000 | 0.166 | 0.000 |
| vit-b16-on-v2 | 10,000 | 40 | 1.80 | 0.740 | 0.160 | 0.506 | 0.160 |
| spiking-resnet-cifar100 | 10,000 | 6 | 2.42 | 0.860 | 0.000 | 1.343 | 0.000 |

### B.2 Router audit: per-condition numbers

Section 5.8 summarizes three audit questions; here we list the per-condition values: row-level margin–utility correlation, AUC for ranking repair candidates, within-class and class-balanced checks for mixture effects, and temperature sensitivity of the operating point. All nine tested conditions behave qualitatively identically.

Table 7: Per-condition router audit. Row-level Spearman between margin and utility is weak (the law is a block-mean statement); the margin nonetheless ranks repair candidates (AUC of low margin predicting a repair); the profile survives class conditioning (fraction of classes with negative within-class median correlation and its exact sign test, and the class-balanced decile-curve Spearman); and the operating point is calibration-insensitive (gain spread across $T \in \{0.5, 2\}$).

| Condition | Rows | Cheap (%) | Row $\rho$ | $\text{AUC}_{\text{repair}}$ | Within-class neg. | Sign $p$ | Class-bal. $\rho$ | Temp. spread (pp) |
|---|---|---|---|---|---|---|---|---|
| imagenet-val-T | 50,000 | 77.47 | −0.102 | 0.714 | 70% | 2.2e−34 | −0.806 | 0.004 |
| imagenet-val-L | 50,000 | 80.82 | −0.080 | 0.719 | 67% | 2.0e−24 | −0.816 | 0.012 |
| imagenetv2-matched-T | 10,000 | 65.43 | −0.087 | 0.665 | 63% | 8.4e−09 | −0.957 | 0.100 |
| imagenetv2-threshold0.7-T | 10,000 | 74.07 | −0.106 | 0.696 | 67% | 1.5e−13 | −0.844 | 0.040 |
| imagenetv2-top-images-T | 10,000 | 79.69 | −0.084 | 0.726 | 61% | 8.6e−06 | −0.472 | 0.020 |
| imagenet-r-T | 30,000 | 37.39 | −0.028 | 0.503 | 62% | 5.3e−04 | −0.818 | 0.113 |
| imagenet-sketch-T | 50,889 | 26.23 | −0.036 | 0.613 | 57% | 8.6e−06 | −0.976 | 0.031 |
| resnet50-v2-on-v2 | 10,000 | 69.95 | −0.035 | 0.809 | 59% | 3.1e−03 | −0.888 | 0.000 |
| vit-b16-on-v2 | 10,000 | 69.81 | −0.059 | 0.698 | 60% | 2.3e−05 | −0.964 | 0.040 |

### B.3 Paired statistics of the classification panel

Table 8 attaches paired image-level statistics to the three classification contrasts – routed26 vs cheap, vs uniform escalation at the same fusion weight, and vs matched random routing – on all ten combinations of the response-law panel. Holm-adjusted exact McNemar $p$-values for the routed26-minus-cheap contrast: val-T $4.6\times10^{-58}$, val-L $8.2\times10^{-33}$, V2 matched $8.1\times10^{-13}$, V2 threshold $9.2\times10^{-15}$, V2 top-images $3.1\times10^{-8}$, R $8.0\times10^{-10}$, Sketch $1.5\times10^{-23}$, ResNet-50 $1.2\times10^{-2}$, ViT-B/16 $8.0\times10^{-6}$, spiking CNN $6.3\times10^{-18}$ (negative). For routed26-minus-uniform the only decisive rows are R ($4.0\times10^{-5}$), Sketch ($2.9\times10^{-5}$) and the spiking CNN ($1.5\times10^{-76}$, where uniform is far worse); all other rows are tie-dominated ($p = 1$). For routed26-minus-random the selection edge is significant on val, all three V2 variants, Sketch and ViT ($p \leq 6.2\times10^{-4}$) and unresolved on R (0.19), ResNet-50 (0.25) and the spiking CNN (0.48), consistent with the scope condition that the pay is in repair rather than selection on shifted or harmful axes. The ViT-B/16 control interpolates its positional embeddings from the 224 pretraining grid for 288 input.

### B.4 One probe, one rule: unified decision audit

This appendix runs the probe protocol of Appendix B.1 as a *single operational rule* on every cached stream: ten classification model/data combinations, nineteen synthetic corruption streams, four detection axes (including the second split), and three segmentation axes (including the second split). Each stream is halved by row index; the even half donates a 2,048-row probe (all rows when the half

Table 8: Paired image-level statistics of the classification panel: mean effect in pp with 2,000-draw bootstrap 95% CI; * marks CIs excluding zero. "Uniform" is fused all-rows escalation at the same weight ($w_{\text{cheap}}$=0.3); routed26 and random26 keep the same row count. McNemar $p$-values (Holm per contrast) are listed in the text.

| Combination | Route26−cheap | Route26−uniform | Route26−random26 |
|---|---|---|---|
| val (QSD-T) | $+1.66\,[+1.46, +1.86]^*$ | $+0.01\,[-0.01, +0.04]$ | $+1.05\,[+0.91, +1.21]^*$ |
| val (QSD-L) | $+1.27\,[+1.06, +1.47]^*$ | $+0.01\,[-0.01, +0.03]$ | $+0.76\,[+0.60, +0.92]^*$ |
| V2 matched | $+1.85\,[+1.38, +2.34]^*$ | $-0.06\,[-0.19, +0.06]$ | $+1.14\,[+0.76, +1.54]^*$ |
| V2 threshold | $+1.99\,[+1.52, +2.48]^*$ | $-0.03\,[-0.12, +0.05]$ | $+1.34\,[+0.97, +1.71]^*$ |
| V2 top-images | $+1.33\,[+0.89, +1.82]^*$ | $+0.03\,[-0.02, +0.09]$ | $+0.81\,[+0.46, +1.14]^*$ |
| ImageNet-R | $+0.68\,[+0.48, +0.89]^*$ | $-0.41\,[-0.59, -0.23]^*$ | $+0.19\,[-0.01, +0.40]$ |
| Sketch | $+0.88\,[+0.72, +1.05]^*$ | $-0.30\,[-0.42, -0.18]^*$ | $+0.30\,[+0.15, +0.45]^*$ |
| ResNet-50 on V2 | $+0.42\,[+0.10, +0.75]^*$ | $+0.00\,[+0.00, +0.00]$ | $+0.21\,[-0.04, +0.47]$ |
| ViT-B/16 on V2 | $+1.08\,[+0.63, +1.53]^*$ | $-0.02\,[-0.10, +0.06]$ | $+0.76\,[+0.41, +1.13]^*$ |
| spiking CNN (CIFAR-100) | $-2.90\,[-3.53, -2.29]^*$ | $+4.67\,[+4.18, +5.19]^*$ | $+0.22\,[-0.38, +0.80]$ |

is smaller) that estimates the blockwise response curve by margin decile, and the odd half scores the shipped cell. Utility units are the net repair share in percentage points (rows whose per-image metric improves minus rows that degrade); on classification this coincides with top-1 pp. On dense tasks this sign-based quantity is a calibration surrogate, distinct from the mean per-image AP/IoU reported in Appendix G.1; we do not treat it as direct optimization of dataset-level AP or IoU. The safety gate falls back to the minimum route width ($k$=6) when the probe predicts the curve to fall by more than $\tau$ between $k$=13 and $k$=52.

Findings. (i) The probe predicts the sign of the measured odd-half slope on $32/36$ streams; the four misses (validation T/L, mild JPEG, strong blur) have flat or non-monotone tails. (ii) The plain argmax rule selects an operating point with median regret $0.008$ and worst regret $0.67$; the $\tau$=0.15 gate reduces the worst case to $0.23$ without changing other selected operating points, and the outcome is insensitive to $\tau$ over $[0.05, 0.5]$. (iii) On the six streams whose measured slope is negative—the self-trained spiking CNN, both pixelate corruptions, the detection 320→384 axis, the RetinaNet 600→800 axis, and strong blur—the gated rule stays within $0.05$ of the best cell on $6/6$ and never pays the uniform price: uniform loses $7.6$ units on the spiking CNN where the gated rule loses only $0.3$, and $3.20$ units on the RetinaNet axis where the gated rule gains $+0.27$. The same probe therefore estimates both the response curve and whether escalation is worthwhile, using only the labeled probe rows for stream-level calibration.

Table 9: One probe and one safety-gated rule ($\tau$=0.15) on representative streams. Utility is the net repair share (pp) on the odd half; "uniform" is the all-rows-escalated price. The rule matches the odd-half optimum on $15/36$ streams, stays within $0.23$ regret everywhere, and prices all six harmful axes within $0.05$ of their best cell.

| Stream | Rows | Gate | $k_\tau$ | Gain | best $k$ | Best | Uniform | Regret |
|---|---|---|---|---|---|---|---|---|
| QSD-T / val | 50,000 | −0.05 | 19 | +1.65 | 32 | +1.68 | +1.66 | 0.03 |
| QSD-L / val | 50,000 | −0.03 | 13 | +0.98 | 26 | +1.13 | +1.12 | 0.15 |
| self-trained spiking CNN (harmful) | 10,000 | −6.30 | 6 | −0.32 | 6 | −0.32 | −7.62 | 0.00 |
| corruption jpeg | 20,000 | −0.01 | 19 | +1.64 | 40 | +1.81 | +1.81 | 0.17 |
| corruption blur-strong | 20,000 | +0.10 | 52 | +0.56 | 13 | +0.60 | +0.56 | 0.04 |
| corruption pixelate-50k (harmful) | 50,000 | −0.74 | 6 | −0.02 | 6 | −0.02 | −0.50 | 0.00 |
| detection flip-merge | 3,000 | +0.17 | 64 | +2.87 | 64 | +2.87 | +2.87 | 0.00 |
| detection flip-merge, 2nd split | 3,000 | +2.22 | 64 | +2.20 | 64 | +2.20 | +2.20 | 0.00 |
| detection RetinaNet scale (harmful) | 3,000 | −1.01 | 6 | +0.27 | 6 | +0.27 | −3.20 | 0.00 |
| detection scale 320→384 (harmful) | 3,000 | −2.46 | 6 | +1.47 | 6 | +1.47 | −1.13 | 0.00 |
| segmentation flip average | 2,500 | +10.18 | 64 | +22.08 | 64 | +22.08 | +22.08 | 0.00 |
| segmentation flip, 2nd split | 2,500 | +14.88 | 64 | +26.08 | 64 | +26.08 | +26.08 | 0.00 |
| segmentation scale 480→640 | 2,500 | +12.55 | 64 | +20.40 | 64 | +20.40 | +20.40 | 0.00 |

# C TRANSFER AND MECHANISM

## C.1 TRANSFER BEHAVIOR OF THE REPAIR PROFILE

The block audit finds maximum block-optimal minus greedy gaps of 0.007/0.012 pp on the two QSD ImageNet scales but 0.157 pp on ImageNet-R. Its label-aware row-oracle minus greedy gaps at $k$ = 26 remain 1.868/2.140 pp on the two scales and 2.807 pp on R. Same-source curve reconstruction

errors at most 0.026 pp are in-sample descriptions; nested even-to-odd and odd-to-even maximum errors are 0.062/0.098 pp on the original-scale validation set and 0.369/0.347 pp on the second scale. Across domains, the source fingerprint can err by 1.373 pp over the evaluated budgets on ImageNet-A. Together, these measurements show that zero-calibration gain forecasts do not transfer across domains. The plotted decile repair rates were estimated using the same dataset labels subsequently used to check each budget. Equation 3 nearly reconstructs the budget curve by design. Fitting repair rates on even ImageNet validation rows and applying them to target cheap-logit margins gives a different result: at $k = 26$, predicted and measured route gains are 2.199 and 1.850 pp on V2-matched, and 3.882 and 0.884 pp on Sketch. The cross-domain quantitative estimate therefore departs sharply from the measured gain, while the within-dataset repair–harm accounting remains exact. Source-to-target prediction needs separate calibration or a validated domain model.

### C.2 Mechanism diagnostics

The error-decomposition identity of Section 3.4 holds at machine precision on five model/data conditions plus the self-trained spiking CNN (maximum residual $9 \times 10^{-17}$). The independence ratio $J/(AB)$ rises from 1.05–1.26 at the lowest margin decile to 40–63 at the top, where $J/A \approx J/B \approx 0.99$; bottom-two-decile shares are 0.257–0.956 for breaks and 0.478–0.923 for repairs (the 0.257 bound is the self-trained CNN's harmful axis). The four-scalar Gaussian witness reproduces the decile utility shape at correlation 0.979 with a 2.4 pp amplitude mismatch.

**Applicability map.** The witness also answers *when* the law should hold. Sweeping the escalated-view advantage $\Delta\mu$ and its noise scale $\sigma_1$ (Figure 6; $\rho$=0.6 grid, robustness at $\rho \in \{0.3, 0.9\}$) labels each cell: **P** the block utility decays with margin and escalation pays; **H** it decays but escalation loses on average; **X** it is non-monotone. The shipped calibration sits deep in P ($\rho_u = -0.73$; decile-1 utility $+0.099$, decile-10 $-0.005$), and every tested classification family with a mildly stronger escalated view falls in P as well. The empirically harmful axes correspond to H: a weaker escalated view at large noise, matching the self-trained spiking 32→40 and pixelate profiles (repair rate still monotone, net utility negative). The X region appears only for a weaker view at low noise and grows with the shared-noise correlation (0 of 99 cells at $\rho$=0.3, 46 at $\rho$=0.9); no tested stream falls there. The segmentation reversal falls outside this binary-correctness caricature because its per-image metric is continuous IoU. The mechanism numbers of Section 3.4 are reported per condition

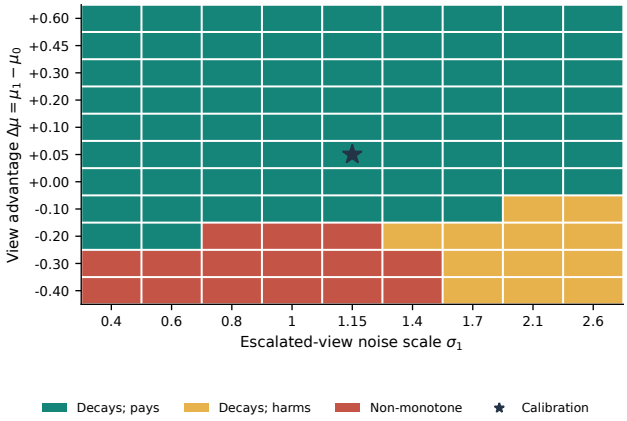


Figure 6: Applicability map of the witness model over the escalated-view advantage and noise scale. The shipped calibration sits in the payoff regime P; the empirically harmful axes (weaker view, large noise) fall into H; non-monotone X appears only for a weaker view at low noise.

below: repair/break totals and bottom-two-decile shares with the recipe-transfer regrets on odd rows (Table 10); the identity itself holds at machine precision on every condition (residuals $\leq 9 \times 10^{-17}$).

Table 10: Repair–break decomposition on the mechanism-audited model/data conditions: totals, the share of breaks and repairs sitting in the bottom two margin deciles, decile-profile Spearman, and the recipe-transfer numbers (shipped-cell regret on odd rows and the odd-half-best $(k, w)$ cell). The self-trained spiking CNN is the condition whose break share drops (0.26) because its harmful axis damages confident rows too.

| Condition | Repairs | Breaks | Break b2 | Repair b2 | $\rho(u, \mathrm{dec.})$ | Shipped regret (pp) | Best cell |
|---|---|---|---|---|---|---|---|
| val-T-official | 1765 | 940 | 0.793 | 0.842 | $-0.795$ | 0.004 | $(k{=}32, w{=}0.3)$ |
| val-T-asymmetric | 1852 | 1012 | 0.779 | 0.826 | $-0.844$ | 0.040 | $(k{=}40, w{=}0.4)$ |
| val-T-mirror | 1807 | 965 | 0.775 | 0.836 | $-0.853$ | 0.052 | $(k{=}32, w{=}0.2)$ |
| val-T-adaptive | 1811 | 966 | 0.807 | 0.828 | $-0.800$ | 0.000 | $(k{=}26, w{=}0.3)$ |
| val-L | 1703 | 1075 | 0.816 | 0.854 | $-0.855$ | 0.108 | $(k{=}26, w{=}0.4)$ |
| v2-matched-T | 430 | 239 | 0.661 | 0.635 | $-0.969$ | 0.320 | $(k{=}40, w{=}0.4)$ |
| resnet50-v2 | 155 | 113 | 0.956 | 0.923 | $-0.888$ | 0.120 | $(k{=}32, w{=}0.2)$ |
| vit-b16-v2 | 334 | 224 | 0.754 | 0.746 | $-0.740$ | 0.020 | $(k{=}40, w{=}0.3)$ |
| spiking-resnet-cifar100 | 504 | 1261 | 0.257 | 0.478 | $+0.152$ | 3.180 | $(k{=}6, w{=}0.7)$ |

# D Distribution shift and prospective evaluation

## D.1 Corruption-stream audit

Each cache in the nineteen-condition registry stores cheap 224@256 and escalated 288@288 logits under the protocol of the main analyses. Table 11 reports the complete registry (all nineteen conditions; pixelate degenerate and strong blur weakened are the documented exceptions), and Figure 7 draws it as a response profile: the exceptions are where the routed gain collapses.

Table 11: The complete nineteen-condition corruption registry (identical protocol: cheap $224@256$ and escalated $288@288$ logits; 15 conditions use 20k rows and four large caches use 50k rows). "4k-probe err" is the worst odd-half curve error over three 4,096-row probe draws; "regret" is the selected-width regret of the 4k probe. The two pixelate levels and strong blur are the weakened or degenerate cases.

| Condition | Rows | Cheap (%) | Routed@26 (pp) | $\rho(u, \mathrm{dec.})$ | Block gap (pp) | 4k-probe err (pp) | Regret (pp) |
|---|---|---|---|---|---|---|---|
| clean-20k | 20,000 | 78.08 | $+1.615$ | $-0.862$ | 0.025 | 0.894 | 0.030 |
| blur | 20,000 | 72.72 | $+1.730$ | $-0.902$ | 0.033 | 0.225 | 0.110 |
| jpeg | 20,000 | 74.73 | $+1.605$ | $-0.730$ | 0.035 | 0.675 | 0.410 |
| noise | 20,000 | 73.70 | $+1.910$ | $-0.844$ | 0.014 | 0.435 | 0.320 |
| brightness | 20,000 | 76.17 | $+1.985$ | $-0.693$ | 0.015 | 0.388 | 0.030 |
| color | 20,000 | 77.33 | $+1.785$ | $-0.938$ | 0.007 | 0.560 | 0.020 |
| contrast | 20,000 | 77.08 | $+2.100$ | $-0.910$ | 0.022 | 0.566 | 0.070 |
| gray | 20,000 | 69.62 | $+2.740$ | $-0.976$ | 0.029 | 0.762 | 0.030 |
| fog | 20,000 | 77.47 | $+1.610$ | $-0.985$ | 0.001 | 0.507 | 0.030 |
| pixelate | 20,000 | 8.06 | $-0.130$ | $-0.421$ | 0.145 | 0.331 | 0.180 |
| rotate | 20,000 | 77.25 | $+1.780$ | $-0.938$ | 0.005 | 0.477 | 0.090 |
| translate | 20,000 | 77.36 | $+1.730$ | $-0.832$ | 0.015 | 0.819 | 0.010 |
| blur-strong | 20,000 | 56.76 | $+0.620$ | $-0.374$ | 0.068 | 0.492 | 0.020 |
| jpeg-mild | 20,000 | 70.50 | $+2.025$ | $-0.966$ | 0.006 | 0.891 | 0.020 |
| noise-strong | 20,000 | 67.94 | $+2.360$ | $-0.681$ | 0.036 | 0.646 | 0.100 |
| fog-50k | 50,000 | 76.88 | $+1.688$ | $-0.957$ | 0.002 | 0.552 | 0.024 |
| gray-50k | 50,000 | 68.50 | $+2.490$ | $-0.976$ | 0.002 | 0.780 | 0.060 |
| pixelate-50k | 50,000 | 7.77 | $-0.176$ | $-0.539$ | 0.092 | 0.409 | 0.000 |
| rotate-50k | 50,000 | 76.53 | $+1.880$ | $-0.994$ | 0.001 | 0.825 | 0.272 |

Two further checks need no table: the shipped cascade re-chains under corruption with the same stopping pattern as the clean stream (stage-2 increments near zero or negative; stage 3 positive wherever repairable structure survives, e.g. grayscale $+0.295$ pp, $p{=}10^{-4}$), and the router's gain survives every information-preserving corruption with retention ratios near one, so the law survives distribution damage as long as the cheap view retains repairable signal.

## D.2 Prospective evaluation on previously unseen streams

To test the unified rule beyond the original registry, we fixed the protocol and success criteria in advance and generated six previously unseen corruption streams: posterize at 4 bits, solarize at threshold 128, histogram equalization, emboss, sharpness 3.0, and red-blue channel swap. Each contains 20,000 rows and follows the same sampling and geometry protocol as the nineteen-stream registry.

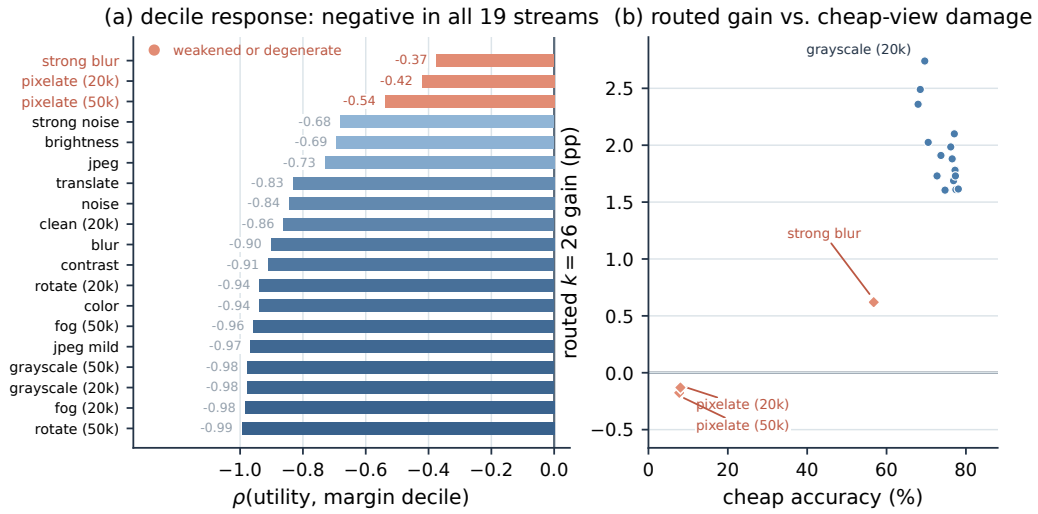


Figure 7: The nineteen-condition corruption registry. (a) Margin-decile/utility correlation is negative in all nineteen conditions; three are weakened or degenerate (strong blur and both pixelate levels), and these are where routed gain collapses. (b) Routed gain is largest under moderate damage and dies only when the cheap view has almost nothing left to repair: pixelate sits at $8.1/7.8\%$ cheap accuracy with a non-positive gain, while strong blur keeps a positive gain at $56.8\%$.

Table 12 reports the outcome. All six streams are escalation-positive ($\rho \in [-0.94, -0.38]$), so this batch evaluates the pricing path and the gate is never triggered. The $k$-only rule stays within $0.07$ units of the odd-half optimum on every stream (median $0.025$); the joint $(k, w)$ rule reaches at most $0.16$ regret. Thus the $k$-only rule meets the $0.10$ regret criterion on all six streams, while the joint rule exceeds it on solarize ($0.16$). Emboss provides an extreme case: cheap accuracy falls to $18.6\%$, yet escalation gains $+2.1$ units at zero regret. Harmful-stream behavior is evaluated separately in the retrospective 36-stream audit and the two harmful dense axes.

Table 12: Prospective evaluation on six previously unseen corruption streams, with protocol and criteria fixed before stream generation. Utility is the net repair share (pp) on the odd half. R1 = $k$-only rule with the $\tau$=0.15 safety gate; R2 = joint $(k, w)$ rule. The gate is inactive because all six streams are escalation-positive.

| Stream | Gate | $k_{R1}$ | Gain | Best $k$ | Best | Uniform | Regret | $\rho$ | R2 pick | Regret R2 |
|---|---|---|---|---|---|---|---|---|---|---|
| posterize | +0.32 | 32 | +1.81 | 52 | +1.83 | +1.83 | 0.02 | −0.80 | (32, 0.3) | 0.02 |
| solarize | +2.01 | 40 | +2.82 | 64 | +2.85 | +2.85 | 0.03 | −0.94 | (40, 0.2) | 0.16 |
| equalize | +0.36 | 26 | +2.12 | 32 | +2.15 | +2.13 | 0.03 | −0.88 | (26, 0.2) | 0.09 |
| emboss | +0.96 | 64 | +2.11 | 64 | +2.11 | +2.11 | 0.00 | −0.82 | (64, 0.3) | 0.00 |
| sharpness | +0.17 | 26 | +1.81 | 19 | +1.82 | +1.79 | 0.01 | −0.38 | (26, 0.3) | 0.09 |
| rgb-swap | +0.90 | 32 | +2.31 | 64 | +2.38 | +2.38 | 0.07 | −0.93 | (32, 0.2) | 0.07 |

# E Matched-cost comparison and allocation audits

## E.1 Matched-cost separation from adaptive test-time policies

Adaptive test-time augmentation selects views from input-dependent statistics (Mocerino et al., 2021; Yang, 2026), so the shipped policy is one member of that family. We separate the contribution by pricing every family under one protocol instead of comparing published operating points: identical candidate views (resize-288, flipped 288, 320), identical probability fusion, the margin statistic wherever a selector uses one, and the *measured* serving cost law of Equation 2 so that no arm is evaluated at a budget it cannot afford. The adaptive side is instantiated twice (a stress-aware top-$k$ selector, "SA-TTS-style", and a confidence-adaptive view selector, "AdapTTA-style"), single-signal selectors (margin, entropy, least-confidence, stress, random) and a label-aware oracle complete the field. Figure 8 shows the shape of the comparison and Table 13 the matched-cost gaps and the paired tests.

The ladder keeps a positive matched-cost gap in 29 of 30 stream×budget cells (+0.01 to +0.88 pp; the only exception is V2-threshold at 80 ms, −0.02 pp), and the paired tests are decisive where it matters most: +0.38/+0.41 pp ($p \leq 4.5\times10^{-6}$) against the adaptive styles on validation, and +0.64/+0.94 pp on V2-matched. The boundary is stated in the same breath: on R and Sketch the families are statistically indistinguishable from SA-TTS-style at the matched budget (non-significant 0.10/0.02 pp; against AdapTTA-style, Sketch significantly favors GEAR at +0.29 pp, $p=2\times10^{-4}$), and on val (QSD-T) at a tight budget the ladder reduces to plain margin ranking (+0.04 pp, $p=0.65$): the separation lives in the cascade and the pricing rule, not in uncertainty beating margin as a ranker. Under heavy corruption the gap can change sign (jpeg-mild −0.14, blur-strong −0.15 pp, non-significant), consistent with the response-physics boundary of the corruption audit. A six-draw re-run leaves the gap signs unchanged except for one budget point (V2 threshold at 80 ms, −0.02 pp), so the conclusion is not a probe artefact.

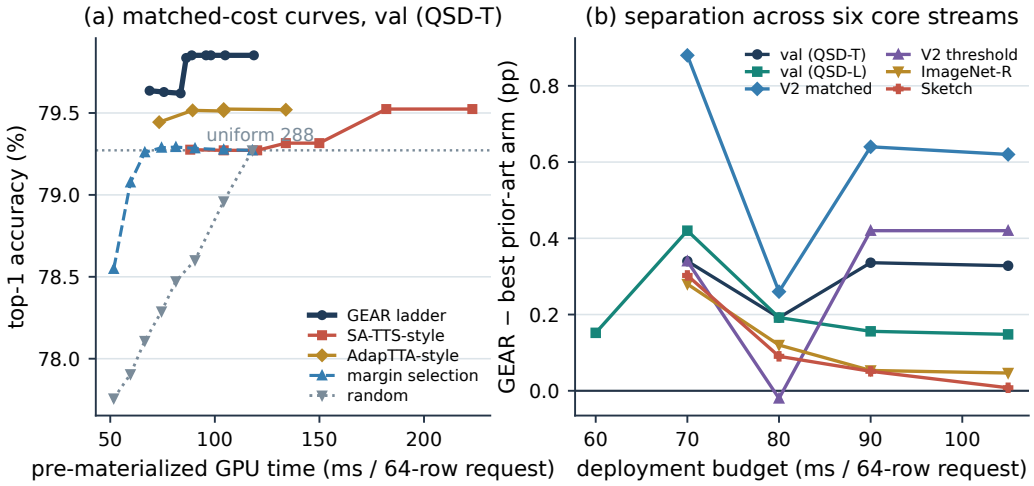


Figure 8: Matched-cost adaptive-policy comparison. (a) Every curve is measured with the same action menu and priced by the measured cost law on val (QSD-T); the shipped ladder dominates the tested adaptive-TTA-style baselines across the served range. (b) Gap to the best matched-cost adaptive baseline at each budget across the six core streams; the ladder is positive almost everywhere, with the largest separation on V2-matched.

Table 13: Exact paired McNemar tests behind the matched-cost comparison. The 26:8:4 ladder is tested against both adaptive-TTA styles; the last pair shows the tight-budget boundary where the ladder reduces to plain margin ranking (13:8:0).

| | vs SA-TTS-style | | vs AdapTTA-style | | 13:8:0 vs SM | |
|---|---|---|---|---|---|---|
| Stream | Δ (pp) | $p$ | Δ (pp) | $p$ | Δ (pp) | $p$ |
| val (QSD-T) | +0.38 | 4.5e−6 | +0.41 | 2.0e−6 | +0.04 | 0.65 |
| val (QSD-L) | +0.19 | 2.4e−2 | +0.20 | 2.4e−2 | +0.20 | 2.3e−2 |
| V2 matched | +0.64 | 1.4e−3 | +0.94 | 1.4e−5 | −0.02 | 1.00 |
| V2 threshold | +0.16 | 0.46 | +0.42 | 4.0e−2 | +0.12 | 0.59 |
| ImageNet-R | −0.10 | 0.31 | +0.12 | 0.25 | −0.19 | 4.0e−2 |
| Sketch | +0.02 | 0.83 | +0.29 | 2.0e−4 | −0.17 | 2.7e−2 |
| noise-strong | +0.23 | 8.1e−2 | +0.41 | 5.5e−3 | −0.35 | 7.6e−3 |
| gray | +0.06 | 0.68 | +0.26 | 6.9e−2 | −0.55 | 9.0e−6 |
| jpeg-mild | −0.14 | 0.25 | −0.03 | 0.88 | −0.09 | 0.54 |
| blur-strong | −0.15 | 0.29 | −0.17 | 0.29 | −0.10 | 4.1e−2 |

### E.2 Audits of the allocation rule

Four checks probe the rule rather than the menu. First, learned utility routers (ridge/MLP/boosted stumps, up to 16,384 labelled rows, held-out scoring at $k=26$) never reach the zero-label margin

reference, although they rank repairable rows above chance (AUC 0.76–0.88): the tested routers do not recover the remaining oracle gap from the available cheap-view observables. Second, the per-request count rule is composition-stable to within 0.03 pp on ordinary batches and keeps the routed row count constant, which CUDA-graph capture requires; a matched threshold rule empties the budget on easy batches but under-escalates hard ones by 0.36 pp. Third, the deeper stages are load-bearing: resetting the stage-2/3 picks to the initial rank or random nested subsets loses 0.11–0.35 pp at equal cost, and spending the stage-3 budget on a wider stage 2 loses 0.15 pp (Table 14). Fourth, monotonicity survives statistical scrutiny in all five streams: bootstrap CIs on the decile correlation exclude zero, no adjacent violation is CI-separated, the isotonic MSE ratio is at least 0.998, and the permutation test rejects its null every time ($p \leq 0.0035$, $z \leq -2.4$; Table 15).

Table 14: Per-stage re-ranking ablation on the full 50,000-row ImageNet val cache. Re-ranking is the shipped behavior; alternative nested selections lose accuracy at equal measured cost.

| Variant | after stage 2 | after stage 3 | vs shipped |
|---|---|---|---|
| re-rank (shipped) | 79.416 | 79.596 | — |
| initial rank | 79.304 | 79.356 | −0.240 |
| random nested | 79.225 | 79.242 | −0.354 |
| replace, no fusion | — | 79.590 | −0.006 |
| deeper stage-2 control | — | 79.444 | −0.152 |

Table 15: Monotonicity audit on five streams. $\rho$ is the margin-decile/net-utility rank correlation; "viol." counts CI-separated adjacent increases; the permutation test reports one-sided $p$ and null-standard-deviation effect $z$.

| Stream | $\rho$ | 95% CI | viol. | MSE ratio | perm. $p$ | $z$ |
|---|---|---|---|---|---|---|
| val (QSD-T) | −0.795 | [−0.91, −0.60] | 0 | 1.000 | 0.0035 | −2.37 |
| val (QSD-L) | −0.856 | [−0.87, −0.42] | 0 | 1.000 | 0.0010 | −2.59 |
| V2 matched | −0.969 | [−0.99, −0.72] | 0 | 0.998 | 0.0005 | −2.90 |
| Sketch | −0.952 | [−0.98, −0.79] | 0 | 1.000 | 0.0010 | −2.95 |
| ImageNet-R | −0.851 | [−0.90, −0.50] | 0 | 1.000 | 0.0005 | −2.61 |

# F Serving, calibration economics, and harmful-stream evaluation

## F.1 End-to-end serving, calibration economics, and harmful streams

The main text prices pre-materialized GPU inference; this appendix adds the data path around it. Through a real JPEG decode/transform pipeline at batch 64, routing converts its GPU saving into wall-clock throughput in every regime tested (+38% with ten prefetch workers, +7% with four, +60% with a predecoded cache) and cuts GPU energy from 0.57 to 0.37 J per image. On the second scale it reaches +55% throughput and 38% lower energy than uniform-288 (Table 16), so the serving Pareto claims hold under the data path.

The width rule needs labelled probe rows once per stream; Table 17 prices that one-off cost. A 512-row probe already stabilises the selected width on the source-like streams, the shipped 2,048-row probe keeps mean selected-width regret within 0.10 pp (worst single draw 0.19 pp), and its calibration cost amortises to 0.54% of total serving time over 10,000 requests and 0.054% over 100,000. Two negative streams are priced at the same time: the $N{=}0$ anchors (no probe) regret 0.004–0.216 pp depending on the rule, while the probe brings every stream within 0.10 pp mean regret.

Finally, the refusal path that was only retrospective in Appendix B.4 is exercised prospectively. Twelve harmful streams were generated after the protocol was fixed; the frozen probe refuses escalation on the most damaged stream (pixelate32, cheap accuracy 2.29%, predicted decline −0.30) and, on the remaining two negative streams, spends at most 0.13 pp where uniform escalation would waste the budget, while paying +1.3 to +9.7 pp on the nine streams worth escalating, including the largest gain anywhere in the registry (occlude-center, +9.73 pp; Table 18). Two criteria set in ad-

vance remain unmet: the refusal does not fire on some negative streams, and the joint $(k, w)$ variant exceeds the regret criterion on one stream.

Table 16: End-to-end serving with the real decode/transform pipeline (JPEG decode, resize, letterbox, normalisation) feeding batch-64 requests on the RTX 4080 SUPER. "Predecoded" replaces decode and transform with a memory-mapped tensor cache to isolate the data pipeline. The routed policy converts its GPU saving into wall-clock throughput in every regime; the advantage shrinks as the CPU side becomes the bottleneck (4 workers) and is largest once data supply is removed. These J/image figures cover the full pipeline and use a wider scope than the GPU power probe of the main text; the two measurements should not be pooled.

| Regime | uniform (img/s) | route (img/s) | ratio | route p50 (ms) | J/img (route / uniform) |
|---|---|---|---|---|---|
| PIL decode, 10 workers | 353.5 | 488.2 | 1.38× | 85.8 | 0.368 / 0.569 |
| PIL decode, 4 workers | 282.9 | 301.9 | 1.07× | 87.3 | 0.461 / 0.591 |
| predecoded cache, 10 workers | 364.3 | 583.9 | 1.60× | 84.5 | 0.356 / 0.547 |
| QSD-L, B=64 | 171.3 | 265.7 | 1.55× | 199.2 | 0.824 / 1.340 |
| QSD-L, B=32 | 179.8 | 265.1 | 1.47× | 90.4 | 0.841 / 1.273 |

Table 17: How many labelled probe rows does the width rule need, and what does it cost? Left block: selected width regret (pp) against the odd-half best width, mean (worst) over three draws, as the probe grows from the $N{=}0$ anchors (no probe: fixed $k{=}26$ or the conservative $k{=}13$) through $N{=}4096$. Right block: the one-off calibration cost of the probe as a fraction of total serving time, amortised over a stream of requests.

| | probe-size regret, mean (worst) | | | | | 2048-row calibration overhead | | |
|---|---|---|---|---|---|---|---|---|
| Stream | 0 (fixed) | 256 | 512 | 2048 | 4096 | 1k req | 10k req | 100k req |
| val (QSD-T) | 0.00 | 0.08 (0.22) | 0.01 (0.03) | 0.01 (0.03) | 0.01 (0.03) | 5.4% | 0.54% | 0.054% |
| val (QSD-L) | 0.00 | 0.20 (0.45) | 0.15 (0.15) | 0.10 (0.15) | 0.05 (0.15) | 5.4% | 0.54% | 0.054% |
| V2 threshold | 0.08 | 0.51 (1.12) | 0.05 (0.08) | 0.07 (0.08) | 0.05 (0.08) | 5.4% | 0.54% | 0.054% |
| Sketch | 0.26 | 0.32 (0.93) | 0.18 (0.26) | 0.07 (0.19) | 0.01 (0.01) | 5.4% | 0.54% | 0.054% |
| V2 matched | 0.02 | 0.46 (1.36) | 0.46 (1.36) | 0.01 (0.04) | 0.00 (0.00) | 5.4% | 0.54% | 0.054% |

Table 18: Twelve prospectively generated harmful streams, evaluated once with the frozen 2,048-row probe rule (protocol fixed before generation). "Gate" is the predicted decline of the probe between $k{=}13$ and $k{=}52$; "ship" is the width the shipped rule spends; "gain" is the held-out utility on the odd half. The gate fires on pixelate32 (cheap accuracy 2.29%) and refuses escalation, which is the prospective exercise of the refusal path; the remaining two negative streams are priced by the probe to cost at most 0.13 pp against gains up to +9.73 pp on streams worth escalating.

| Stream | rows | cheap (%) | gate | ship | gain | best $k$ | best | uniform |
|---|---|---|---|---|---|---|---|---|
| blur_extreme | 20,000 | 24.48 | +0.31 | 40 | −0.13 | 19 | +0.22 | −0.23 |
| pixelate24 | 20,000 | 0.95 | +0.14 | 32 | −0.08 | 19 | −0.02 | −0.13 |
| pixelate32† | 20,000 | 2.29 | −0.30 | 6 | −0.03 | 19 | +0.01 | −0.09 |
| color_strong | 20,000 | 77.91 | +0.45 | 32 | +1.26 | 32 | +1.26 | +1.26 |
| brightness_low | 20,000 | 77.85 | +0.56 | 32 | +1.35 | 32 | +1.35 | +1.34 |
| invert | 20,000 | 76.82 | −0.12 | 13 | +1.53 | 26 | +1.91 | +1.88 |
| downscale_mild | 20,000 | 77.32 | +0.36 | 32 | +1.71 | 32 | +1.71 | +1.70 |
| contrast_high | 20,000 | 76.15 | +0.25 | 19 | +2.00 | 64 | +2.06 | +2.06 |
| jpeg_severe | 20,000 | 48.55 | +0.62 | 40 | +2.08 | 40 | +2.08 | +2.02 |
| speckle | 20,000 | 67.88 | +0.91 | 32 | +2.29 | 26 | +2.34 | +2.29 |
| noise_extreme | 20,000 | 58.86 | +1.09 | 32 | +2.77 | 40 | +2.82 | +2.80 |
| occlude_center | 20,000 | 42.33 | +5.69 | 52 | +9.73 | 64 | +9.77 | +9.77 |

† safety gate fires: the rule refuses escalation and falls back to $k{=}6$.

# G Dense tasks, the second spiking family, and reproducibility

## G.1 Detection and segmentation protocol and arms

Per-image AP is computed within each image (macro over the classes present, IoU 0.5, all score thresholds); the margin is the mean top-5 cheap-view score unless stated otherwise. VOC2012 val detection uses 3,000 images; PASCALContext val segmentation uses 2,500 images and per-image micro IoU with the 21-class protocol. The complete paired statistics are summarized below.

Table 19: Dense-task arms and paired statistics (Route13/Route40 arms omitted for space; the full sweep is in the artifacts). Detection: mean per-image AP (SSD-Lite/RetinaNet COCO weights); segmentation: mean per-image micro IoU (DeepLabV3). The last column is the paired Route26−Rand26 effect with a 2,000-draw bootstrap 95% CI ($^*$ excludes zero; effects are computed from unrounded per-image values).

| Task / axis | | Cheap | Uniform | Route26 | Rand26 | Route26−Rand26 |
|---|---|---|---|---|---|---|
| det flip-merge 320 (SSDLite) | AP | 84.74 | 85.93 | 85.61 | 85.21 | $+0.40[+0.11, +0.70]^*$ |
| det scale 320→384 (SSDLite) | AP | 84.74 | 83.87 | 84.54 | 84.39 | $+0.15[-0.33, +0.65]$ |
| det scale 600→800 (RetinaNet) | AP | 94.40 | 93.99 | 94.26 | 94.36 | $-0.10[-0.41, +0.18]$ |
| seg flip average (DeepLabV3) | IoU | 87.24 | 87.42 | 87.38 | 87.35 | $+0.03[-0.05, +0.12]$ |
| seg scale 480→640 (DeepLabV3) | IoU | 87.24 | 87.34 | 87.30 | 87.35 | $-0.05[-0.19, +0.09]$ |
| det flip-merge 320, 2nd split | AP | 83.45 | 85.07 | 84.52 | 84.20 | $+0.31[+0.004, +0.64]^*$ |
| seg flip average, 2nd split | IoU | 86.39 | 86.62 | 86.57 | 86.46 | $+0.10[+0.04, +0.17]^*$ |

Holm-adjusted escalation $p$-values: 0.023 (det flip-merge), 0.15 (retina), 0.15 (scale), 0.0029 (det flip-merge, 2nd split), $6\times10^{-24}$ (seg flip), $4.7\times10^{-38}$ (seg flip, 2nd split), $1.2\times10^{-26}$ (seg scale); the selection contrast is significant by CI on the two detection splits and the segmentation second split, where the sign test is tie-dominated (Holm $p = 1$).

**Probe size on the dense axes.** The dense caches hold 2,500–3,000 images, so their even half donates at most 1,250–1,500 probe rows, which is already the size the unified audit ships; a five-draw sweep shows the worst regret falling with $N$, from 2.6 units at $N$=512 to 0.67 at the full even half, with five of seven axes at regret 0.00. With only a few thousand images, these dense-task caches limit how precisely probe-size scaling can be assessed beyond classification.

Routing-statistic audit on the detection axes (Spearman of decile utility): flip-merge — mean top-5 $+0.330$, max score $-0.163$, number of kept boxes $-0.239$; scale 320→384 — $-0.547$, $-0.018$, $-0.377$; RetinaNet 600→800 — $-0.337$, $+0.261$, $-0.524$. The statistic that orders escalation value is therefore axis-dependent on dense tasks, unlike the classification panel where the probability margin is uniformly the best of the eight tested signals. Segmentation replicates in the opposite direction on both axes ($\rho = +0.79$ flip, $+0.91$ scale; the flip axis confirms at $+0.60$ on the second split), which is the evidence for the scope condition stated in the main text.

### G.2 Self-trained spiking CNN details

Two neuron recipes were compared on identical data and an identical architecture; both are direct-trained with four time steps, ATan surrogate gradients, and AMP or bf16. The hard-reset recipe (detached reset, threshold 0.5) reaches in eight epochs the accuracy the soft-reset recipe (threshold 1.0) never reaches in sixty, so the family is trained with the hard-reset recipe and released as an artifact; the recipe comparison is an implementation note and is not part of the response analysis. Cached two-geometry accuracies: 32 inputs 60.13%, 40 inputs 38.42%, 0.3/0.7 fusion 52.56%. On the harmful 32→40 axis the block audit gives $\rho(\text{repair rate}) = -0.952$ and $\rho(u) = +0.152$, block-optimal minus greedy 2.46 pp, oracle minus greedy at $k$=26 7.94 pp, nested errors 1.215/1.207 pp, probe selected-width regret 0.000 pp at $N \geq 2{,}048$, identity residual $8\times10^{-17}$, and $J/(AB) = 1.05$ at the lowest decile. Recipe transfer: the shipped ($k$=26, $w$=0.3) cell sits 3.18 pp below the odd-half optimum ($k$=6, $w$=0.7), while the even-half-selected cell ($k$=19, $w$=0.7) is 0.08 pp from that optimum.

### G.3 Evidence provenance and reproducibility

The evidence layers used in this paper are: deployment-facing measurements (locked system and frozen default, fixed external evaluation without target retuning) versus interpretive analyses (held-out even/odd splits, label-aware descriptive audits, retrospective replays), with serving restricted to pre-materialized GPU inference and routing. The numerical results come from four evidence groups: the primary QSD geometry and six-distribution analyses; second-scale service and calibration-economics audits; mechanism, corruption, statistic, self-trained SNN, and dense-task studies; and the paired-statistics, unified-probe, prospective-corruption, dense probe-size, and witness-map evaluations. Experiments built from different caches or model scales retain their own cheap baselines and are reported separately. The accompanying artifact contains the analysis code and manuscript-level aggregate tables, while datasets, checkpoints, prediction caches, and raw result files remain

external because of size or access constraints. Reproducing the numerical results therefore requires matching the dataset and preprocessing identifiers, checkpoint and cache hashes, latency configuration, and table-generation scripts associated with each experiment.