WHEN ATTENTION GEOMETRY SHIFTS: CALIBRATED STRUCTURAL EVALUATION OF VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision transformers route information through attention, yet shift detection often relies only on pooled features or classifier confidence. This paper studies whether a query's routing pattern can be compared with nominal executions to characterize visual shift. **Attention-Geometry Tracking (ATH)** forms a calibration bank of clean embeddings and attention summaries, retrieves a coverage-qualified Local reference or falls back to an Aggregate reference, and scores shifts in key marginal, stable rank, entropy dispersion, and mean entropy. It requires neither labels nor classifier logits at scoring time and records the reference branch for every query.

On a 75-condition window-averaged Swin-B corruption grid, mean-entropy shift (MES) is the strongest tested standalone score (mean AUROC 0.8049); the evaluated transferred-weight fusion reaches 0.6407. In a four-model OOD experiment—ViT-B/16, Swin-B, DINOv2-B/14, and CLIP ViT-B/16—Local references improve ImageNet-R AUROC over Aggregate references for every model, whereas DTD effects are mixed. Reference conditioning can therefore alter rankings but is shift dependent; high ImageNet-R FPR@95 precludes an operating-point claim. The corruption grids are explicitly exploratory because their selection slices overlap evaluation sources.

1 INTRODUCTION

Distribution shift can alter attention geometry; whether this provides useful signal beyond output confidence is empirical. This is particularly relevant for encoders without a task-specific classifier head, where logit-based scores such as MSP and Energy are not defined. For encoders with logits, this paper treats confidence scores as comparators rather than assuming that they represent routing change or a deployment operating point.

This work studies a focused question: **does a query induce routing geometry consistent with structurally similar nominal executions?** ATH separates two problems that are often conflated. It derives specified attention summaries and selects a compatible nominal reference from a calibration reference bank. A query uses a softmax-weighted Local reference when its nearest calibration anchors have sufficient coverage; otherwise it uses an Aggregate reference.

This formulation makes negative findings meaningful. A Local reference can help, tie, or hurt relative to an Aggregate reference. ATH can report either an individual component or a weighted fusion of components. Fusion performance can depend on whether its weights and component directions were selected for the receiving architecture. This paper reports that dependency rather than interpreting a transferred fusion as an architecture-independent test of fusion.

The contributions are:

1. **Calibrated structural evaluation.** A calibration reference bank of clean anchors supplies Local or Aggregate reference statistics for attention-geometry comparison without requiring labels or classifier logits at inference.
2. **A reference-conditioned structural score.** ATH applies established routing summaries—marginal divergence, rank-related deviation, entropy dispersion, and entropy shift—against

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.

a query-conditioned or aggregate calibrated reference. Reported component evidence covers ViT-B/16 and the window-averaged Swin-B operator.

3. **Coverage-qualified locality.** The method makes Local, Aggregate, and fallback reference modes explicit and evaluates their domain-shift behavior.
4. **Evidence with explicit scope.** This paper reports exploratory 75-condition ViT-B/16 and Swin-B corruption ablations, an exploratory ViT feature/logit comparison, four-model OOD evaluation, and a second Swin OOD seed. It does not claim all-architecture or all-condition superiority over feature or logit baselines.

2 Calibrated Attention-Geometry Tracking

2.1 Routing representation and calibration anchors

For layer ℓ and head h, let $A_{\ell h} \in [0,1]^{N^\ell \times N^\ell}$ be a row-normalized attention operator (Vaswani et al., 2017). Where a model does not expose probabilities directly, its native attention-side tensors may be used only to form this forward attention operator. In the reported Swin experiment (Liu et al., 2021), the hook captures post-normalization window attention, reshapes it as a batch of windows, and averages attention probabilities across windows before any statistic is computed. Consequently, the reported Swin score is a **window-averaged structural summary**, not a per-window score and not a dense image-wide attention matrix.

For each clean calibration input $a \in \{1, \dots, M\}$, the calibration reference bank retains a normalized state embedding z_a and reference statistics

$$\mathcal{B}_a = \{c_a, r_a, q_a, \mu_a\}, \tag{1}$$

where c_a, r_a, q_a, and μ_a denote, respectively, the key-marginal, stable-rank, entropy-IQR, and mean-entropy statistics for anchor a. An **anchor** is one clean calibration input together with this stored embedding and its reference statistics. Calibration anchors and evaluation samples are index-disjoint.

2.2 Local, Aggregate, and coverage-gated references

Let M be the number of clean anchors in the calibration reference bank, and choose a retrieval count $k \in \{1, \dots, M\}$ for each query. Given a query embedding z and stored anchor embedding z_a, cosine similarity is

$$s_a = \frac{\langle z, z_a \rangle}{\|z\|_2 \|z_a\|_2} \tag{2}$$

The stored embedding is the anchor's clean-model embedding; the query embedding is produced for the input currently being scored. The implementation L2-normalizes both, then evaluates these inner products with a matrix–vector multiply. Let $\mathcal{N}_k(z)$ denote the k most similar calibration anchors and define coverage

$$\gamma(z) = \frac{1}{k} \sum_{a \in \mathcal{N}_k(z)} s_a \tag{3}$$

When $\gamma(z) \geq \tau_{\min}$, ATH uses a Local reference:

$$w_a = \frac{\exp(s_a)}{\sum_{a' \in \mathcal{N}_k(z)} \exp(s_{a'})}, \qquad \mathcal{B}_{\text{local}}(z) = \sum_{a \in \mathcal{N}_k(z)} w_a \mathcal{B}_a \tag{4}$$

Otherwise it uses an Aggregate reference:

$$\mathcal{B}_{\text{agg}} = \frac{1}{M}\sum_{a=1}^{M}\mathcal{B}_a \tag{5}$$

The coverage threshold $\tau_{\min}$ is a **reference-selection threshold**, not an anomaly threshold. It is estimated from clean validation coverage for each model, calibration size, representation, and k. Section 4.1 states which selection slices overlap evaluation sources in each run. In the reported $M = 500$, $k = 10$ configuration, $\tau_{\min}$ is 0.1409 for ViT-B/16, 0.1564 for Swin-B, 0.1170 for DINOv2-B/14, and 0.6033 for CLIP ViT-B/16.

2.3 Structural components

For layer ℓ and head h, define the key marginal over a fixed key position j as

$$c_j^{\ell h} = \frac{1}{N^\ell}\sum_{i=1}^{N^\ell} A_{ij}^{\ell h} \tag{6}$$

The scorer renormalizes both $c^{\ell h}$ and the reference $\hat{c}^{\ell h}$ before comparison as a numerical safeguard. Let $\bar{c}^{\ell h}$ and $\bar{\hat{c}}^{\ell h}$ denote these normalized vectors, and let $m^{\ell h} = \frac{1}{2}(\bar{c}^{\ell h} + \bar{\hat{c}}^{\ell h})$. **RMD** is the Jensen–Shannon divergence implemented as two KL terms:

$$\begin{aligned}\mathrm{RMD}_{\ell h} = \mathrm{clip}_{[0,1]}\Big\{\tfrac{1}{2}\Big[&\sum_j \bar{c}_j^{\ell h}\big(\log_2(\bar{c}_j^{\ell h}+\epsilon) - \log_2(m_j^{\ell h}+\epsilon)\big) \\ &+ \sum_j \bar{\hat{c}}_j^{\ell h}\big(\log_2(\bar{\hat{c}}_j^{\ell h}+\epsilon) - \log_2(m_j^{\ell h}+\epsilon)\big)\Big]\Big\}\end{aligned} \tag{7}$$

SRD compares stable rank $\mathrm{sr}(A) = \|A\|_F^2/(\|A\|_2^2+\epsilon)$ with reference $\hat{r}$:

$$\mathrm{SRD}(A,\hat{r}) = \tanh\left(\left|1 - \frac{\mathrm{sr}(A)}{\hat{r}+\epsilon}\right|\right) \tag{8}$$

For row entropy, the implementation first applies $\tilde{A}_{ij}^{\ell h} = \max\{A_{ij}^{\ell h}, \epsilon\}$ and then computes $H_i^{\ell h} = -\sum_j \tilde{A}_{ij}^{\ell h}\log_2 \tilde{A}_{ij}^{\ell h}$. Define $\mathrm{IQR}(H^{\ell h}) = Q_{0.75}(H^{\ell h}) - Q_{0.25}(H^{\ell h})$, with both the quantiles and $\mathrm{mean}(H)$ below taken over query position i. **ARD** is the one-sided relative excess in entropy interquartile range and **MES** is normalized mean-entropy shift:

$$\mathrm{ARD}(A,\hat{q}) = \tanh\left(\max\left\{0, \frac{\mathrm{IQR}(H)}{\hat{q}+\epsilon} - 1\right\}\right), \qquad \mathrm{MES}(A,\hat{\mu}) = \tanh\left(\left|1 - \frac{\mathrm{mean}(H)}{\hat{\mu}+\epsilon}\right|\right) \tag{9}$$

For layer ℓ, let H_ℓ be its observed head count and let $d_{\ell h}$ denote the vector of four component values for head h. The evaluated fusion is

$$\mathrm{ATH} = \frac{1}{L}\sum_{\ell=1}^{L}\left\{\frac{1}{H_\ell}\sum_{h=1}^{H_\ell}\left[\alpha\max(d_{\ell h}) + (1-\alpha)\sum_u w_u d_{\ell h}^u\right]\right\} \tag{10}$$

where $u \in \{\mathrm{RMD}, \mathrm{SRD}, \mathrm{ARD}, \mathrm{MES}\}$, $\alpha = 0.5$, and $(w_{\mathrm{RMD}}, w_{\mathrm{SRD}}, w_{\mathrm{ARD}}, w_{\mathrm{MES}}) = (0.0255, 0.3858, 0.2233, 0.3654)$. These weights were derived on a 200-image ViT validation slice at severity 3 and seed 42 by normalizing each component's pooled layer-level absolute clean-to-corrupted mean shift divided by its clean-sample standard deviation. They were transferred unchanged to the reported Swin ablation and are not claimed to be Swin-optimal. This ViT slice

overlaps the existing ViT corruption evaluation sources, so it is configuration provenance rather than held-out weight validation. MES is also evaluated alone.

This is a **mean of per-layer head means**, not a flat mean over all layer–head pairs. The distinction matters for hierarchical architectures such as Swin-B, whose H_ℓ varies by stage. It exactly matches the scorer used for Table 1: the active heads are averaged within a layer before layer means are averaged. Padding is a rectangular-storage convention; scoring slices to each layer's actual head count, so padded heads contribute no score.

2.4 Implementation path and complexity

Row-entropy, key-marginal, entropy-IQR, and stable-rank summaries are computed from attention blocks. In the reported implementation, stable rank is computed from the retained attention operator. A deployment that cannot retain that operator would need a bounded-pass spectral estimator or another explicitly stated approximation; neither alternative is evaluated here. The calibration reference bank retains embeddings and per-anchor reference statistics rather than full calibration attention operators.

The brute-force similarity calculation costs $O(MD)$ per query for M anchors of embedding dimension D; top-k selection and synthesis add implementation-dependent work. This describes an implementation interface only: no throughput, latency, area, energy, or synthesis result is reported.

2.5 Selection of neighborhood size

For ViT-B/16 with $M = 500$, the saved selection sweep reports AUROCs of 0.7073, 0.8638, 0.8758, 0.8758, and 0.8743 for $k \in \{1, 5, 10, 20, 50\}$. Although $k = 20$ is the numerical maximum, $k \in \{10, 20, 50\}$ is statistically indistinguishable from it by paired DeLong tests; the workflow retained $k = 10$, the tied value with the lowest reported fallback rate (12.99%). This sweep ran at severity 3 and its selected value was applied unchanged across severities 1–5, so 0.8758 is a single mid-severity value rather than a grid average. This sweep used a 1,000-image validation slice. The coverage thresholds are a separate measurement: the clean-side tenth percentile of top-k mean similarity on the 200-image slice of Section 2.3, so they carry no severity dependence. However, the selection slice overlaps the existing ViT corruption evaluation slice. These values therefore document the configuration history and do **not** validate $k = 10$ on held-out ViT corruption data. The completed Swin-B, DINOv2-B/14, and CLIP ViT-B/16 evaluations also use the primary value $k = 10$; they do not each have a separate corrected k-selection result. Main experiments fix $M = 500$ to hold reference-bank size and calibrated thresholds constant across conditions. This is a protocol choice, not evidence that $M = 500$ is optimal. The ViT result is not presented as a parabola or an architecture fingerprint. No comparable cross-architecture ATH-k sweep is complete, so the paper draws no conclusion about training objectives from the direction of a k curve.

3 Related Work

ATH is positioned as a reference-conditioning procedure, not as a claim that entropy or rank statistics are new. Attention operators originate in the Transformer architecture (Vaswani et al., 2017) and are central to vision-transformer families including ViT (Dosovitskiy et al., 2021) and Swin (Liu et al., 2021). Rank degeneration and token-uniformity in attention have been analyzed previously (Dong et al., 2021); ATH uses rank-related quantities as measurements against calibrated references rather than as a new explanation of rank collapse.

OOD detection commonly scores logits, features, or both. MSP is a standard confidence baseline (Hendrycks and Gimpel, 2017); ODIN modifies logit scoring with input perturbation and temperature scaling (Liang et al., 2018); ReAct rectifies activation-based scores (Sun et al., 2021); energy scores provide another logit-based formulation (Liu et al., 2020); and DOCTOR scores softmax disagreement (Granese et al., 2021). The comparator reported here is the DOCTOR ranking transform $1 - \sum_c p_c^2$, which is order-equivalent to DOCTOR-α and therefore leaves AUROC and FPR@95 unchanged. Feature-space approaches include class-conditional Mahalanobis distance (Lee et al., 2018), deep nearest-neighbor distance (Sun et al., 2022), and ViM's feature-residual and logit construction (Wang et al., 2022). KNN-OOD is particularly close in retrieval structure: both methods

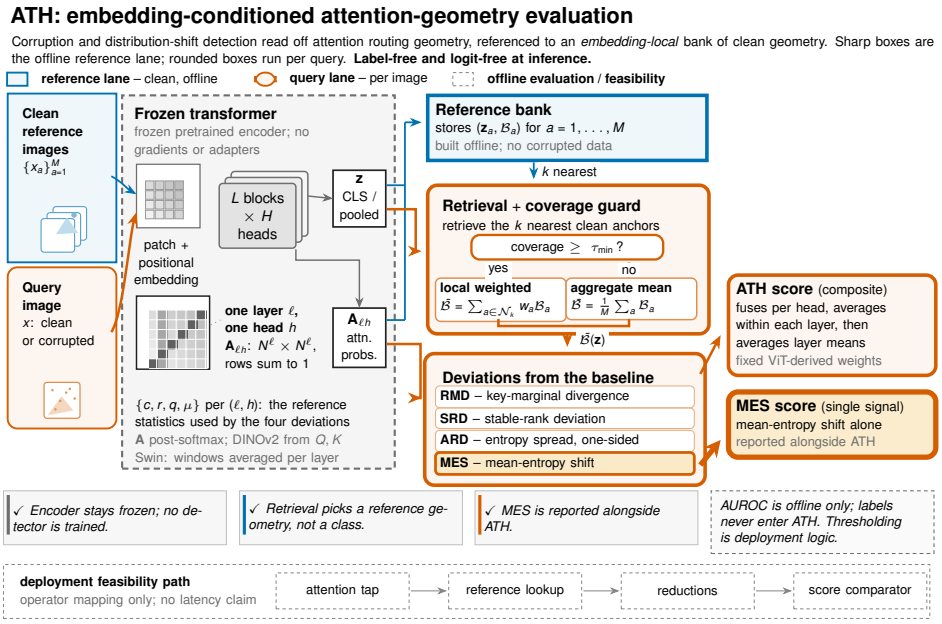

Figure 1: **ATH pipeline.** The scoring side receives attention summaries, selects a Local or Aggregate reference from the calibration bank through coverage, and produces individual and fused scores. The Swin-B path averages windows before these summaries. This is an implementation path, not a measured hardware result; no fusion weights are shown (Section 2.3).

retrieve nominal examples in embedding space. They differ in the final score: KNN-OOD scores feature-space proximity itself, whereas ATH uses retrieved embeddings only to synthesize a reference for attention-geometry comparison. This distinction is methodological. Section 5.2 reports an exploratory same-example comparison for a declared ViT subset; it is not evidence of superiority on every corruption or architecture.

4 Experimental Protocol

4.1 Corruption ablations

The saved corruption ablations use ViT-B/16 (Dosovitskiy et al., 2021) and Swin-B (Liu et al., 2021) on 15 ImageNet-C-style corruption types at five severity levels (Hendrycks and Dietterich, 2019): 75 conditions per model. Each condition compares 500 clean evaluation images with 500 altered counterparts. The same clean cohort is reused across conditions; condition averages are therefore not independent samples. Calibration-bank and evaluation indices are disjoint in both runs. However, the threshold-calibration slice used in those runs overlaps clean evaluation sources; the ViT run additionally overlaps k and fusion-weight selection. Both corruption grids are therefore exploratory.

Each calibration reference bank contains 500 clean images and is index-disjoint from evaluation images. Clean examples have label 0 and altered examples label 1. A larger score indicates stronger evidence of alteration. All 37,500 altered-image calls in each completed grid used the requested corruption; none fell back to synthetic Gaussian noise.

4.2 Completed OOD protocol

For each of ViT-B/16 (Dosovitskiy et al., 2021), Swin-B (Liu et al., 2021), DINOv2-B/14 (Oquab et al., 2024), and CLIP ViT-B/16 (Radford et al., 2021), ImageNetV2 (Recht et al., 2019) supplies 500 calibration anchors, a disjoint 1,000-example validation reservation, and a disjoint 1,000-image in-domain control set. ImageNet-R (Hendrycks et al., 2021) and DTD (Cimpoi et al., 2014) each supply 1,000 shifted examples. Local and Aggregate references are scored on identical control and

Table 1: **Component macro-averages over 75 corruption–severity conditions per model (exploratory: selection slices overlap evaluation sources; see Section 4.1).** Best value per column in bold.

Method	ViT AUROC	ViT FPR@95	Swin AUROC	Swin FPR@95
MES only	**0.7316**	**0.5811**	**0.8049**	**0.4973**
ATH minus ARD	0.7076	0.6166	0.7669	0.5298
SRD only	0.7097	0.6298	0.7639	0.5411
MSP	0.6855	0.7816	0.7459	0.6782
ATH full	0.6965	0.6572	0.6407	0.7277
RMD only	0.5331	0.9084	0.5331	0.9243
ARD only	0.5973	0.8448	0.4556	0.9372

shifted examples. The four-model run uses seed 42. Swin-B is repeated with seed 43, using the same pre-specified $M = 500$, $k = 10$, and $\tau_{\min} = 0.1564$ protocol; this is a split-and-sampling-seed replication, not an independent dataset replication.

4.3 Scope discipline

The paper reports AUROC and FPR@95 for corruption detection. For OOD, it reports paired Local-minus-Aggregate AUROC differences, paired DeLong Wald intervals (DeLong et al., 1988), and Holm-adjusted tests (Holm, 1979) over the declared two-shift family for each model and seed. A same-example ViT feature/logit comparison is reported only as exploratory because its 500 source identifiers are the selection-overlapping ViT evaluation slice. It uses four corruptions (Gaussian noise, defocus blur, frost, and JPEG compression) at severities 3 and 5. Source-blocked bootstrap resampling preserves the nine records per source (one clean and eight altered). It compares MES-Local with a predeclared 13-comparator family and uses a maximum-standardized family-wise interval. The intervals quantify resampling variation conditional on the selected slice; they are not held-out confirmation. No all-model or full-75-condition feature-baseline claim is made.

5 Results

5.1 Exploratory component results

Table 1 gives descriptive macro-averages over 75 corruption–severity conditions for each model. All ViT entries are sourced consistently from the same component-ablation run. Within the Swin grid, MES is the strongest tested standalone structural component (0.8049 AUROC). The result is descriptive because the same 500 clean images recur across conditions. It supports neither cross-architecture superiority nor a claim that MES is universally optimal.

Figure 2 plots the MES severity trajectories for both grids. MES rises from severity 1 to 5: 0.6031, 0.6787, 0.7407, 0.7980, and 0.8374 for exploratory ViT-B/16; and 0.6822, 0.7643, 0.8135, 0.8668, and 0.8975 for Swin-B. These trajectories are consistent with increasing structural deviation, but do not establish causality. For Swin-B, removing ARD from the evaluated fusion improves mean AUROC from 0.6407 to 0.7669. Because its fusion weights were selected in a separate ViT procedure and transferred unchanged, this identifies a **cross-architecture weight-transfer failure**, not a general result that fusion is unhelpful.

ARD deserves a narrower interpretation. Its one-sided form responds only to expansion of entropy interquartile range. It has AUROC 0.4556 in this Swin experiment and its rectified output is nearly unchanged in aggregate between clean and altered records (means 0.0474 and 0.0472, respectively). The saved records do not contain the signed, pre-rectification IQR ratio needed to establish whether a systematic contraction caused this result. The paper therefore treats contraction as a testable hypothesis, not as an established mechanism.

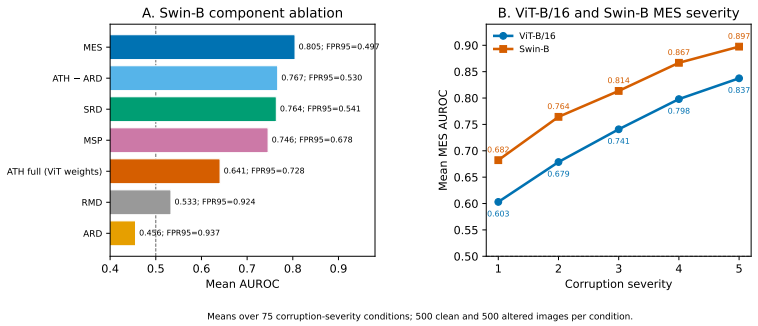

Figure 2: **Corruption evidence.** Panel A: Swin-B component means over 75 corruption–severity conditions, all seven methods of Table 1. Panel B: MES mean by severity for both grids. No uncertainty bars, because the same clean cohort recurs across conditions. "ATH full" uses the transferred ViT-derived weights.

5.2 Exploratory, source-blocked ViT comparison

The exploratory ViT comparison uses four named corruptions at severities 3 and 5, giving eight conditions and 500 shared source images. Table 3 in Appendix A reports the full declared 13-comparator family, raw-direction macro AUROC, and the family-wise interval for MES-Local minus each comparator. MES-Local scores 0.7934 on this subset and exceeds every comparator; the intervals exclude zero for eleven of the thirteen, the exceptions being MES-Aggregate ($+0.0087$ $[-0.0031, +0.0205]$) and Entropy ($+0.0277$ $[-0.0009, +0.0563]$). The largest margins are over KNN-OOD ($+0.2150$) and the normalized-feature Residual variant ($+0.2133$); the nearest structural comparator is ATH-Local at $+0.0213$ $[+0.0028, +0.0398]$. The source-blocked resampling is correct, but the table is descriptive and does not provide held-out support for a MES, fusion, or feature-baseline claim.

The table intentionally does not relabel oriented scores as baselines, does not pool repeated source records as independent samples, and does not generalize this result to the 75-condition grid or to Swin, DINOv2, or CLIP. In particular, it supplies no held-out evidence that ATH as a family beats every feature-space detector; the tested ATH fusion is lower than MES-Local in this selected subset.

5.3 Local references change OOD ranking in a shift-dependent way

Table 2 compares Local and Aggregate reference synthesis. In the seed-42 four-model run, Local improves ImageNet-R AUROC for every tested architecture, but that relative ranking advantage does not establish a usable ImageNet-R detector: Local FPR@95 ranges from 0.847 to 0.909. It is not universally preferable: on DTD, where absolute AUROC is higher and FPR@95 is lower for most cells, Aggregate ties or outperforms Local depending on the architecture. The Swin seed-43 repetition preserves the positive ImageNet-R and negative DTD AUROC directions, but its DTD FPR@95 direction reverses.

Intervals are **unadjusted** paired 95% DeLong Wald intervals recovered from the saved difference and z statistic. FPR@95 is the false-positive rate on the control set at the threshold achieving 95% true-positive rate on the shifted set, recomputed from the same-example score records; it is descriptive here and has no paired interval or hypothesis test. Holm correction applies only to the two p-values for ImageNet-R and DTD within each model and seed. The Swin/ImageNet-R AUROC direction repeats at seed 43, but these two seed-specific analyses are not pooled into a cross-seed p-value. The DTD AUROC direction also repeats, whereas Local's relative DTD FPR@95 changes from worse at seed 42 (0.742 versus 0.707) to better at seed 43 (0.615 versus 0.636); the operating-point direction is therefore not replicated.

The operational pattern reverses the simple "Local helps" narrative. On ImageNet-R, Local improves both AUROC and the reported FPR@95 relative to Aggregate in all four models, yet its FPR@95 remains 0.847–0.909. Accordingly, the paper makes no usability or deployed-threshold claim on ImageNet-R and does not claim that the advantage persists at a low false-positive rate;

Table 2: **Paired Local-minus-Aggregate OOD results.** Models are ViT-B/16, Swin-B, DINOv2-B/14 and CLIP ViT-B/16, with the seed in parentheses; per-row FPR@95 and fallback rates are in Table 4. The OOD protocol reserves its validation slice, so these rows are not selection-overlapping.

Model	Shift	Local	Aggregate	Δ [95% CI]	Holm p
ViT (42)	ImageNet-R	0.6201	0.5787	+0.0414 [+0.0282, +0.0545]	1.35×10^{-9}
ViT (42)	DTD	0.8950	0.8965	-0.0015 [-0.0091, +0.0060]	0.6895
Swin (42)	ImageNet-R	0.5875	0.5687	+0.0188 [+0.0065, +0.0311]	0.0027
Swin (42)	DTD	0.8325	0.8467	-0.0143 [-0.0215, -0.0071]	0.0002
DINOv2 (42)	ImageNet-R	0.6937	0.6362	+0.0576 [+0.0452, +0.0699]	$< 10^{-12}$
DINOv2 (42)	DTD	0.8506	0.8513	-0.0007 [-0.0085, +0.0072]	0.8690
CLIP (42)	ImageNet-R	0.6971	0.6397	+0.0574 [+0.0448, +0.0699]	$< 10^{-12}$
CLIP (42)	DTD	0.7544	0.8373	-0.0829 [-0.0954, -0.0703]	$< 10^{-12}$
Swin (43)	ImageNet-R	0.5960	0.5705	+0.0256 [+0.0148, +0.0364]	7.22×10^{-6}
Swin (43)	DTD	0.8339	0.8416	-0.0077 [-0.0143, -0.0011]	0.0213

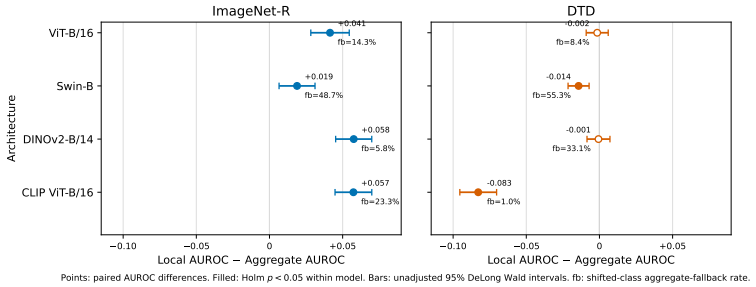

Figure 3: **Local versus Aggregate reference synthesis, seed 42.** Points are paired AUROC differences; bars are unadjusted 95% paired DeLong Wald intervals. Filled points have Holm $p < 0.05$ within that model's two-shift family. The label is the shifted-class aggregate-fallback rate, making the reference-selection branch visible. Holm adjustment appears only in Table 2; the seed-43 replication is tabulated separately.

that requires a pre-specified operating-point analysis. On DTD, the scores separate the two sets more strongly in absolute terms, while Local is approximately neutral on AUROC for ViT-B/16 and DINOv2-B/14 and worse for Swin-B and CLIP. Thus the completed evidence supports a narrower mechanism result: Local reference conditioning can change rankings, but its direction and operational utility depend on the shift and coverage regime.

5.4 Coverage is an observable branch, not a hidden failure

Coverage controls whether a Local reference is trusted. In the Swin corruption grid, 16,259 of 75,000 condition records (21.7%) used the aggregate branch; this accounting repeats the 500 clean records for each of 75 conditions. Per unique scored record, 12,633 of 38,000 instances used that branch (33.24%): 49 of 500 clean records (9.8%) and 12,584 of 37,500 altered records (33.56%). Per-condition fallback spans 8.3% to 40.2%, and the coverage warning is triggered in 74 of 75 conditions. In the reported OOD evaluations, shifted-set fallback spans 1.0% (CLIP/DTD, seed 42) to 55.6% (Swin/DTD, seed 43). This is declared reference-selection behavior, not an alteration-generation failure, but it materially limits how far a Local score differs from Aggregate and must be reported alongside every Local result.

6 Discussion: deployment implications

ATH requires that model execution make available either routing operators from which the required summaries can be computed, or the complete score-ready summaries themselves, together with a calibration reference bank and reference-selection logic. Row entropy, key marginal, entropy-IQR,

and stable-rank summaries can be accumulated from attention blocks. This describes an implementation path only; no throughput, latency, area, energy, or synthesis result is reported. The reported Swin implementation averages windows before statistic computation and therefore does not demonstrate a per-window memory advantage.

7 LIMITATIONS

- The existing ViT and Swin corruption grids reuse clean evaluation sources in the threshold-calibration slice used by those runs; ViT additionally reuses images employed in prior k and fusion-weight selection. Source-blocked intervals handle repeated source records but do not repair selection overlap; these results are exploratory. A new three-way split is implemented for future runs, but no replacement corruption run is reported here.
- The paper has no completed corruption grid for DINOv2 or CLIP. It makes no all-architecture or all-condition superiority claim over KNN-OOD, Mahalanobis, residual distance, logit scores, or other baselines.
- The reported Swin fusion weights were transferred from a ViT validation procedure. Per-model weight selection on a disjoint split remains a necessary follow-up.
- The saved Swin records omit ARD's signed pre-rectification IQR ratio. They establish that rectified ARD has negligible aggregate class separation here, but cannot establish whether contraction, expansion, or another effect explains that failure.
- Local references improve ImageNet-R ranking in this protocol, but FPR@95 remains 0.847–0.909; this study does not establish a usable low-false-positive operating point or evaluate one.
- Local references do not uniformly improve DTD; the stronger absolute DTD separation coexists with neutral or negative Local-versus-Aggregate differences. Although the DTD AUROC direction repeats for Swin across seeds, its relative FPR@95 direction does not.
- Within the exploratory window-averaged Swin-B grid, MES is the strongest tested standalone component. Whether MES predominates under a held-out protocol for ViT, Swin, DINOv2, CLIP, or other attention operators is untested.
- Exact rank-related statistics can require more buffering or passes than entropy and marginal reductions; each deployment must state its retained-operator or approximation policy.

8 CONCLUSION

ATH turns model-internal attention summaries into an auditable, reference-conditioned shift score. Across four held-out OOD evaluations, Local references improve ImageNet-R AUROC over Aggregate references, while DTD shows that this direction is not universal; the high ImageNet-R FPR@95 values rule out a deployment operating-point claim. In the exploratory corruption grids, MES is the strongest tested standalone component, and ViT-derived fusion weights do not transfer cleanly to Swin. These findings do not establish a universally superior OOD detector. They establish a narrower mechanism result: exposing the reference branch, component scores, and fallback rate reveals when locality changes ranking—and when it does not. Testing this mechanism under disjoint, model-specific selection and held-out corruption protocols is the necessary next step.

ETHICS STATEMENT

This work uses public pretrained models and established image datasets and does not involve human-subject interaction or the collection of personal data. It evaluates whether an input departs from calibrated reference executions; it does not establish that a structural score is a safety guarantee or a sufficient basis for an automated decision about people. A score may be misused if it is treated as a universal reliability signal despite the shift dependence and operating-point limitations reported here. Any operational use should be validated on its intended domain, report its fallback behavior, and preserve appropriate human oversight for consequential decisions.

Reproducibility Statement

Sections 2 and 4 specify the models, datasets, reference-bank size, retrieval count, thresholds, seeds, source-blocked analysis, and split limitations used for each reported result. Tables 1–4 and the figures report the completed analyses, while the manuscript states the selection overlap and other limitations that constrain their interpretation. The corruption experiments use a vendored `imagecorruptions` 1.1.2 implementation, pinned by version and content hash in each run configuration.

AI Use Statement

Generative AI tools were used to assist with drafting and polishing prose, drafting portions of the paper, and checking mathematical statements against the implementation. The author reviewed every AI-assisted output and independently checked reported claims against the implementation, saved experimental artifacts, and direct calculations. Generative AI did not generate experimental data or substitute for the author's responsibility for the paper's methods, results, or conclusions.

References

Cimpoi, M. et al. 2014. Describing Textures in the Wild. CVPR.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics, 44(3):837-845.

Dong, Y., Cordonnier, J.-B., and Loukas, A. 2021. Attention Is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth. ICML.

Dosovitskiy, A. et al. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., and Piantanida, P. 2021. DOCTOR: A Simple Method for Detecting Misclassification Errors. NeurIPS.

Hendrycks, D. and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR.

Hendrycks, D. and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. ICLR.

Hendrycks, D. et al. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. ICCV.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6(2):65-70.

Lee, K. et al. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. NeurIPS.

Liang, S., Li, Y., and Srikant, R. 2018. Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. ICLR.

Liu, W. et al. 2020. Energy-based Out-of-distribution Detection. NeurIPS.

Liu, Z. et al. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. ICCV.

Oquab, M. et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. TMLR.

Radford, A. et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. ICML.

Recht, B. et al. 2019. Do ImageNet Classifiers Generalize to ImageNet? ICML.

Sun, Y., Guo, C., and Li, Y. 2021. ReAct: Out-of-distribution Detection With Rectified Activations. NeurIPS.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. ICML.

Vaswani, A. et al. 2017. Attention Is All You Need. NeurIPS.

Wang, H. et al. 2022. ViM: Out-of-Distribution with Virtual-logit Matching. CVPR.

A Full comparator family for Section 5.2

Table 3: **Source-blocked comparison of MES-Local against the declared 13-comparator family on an eight-condition ViT subset (exploratory: the 500 source images are the selection-overlapping evaluation slice).** MES-Local scores 0.7934 on this subset; positive differences favour it.

Comparator	AUROC	MES − comp.	95% FW interval
ATH-Local	0.7724	+0.0213	[+0.0028, +0.0398]
ATH-Aggregate	0.7512	+0.0423	[+0.0190, +0.0655]
ATH-EqualWeight-Local	0.7601	+0.0335	[+0.0141, +0.0529]
MES-Aggregate	0.7846	+0.0087	[-0.0031, +0.0205]
KNN-OOD	0.5787	+0.2150	[+0.1857, +0.2442]
Mahalanobis (raw)	0.6945	+0.0989	[+0.0682, +0.1297]
Residual distance (raw)	0.6960	+0.0975	[+0.0667, +0.1282]
Mahalanobis (normalized)	0.5968	+0.1965	[+0.1669, +0.2261]
Residual distance (normalized)	0.5801	+0.2133	[+0.1850, +0.2415]
MSP	0.7356	+0.0579	[+0.0276, +0.0882]
Energy	0.7519	+0.0418	[+0.0116, +0.0719]
DOCTOR ranking transform	0.7449	+0.0485	[+0.0191, +0.0779]
Entropy	0.7657	+0.0277	[-0.0009, +0.0563]

B Operating-point and coverage detail for Table 2

Table 4: **FPR@95 and aggregate-fallback rates for every row of Table 2.** FPR@95 is defined in Section 5.3; fallback is the share of queries routed to the Aggregate reference.

Model (seed)	Shift	FPR@95 Local / Aggregate	Fallback control / shifted
ViT (42)	ImageNet-R	0.905 / 0.958	9.2% / 14.3%
ViT (42)	DTD	0.459 / 0.509	9.2% / 8.4%
Swin (42)	ImageNet-R	0.909 / 0.923	10.1% / 48.7%
Swin (42)	DTD	0.742 / 0.707	10.1% / 55.3%
DINOv2 (42)	ImageNet-R	0.847 / 0.901	10.6% / 5.8%
DINOv2 (42)	DTD	0.531 / 0.557	10.6% / 33.1%
CLIP (42)	ImageNet-R	0.853 / 0.929	13.6% / 23.3%
CLIP (42)	DTD	0.619 / 0.506	13.6% / 1.0%
Swin (43)	ImageNet-R	0.899 / 0.924	9.9% / 45.9%
Swin (43)	DTD	0.615 / 0.636	9.9% / 55.6%