

Single-Pass Conformal Cross-Modal Anomaly Screening with Tabular Foundation Models

José Lucas De Melo Costa, Seong Woo Ahn, Fabrice Popineau, Arpad Rimmel,
and Bich-Liên Doan

Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay,
CNRS, CentraleSupélec, 91190 Gif-sur-Yvette, France

Abstract. A pretrained tabular foundation model already encodes the normal-data manifold; we turn it into a calibrated cross-modal screen with no training. Our target is a rare subpopulation of records that look ordinary in each modality yet sit off the rule binding them. We probe it by reconstructing each feature from the others under an in-context model (TabPFN-v2) whose fusion keeps the cross-modal correlations in the score. The obstacle is that the reference pool is at once the normality model and the calibration set, so exchangeability fails and the realized false-positive rate reaches 0.49-0.68 at a nominal 0.10. Masking the diagonal of the model’s sample-attention matrix largely corrects this in one forward pass, approximating all n leave-one-out calibration scores. This approximation does not exactly restore exchangeability, so the guarantee we report is empirical: a realized false-positive-rate certificate rather than an exact conformal one. Across a six-pair medical evaluation suite this single-pass calibration holds the realized false-positive rate near the nominal level (within about one standard error, $\text{FPR} \leq 0.125$) on five of six pairs, with no held-out split and no refitting, at power on par with the strongest training-free baselines. The fused screen beats both unimodal screens on rare cell states neither margin resolves. Code is available at: <https://github.com/jose-melo/single-pass-conformal-calibration>

Keywords: Multimodal fusion · Cross-modal screening · Conformal calibration · Tabular foundation models · Missing modality · In-context learning

1 Introduction

Co-measuring several modalities pays off only because they are not independent: a learned cross-modal rule ties them together, the transcriptome predicting surface protein and an image agreeing with the record [15,7,13]. Each record respects every margin, yet a small group sits off the joint rule, ordinary in each modality but atypical across them, and no unimodal screen separates it. Our aim is to surface this rare biological subpopulation, an aberrant cell state or clinical stage rather than a corrupted record, before a downstream fusion model absorbs it silently.

We cast the task as anomaly screening over the fused record (Fig. 1). An in-context tabular foundation model makes it training-free: TabPFN-v2 scores a record against a reference pool of assumed-clean inliers [14,19], and reconstructing each feature from the others probes the cross-modal rule, so an inconsistency inflates the standardized residual [17,6].

A screen is useful only if it controls its error rate. Conformal inference supplies finite-sample false-positive control *under exchangeability*, a premise our single-pass detector meets only approximately, so its control is empirical rather than certified, and Benjamini–Hochberg extends it toward false-discovery-rate control across a batch [3,4,18]. To test this across modality pairs, we assemble an evaluation suite of paired medical tables whose unimodal backbone separates calibration validity from cross-modal signal.

The difficulty is the reference pool’s double duty, which breaks exchangeability and makes the gate over-flag (Sec. 3). Our fix masks the sample-attention diagonal, returning all n leave-one-out scores in one forward pass rather than n refits (Sec. 4), while a correlation-preserving fusion keeps the screen cross-modal.

Main contributions:

1. **Single-pass leave-one-out conformal calibration.** Masking the sample-attention diagonal largely corrects the dual-role bias and recovers all n calibration scores in one forward pass, its validity certified empirically rather than by an exact conformal guarantee.
2. **A cross-modal anomaly-screening evaluation suite.** We assemble an evaluation suite for *empirically calibrated* cross-modal anomaly detection on medical multimodal tables: six orthogonal instrument pairs, a nine-table clinical validity backbone, and out-of-scope pairs.
3. **Correlation-preserving fusion and a scope analysis.** A raw-feature fusion ties the score to the cross-modal coupling: it clears *both* unimodal screens on rare cell states neither margin resolves.

2 Related Work

Multimodal fusion and cross-modal anomalies. The strongest medical fusers for image–tabular, histology–omics, and transcriptome–protein data (TIP [7], HEALNet [13], RoVTL [11], totalVI [8], Seurat-WNN [10,15]) return a label or embedding, never a calibrated, error-controlled screen for the event we target: a record plausible in every margin yet jointly atypical. Cross-modal AD is developed mostly for industrial inspection: crossmodal feature mapping [5] and M3DM [24] fuse the two modalities per dataset. Benchmarks like MVTec-3D and a preliminary

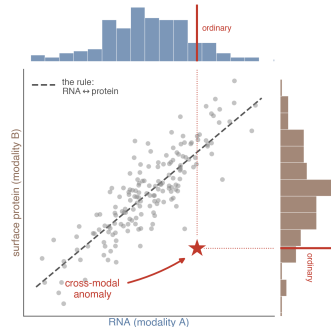


Fig. 1. A cross-modal anomaly. The starred record is ordinary in each margin, yet off the RNA↔protein rule.

e-commerce error-detection set collect such cases, none distribution-free, medical, or tabular, and none says which pairs its residual separates. ADBench [9] places reconstruction among the dependency detectors, predicting those pairs; we make the prediction quantitative.

In-context detection and conformal error control. TabPFN [14] infers in one forward pass by conditioning on a reference set, and FoMo-0D [19] repurposes this for outlier scoring; both emit raw scores with no error control [17,6]. Our contribution is the error control around it: conformal inference turns any score into a calibrated p -value under exchangeability [1,3,16], Benjamini–Hochberg targets FDR control [4,21], and AdaDetect [18] learns the score under it. The refit-free routes [23,2] cost K or n passes, and in-context calibration is silently invalid because the pool is both normality model and calibration set. No prior method calibrates one across fused modalities.

3 Background and Problem Setup

Multimodal records and the cross-modal anomaly. A co-measured record concatenates blocks $x = (x^{(1)}, \dots, x^{(M)})$, $x^{(m)} \in \mathbb{R}^{d_m}$, $d = \sum_m d_m$ (highly-variable genes, protein readouts, image embeddings) into one joint table that TabPFN-v2 consumes [20]. A reference pool $R = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ collects assumed-clean inliers; pathology is never fitted. Our target $x_\star$ is plausible in every block yet sits off the joint rule: a *dependency* anomaly in the ADBench taxonomy [9], as opposed to a *clustered* anomaly forming a second mode.

An in-context reconstruction score. A pretrained tabular foundation model predicts *in context*, so R acts as a frozen model of normality [14]. We turn this into a dependency tester by *masked reconstruction*: for each coordinate j we predict x_j from x_{-j} under the model conditioned on R , scoring the record by its mean squared standardized residual,

$$s(x) = \frac{1}{d} \sum_{j=1}^d \frac{(x_j - \hat{x}_j)^2}{\hat{\sigma}_j^2}, \quad \hat{x}_j = \mathbb{E}[x_j \mid x_{-j}, R]. \quad (1)$$

Here $\hat{\sigma}_j^2$ is the residual variance fit on reference inliers under the same masked context, so calibration and test standardization stay consistent. Categorical clinical covariates are one-hot encoded and enter Eq. (1) as numeric coordinates. A broken inter-block rule inflates s , so Eq. (1) answers to cross-modal inconsistency rather than marginal extremity, blind to coherent subgroups that reconstruct well.

Turning a score into a distribution-free gate. Given calibration inlier scores $\{s_i\}$ ($i = 1, \dots, n_{\text{cal}}$) and a query score $s(x_\star)$, the conformal p -value is the calibration rank

$$p(x_\star) = \frac{1 + \#\{i : s_i \geq s(x_\star)\}}{n_{\text{cal}} + 1}, \quad (2)$$

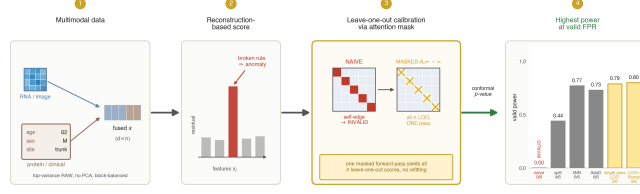


Fig. 2. Single-pass conformal cross-modal screen. (1) Blocks are concatenated into one joint table. (2) An in-context model reconstructs each feature from the rest. (3) Masking the sample-attention diagonal ($A_{ii} \leftarrow -\infty$) yields all n leave-one-out scores in one pass. (4) Conformal p -values with Benjamini-Hochberg give power at a valid FPR.

which under *exchangeability* of the calibration scores with an inlier query’s is super-uniform, so thresholding at α controls the FPR, and Benjamini-Hochberg then targets FDR control at q [1,2,4,18], with Storey’s inlier-fraction estimator sharpening the threshold [21]. Both guarantees are exact only under exchangeability, which an in-context detector forfeits: the pool R both *defines* the score in Eq. (1) and calibrates it, so a calibration point $x_i \in R$ is scored from inside its own context while a query sits outside. Its residual deflates and the gate over-flags, the *dual-role bias* Sec. 4 removes.

An ordinary mole in the wrong context. The lesion looks benign on its own and the metadata (age, body site) is unremarkable on its own; what is off is the combination, since a lesion like this, at this site, in a patient this age, is rare. Fusing image and record under our method catches it at power 0.93 (vs. k -NN 0.75) at a controlled FPR of 0.10, against 0.56 for the uncalibrated screen (ISIC row, Table 1).

4 Method

Three design choices make the detector valid and cross-modal: refit-free calibration, dependency-preserving fusion, and per-pattern recalibration for missing modalities.

4.1 Single-pass leave-one-out calibration

Scoring every calibration point out of its own context removes the dual-role bias, and two textbook routes do so at a cost: *fixed-context calibration* scores a disjoint held-out set for exact control [16,1] but spends that context, and *jackknife* calibration scores each x_i against $R \setminus \{x_i\}$ [23,2] at n refits, prohibitive for a foundation model. For a model with full quadratic sample attention these n refits collapse into one (Fig. 2). Placing R as both context and query block, the model forms a sample-attention matrix $A \in \mathbb{R}^{n \times n}$ at each layer whose diagonal A_{ii} carries query i ’s self-edge to its own context token. Masking that diagonal before every softmax gives an approximate leave-one-out score,

$$s_i \approx f(x_i \mid R \setminus \{x_i\}) \quad \text{after} \quad A_{ii} \leftarrow -\infty, \quad (3)$$

so each query attends to every reference row but itself, and one pass returns all n leave-one-out scores $\{s_i\}$. Test points keep the full pool. Both sets enter Eq. (2), and Benjamini–Hochberg, or its power-recovering Storey variant, is then applied at level q [4,21,3,18]. We instantiate the construction on TabPFN-v2 alone, and every number we report is for that backbone. In principle the argument needs only three properties of the scorer: (i) full quadratic attention over the context samples, exposed and maskable at every layer; (ii) in-context conditioning with no per-query refit; and (iii) a per-feature predictive mean and residual scale to feed Eq. (1). Whether other in-context tabular models with these properties behave the same way is untested here.

Two approximations separate our scores from exact exchangeability. Masking the diagonal still leaves a held-in point reconstructable through correlated neighbours, and across layers its value mixes back through other tokens; both channels deflate its residual and skew ranks small, an anti-conservative leakage that grows with feature redundancy and depth. Conversely, each calibration score sees a context of size $n-1$ against a test point’s full n , a conservative $+1$ offset consistent with the masked/exact scale ratio ≈ 1.02 . A second masked pass on the query would restore exact exchangeability; we keep the single-pass form. Its p -values are therefore not guaranteed super-uniform. The validity certificate we report is accordingly the per-inlier false-positive rate at the fixed level $\alpha=0.10$, which super-uniformity controls directly. Benjamini–Hochberg at $q=0.10$ characterizes batch-discovery FDR and is not the certified quantity (Sec. 6). Where redundancy inflates the leakage, a reference-only prune restores exchangeable calibration at single-pass cost, with CV+ [23,2] the exact fallback.

Correlation-preserving fusion. Eq. (1) predicts each x_j from x_{-j} , reading only conditional structure, so PCA is harmful: decorrelating a modality collapses $\hat{x}_j = \mathbb{E}[x_j \mid x_{-j}]$ toward the marginal mean and $s(x)$ becomes a whitened marginal distance blind to the dependency. We keep each modality’s top-variance *raw* features, so within- and cross-block correlations survive into Eq. (1), and rescale each block so a low-dimensional modality is not swamped.

Per-missingness-pattern calibration. At deployment a query is often observed on only some modalities $m \subseteq \{1, \dots, M\}$. We recalibrate under that pattern, restricting the pool to the same blocks and scoring through one map $f(\cdot \mid R; m)$: masked leave-one-out calibration carries over within a pattern, and power tracks whichever observed modality holds the anomaly.

5 Experimental Setup

Datasets and data use. Our primary suite contains six *orthogonal instrument pairs*, two channels probing the same unit: three CITE-seq panels pairing transcriptome with surface protein (CBMC, 10x PBMC, BMMC) [20], TCGA-BRCA expression with clinical covariates, ISIC dermoscopy embeddings with clinical metadata, and 10x Visium H&E histology with spatial expression. The reference pool is

the majority phenotype ($n_{\text{ref}} \sim 10^2$; ISIC 300) and the target is a rare minority phenotype defined by an independent annotation (cell type, clinical stage, or diagnosis), never by the features scored. The cross-modal necessity test adds two deeper CITE-seq panels, Hao (228 proteins) [10] and Triana (197) [22]; a nine-table clinical backbone [9] certifies calibration validity without cross-modal structure; and four out-of-scope pairs, Multiome RNA+ATAC, TCGA glioma, MOFA CLL, and spatial CosMx NSCLC [12], probe the boundary. All datasets are public, secondary-use, and de-identified; no new human-subject data were collected, no individual-level clinical decision was made, and ethics/consent remain those of the original studies or providers. Raw dermoscopic images are not redistributed: the released code ships precomputed ISIC embeddings and metadata only, and fetches large public sources from their original providers under the corresponding terms of use.

Methods and baselines. Every method shares one conformal scaffold and one score, differing only in calibration. Against our single-pass LOO (Sec. 4.1) and its Storey variant [21] we place naive in-context conformalization, split conformal, cross-conformal (CV+) [2,23], AdaDetect [18], and the strongest of eight classical detectors (k -NN, LOF, and others). Trained fusers (totalVI [8], Seurat-WNN [10,15], TIP [7], HEALNet [13]) return embeddings, not p -values, and wrapping one would require per-dataset training, so we scope the table to the training-free regime.

Pipeline. The frozen scorer is TabPFN-v2 [14], at one operating point $\alpha=q=0.10$ under Benjamini–Hochberg [4]. Realized FPR is our validity certificate: at $n \sim 10^2$ its binomial SE is about 0.03, so we mark a cell *valid* only inside $\text{FPR} \leq 0.125$ (nominal plus about one SE) and read borderline verdicts as noisy. In operational terms a batch of 100 screened inliers then yields roughly 12–13 false flags instead of 10. That tolerance is acceptable for triage, where a few extra records enter manual review, and not for diagnosis: a screen used to rule out disease would need the exact finite-sample guarantee rather than this certificate. Our headline metric, *valid power*, is the mean raw power over all six pairs, with each invalid pair contributing zero. The ISIC valid arm uses a reduced d/n . Masked-LOO rank fidelity to exact n -refit LOO is measured on the one tractable panel (CBMC pure-null, $d=8$, $n=45$). Suite results use 15 seeds with bootstrap FPR intervals.

6 Results

Validity and power. Masked LOO tracks exact n -refit LOO on the low-dimensional CBMC null (Spearman 0.975, scale ratio ≈ 1.02); this rank check is feasible only in that low-redundancy regime, so on the high- d pairs validity rests on the FPR certificate alone. Pooled across the suite, the 8400 inlier p -values exceed the uniform diagonal by at most 0.027, empirically super-uniform across thresholds rather than only at α , and realized FDR is 0.05/0.12/0.23 at $q=0.05/0.10/0.20$. On the six orthogonal pairs (Table 1, top), Storey-adjusted single-pass LOO controls

Table 1. One suite, one conformal scaffold, two regimes. Each cell: raw power (top) over realized per-inlier FPR (bottom), with 95% CIs. Cells with realized FPR > 0.125 are marked **X** and their power greyed, for every method *including ours*, so the table reads in greyscale. An invalid pair contributes 0 to *valid power*, and **bold** marks the best valid construction. *Top*: six orthogonal pairs. *Bottom*: nine clinical unimodal tables, both summary rows reporting *valid* power; Table 2 gives the per-table breakdown and the raw mean. *best classical* = strongest of eight AD detectors; CV+ shown only where its K/n refits were tractable.

dataset	modalities	conformal constructions				AD baselines		
		naive	split	single-pass	LOO	LOO+Storey	CV+	AdaDetect
Orthogonal instrument pairs (cross-modal)								
CBMC CITE-seq	RNA+prot	1.00 [97.1,100]	1.00 [97.1,100]	0.99 [95.1,100]	1.00 [97.1,100]	—	0.96 [91.98]	1.00 [97.1,100]
		0.49 [40.58]✗	0.07 [04.13]	0.11 [09.13]	0.11 [09.13]	—	0.08 [04.14]	0.08 [04.14]
10x PBMC CITE-seq	RNA+14 prot	1.00 [97.1,100]	1.00 [97.1,100]	1.00 [97.1,100]	1.00 [97.1,100]	—	0.99 [95.1,100]	1.00 [97.1,100]
		0.53 [44.62]✗	0.13 [08.20]✗	0.10 [10.14]	0.12 [10.14]	—	0.05 [02.10]	0.11 [07.18]
BMMC CITE-seq	RNA+134 prot	1.00 [97.1,100]	0.93 [87.96]	0.99 [95.1,100]	0.99 [95.1,100]	—	0.98 [94.99]	1.00 [97.1,100]
		0.57 [48.66]✗	0.08 [04.14]	0.12 [09.13]	0.11 [09.13]	—	0.07 [04.13]	0.12 [07.19]
TCGA-BRCA	expr+clin	0.94 [88.97]	0.68 [59.76]	0.83 [75.89]	0.88 [81.93]	—	0.77 [69.84]	0.85 [78.90]
		0.61 [52.69]✗	0.09 [05.15]	0.11 [09.13]	0.11 [09.13]	—	0.11 [07.18]	0.05 [02.10]
ISIC	image+clin	0.99 [95.1,100]	0.00 [00.03]	0.90 [83.94]	0.93 [87.96]	0.92 [86.96]	0.50 [41.59]	0.75 [67.82]
		0.56 [47.65]✗	0.13 [08.20]✗	0.10 [08.11]	0.10 [08.11]	0.09 [05.15]	0.10 [06.17]	0.08 [04.14]
10x Visium (spatial)	H&E+expr	0.94 [88.97]	0.00 [00.03]	0.14 [09.21]	0.56 [47.65]	—	0.17 [11.25]	0.00 [00.03]
		0.68 [59.76]✗	0.08 [04.14]	0.13 [10.13]✗	0.13 [10.13]✗	—	0.08 [04.14]	0.09 [05.15]
valid power ↑		0.00	0.43	0.79	0.80	—	0.73	0.77
controls FPR		0/6	4/6	5/6	5/6	—	6/6	6/6
Validity backbone (unimodal, 9 clinical+omics tables)								
valid power ↑		0.00	0.25	0.37	0.45	0.41	0.18	0.29
controls FPR		0/9	7/9	8/9	8/9	9/9	9/9	8/9

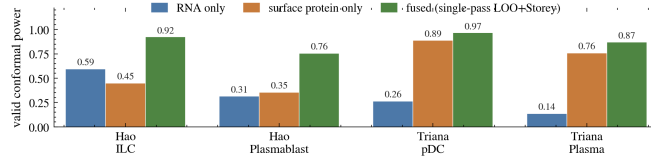


Fig. 3. Cross-modal necessity. Valid power (LOO+Storey) for RNA alone, protein alone and the fused record on four rare cell states; the fused screen matches or exceeds *both* margins on every state.

FPR on five of six at valid power 0.80 (plain LOO 0.79), naive conformalization is invalid on all six (FPR 0.49–0.68), and split conformal reaches only 0.44. AdaDetect and k -NN control FPR 6/6 at 0.73 and 0.77. The largest valid margin is ISIC; on boundary-uncertain Visium raw power 0.56 tops every baseline (≤ 0.17) but is uncertified. Visium is also the most redundant pair in the suite, as H&E morphology and spatial expression measure the same tissue section; the anti-conservative leakage of Sec. 4.1 grows with feature redundancy, and it is on this pair that the certificate fails. On the nine clinical unimodal tables, single-pass LOO+Storey leads at valid power 0.45 (8/9) and plain LOO at 0.37 (8/9), above CV+’s 0.41 (which costs K refits), AdaDetect’s 0.18 and the best classical detector’s 0.29 (Table 2).

Cross-modal necessity, fusion, and dropout. On the deep Hao and Triana panels the fused screen clears *both* unimodal screens on all four rare states (Fig. 3). Per-modality PCA fusion collapses ISIC power to 0.18. Per-pattern recalibration

Table 2. Per-table backbone: power over realized FPR, each with a seed-bootstrap 95% CI. Marking follows Table 1, for every method *including ours*. *mean power (raw)* averages the top number over all nine tables; *valid power* scores an invalid table as 0 and is the quantity Table 1 reports. Bold marks the best *valid* construction: naive attains the highest raw mean (0.63) but controls FPR on 0/9 tables, so it is never bolded.

table	conformal constructions				AD baselines		
	naive	split	single-pass LOO	LOO+Storey	CV+	AdaDetect	best classical
annthyroid	0.52 [23, 80] 0.17 [14, 20] ✗	0.26 [00, 53] 0.11 [08, 13]	0.44 [20, 70] 0.10 [08, 12]	0.55 [27, 83] 0.10 [08, 12]	0.47 [26, 68] 0.10 [07, 12]	0.23 [00, 46] 0.07 [04, 09]	0.29 [11, 48] 0.10 [07, 14]
thyroid	0.82 [74, 89] 0.17 [14, 19] ✗	0.45 [18, 72] 0.10 [08, 12]	0.77 [66, 86] 0.10 [08, 12]	0.81 [70, 90] 0.10 [08, 12]	0.75 [66, 84] 0.10 [07, 12]	0.41 [16, 66] 0.08 [05, 11]	0.38 [09, 67] 0.09 [07, 10]
cardio	0.82 [75, 87] 0.37 [34, 40] ✗	0.35 [13, 57] 0.10 [07, 13]	0.48 [29, 66] 0.11 [09, 13]	0.68 [53, 80] 0.11 [09, 13]	0.46 [33, 57] 0.08 [05, 11]	0.24 [06, 44] 0.10 [06, 14]	0.37 [13, 60] 0.12 [09, 16]
Cardiotocography	0.35 [23, 48] 0.32 [26, 39] ✗	0.05 [00, 14] 0.08 [06, 10]	0.09 [02, 16] 0.11 [08, 13]	0.20 [14, 26] 0.11 [08, 13]	0.11 [04, 19] 0.09 [08, 10]	0.14 [04, 28] 0.08 [06, 11]	0.09 [00, 23] 0.11 [08, 14]
mammography	0.60 [44, 73] 0.23 [18, 29] ✗	0.45 [29, 58] 0.09 [07, 11]	0.40 [24, 55] 0.12 [09, 14]	0.51 [38, 62] 0.12 [09, 14]	0.36 [22, 48] 0.10 [07, 13]	0.09 [00, 21] 0.08 [06, 10]	0.17 [00, 35] 0.11 [07, 15]
breastw	0.98 [97, 99] 0.16 [12, 19] ✗	0.68 [38, 95] 0.10 [07, 13]	0.93 [84, 98] 0.11 [09, 14]	0.99 [99, 1.00] 0.11 [09, 14]	0.97 [93, 98] 0.10 [08, 13]	0.14 [00, 35] 0.04 [01, 08]	0.93 [87, 98] 0.10 [06, 13]
Pima	0.07 [00, 14] 0.17 [14, 21] ✗	0.00 [00, 00] 0.08 [06, 10]	0.00 [00, 00] 0.10 [08, 13]	0.01 [00, 03] 0.10 [08, 13]	0.00 [00, 00] 0.10 [08, 11]	0.00 [00, 00] 0.07 [05, 10]	0.04 [00, 11] 0.14 [11, 17] ✗
uci.ctg	0.50 [36, 60] 0.33 [30, 38] ✗	0.11 [00, 28] 0.13 [09, 19] ✗	0.35 [20, 49] 0.13 [10, 16] ✗	0.41 [28, 53] 0.13 [10, 16] ✗	0.44 [28, 56] 0.11 [09, 14]	0.28 [11, 45] 0.09 [06, 11]	0.28 [06, 51] 0.08 [04, 12]
uci.wdbc	0.97 [93, 1.00] 0.70 [65, 76] ✗	0.08 [00, 24] 0.14 [11, 17] ✗	0.25 [10, 41] 0.12 [10, 14]	0.26 [10, 43] 0.12 [10, 14]	0.17 [00, 34] 0.08 [05, 10]	0.07 [00, 22] 0.09 [07, 12]	0.09 [00, 26] 0.09 [07, 11]
mean power (raw) ↑	0.63	0.27	0.41	0.49	0.41	0.18	0.29
valid power ↑	0.00	0.25	0.37	0.45	0.41	0.18	0.29
controls FPR	0/9	7/9	8/9	8/9	9/9	9/9	8/9

holds on all 21 ISIC dropout regimes ($\text{FPR} \leq 0.111$): power moves $0.81 \rightarrow 0.79$ when the image is dropped and $0.81 \rightarrow 0.45$ when the clinical block is dropped.

7 Discussion

Once validity is enforced our method beats the best classical detector, 0.80 against 0.77. The gain comes from the fusion rather than from raw ranking: the construction is split-free and single-pass, and it reads a cross-modal signal a marginal k -NN cannot. We default to plain single-pass LOO, using Storey only when power-limited. Beating *both* unimodal screens places the signal in the coupling, and the effect reverses under PCA, which destroys the conditional structure the score reads. Two conditions gate the method: first, the blocks must be *orthogonal*, correlated layers keep a subtype reconstructable; second, the anomaly must break a *dependency*, not sit in a low-density pocket: CosMx NSCLC’s discordant cells are local outliers, reconstruction reading AUROC 0.65 at zero valid power where LOF (0.81) and a supervised probe (0.89) lead, a *clustered* anomaly [9].

Limitations. Our validity is empirical: the per-inlier FPR is certified at one operating point ($\alpha=0.10$, $n \sim 10^2$), and the masked-pass $\approx$ exact-refit equivalence only on the low-dimensional CBMC panel. The clean-pool assumption is untested, and contamination acts through two channels that both cost power rather than validity. Contaminants in the *calibration* set fatten the right tail of $\{s_i\}$, so an inlier query’s rank in Eq. (2) rises, p -values grow and the gate turns *conservative*.

Contaminants in the *context* widen the model of normality and inflate $\hat{\sigma}_j^2$, so a true anomaly’s standardized residual deflates and its score moves toward the reference bulk. Neither channel pushes the realized FPR up, so we expect it to hold or fall under contamination, with the loss showing as missed detections. Measuring where that breakdown sets in, by injecting a known fraction of target-phenotype records into the reference pool, is the experiment this limitation calls for.

8 Conclusion

A single masked forward pass turns a frozen tabular foundation model into a calibrated, split-free screen for rare cross-modal subpopulations. Setting each query’s self-attention edge to $-\infty$ removes the dual-role bias that drives the naive false-positive rate to 0.49–0.68, and returns every leave-one-out calibration score at once. On this suite it is the only single-pass in-context route that stays valid, at power on par with the strongest training-free baselines, and the fused score flags rare cell states neither modality resolves alone. Future work targets local contradictions and a finite-sample guarantee in place of the empirical certificate.

Acknowledgments. This work was performed using HPC resources from the “Mésocentre” computing center of CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France (<http://mesocentre.centralesupelec.fr/>). This work was also granted access to the HPC resources of IDRIS under the allocation AD011011828R6 made by GENCI. This research publication is supported by the Chair “Artificial intelligence applied to credit card fraud detection and automated trading” led by CentraleSupélec and sponsored by the LUSIS company, and by a CIFRE fellowship. The authors thank the International Skin Imaging Collaboration (ISIC) and 10x Genomics for their public resources. The results published here are in whole or part based upon data generated by the TCGA Research Network (<https://www.cancer.gov/tcga>).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning* **16**(4), 494–591 (2023)
2. Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Predictive inference with the jackknife+. *The Annals of Statistics* **49**(1), 486–507 (2021)
3. Bates, S., Candès, E., Lei, L., et al.: Testing for outliers with conformal p-values. *The Annals of Statistics* **51**(1), 149–178 (2023)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**(1), 289–300 (1995)

5. Costanzino, A., Ramirez, P.Z., Lisanti, G., Di Stefano, L.: Multimodal industrial anomaly detection by crossmodal feature mapping. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
6. Ding, X., Wen, H., Klüttermann, S., Akoglu, L.: From zero to hero: Advancing zero-shot foundation models for tabular outlier detection. arXiv preprint arXiv:2602.03018 (2026)
7. Du, S., Zheng, S., Wang, Y., et al.: TIP: Tabular-image pre-training for multimodal classification with incomplete data. In: European Conference on Computer Vision (ECCV) (2024)
8. Gayoso, A., Steier, Z., Lopez, R., et al.: Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods* **18**(3), 272–282 (2021)
9. Han, S., Hu, X., Huang, H., et al.: ADBench: Anomaly detection benchmark. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
10. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., et al.: Integrated analysis of multimodal single-cell data. *Cell* **184**(13), 3573–3587 (2021)
11. Hasny, M., et al.: No data? no problem: Robust vision-tabular learning with missing values. arXiv preprint arXiv:2512.19602 (2025), roVTL
12. He, S., Bhatt, R., Brown, C., et al.: High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology* **40**(12), 1794–1806 (2022)
13. Hemker, K., Simidjievski, N., Jamnik, M.: HEALNet: Multimodal fusion for heterogeneous biomedical data. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
14. Hollmann, N., Müller, S., Purucker, L., et al.: Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025)
15. Huang, S.C., Pareek, A., Seyyedi, S., et al.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine* **3**(1), 136 (2020)
16. Laxhammar, R., Falkman, G.: Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence* **74**, 67–94 (2015)
17. Lu, S., Liu, J., Peters, S., et al.: uLEAD-TabPFN: Uncertainty-aware dependency-based anomaly detection with TabPFN. arXiv preprint arXiv:2604.20255 (2026)
18. Marandon, A., Lei, L., Mary, D., Roquain, E.: Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics* **52**(1), 157–183 (2024), adaDetect
19. Shen, Y., Akoglu, L., et al.: FoMo-0D: A foundation model for zero-shot tabular outlier detection. arXiv preprint arXiv:2409.05672 (2024)
20. Stoeckius, M., Hafemeister, C., Stephenson, W., et al.: Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**(9), 865–868 (2017)
21. Storey, J.D.: A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B* **64**(3), 479–498 (2002)
22. Triana, S., Vonficht, D., Jopp-Saile, L., et al.: Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nature Immunology* **22**(12), 1577–1589 (2021)
23. Vovk, V.: Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* **74**, 9–28 (2015)
24. Wang, Y., Peng, J., Zhang, J., et al.: Multimodal industrial anomaly detection via hybrid fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023), m3DM