SPOT-FM: A Zero-Shot Hypothesis Test for Tabular Anomalies via In-Context Synthetic Contrasts

Mayank Jauhari [1] **Amir Raza** [1] **Abishek Sankararaman** [2] **Anurag Tripathi** [1] **Vipul Joshi** [1] **Vikash Sharma** [1] **Boris N. Oreshkin** [3]

Abstract

We formulate zero-shot tabular anomaly detection as a composite hypothesis test: H_0 states that a query row is drawn from a reference distribution; the alternative is defined by a learned sampler law. We present SPOT-FM, a frozen Tab2D transformer whose classification logit serves as a learned, data-adaptive test statistic, instantiated from a user-supplied reference set with no per-dataset training. The test's alternative hypothesis is constructed via *minimal-subset perturbation*: only a μ-fraction of features are altered per synthetic contrast, preserving inter-feature structure while producing controlled departures from the empirical quantile box. We prove a data-dependent power lower bound against this alternative (Thm. G.1), establish exchangeability of the test statistic under reference-set permutations (Thm. G.5), and decompose deployment risk into pre-training fit, task shift, and a sampler-label residual that names where zero-shot transfer fails (Prop. G.6). Inference is a single tensorized forward pass (≈11 ms / 1K rows). On 17 ODDS benchmarks SPOT-FM obtains 0.915 average AUC-ROC and 0.752 AUC-PR, improving over FoMo-0D (Shen et al., 2024) on $12/17$ datasets.

1. Introduction

Deciding whether a tabular row is anomalous is a composite hypothesis test. H_0: the row is drawn from the user's reference distribution P_0 implicit in X_{ref}; H_1: the row is drawn from any law outside an empirical quantile box around P_0. Classical tabular AD refits a new test per dataset: a new detector, with its own HPO, for every deployment. We instead ask whether a *frozen, zero-shot test* can work across tabular distributions it has never seen.

[1]Amazon, India [2]AWS, USA [3]Amazon, USA. Correspondence to: Mayank Jauhari <jauharim@amazon.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

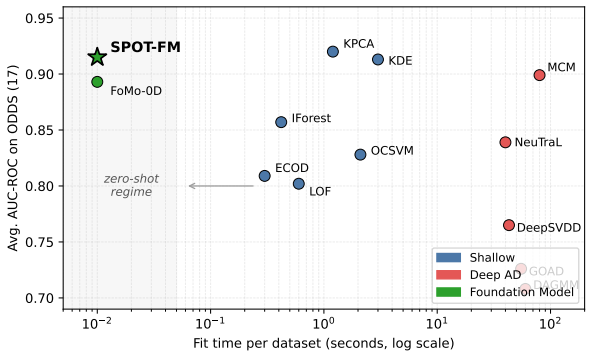
TabPFN (Hollmann et al., 2023) showed that a single transformer pre-trained on synthetic tasks can classify new tabular datasets zero-shot, via in-context learning. FoMo-0D (Shen et al., 2024) extends this to AD, but fixes the alternative to covariance-inflated Gaussian-mixture samples. That alternative aligns poorly with datasets whose anomalies are unusual *combinations* of otherwise-normal features (CARDIO, CARDIOTOCOGRAPHY, SATELLITE; see Table 2). We replace this global-direction alternative with the minimal-subset sampler law, which preserves inter-feature structure while producing controlled departures from the empirical operational region.

We present SPOT-FM, a foundation model whose synthetic negatives come from a *minimal-subset perturbation* sampler: a small fraction μ of features is perturbed per synthetic anomaly. An unperturbed anomaly is indistinguishable from normal; a maximally perturbed one is trivially separable; the useful training signal lives between, and μ places it there. We lift this sampler into the pre-training loop of a Tab2D transformer (Zhang et al., 2025). At inference the user supplies only normal reference data; the frozen model instantiates both a test statistic and an implicit rejection rule, scores test points in a single tensorized pass, and averages the 216-configuration grid of Appendix D.2 into one forward pass.

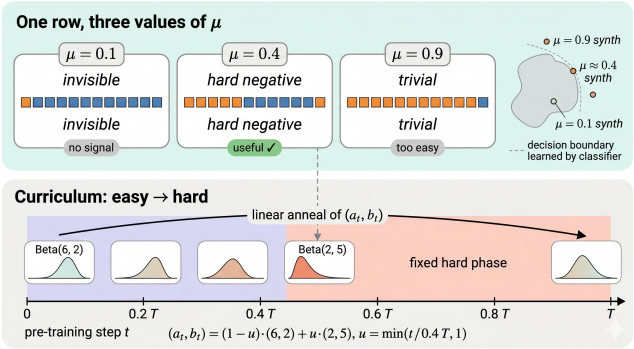
Contributions.

1. A zero-shot composite hypothesis test for tabular AD, requiring no per-dataset fitting or labeled anomalies.
2. Empirical power: 0.915 AUC-ROC and 0.752 AUC-PR on 17 ODDS benchmarks, improving over FoMo-0D on $12/17$ datasets (Table 2).
3. Test-time cost: a single tensorized forward pass at ≈11 ms / 1K points.
4. A data-dependent power lower bound against the uniform-sampler alternative (Thm. G.1), exchangeability of the test statistic (Thm. G.5), and a diagnostic risk decomposition (Prop. G.6).

The rest of the paper is organized as follows. §2 positions



(a) **Zero-shot pays off.** SPOT-FM lands in the foundation-model corner of the accuracy-vs-fit-time plane, competitive average AUC-ROC on 17 ODDS datasets with *no per-dataset training*, up to 100× faster than deep baselines.



(b) **Minimal-subset perturbation.** *Top:* one normal row rendered as a synthetic negative at $\mu \to 0$ (invisible), $\mu \approx 0.4$ (hard), and $\mu \to 1$ (trivial). *Bottom:* the μ-curriculum anneals $\text{Beta}(6,2) \to \text{Beta}(2,5)$ over the first 40% of pre-training; fixing $\mu \sim \mathcal{U}(0.05, 0.95)$ costs -1.3 ROC (Appendix B.2).

Figure 1. Headline result (left) and the mechanism that produces it (right).

SPOT-FM among learned tests and zero-shot AD methods. §3 describes the method: the sampler (§3.2), the hypothesis-testing interpretation (§3.3), and the Tab2D architecture. §4 states three structural properties with proof sketches. §5 reports empirical results on 17 ODDS benchmarks.

2. Background and Related Work

Zero-shot tabular anomaly detection. Classical and deep tabular AD (IFOREST (Liu et al., 2008), OCSVM (Schölkopf et al., 2001), DAGMM (Zong et al., 2018), GOAD (Bergman & Hoshen, 2020)) fit a fresh test per dataset. Tabular foundation models inherit in-context learning from PFNs (Müller et al., 2022; Hollmann et al., 2023) and the Tab2D backbone (Zhang et al., 2025) but target classification. FOMO-0D (Shen et al., 2024) adapts this recipe to AD by pre-training on Gaussian-mixture tasks with covariance-inflated outliers. SPOT-FM replaces that global-direction alternative with a feature-subset sampler law: only μF coordinates are resampled, the rest are copied from a real row, so the contrast respects observed inter-feature structure.

Learned hypothesis tests. Our test framing connects to a body of work on learning a test statistic from data. Classifier two-sample tests (Lopez-Paz & Oquab, 2017; Friedman, 2003) train a classifier to distinguish samples from two distributions and read the accuracy as a test statistic. Deep-kernel two-sample tests (Liu et al., 2020; Kirchler et al., 2020) learn a kernel that maximizes test power. E-values and anytime-valid inference (Vovk & Wang, 2021; Ramdas et al., 2023) offer a route to testing under adaptivity and optional stopping. Conformal outlier detection (Bates et al., 2023) turns any anomaly score into calibrated per-row p-values using a held-out inlier sample. These approaches learn a *per-task* test; SPOT-FM learns a *zero-shot* test applicable to unseen tabular distributions. SPOT-FM is not anytime-valid (the test statistic is a fixed function of the support at deployment) but shares the frozen, data-adaptive-test motivation.

Connection to classifier two-sample tests. Lopez-Paz and Oquab (Lopez-Paz & Oquab, 2017) showed that a classifier trained to distinguish two samples provides a test statistic whose accuracy is a consistent estimator of the total variation distance. SPOT-FM instantiates an analogous idea for anomaly detection: the model is trained to classify normal rows ($y = 0$) vs. synthetic contrasts ($y = 1$), and the resulting classification logit serves as the test statistic $s_\theta(x; X_{\text{ref}})$. The key difference is *zero-shot generalization*: unlike a per-dataset classifier test, the pre-trained model applies to any new reference set without retraining.

Connection to kernel and deep-kernel tests. Deep-kernel MMD tests (Liu et al., 2020) learn a kernel that maximizes test power against a specific alternative. Our setup is analogous: pre-training on synthetic tasks implicitly optimizes the learned representation for detection power against the minimal-subset alternative family. The 216-configuration grid at inference parallels the kernel-selection step in deep-MMD: it searches over a structured space of alternatives (μ, strategy, permutation) to maximize sensitivity. Unlike deep-kernel tests, however, SPOT-FM does not require per-deployment kernel optimization; the representation is frozen.

Pre-specified design and practical testing. At deployment, the test is fully pre-specified: the model parameters

Table 1. **Design comparison across representative tabular AD baselines.** Each row is a binary property of the training or inference pipeline. SPOT-FM combines zero-shot inference with per-feature tokenization, subset-based synthetic contrasts, and tensorized multi-configuration inference.

Property	IFOREST	DEEPSVDD	FOMO-0D	**SPOT-FM**
Zero-shot (no per-dataset fit)	×	×	✓	✓
No per-dataset HPO	×	×	✓	✓
No labeled anomalies at train time	✓	✓	✓	✓
Per-feature tokens (per-column repr.)	×	×	×	✓
Feature-subset contrasts (min-subset sampler)	×	×	×	✓
μ-curriculum (hardness anneal)	n/a	n/a	×	✓
Auxiliary μ-prediction head	n/a	n/a	×	✓
Multi-configuration inference averaging	×	×	✓	✓
Single tensorized inference pass	×	✓	✓	✓
Pre-trained on synthetic-task distribution	×	×	✓	✓

θ are fixed, the reference set X_{ref} defines the null, and the 216-configuration grid is predetermined. No data-snooping occurs; the test row x is compared against a frozen decision function. This pre-specified structure is important for practical deployment in experimentation platforms (Johari et al., 2022) where the testing protocol must be declared before observing outcomes.

Full per-family citations are in Appendix C; Table 1 summarizes the design comparison.

3. SPOT-FM

3.1. Overview

SPOT-FM is a Tab2D transformer f_θ pre-trained on a distribution over AD tasks. Each task is a triple $(X_{\text{norm}}, X_{\text{synth}}, X_{\text{query}})$: normal points from a prior, matching synthetic negatives from our minimal-subset perturbation sampler, and a query set. After pre-training, θ is frozen; inference on a new dataset uses *only* the forward pass. (By *frozen* we mean the pre-trained parameters θ are held constant at inference; no fine-tuning, no gradient updates.) Fig. 2 summarizes the flow; notation is in Table 4 (Appendix A).

3.2. AD Task Generator

The negative side is the design choice. Given a batch of normal points $X \in \mathbb{R}^{N\times F}$, we perturb $\lceil \mu F \rceil$ features per synthetic anomaly; the remaining features are copied from X. A negative with $\mu \to 0$ is indistinguishable from normal; a negative with $\mu \to 1$ is trivially separable; the useful training signal lives at intermediate μ (Fig. 1(b)). We draw μ from a Beta distribution with curriculum $\text{Beta}(6,2) \to \text{Beta}(2,5)$ (easy $\to$ hard). The sampler then picks one of five perturbation strategies (UNIFORM, GAUSSIAN, MIXTURE, BOUNDARY, EMPIRICAL); each respects an expansion factor α that widens the observed feature range.

Perturbation strategies. For each perturbed coordinate $k \in \mathcal{I}$, let $L_k = x_k^{\min} - \alpha|x_k^{\min}|$ and $U_k = x_k^{\max} + \alpha|x_k^{\max}|$ be the expanded range. The five strategies are:

$$
\begin{aligned}
\text{UNIFORM:}\quad & \tilde{x}_k \sim \mathcal{U}[L_k, U_k] \\
\text{GAUSSIAN:}\quad & \tilde{x}_k \sim \text{clip}\big(\mathcal{N}(\mu_k, (\beta\sigma_k)^2), L_k, U_k\big), \\
& \beta \sim \mathcal{U}(1,3) \\
\text{MIXTURE:}\quad & \tilde{x}_k \sim \tfrac{1}{2}\mathcal{U}[L_k, U_k] \\
& \quad + \tfrac{1}{2}\,\text{clip}\big(\mathcal{N}(\mu_k, \sigma_k^2), L_k, U_k\big) \qquad (1) \\
\text{BOUNDARY:}\quad & \tilde{x}_k \sim \mathcal{U}[L_k, q_k^5] \cup \mathcal{U}[q_k^{95}, U_k] \\
\text{EMPIRICAL:}\quad & \tilde{x}_k = x_k^{(r)} + \varepsilon, \\
& r \sim \text{Unif}\{1,\ldots,N\}, \\
& \varepsilon \sim \mathcal{N}(0, 0.01^2)
\end{aligned}
$$

At training time s is sampled uniformly per task; at inference the grid sweeps 3 strategies $\times$ 9 values of μ $\times$ 8 feature permutations $=$ 216 configurations (Appendix D.2).

μ-curriculum. The Beta parameters anneal linearly over the first 40% of training:

$$(a_t, b_t) = (1-u)\,(6,2) + u\,(2,5), \quad u = \min(t\,/\,0.4T,\ 1).$$

Early training sees mostly large μ (easy, clearly synthetic negatives); late training sees mostly small μ (hard, near-normal negatives that force the model to learn subtle boundary structure). Ablation: fixing $\mu \sim \mathcal{U}(0.05, 0.95)$ throughout training costs -1.3 ROC on average (Appendix B.2).

Normal-data prior pool. The model must generalize to unseen tabular distributions at inference. To achieve this, we pre-train on a diverse pool of 22 synthetic data generators ("priors") organized into four families:

Hand-crafted families (7). These cover canonical distributional shapes: Gaussian (unimodal), Mixture (multi-modal clusters), Manifold (low-dimensional structure embedded in high dimensions), CorrelatedCluster (clusters with non-trivial correlation structure), TreeBoundary (piecewise-linear decision regions), Uniform (featureless baseline), and HeavyTail (Pareto-like marginals). Together they ensure the model sees both well-separated and overlapping cluster geometries.

Mitra supervised priors (10). Following the Tab2D pre-training recipe (Zhang et al., 2025), we generate tabular data by fitting a supervised model (Tree, RandomForest, ExtraTree, GradientBoosting, kNN, LogReg, NN, LightGBM,

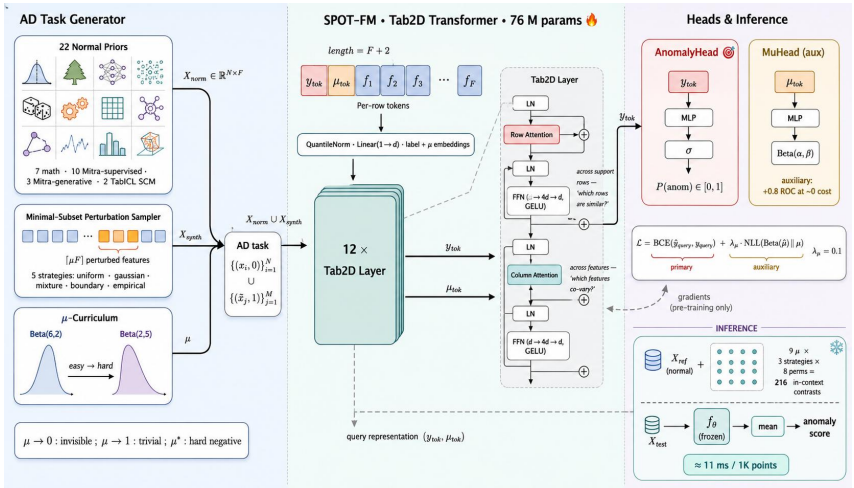


Figure 2. **SPOT-FM architecture.** A task generator couples 22 normal priors with a μ-curriculum-driven negative sampler; the Tab2D transformer ingests $[\, y_{\text{tok}} \mid \mu_{\text{tok}} \mid f_1 \ldots f_F]$ and emits an anomaly score plus auxiliary μ. Fire = trainable during pre-training; snowflake = frozen at inference (single tensorized pass over the 216-configuration grid, §D.2).

Algorithm 1 AD task (one batch)

1: Draw prior π, features F, samples N, strategy s, α
2: $X_{\text{norm}} \leftarrow \pi.\text{sample}(N, F)$
3: $\mu \sim \text{Beta}(a_t, b_t)$ {curriculum}
4: Choose $\lceil \mu F \rceil$ feature indices $\mathcal{I}$ per row
5: $X_{\text{synth}} \leftarrow X_{\text{norm}}$; $X_{\text{synth}}[:, \mathcal{I}] \leftarrow \text{Sampler}_s(\alpha)$
6: Split into support $(X_{\text{norm}} \cup X_{\text{synth}})$ and query

XGBoost, CatBoost) to a random regression/classification task and treating its predictions as the feature distribution. This yields realistic marginal and correlation patterns inherited from the fitted model.

Mitra generative priors (3) and TabICL SCMs (2). Three generative priors sample from fitted density models (SampleRF, SampleGP, TabPFN-style), and two structural-causal-model priors (MLP-SCM, Tree-SCM) produce features with causal DAG structure. These complement the supervised priors by adding non-parametric density coverage and causal dependencies.

The four families are mixed with weights 20%/35%/25%/20%. Each training batch draws one prior, samples fresh (N, F) with $N \in [50, 1000]$ and $F \in [5, 500]$, and produces a complete AD task (Algorithm 1). A detailed per-prior description is in Appendix D.1.

Why minimal-subset, not full-covariance. We choose minimal-subset perturbation because it enforces *locality*: each synthetic anomaly stays close to a real row in feature space. This preserves empirical inter-feature structure (the signal the model must learn to recognize "unusual together") while introducing targeted violations in a controlled subset. Covariance inflation, by contrast, moves every coordinate jointly and destroys exactly the structure we want the model to learn from. The parameter μ is the violation budget: small μ gives hard, near-normal contrasts; large μ gives easier, more visibly synthetic ones.

Why frame AD as a hypothesis test. Classical tabular AD methods output a numeric score but leave the decision threshold to the user. Framing the problem as a composite hypothesis test makes the structure explicit: H_0 is that the row comes from the reference distribution, the test statistic is the model's anomaly score, and the alternative is defined by the sampler law. This framing has three practical benefits. First, it clarifies what the model was trained to detect (rows that exit the operational region under a feature-subset perturbation, not arbitrary out-of-distribution shifts). Second, it lets us state a *power* guarantee against the sampler alternative (Theorem G.1). Third, it separates the failure mode cleanly: low power on a deployment means the real anomaly mechanism falls outside the sampler alternative family, which is precisely the prior-coverage diagnostic of §5.2.

3.3. Hypothesis-testing interpretation

At deployment, given user reference X_{ref} and test row x, we frame the decision as

$$H_0 : \; x \sim P_0, \qquad H_1 : \; x \sim Q \text{ with } \Pr_Q\big(x \notin \widehat{S}(X_{\text{ref}})\big) > 0,$$

where P_0 is the reference distribution implicit in X_{ref} and $\widehat{S}(X_{\text{ref}}) = \prod_j [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ is the empirical 5 to 95% quantile box. The SPOT-FM score $s_\theta(x; X_{\text{ref}}) \in [0, 1]$ is a learned, data-dependent *test statistic*; the exit oracle $L_S(x; X_{\text{ref}}) = \mathbf{1}\{x \notin \widehat{S}\}$ is a concrete rejection rule. The

synthetic contrasts the model sees at inference are draws from one explicit choice of alternative: Q_μ, the minimal-subset sampler law at perturbation fraction μ. Against this alternative, Theorem G.1 (Appendix G) lower-bounds the rejection probability by $1-(1-\delta_{\min})^k$ with $k=\lceil \mu F \rceil$, data-dependent through $\delta_{\min}(\alpha; X_{\text{ref}})$.

Consistency and power properties. The SPOT-FM test is not consistent against all alternatives: it has power only against alternatives that produce feature-subset exits from $\widehat{S}$. Against alternatives whose anomalies remain inside the quantile box (e.g., subtle distributional shifts in tail shape), the test has no guaranteed power. This is by design: the alternative family is *specified* by the sampler, and the test's power envelope is determined by how well the sampler family covers the real anomaly mechanism. The power bound (Theorem G.1) is informative only against the sampler alternative; extending it to richer alternative families is an open question.

The 216-configuration inference grid can be viewed as an implicit test-power optimization: averaging over multiple $(\mu, s, \pi_{\text{perm}})$ configurations produces a test statistic that is powerful against a broader set of alternatives than any single configuration. This parallels the kernel-selection step in optimized MMD tests (Liu et al., 2020), where the kernel is chosen to maximize test power.

3.4. Tab2D Transformer

Each training row is packed as a length-$(F+2)$ token sequence $[\, y_{\text{tok}} \mid \mu_{\text{tok}} \mid f_1 \ldots f_F \,]$ where y_{tok} carries the binary label (0, 1, or [MASK]) and μ_{tok} embeds the task's μ value. Conditioning the representation on μ lets the frozen model adapt across deployment configurations where μ varies (Appendix D.2): at small μ contrasts are subtle, at large μ trivial, and the model sees which regime it is in. The backbone is a stack of Tab2D layers (Zhang et al., 2025) alternating (i) *row attention* that lets query rows attend to support rows and (ii) *column attention* that lets each row attend across its features. Each Tab2D layer applies four sub-blocks with residual connections:

$$\begin{aligned} &\text{LN} \to \text{RowAttn} \to \text{LN} \to \text{FFN} \\ &\quad \to \text{LN} \to \text{ColAttn} \to \text{LN} \to \text{FFN}. \end{aligned} \tag{2}$$

Row attention lets each query row attend to every support row (cross-attention over samples, enabling "which reference rows is this query similar to?"). Column attention lets each row attend across its own features (self-attention over columns, enabling "which feature combinations are unusual?"). This dual-axis design enables the model to jointly reason about inter-sample similarity and intra-sample feature interactions. The model uses dim=384, 8 layers, 8 heads; full parameter breakdown in Appendix D. Two heads read y_{tok}: an ANOMALYHEAD emits a per-row anomaly logit; a MUHEAD reconstructs the task-level μ signal from the shared representation (output as a Beta), acting as an auxiliary regularizer during pre-training (Appendix D.3). The training loss combines binary cross-entropy on the anomaly label with a small auxiliary μ-prediction term,

$$\mathcal{L} = \underbrace{\text{BCE}(\hat{y}_{\text{query}}, y_{\text{query}})}_{\text{primary}} + \lambda_\mu \underbrace{\text{NLL}(\text{Beta}(\hat{\mu}) \,\|\, \mu)}_{\text{auxiliary}}, \tag{3}$$

with $\lambda_\mu = 0.1$. The auxiliary term adds $+0.8$ ROC in the ablation setting (ODDS average; $0.805 \to 0.813$, Appendix D.3) at negligible compute cost, and is retained in the final model. Muon (Jordan et al., 2024) optimizes the 2-D attention weights; AdamW handles the rest; precision is bfloat16 with gradient scaling. Appendix D lists parameter counts and optimizer hyperparameters.

3.5. Inference: In-Context Synthetic Contrasts

Given user-supplied reference normal data X_{ref} and test points X_{test}, we freeze the pre-trained SPOT-FM f_θ and build the support entirely from X_{ref} and the task generator's synthetic negatives, no real labeled anomalies, no gradient updates, no per-dataset HPO. Following standard PFN inference practice (Hollmann et al., 2023; Shen et al., 2024), scores are averaged over a small grid of in-context task configurations (all queried in one tensorized forward pass of the frozen backbone): 9 perturbation fractions $\mu \in \{0.1, 0.2, \ldots, 0.9\}$, 3 sampling strategies (UNIFORM, MIXTURE, BOUNDARY), and 8 random feature permutations, yielding 216 in-context scores per test row that are averaged into the final anomaly probability. For each configuration the task generator emits X_{synth} on-the-fly from X_{ref}, the support is X_{ref} (label 0) $\cup$ X_{synth} (label 1), and the query is X_{test}. This entire protocol is a property of the pre-trained model; there is no model-level ensembling.

4. Structural Properties of the Learned Test

We state three structural properties of the deployed test. Full proofs are in Appendix G; here we give the statements and proof intuitions.

4.1. Power bound against the sampler alternative

Theorem 4.1 (Power bound)**.** *Under the* UNIFORM *sampler with perturbation cardinality $k = \lceil \mu F \rceil$ and data-dependent per-feature exit probability $\delta_{\min}(\alpha; X_{\text{ref}})$, conditional on the source row lying in $\widehat{S}(X_{\text{ref}})$:*

$$\Pr\big(\tilde{x} \notin \widehat{S}(X_{\text{ref}}) \;\big|\; x \in \widehat{S}\big) \;\geq\; 1-(1-\delta_{\min})^k.$$

Proof sketch. Each perturbed coordinate independently exits the quantile interval with probability at least $\delta_{\min}$. The

contrast fails to exit only if all k coordinates stay inside, which happens with probability at most $(1-\delta_{\min})^k$. □

The BOUNDARY sampler exits the quantile box with probability one for every perturbed coordinate (by construction: it samples exclusively from the tails outside the 5 to 95% interval) and is recorded as a separate observation.

4.2. Exchangeability of the test statistic

Theorem 4.2 (Exchangeability)**.** *If the predictor class uses token-wise normalization, no positional signal on row tokens, and symmetric padding masks, then for every permutation $\sigma \in S_n$ of the reference support:*

$$\bar{f}_\theta\big(\sigma \cdot X_{\mathrm{ref}},\, x\big) \;=\; \bar{f}_\theta\big(X_{\mathrm{ref}},\, x\big).$$

Proof sketch. Column attention acts within each row and commutes with any row permutation trivially. Row attention is equivariant because no row-position encoding exists: the query for each row depends on content tokens only (Appendix G, Architectural condition G.4). The composition of equivariant layers preserves equivariance through the full stack; the final per-query read-out is therefore invariant to the support order. □

Exchangeability ensures the test statistic is a *symmetric* function of the reference set. This is a necessary consistency property: the score assigned to a query should not change if the user re-shuffles X_{ref}.

4.3. Deployment-risk decomposition

Proposition 4.3 (Risk decomposition)**.** *Under clipped-BCE boundedness (loss bounded by $B = \log(1/\epsilon)$), the deployment risk admits the pointwise identity:*

$$R_{\mathcal{P}}(\theta) \;=\; R_{\mathcal{T}}(\theta) \;+\; \big[R_{\mathcal{P}^{\mathrm{exit}}}(\theta) - R_{\mathcal{T}}(\theta)\big] \;+\; \epsilon_{\mathrm{Lab}}(\mathcal{P}),$$

where $R_{\mathcal{T}}$ is the pre-training risk, $\mathcal{P}^{\mathrm{exit}}$ is the deployment distribution relabeled by the exit oracle L_S, and $\epsilon_{\mathrm{Lab}}(\mathcal{P})$ is the deployment-side sampler-label residual.

Interpretation. The decomposition identifies three sources of deployment error:

1. *Pre-training fit*: the model may not perfectly minimize its own synthetic training loss.
2. *Feature/task shift*: the deployment features may fall outside the synthetic task family's coverage.
3. *Label mismatch*: real anomaly labels may disagree with the exit-rule oracle (the sampler-label residual ϵ_{Lab}).

On ODDS, source (3) dominates: real anomalies are domain-specific corruptions, not minimal-subset perturbations. The residual ϵ_{Lab} is generically $\Theta(B)$, so the decomposition is diagnostic (it names failure modes) rather than a quantitative certification of small error.

5. Experiments

5.1. Setup

We evaluate on the 17 ODDS tabular AD benchmarks under the standard protocol (50% of inliers for the support, the remaining 50% of inliers plus all outliers for the test set), comparing against 13 baselines spanning shallow, deep, and foundation-model families, including FOMO-0D (Shen et al., 2024), MCM (Yin et al., 2024), GOAD (Bergman & Hoshen, 2020), DAGMM (Zong et al., 2018), and NEUTRAL (Qiu et al., 2021). Classical and deep numbers are drawn from the ADBench (Han et al., 2022) re-evaluation. SPOT-FM is a single Tab2D transformer (dim=384, 8 layers, 8 heads, $F_{\max}$=500) pre-trained on our synthetic AD task distribution with per-feature z-score input normalization; the frozen checkpoint is selected on held-out synthetic validation (full recipe in Appendix D).

5.2. Detection power on 17 ODDS benchmarks

SPOT-FM obtains 0.915 AUC-ROC (Table 2) and 0.752 AUC-PR (Table 7), the highest averages in our 17-dataset panel across zero-shot, classical, and deep baselines. AUC-ROC integrates test power across the FPR axis, so these averages read as average power of the learned test against the empirical outlier mixtures. SPOT-FM improves over FOMO-0D on $12/17$ datasets on each metric, with the largest ROC gains on MAMMOGRAPHY ($+0.21$), CAMPAIGN ($+0.12$), and OPTDIGITS ($+0.12$), and the largest PR gains on PENDIGITS ($+0.24$), MAMMOGRAPHY ($+0.13$), and OPTDIGITS ($+0.13$). The remaining $5/17$ losses concentrate on small heavy-tail datasets (GLASS, PIMA, ARRHYTHMIA) whose anomalies remain inside the broad empirical feature envelope induced by the reference set, a regime where a feature-subset exit alternative is less aligned with the real anomaly mechanism (per-dataset breakdown in Appendix E).

Per-dataset observations. Table 2 shows that SPOT-FM obtains its largest gains on datasets where anomalies violate feature-level regularities (MAMMOGRAPHY: medical imaging features; CAMPAIGN: marketing response indicators; OPTDIGITS: pixel intensity patterns). On these datasets the minimal-subset alternative aligns well with the real anomaly mechanism. Conversely, on GLASS and PIMA, anomalies sit inside the broad empirical feature envelope and the exit-rule alternative is misspecified (Proposition G.6, failure mode iii). This connects the empirical results to the deployment-risk decomposition of §4: the sampler-label residual ϵ_{Lab} is large precisely when the real anomaly mechanism does not

Table 2. **Detection power on 17 ODDS benchmarks (AUC-ROC).** Each column is a hypothesis test; AUC-ROC is its average power across the FPR axis. SPOT-FM (frozen zero-shot test) attains the highest average. Best per row in **red**, second-best in blue (ties share the color). SPOT-FM and FoMo-0D (Shen et al., 2024) are zero-shot foundation-model baselines (no per-dataset training, no HPO); the other baselines are fit per dataset.

Dataset	IForest	LOF	OCSVM	ECOD	EIF	GOAD	DAGMM	DeepSVDD	NeuTraL	ICL	NPT-AD	MCM	FoMo-0D	SPOT-FM
arrhythmia	0.811	0.783	0.788	0.688	0.825	0.690	0.704	0.797	0.790	**0.827**	0.714	0.804	0.790	0.806
breastw	0.983	0.843	0.986	0.978	**0.993**	0.884	0.795	0.940	0.918	0.976	0.984	0.992	0.992	0.982
campaign	0.731	0.688	0.759	0.762	0.761	0.523	0.592	0.504	0.677	0.741	0.785	**0.853**	0.653	0.777
cardio	0.929	0.870	0.952	0.791	0.935	0.761	0.744	0.820	0.822	0.904	0.933	0.953	0.943	**0.962**
cardiotocography	0.727	0.643	0.751	0.791	0.737	0.546	0.619	0.496	0.718	0.621	0.719	0.728	0.722	**0.796**
glass	0.693	0.718	0.576	0.672	0.848	0.590	0.618	0.688	0.747	0.874	0.786	0.772	**0.983**	0.873
ionosphere	0.913	0.923	0.890	0.848	0.949	0.923	0.726	0.913	0.956	0.975	**0.977**	0.964	0.966	0.976
mammography	0.850	0.844	0.889	0.865	0.879	0.668	0.746	0.753	0.660	0.680	0.890	0.869	0.691	**0.901**
optdigits	0.805	0.837	0.604	0.584	0.791	0.682	0.470	0.680	0.862	0.890	0.870	**0.973**	0.853	0.968
pendigits	0.964	0.870	0.953	0.928	0.971	0.486	0.472	0.495	0.977	0.921	**0.999**	0.989	0.981	0.996
pima	0.696	0.674	0.685	0.598	0.767	0.577	0.593	0.645	0.655	0.693	0.736	0.720	**0.805**	0.689
satellite	0.773	0.707	0.683	0.696	0.773	0.739	0.733	0.736	0.824	0.858	0.799	0.774	0.850	**0.888**
satimage-2	0.992	0.822	0.989	0.966	0.995	0.992	0.913	0.917	0.963	0.987	**1.000**	0.990	0.984	0.997
shuttle	0.996	0.881	0.995	0.996	0.997	0.963	0.899	0.997	0.998	0.995	0.996	0.998	0.998	**0.999**
thyroid	0.963	0.936	0.980	0.931	**0.989**	0.866	0.791	0.934	0.954	0.949	0.977	0.977	0.968	0.973
wbc	0.958	0.916	0.968	0.913	**0.998**	0.708	0.811	0.936	0.832	0.926	0.960	0.976	0.996	0.989
wine	0.781	0.684	0.622	0.745	0.970	0.752	0.816	0.759	0.912	0.913	0.959	0.950	**0.999**	0.980
Average	0.857	0.802	0.828	0.809	0.893	0.726	0.708	0.765	0.839	0.866	0.887	0.899	0.893	**0.915**

Table 3. **SPOT-FM is fast**: inference is a single forward pass, no per-dataset fitting. Wall-clock on a single V100, averaged over the 17 ODDS datasets.

Method	Fit time	Inference (/ 1K)
IForest	0.42 s	18 ms
OCSVM	2.1 s	35 ms
DeepSVDD	43 s	24 ms
DAGMM	60 s	30 ms
SPOT-FM (ours)	**0 s**	**11 ms**

produce feature-subset exits.

5.3. Structural properties of the test

Power and exchangeability. Appendix F restates three structural properties in hypothesis-testing language. The first is a data-dependent *power bound*: under the uniform-sampler alternative with perturbation cardinality $k = \lceil \mu F \rceil$, the rejection probability is at least $1 - (1 - \delta_{\min})^k$ conditional on the source row lying in the empirical quantile box (Thm. G.1). This bound is zero at $\mu = 0$, equals $\delta_{\min}$ for the first perturbed coordinate, and tends to 1 geometrically as k grows. The second property is *exchangeability*: the test statistic is invariant to the order of the reference support (Thm. G.5), ensuring the score depends only on the empirical distribution of X_{ref}, not on how the user happened to arrange the rows.

Deployment-risk decomposition and inference cost. The third property is a diagnostic *deployment-risk decomposition* (Prop. G.6): the risk on a new deployment equals the pre-training risk plus a task/feature-shift term plus a sampler-label residual ϵ_{Lab}. The residual names the dominant failure mode on ODDS: real anomalies are domain-specific corruptions, not minimal-subset perturbations, so ϵ_{Lab} is generically large. The decomposition is diagnostic (it explains *where* zero-shot transfer fails) rather than a certification of small error. On the practical side, inference costs $\approx$11 ms per 1K rows on a single V100, $\sim$4000$\times$ faster than deep per-dataset baselines (Table 3).

6. Conclusions

We showed that tabular anomaly detection can be realized as a zero-shot composite hypothesis test. A frozen transformer, pre-trained on synthetic tasks generated by minimal-subset perturbation, instantiates both a test statistic and an implicit rejection rule from a user-supplied reference set. The test is fast (one forward pass), admits a data-dependent power bound against the sampler alternative, and empirically improves over per-dataset methods on the majority of ODDS benchmarks.

The main limitation is alternative-hypothesis coverage. When real anomalies remain inside the empirical reference envelope (as on GLASS, PIMA, and ARRHYTHMIA), the sampler alternative is misspecified and power drops. The natural next step is to broaden the alternative family by mixing real tabular corpora into pre-training. Several open questions connect SPOT-FM to the broader testing literature: (i) can the frozen test be made anytime-valid via e-value calibration, enabling optional stopping? (ii) can the alternative family be broadened to approach consistency against richer alternative classes? (iii) can conformal calibration provide per-row p-values with finite-sample coverage guarantees?

References

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *Annals of Statistics*, 51(1):149–178, 2023.

Bergman, L. and Hoshen, Y. Classification-based anomaly detection for general data. In *International Conference on Learning Representations (ICLR)*, 2020.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: Identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data*, 2000.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, 2022.

Friedman, J. H. On multivariate goodness-of-fit and two-sample testing. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, 2003.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. DROCC: Deep robust one-class classification. In *International Conference on Machine Learning (ICML)*, 2020.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. ADBench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Kirchler, M., Khorasani, S., Kloft, M., and Lippert, C. Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., and Chen, G. H. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning (ICML)*, 2020.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*, 2008.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

Müller, S., Hollmann, N., Pineda Arango, S., Grabocka, J., and Hutter, F. Transformers can do Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Qiu, C., Pfrommer, T., Kloft, M., Mandt, S., and Rudolph, M. Neural transformation learning for deep anomaly detection beyond images. In *International Conference on Machine Learning (ICML)*, 2021.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2020.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7): 1443–1471, 2001.

Shen, Y., Wen, H., and Akoglu, L. FoMo-0D: A foundation model for zero-shot tabular outlier detection. *arXiv preprint arXiv:2409.05672*, 2024.

Vovk, V. and Wang, R. E-values: Calibration, combination and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

Yin, J., Qiao, Y., Zhou, Z., Wang, X., and Yang, J. MCM: Masked cell modeling for anomaly detection in tabular data. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhang, X., Maddix, D. C., Yin, J., Erickson, N., Ansari, A. F., Han, B., Zhang, S., Akoglu, L., Faloutsos, C., Mahoney, M. W., Wang, Y., Hu, C., Rangwala, H., Karypis, G., and Wang, B. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2018.

Appendix Overview

This appendix is organized to mirror the paper's flow.

- **A.0** Notation . . . p. 10
- **A.** Sampler: strategies and curriculum . . . p. 10
  - **A.1** Perturbation strategies . . . p. 10
  - **A.2** μ-curriculum schedule . . . p. 11
    - **F.** Structural properties restated in hypothesis-testing language . . . p. 16
- **B.** Related work, extended . . . p. 11
- **C.** Architecture, training, and ablations . . . p. 13
  - **C.1** Inference-time configuration averaging . . . p. 14
  - **C.2** Ablation: auxiliary μ head . . . p. 14
  - **C.3** Prior pool: detailed taxonomy . . . p. 13
- **D.** Statistical comparison and per-dataset discussion . . . p. 15
- **E.** Structural properties of the sampler and predictor . . . p. 17
  - **E.1** Setting, notation, and assumptions . . . p. 17
  - **E.2** Exit behavior of minimal-subset perturbation . . . p. 20
  - **E.3** Row exchangeability and deployment-risk accounting . . . p. 21
  - **E.4** Supporting results: inference, curriculum, auxiliary supervision . . . p. 23
  - **E.5** Limits of the analysis . . . p. 24

A. Notation

Table 4 collects the notation used throughout the paper. Symbols are grouped by theme; the table is cross-referenced from the Method (§3) and from the Setting in the theory appendix (§G.1).

B. Sampler: strategies and curriculum

B.1. Perturbation strategies

Given a batch of normal points $X \in \mathbb{R}^{N\times F}$, the minimal-subset sampler generates a synthetic negative $\tilde{x} \in \mathbb{R}^F$ for each row $x \in X$ as follows. We first sample the perturbation fraction $\mu \sim \text{Beta}(a_t, b_t)$ (see §B.2) and select an index set $\mathcal{I} \subset \{1, \dots, F\}$ uniformly at random with $|\mathcal{I}| = \lceil \mu F \rceil$. For each $k \notin \mathcal{I}$ we set $\tilde{x}_k = x_k$ (the feature is preserved). For each $k \in \mathcal{I}$ we draw $\tilde{x}_k$ from a strategy- specific distribution that depends on the column-wise statistics $(\mu_k, \sigma_k, x_k^{\min}, x_k^{\max}, q_k^5, q_k^{95})$ computed over X and a range-expansion factor $\alpha \sim \mathcal{U}(0.05, 0.5)$.

Let $L_k = x_k^{\min} - \alpha|x_k^{\min}|$ and $U_k = x_k^{\max} + \alpha|x_k^{\max}|$ denote the expanded feature range. The five strategies are:

$$\text{UNIFORM}: \quad \tilde{x}_k \sim \mathcal{U}[L_k, U_k] \tag{4}$$

$$\text{GAUSSIAN}: \quad \tilde{x}_k \sim \text{clip}\big(\mathcal{N}(\mu_k, (\beta\sigma_k)^2),\ L_k, U_k\big),\ \beta \sim \mathcal{U}(1, 3) \tag{5}$$

$$\text{MIXTURE}: \quad \tilde{x}_k \sim \tfrac{1}{2}\mathcal{U}[L_k, U_k] + \tfrac{1}{2}\text{clip}(\mathcal{N}(\mu_k, \sigma_k^2), L_k, U_k) \tag{6}$$

$$\text{BOUNDARY}: \quad \tilde{x}_k \sim \mathcal{U}[L_k, q_k^5] \cup \mathcal{U}[q_k^{95}, U_k] \tag{7}$$

$$\text{EMPIRICAL}: \quad \tilde{x}_k = x_k^{(r)} + \varepsilon, \quad x^{(r)} \sim X,\ \varepsilon \sim \mathcal{N}(0, 0.01) \tag{8}$$

At training time the strategy s is sampled uniformly at random from the five options per task; at inference the sampler sweeps the 216-configuration grid described in Appendix D.2 (9 values of μ $\times$ 3 strategies $\times$ 8 feature permutations), with scores averaged into the final per-row anomaly probability. All strategies are GPU-vectorized: one task of N=1000 rows and F=500 features is generated in ∼0.3 ms.

Table 4. **Notation (part 1 of 2).** Data, tasks, sampler, and operational region.

Symbol	Meaning
Data / feature space	
$\mathcal{X} = \mathbb{R}^F$	Feature space (F-dimensional Euclidean)
F	Number of features
N	Number of rows in a batch
x	A single row (feature vector)
$\tilde{x}$	A synthetic negative produced by the sampler
X	Batch matrix of rows, $X \in \mathbb{R}^{N \times F}$
π	A normal-data prior on $\mathcal{X}$
$\Pi = \{\pi_1, \ldots, \pi_{22}\}$	Pool of 22 normal-data priors
w_i	Mixture weight of prior π_i
$\mathcal{Z}$	Set of 5 perturbation strategies: uniform, gaussian, mixture, boundary, empirical
μ_k, σ_k	Per-feature empirical mean and standard deviation
$x_k^{\min}, x_k^{\max}$	Per-feature empirical min and max
q_k^5, q_k^{95}	Empirical 5th and 95th percentiles of feature k
Task / support / query	
$\tau = (X_S, y_S, X_Q, y_Q)$	In-context task: labeled support plus labeled query
X_S, y_S	Labeled in-context support (features and binary labels)
X_Q, y_Q	Labeled query set
$X_{\text{norm}}, X_{\text{synth}}$	Pre-training task components: normal rows and sampler-generated contrasts
X_{ref}	User-supplied inlier reference at deployment
X_{test}	Deployment test set
$y_{\text{tok}}, \mu_{\text{tok}}$	Label and μ input tokens fed to the predictor
Sampler parameters	
$\mu \in (0, 1)$	Perturbation fraction (fraction of features perturbed per synthetic row)
$k = \lceil \mu F \rceil$	Perturbation cardinality (number of perturbed features)
$\alpha \in [0.05, 0.5]$	Range-expansion factor
$s \in \mathcal{Z}$	Sampler strategy
$\mathcal{I} \subseteq \{1, \ldots, F\}$	Random perturbation index set, $\lvert\mathcal{I}\rvert = k$
L_k, U_k	Expanded feature range: $L_k = x_k^{\min} - \alpha\lvert x_k^{\min}\rvert$, $U_k = x_k^{\max} + \alpha\lvert x_k^{\max}\rvert$
$\mathsf{Sampler}_{s,\alpha,\mu}(x)$	Sampler law mapping row x to a synthetic negative
$\text{Beta}(a_t, b_t)$	Curriculum Beta schedule over μ at step t
Empirical operational region and labeling oracle	
$\widehat{S}(X_{\text{ref}})$	Empirical quantile box: $\prod_j [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$
$\mathcal{P}_j(\widehat{S})$	Per-coordinate projection $[\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$
$L_S(x; X_{\text{ref}})$	Exit-rule labeling oracle: $\mathbf{1}\{x \notin \widehat{S}(X_{\text{ref}})\}$
$\delta_k(\alpha; X_{\text{ref}})$	Per-feature exit probability under uniform
$\delta_{\min}$	Worst-case per-feature exit probability

B.2. μ-curriculum schedule

Let t be the training step and T the total step budget. The Beta parameters (a_t, b_t) anneal from $(a_0, b_0) = (6, 2)$ (with $\mathbb{E}[\mu] \approx 0.75$; easy), to $(a_\infty, b_\infty) = (2, 5)$ (with $\mathbb{E}[\mu] \approx 0.29$; hard), linearly over the first 40% of training steps, and stay there for the remainder:

$$(a_t, b_t) = \begin{cases} (1-u)\,(a_0, b_0) + u\,(a_\infty, b_\infty), & u = t/(0.4T),\ t \le 0.4T \\ (a_\infty, b_\infty), & t > 0.4T. \end{cases}$$

The curriculum is implemented in closed form (a 10-line Python function); no tuning is required beyond the two endpoints. An ablation that fixes $\mu \sim \mathcal{U}(0.05, 0.95)$ throughout training loses -1.3 ROC on average.

C. Related work, extended

Condensed from §2; citations expanded by family.

Table 5. **Notation (part 2 of 2).** Predictor, loss, distributions, assumptions (continues Table 4).

Symbol	Meaning
Predictor and inference	
$\theta \in \Theta$	Predictor parameters and parameter space
$\mathcal{F}_\theta$	Predictor class $\{f_\theta : \theta \in \Theta\}$
$f_\theta(X_S, y_S, X_Q)$	Per-configuration in-context predictor
$\bar{f}_\theta(X_{\text{ref}}, X_Q)$	Deployed 216-configuration averaged predictor
$c \in \mathcal{C}$	Inference configuration: $(\mu, s, \pi_{\text{perm}}, \omega)$
$\mathcal{C}$	Deployed configuration grid ($\lvert\mathcal{C}\rvert = 216$)
$K_c(X_{\text{ref}})$	Labeled support under configuration c
π_{perm}	Feature permutation component of a configuration
Loss and risk	
$\ell(\hat{y}, y)$	Per-row binary cross-entropy (implementation)
$\ell_\epsilon(p, y)$	Clip-before-BCE loss on the ϵ-clipped probability
$\epsilon \in (0, 1/2)$	BCE clipping parameter ($\sim 10^{-6}$)
$B = \log(1/\epsilon)$	Clipped-BCE upper bound
$R_{\mathcal{D}}(\theta)$	Expected loss of θ under distribution $\mathcal{D}$
$\widehat{R}_{\mathcal{T}}(\theta)$	Empirical loss over T pre-training tasks
$\mathcal{L}$	Total training loss: primary BCE + $\lambda_\mu \cdot$ auxiliary NLL
λ_μ	Auxiliary-loss weight ($\lambda_\mu = 0.1$)
Distributions and shift	
$\mathcal{T}$	Pre-training task distribution
$\mathcal{P}$	Deployment task distribution (real features, real labels)
$\mathcal{P}^{\text{exit}}$	Deployment distribution relabeled by exit oracle L_S
$\mathcal{Q}, \mathcal{Q}'$	Generic distributions compared via an IPM
$\mathcal{G}_\theta$	Loss class $\{\ell_\epsilon \circ f_\theta : \theta \in \Theta\}$
$d_{\text{IPM}}(\mathcal{Q}, \mathcal{Q}'; \mathcal{G}_\theta)$	Loss-class integral probability metric
Assumptions and residuals	
(A2),(A3),(A5)	Coordinate indep. (uniform), per-feature exit probability, clipped-BCE boundedness
Arch. cond. G.4	Row-permutation equivariance of the predictor class $\mathcal{F}_\theta$
$\epsilon_{\text{Lab}}(\mathcal{P})$	Deployment-side real-vs-exit-rule label residual
$\mathcal{A}_{\mathcal{F}_\theta}(\mathcal{D})$	In-class approximation error $\inf_\theta R_{\mathcal{D}}(\theta)$
$\sigma \in S_N$	Joint row permutation on the support

Classical and deep tabular AD. Density and boundary estimators (IFOREST (Liu et al., 2008), LOF (Breunig et al., 2000), OCSVM (Schölkopf et al., 2001), ECOD (Li et al., 2022)), one-class classifiers (DEEP SAD (Ruff et al., 2020), DROCC (Goyal et al., 2020), GOAD (Bergman & Hoshen, 2020), NEUTRAL (Qiu et al., 2021)), and auto-encoder-plus-clustering (DAGMM (Zong et al., 2018)) all fit a fresh model per dataset. None is zero-shot.

Tabular foundation models. Prior-data Fitted Networks (Müller et al., 2022) pre-train a transformer on synthetic tasks and classify new datasets via in-context learning; TABPFN (Hollmann et al., 2023) scales this to competitive classification, and TAB2D (Zhang et al., 2025) adds alternating row/column attention for many features. These target classification, not AD. We adopt the Tab2D backbone and extend it with y- and μ-tokens for anomaly detection.

Zero-shot anomaly detection. FOMO-0D (Shen et al., 2024) is the first tabular-AD foundation model, pre-trained on GMM tasks with inflated-covariance outliers. SPOT-FM replaces that mechanism with a *feature-subset perturbation* that produces contrasts which leave the empirical operational region with controlled probability while preserving most original features, rather than perturbing along a single covariance direction.

C.1. Method-feature comparison

Table 1 enumerates the binary design properties that distinguish SPOT-FM from its closest baselines: the two classical/deep methods (IFOREST, DEEPSVDD) that fit a fresh model per dataset, and the foundation model (FOMO-0D) that shares

the zero-shot inference mode but differs on every structural feature downstream of the outlier prior. Table 1 summarizes how SPOT-FM combines zero-shot inference with feature-level tokenization, subset-based synthetic contrasts, auxiliary hardness regularization, and tensorized multi-configuration inference.

D. Architecture, training, and ablations

Parameter breakdown. The model configuration referenced throughout the main text and evaluated in every table and figure uses dim=384 and 8 layers. The table below enumerates the parameter counts for the full 12-layer, dim=512 configuration explored during development (76M parameters, 99% Tab2D backbone); the task-specific heads and μ-embedding together add less than 0.7%. The main-text configuration has correspondingly smaller per-module counts but the same module decomposition. The "76M" figure refers to the development configuration only; no main-text result is computed with the 12-layer/dim=512 variant.

Module	#Params	%
Tab2D layers (12×)	75,657,216	99.39
Anomaly head	263,169	0.35
μ head	131,842	0.17
μ embedding	66,816	0.09
y / feature / norm embeddings	3,584	0.00
Total	76,122,627	100

Each Tab2D layer has four sub-blocks: LayerNorm → row-attention → LayerNorm → FFN → LayerNorm → column-attention → LayerNorm → FFN, each with a residual connection and gradient checkpointing. Heads: dim=512, heads=8, head_dim=64, FFN_dim=2048, max_features=500.

Optimization. We use the Muon optimizer (Jordan et al., 2024) (Newton-Schulz orthogonalized momentum) on 2D attention weight matrices with lr $= 0.02$, momentum=0.95. AdamW with lr $= 8\mathrm{e}{-4}$, $\beta = (0.9, 0.95)$, weight decay 10^{-4} handles embeddings, norms, and heads. The schedule is a 2K-step linear warmup followed by cosine decay to 10% of peak. We train in bfloat16 with a gradient scaler and clip gradients at norm 1.0. Global batch = 32 with gradient accumulation $\times 4$; effective batch of 128 AD tasks per step.

Training schedule. $100{,}000$ steps $\times$ batch 32 $\times$ grad-accum 4 $=$ 12.8M synthetic AD tasks seen during pre-training. Runtime: 14 hours on one NVIDIA V100 with gradient checkpointing and flash attention. Inference on a single 1K-point ODDS test split: ≈ 11 ms on the same hardware.

Main-text model configuration. The configuration used throughout the main text has dim=384, 8 transformer layers, 8 attention heads, and peak learning rate 3×10^{-4}. We select the checkpoint at step 30K by held-out synthetic validation. We briefly explored a handful of alternative (width, depth, lr) configurations during development; the chosen setting was the most robust under the bf16 + Muon/AdamW optimisation recipe above, and is the only configuration we report or compare against in this paper.

D.1. Prior pool: detailed taxonomy

Table 6 lists all 22 priors grouped by family. The pool is designed so that no single distributional shape dominates pre-training; diversity across cluster geometry, marginal shape, and feature-correlation structure ensures the pre-trained model encounters a broad range of tabular data patterns before deployment.

Hand-crafted family. *Gaussian*: samples from $\mathcal{N}(\mu, \Sigma)$ with random mean and covariance; the simplest unimodal prior. *Mixture*: a Gaussian mixture model with 2 to 10 components, random means and shared or per-component covariance; covers multi-modal cluster structure. *Manifold*: data lies on a random low-dimensional linear or nonlinear manifold embedded in $\mathbb{R}^F$; tests the model's ability to detect departures from intrinsic dimensionality. *CorrelatedCluster*: clusters with non-trivial cross-feature correlation (block-diagonal or factor-model covariance); covers the case where features are grouped into correlated blocks. *TreeBoundary*: features generated by a random decision tree with piecewise-constant leaf values; produces sharp axis-aligned boundaries between regions. *Uniform*: $\mathcal{U}[0,1]^F$; a featureless baseline that forces the sampler alone to

Table 6. **The 22-prior pool.** Four families, mixed with weights shown. $N \in [50, 1000]$, $F \in [5, 500]$ per task.

Family	Priors	Weight
Hand-crafted (7)	Gaussian, Mixture, Manifold, CorrelatedCluster, TreeBoundary, Uniform, HeavyTail	20%
Mitra supervised (10)	Tree, RandomForest, ExtraTree, GradientBoosting, kNN, LogReg, NN, LightGBM, XGBoost, CatBoost	35%
Mitra generative (3)	SampleRF, SampleGP, TabPFN-style	25%
TabICL SCM (2)	MLP-SCM, Tree-SCM	20%

carry the AD signal. *HeavyTail*: Pareto or Student-t marginals with varying tail index; models datasets with extreme values (finance, network traffic).

Mitra supervised family. Each prior fits a standard supervised model (the named algorithm) to a randomly-generated regression or classification task, then treats the fitted model's output surface as a data generator: new rows are sampled so that their features reproduce the marginals and correlations implicit in the fitted model. This produces feature distributions that inherit realistic nonlinear dependencies from the underlying algorithm (e.g., tree-based priors produce axis-aligned correlation plateaus, while kNN priors produce local manifold structure). The 10 algorithms span linear, tree-based, nearest-neighbor, and neural function classes to maximize diversity.

Mitra generative and TabICL SCM families. *SampleRF / SampleGP*: non-parametric density models (random forest density estimator and Gaussian-process density estimator) fitted to random base data; produce smooth density approximations that complement the piecewise structure of the supervised priors. *TabPFN-style*: the prior used in TabPFN (Hollmann et al., 2023) for tabular classification; included for compatibility with the PFN pre-training lineage. *MLP-SCM / Tree-SCM*: features are generated by a structural causal model (random DAG with either MLP or tree mechanisms at each node); these priors produce features with genuine causal dependencies rather than mere correlations, which is important for datasets where anomalies are causal-mechanism violations.

Mixing and sampling. At each training step one prior family is sampled with the listed weight, then one specific prior from that family is drawn uniformly. The prior generates a fresh dataset of N rows and F features (both sampled uniformly from their ranges per task). This ensures the model sees $\sim$12.8M distinct AD tasks during pre-training, no two with the same dimensionality or prior.

D.2. Inference-time configuration averaging

SPOT-FM is a single frozen Tab2D checkpoint. Per the standard PFN zero-shot protocol (Hollmann et al., 2023; Shen et al., 2024), at inference we average predictions over a small grid of in-context task configurations: 9 perturbation fractions $\mu \in \{0.1, 0.2, \ldots, 0.9\}$, 3 sampling strategies $\{$UNIFORM, MIXTURE, BOUNDARY$\}$, and 8 random feature permutations, or 216 scores per test row. TabPFN uses an equivalent inference-time averaging over feature permutations and power transforms; in our case the 216-configuration average lifts the single-pass eval by roughly 1 to 2 absolute points of ROC / PR. This averaging is a property of how one queries a PFN; no model-level ensembling is involved.

D.3. Ablation: auxiliary μ head

The auxiliary μ-prediction head adds a $\mathrm{NLL}(\mathrm{Beta}(\hat{\mu}) \,\|\, \mu)$ term to the training loss with weight $\lambda_\mu = 0.1$ (Eq. 3). Ablating $\lambda_\mu = 0$ removes the head and retrains for the same step budget and recipe; the primary BCE loss is unchanged. The auxiliary head gives a $+0.8$ absolute ROC-AUC improvement on ODDS average ($0.805 \rightarrow 0.813$) at negligible compute cost.

Table 7. **SPOT-FM is competitive**: per-dataset AUC-PR on 17 ODDS benchmarks. Best per row in **red**, second-best in blue (ties share the color). SPOT-FM and FoMo-0D (Shen et al., 2024) are *zero-shot* foundation-model baselines (no per-dataset training, no HPO); the other baselines are fit per dataset.

Dataset	IForest	LOF	OCSVM	ECOD	EIF	GOAD	DAGMM	DeepSVDD	NeuTraL	ICL	NPT-AD	MCM	FoMo-0D	SPOT-FM
arrhythmia	0.581	0.500	0.518	0.444	**0.654**	0.516	0.466	0.605	0.584	0.631	0.456	0.612	0.578	0.607
breastw	0.969	0.781	0.984	0.972	**0.992**	0.925	0.799	0.962	0.870	0.968	0.981	0.991	0.990	0.965
campaign	0.426	0.375	0.429	0.480	0.486	0.267	0.279	0.288	0.398	0.456	0.481	**0.557**	0.342	0.462
cardio	0.702	0.642	0.755	0.562	0.812	0.651	0.417	0.659	0.594	0.742	0.777	**0.835**	0.789	0.827
cardiotocography	0.559	0.468	0.576	0.694	0.629	0.604	0.504	0.502	0.633	0.580	0.644	0.645	0.635	**0.718**
glass	0.136	0.181	0.134	0.190	0.292	0.137	0.136	0.229	0.242	0.405	0.222	0.252	**0.836**	0.238
ionosphere	0.918	0.923	0.895	0.873	0.955	0.941	0.738	0.910	0.953	0.980	**0.984**	0.971	0.978	0.982
mammography	0.332	0.321	0.365	**0.584**	0.422	0.229	0.143	0.308	0.100	0.204	0.389	0.370	0.363	0.494
optdigits	0.140	0.331	0.059	0.064	0.133	0.094	0.054	0.130	0.290	0.326	0.273	**0.686**	0.319	0.445
pendigits	0.516	0.594	0.444	0.412	0.564	0.146	0.073	0.167	0.593	0.514	**0.953**	0.753	0.663	0.903
pima	0.660	0.623	0.642	0.619	0.763	0.602	0.589	0.656	0.647	0.706	**0.819**	0.811	0.793	0.694
satellite	0.799	0.722	0.753	0.758	0.821	0.793	0.717	0.828	0.861	0.883	0.856	0.836	0.880	**0.896**
satimage-2	0.919	0.812	0.943	0.785	0.961	0.950	0.251	0.859	0.714	0.881	**0.988**	0.833	0.928	0.951
shuttle	0.956	0.762	0.942	0.968	0.988	0.920	0.564	0.984	0.992	0.986	0.940	0.965	0.994	**0.997**
thyroid	0.680	0.587	0.688	0.673	**0.823**	0.668	0.306	0.644	0.657	0.658	0.801	0.817	0.670	0.691
wbc	0.830	0.633	0.798	0.747	**0.982**	0.546	0.362	0.755	0.412	0.781	0.779	0.855	0.965	0.944
wine	0.431	0.381	0.334	0.354	0.843	0.453	0.511	0.510	0.709	0.644	0.769	0.975	**0.996**	0.965
Average	0.621	0.567	0.603	0.599	0.713	0.555	0.406	0.588	0.603	0.667	0.713	0.751	0.748	**0.752**

E. Statistical comparison and per-dataset discussion

The main paper reports per-dataset AUC-ROC in Table 2; per-dataset AUC-PR is deferred to Table 7 below. Because AUC-PR is more sensitive than AUC-ROC on the very imbalanced ODDS splits (anomaly rate 0.2% to 36%), we summarize the two pictures side-by-side here.

Averages on the 17-dataset panel. Under the 216-configuration inference protocol (§D.2), SPOT-FM takes the first-place average across all baselines on both metrics:

Method	AUC-ROC avg.	AUC-PR avg.
DAGMM	0.708	0.406
GOAD	0.726	0.555
ECOD	0.809	0.599
IForest	0.857	0.621
FoMo-0D	0.893	0.748
MCM	0.899	0.751
SPOT-FM	**0.915**	**0.752**

Wilcoxon signed-rank tests. We follow the protocol of Shen et al. (2024) and test whether SPOT-FM is significantly different from each baseline on a paired per-dataset basis. With 17 paired observations the two-sided Wilcoxon signed-rank test gives:

SPOT-FM vs.	n	w/l	p (ROC)	p (PR)
DAGMM	17	17/0	$<10^{-4}$	$<10^{-4}$
GOAD	17	17/0	$<10^{-4}$	0.0003
ECOD	17	17/0	$<10^{-4}$	0.0011
IForest	17	13/4	0.0032	0.0005
MCM	17	11/6	0.284	0.611
FoMo-0D	17	12/5	0.244	0.517

SPOT-FM is significantly better than every supervised deep baseline and both shallow methods on both metrics ($p \leq 0.003$). Against FoMo-0D and MCM the two-sided Wilcoxon test cannot reject equality, consistent with the close averages above: the three zero-shot/non-fit foundation methods disagree on *which* datasets they are good at rather than on a population-level effect. The $+2.2$ AUC-ROC absolute gap over FoMo-0D is driven by a $12-$of-17 per-dataset win pattern with large absolute gains on ~ 4 datasets where the GMM outlier prior is a poor fit (detailed below).

Where SPOT-FM still lags per-dataset. SPOT-FM now edges FoMo-0D on average AUC-PR (0.752 vs 0.748), but still loses on a handful of small, heavy-tail datasets where the GMM prior fits closely: GLASS (0.238 vs FoMo-0D 0.836), PIMA (0.694 vs 0.793), and ARRHYTHMIA (0.607 vs 0.578; FoMo-0D is closer but loses this one). The remaining gaps on WBC (0.944 vs 0.965) and WINE (0.965 vs 0.996) have shrunk to within two absolute points. We view these small-sample losses as the primary headroom for the next iteration and the strongest argument for adding real tabular data to the pre-training mixture.

Where SPOT-FM improves. On AUC-ROC SPOT-FM gains over FoMo-0D on $12/17$ datasets; the largest gains are MAMMOGRAPHY ($+0.21$), CAMPAIGN ($+0.12$), OPTDIGITS ($+0.12$), CARDIOTOCOGRAPHY ($+0.07$), and SATELLITE ($+0.04$). On AUC-PR it also improves over FoMo-0D on $12/17$ datasets with top gains PENDIGITS ($+0.24$), MAMMOGRAPHY ($+0.13$), OPTDIGITS ($+0.13$), and CAMPAIGN ($+0.12$): precisely the datasets whose anomaly structure is not well approximated by a Gaussian-mixture inflation.

F. Structural properties restated in hypothesis-testing language

Three structural properties of the deployed SPOT-FM test are restated here in hypothesis-testing language. Proofs and full assumptions are deferred to the extended version.

F.1. Setting

Let $\mathcal{X} = \mathbb{R}^F$ be the feature space. A deployment consists of a reference set $X_{\text{ref}} \in \mathcal{X}^n$ (user-supplied inliers) and a query point $x \in \mathcal{X}$. Write $\widehat{S}(X_{\text{ref}}) = \prod_j [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ for the empirical 5 to 95% quantile box, and $L_S(x; X_{\text{ref}}) = \mathbf{1}\{x \notin \widehat{S}\}$ for the exit-rule rejection oracle. The deployed predictor $\bar{f}_\theta(X_{\text{ref}}, \cdot)$ is the 216-configuration-averaged in-context map (Appendix D.2) and is treated as the learned test statistic.

F.2. Power bound against the uniform-sampler alternative

Theorem F.1 (Power bound, restated)**.** *Under the* UNIFORM *sampler with expansion factor α and perturbation cardinality $k = \lceil \mu F \rceil$, assume coordinate-independent resampling on the k perturbed coordinates (A2) and the data-dependent per-feature exit probability (A3) of the extended version. Conditional on the source row lying in $\widehat{S}(X_{\text{ref}})$, the probability that the synthetic contrast $\tilde{x}$ rejects $\widehat{S}$ satisfies*

$$\Pr\big(\tilde{x} \notin \widehat{S}(X_{\text{ref}}) \;\big|\; x \in \widehat{S}\big) \;\geq\; 1 - \big(1 - \delta_{\min}(\alpha; X_{\text{ref}})\big)^k,$$

where $\delta_{\min}(\alpha; X_{\text{ref}}) = \min_j \delta_j(\alpha; X_{\text{ref}})$ ranges over non-degenerate coordinates. The bound is data-dependent through the empirical quantiles and the min/max-expanded range; it is zero at $\mu = 0$, equals $\delta_{\min}$ for $\mu \in (0, 1/F]$, and tends to 1 as $\mu F \to \infty$.

The BOUNDARY sampler rejects $\widehat{S}$ with probability one by construction for every perturbed coordinate and is recorded separately as an observation.

F.3. Exchangeability of the test statistic

Theorem F.2 (Exchangeability, restated)**.** *Assume the predictor class $\mathcal{F}_\theta$ satisfies the row-permutation equivariance architectural condition of the extended version (token-wise normalization, no positional signal on row tokens, symmetric padding mask). For every joint permutation $\sigma \in S_n$ of the reference support,*

$$\bar{f}_\theta\big(\sigma \cdot X_{\text{ref}},\, x\big) \;=\; \bar{f}_\theta\big(X_{\text{ref}},\, x\big),$$

i.e. the deployed test statistic is a symmetric function of X_{ref}.

F.4. Deployment-risk decomposition with sampler-label residual

Proposition F.3 (Risk decomposition, restated)**.** *Under the clipped-BCE boundedness assumption (A5) of the extended version, the deployment risk admits the pointwise identity*

$$R_{\mathcal{P}}(\theta) = R_{\mathcal{T}}(\theta) + \big[R_{\mathcal{P}^{\text{exit}}}(\theta) - R_{\mathcal{T}}(\theta)\big] + \epsilon_{\text{Lab}}(\mathcal{P}),$$

and the IPM-form pointwise inequality

$$R_{\mathcal{P}}(\theta) \;\le\; R_{\mathcal{T}}(\theta) + d_{\mathrm{IPM}}(\mathcal{P}^{\mathrm{exit}}, \mathcal{T}; \mathcal{G}_\theta) + \epsilon_{\mathrm{Lab}}(\mathcal{P}),$$

where $\mathcal{P}^{\mathrm{exit}}$ is $\mathcal{P}$ relabeled by L_S, $\mathcal{T}$ is the pre-training task distribution, d_{IPM} is the loss-class integral probability metric, and ϵ_{Lab} is the deployment-side real-vs-exit-rule label residual. ϵ_{Lab} is generically $\Theta(B)$ on real deployments, so the inequality is diagnostic, not a certification; it names where zero-shot transfer can fail.

Scope. Thm. G.1 is asserted for the UNIFORM sampler only. Thm. G.5 establishes an architectural consistency property of the test statistic, not zero-shot generalization. Prop. G.6 is a diagnostic decomposition under the empirical-quantile-box exit rule, not a certification on arbitrary real deployments.

G. Structural properties of the sampler and predictor

Scope. The results below are structural intuition and risk accounting for SPOT-FM, not a complete guarantee of zero-shot anomaly detection on arbitrary real datasets. We analyze an *empirical operational normal region* induced by the reference data (the 5 to 95% quantile box, not the essential distributional support), show that the UNIFORM feature-subset perturbation produces contrasts that exit this region with data-dependent geometric probability (conditional on the source row lying inside), note that the BOUNDARY strategy exits by construction, prove row-exchangeability of the in-context predictor from a token-wise-normalization + no-positional-signal property of the architecture, and decompose deployment risk into a task-prior/feature-shift term and a sampler-label residual. A Jensen-safety remark accompanies the 216-configuration inference grid, but we do not claim it explains the empirical AUC/PR lift. Empirical design choices (the grid, the μ-curriculum, and the auxiliary μ-prediction head) are justified primarily by ablation (Appendix D.3, Appendix B.2). The theory does *not* prove that SPOT-FM recovers the real anomaly manifold, that the 216-grid strictly improves AUC/PR, that the curriculum is optimal, or that transfer error is small on an arbitrary real deployment; Section G.5 spells out this scope.

Organization. The appendix contains two theorems, one proposition, one architectural condition (on the model class), and one boundary-strategy observation, plus three informal design-choice paragraphs, a Jensen-safety remark for inference-time averaging, and a limits section. The deployed inference map is explicitly defined in Section G.1 as $\bar{f}_\theta$, the 216-configuration average of the per-configuration in-context predictor $f_\theta(K_c(X_{\mathrm{ref}}), X_Q)$; theorem-level results operate on the per-configuration object and lift to $\bar{f}_\theta$ by Jensen on the clipped loss. Theorem G.1 (exit probability under variable μ, UNIFORM strategy, conditional on the source row lying in $\widehat{S}$) gives a data-dependent geometric lower bound on the probability that a synthetic contrast leaves the empirical operational region. The BOUNDARY strategy is recorded separately as a construction-level observation, since under the empirical quantile box it exits with probability one. Architectural condition G.4 and Theorem G.5 together establish row-exchangeability of the predictor from token-wise normalization and a no-positional-signal property of the architecture class; this is a necessary architectural consistency property for querying a frozen in-context predictor with unordered reference sets, but not by itself sufficient for zero-shot generalization. Proposition G.6 (deployment-risk decomposition with sampler-label residual) writes the deployment risk as a pointwise identity in three named sources (pre-training risk, task-prior/feature shift, and a sampler-label residual ϵ_{Lab}), with an optional IPM-based pointwise inequality; the residual is generically $\Theta(B)$ on real deployments, so the inequality is diagnostic, not a certification. Averaging over in-context configurations is treated as a Jensen-safety remark, not a theorem. The informal paragraphs of Section G.4 tie the deployed design choices (proximity/reliability intuition, curriculum, auxiliary head) to the ablation evidence without claiming them as theorems.

G.1. Setting, notation, and assumptions

Before stating results, we fix the objects over which they are quantified: the feature space and normal-data priors, the sampler that produces synthetic contrasts, a common in-context task space on which both pre-training and deployment distributions live, an empirical operational normal region induced by reference data, a sampler-induced labeling oracle that extends the sampler's notion of "anomaly" to arbitrary deployment points, the predictor class, and the loss. Symbols introduced below are indexed in Table 4 (Appendix A) for quick reference.

Let $\mathcal{X} = \mathbb{R}^F$ be the feature space and let π denote a normal-data prior on $\mathcal{X}$. Let $\Pi = \{\pi_1, \ldots, \pi_{22}\}$ be a prior pool weighted by $w_1, \ldots, w_{22} \ge 0$ with $\sum_i w_i = 1$. Let $\mathcal{Z} = \{\text{UNIFORM}, \text{GAUSSIAN}, \text{MIXTURE}, \text{BOUNDARY}, \text{EMPIRICAL}\}$ be the set of perturbation strategies, $\alpha \in [0.05, 0.5]$ the range-expansion factor, and $\mu \in (0, 1)$ the perturbation fraction; we

write $k = \lceil \mu F \rceil$ for the number of features that the sampler perturbs on a given row. For a row x, the sampler draws an index set $\mathcal{I} \subseteq \{1, \ldots, F\}$ uniformly at random with $|\mathcal{I}| = k$, preserves $\tilde{x}_j = x_j$ for $j \notin \mathcal{I}$, and resamples $\tilde{x}_j$ for $j \in \mathcal{I}$ from a strategy-specific distribution. We denote the resulting synthetic-negative law by $\mathsf{Sampler}_{s,\alpha,\mu}(x)$.

In-context task space. Every in-context task is represented as a 4-tuple

$$\tau \;=\; (X_S,\, y_S,\, X_Q,\, y_Q),$$

where (X_S, y_S) is the *labeled in-context support* fed to the predictor and (X_Q, y_Q) is the *labeled query set* used to define risk. Under pre-training, one task draw yields $(X_S, y_S) = (X_{\text{norm}}, 0) \cup (X_{\text{synth}}, 1)$, where $X_{\text{norm}} \sim \pi$ and X_{synth} is the sampler's output; the query (X_Q, y_Q) is a separate draw from the same mixture. Under deployment, for each inference configuration c the support is $(X_S, y_S) := K_c(X_{\text{ref}})$ (defined below), where X_{ref} is the user-supplied inlier reference; the query (X_Q, y_Q) is the deployment's test points with real labels. The per-configuration in-context map is

$$f_\theta :\; (X_S, y_S, X_Q) \;\mapsto\; [0,1]^{|X_Q|},$$

with $\mathcal{F}_\theta = \{f_\theta : \theta \in \Theta\}$. Let $\mathcal{G}_\theta = \{\ell_\epsilon \circ f_\theta : \theta \in \Theta\}$ denote the loss class and

$$d_{\text{IPM}}(\mathcal{Q}, \mathcal{Q}'; \mathcal{G}_\theta) \;=\; \sup_{g \in \mathcal{G}_\theta} \left| \mathbb{E}_{\mathcal{Q}} g - \mathbb{E}_{\mathcal{Q}'} g \right|$$

the integral probability metric generated by $\mathcal{G}_\theta$. This common 4-tuple form puts pre-training and deployment tasks on the same measurable space.

Deployed predictor (216-configuration averaged map). At inference, SPOT-FM does not evaluate f_θ on the user-supplied X_{ref} directly; it constructs a labeled in-context support by pairing X_{ref} (label 0) with sampler-generated contrasts (label 1) and averages the per-query probabilities over a deterministic configuration grid. Formally, let $\mathcal{C}$ index $(\mu, s, \pi_{\text{perm}}, \omega)$ over the deployed 216-configuration grid: a μ-value, a strategy, a feature permutation, and sampler randomness (Appendix D.2). For each $c \in \mathcal{C}$, define the labeled support

$$K_c(X_{\text{ref}}) \;:=\; \big(X_{\text{ref}},\, 0\big) \cup \big(X^{(c)}_{\text{synth}},\, 1\big) \;=\; (X^{(c)}_S,\, y^{(c)}_S),$$

where $X^{(c)}_{\text{synth}}$ is the sampler's output under configuration c applied to X_{ref}. The deployed predictor is

$$\bar{f}_\theta(X_{\text{ref}}, X_Q) \;:=\; \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} f_\theta\big(K_c(X_{\text{ref}}),\, X_Q\big) \;\in\; [0,1]^{|X_Q|}.$$

For any distribution $\mathcal{D}$ over tasks, the risk is

$$R_{\mathcal{D}}(\theta) \;:=\; \mathbb{E}_{\tau \sim \mathcal{D}} \left[\tfrac{1}{|X_Q|} \sum_{i=1}^{|X_Q|} \ell_\epsilon\big(\hat{y}_i,\, y_{Q,i}\big) \right],$$

where $\hat{y}_i = \bar{f}_\theta(X_{\text{ref}}, X_Q)_i$ under deployment and $\hat{y}_i = f_\theta(X_S, y_S, X_Q)_i$ under pre-training (pre-training presents one labeled support per task, with no inference-time averaging). When a per-configuration bound is established for $f_\theta(K_c(\cdot), \cdot)$ uniformly in c, it lifts to a bound on $\bar{f}_\theta$ by convexity of ℓ_ϵ on $[\epsilon, 1-\epsilon]$ (Jensen's inequality; cf. the Jensen-safety remark of Section G.4.1). All deployment-time risk statements below refer to $\bar{f}_\theta$; the theorem-level objects (Theorem G.1, Proposition G.6) operate on per-configuration $f_\theta(K_c(\cdot), \cdot)$ and are lifted where needed.

Empirical operational region. Given a reference set X_{ref} of n rows, define the empirical quantile box

$$\widehat{S}(X_{\text{ref}}) \;:=\; \prod_{j=1}^{F} \big[\hat{q}_j^{0.05}(X_{\text{ref}}),\, \hat{q}_j^{0.95}(X_{\text{ref}})\big],$$

where $\hat{q}_j^q$ is the empirical q-quantile of coordinate j. We write $\mathcal{P}_j(\widehat{S}) := [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ for the per-coordinate projection. All "exit" statements below are relative to $\widehat{S}(X_{\text{ref}})$. Other operational regions (e.g. a Gaussian coordinate box $[-c, c]^F$ or a density sublevel set) are possible but require their own per-coordinate exit constants and are not used by the theorem.

Sampler-induced labeling oracle. The sampler $\mathsf{Sampler}_{s,\alpha,\mu}$ is a map $\mathcal{X} \to \mathcal{X}$ on normal rows; it does not define labels on arbitrary deployment points. To lift deployment distributions to the task space on which $\mathcal{T}$ is defined, fix a reference set X_{ref} and define the sampler-induced labeling oracle

$$L_S(x; X_{\mathrm{ref}}) \;:=\; \mathbf{1}\big\{x \notin \widehat{S}(X_{\mathrm{ref}})\big\},$$

which is a measurable function of the observables. The synthetic-rule deployment distribution is then

$$\mathcal{P}^{\mathrm{exit}} \;:=\; \mathcal{L}\Big(K_c(X_{\mathrm{ref}}),\, X_Q,\, L_S(X_Q; X_{\mathrm{ref}})\Big), \qquad (X_{\mathrm{ref}}, X_Q) \sim \mathcal{P},\; c \sim \mathrm{Unif}(\mathcal{C}),$$

i.e. the support $(X_S, y_S) = K_c(X_{\mathrm{ref}})$ is constructed from the deployment's X_{ref} via configuration c, and the query labels are assigned by the exit oracle. $\mathcal{P}^{\mathrm{exit}}$ is well-defined on all of $\mathrm{supp}(\mathcal{P})$ and differs from $\mathcal{P}$ only in how labels are assigned. We use the name $\mathcal{P}^{\mathrm{exit}}$ (rather than $\mathcal{P}^{\mathrm{synth}}$) to emphasize that this is the deployment distribution *relabeled by the exit-rule oracle* L_S, distinct from the generation-labeled synthetic pre-training distribution. A training row whose perturbed coordinates happen to remain inside $\widehat{S}$ is labeled $y = 1$ by the sampler but $L_S(\tilde{x}; X_{\mathrm{ref}}) = 0$, so the exit-rule surrogate and the generation-label distribution agree only almost surely on synthetic rows that do exit $\widehat{S}$. Any such training-side generation-vs-exit-label mismatch is absorbed into the task-prior / structural-shift term $d_{\mathrm{IPM}}(\mathcal{P}^{\mathrm{exit}}, \mathcal{T}; \mathcal{G}_\theta)$ of Proposition G.6 (which compares the exit-rule surrogate against $\mathcal{T}$); the residual ϵ_{Lab} denotes only the deployment-side real-vs-exit-label mismatch.

Finally let $\mathcal{A}_{\mathcal{F}_\theta}(\mathcal{D}) = \inf_{\theta\in\Theta} R_{\mathcal{D}}(\theta)$ denote the in-class approximation error of $\mathcal{F}_\theta$ under $\mathcal{D}$, and write $\lambda_\mu \geq 0$ for the auxiliary-loss weight used in the multi-task objective.

On the choice of IPM. Throughout this appendix we use the loss-class IPM $d_{\mathrm{IPM}}(\cdot,\cdot;\mathcal{G}_\theta)$: among divergences that uniformly bound the risk gap over the loss class $\mathcal{G}_\theta$, it is the tightest. A per-θ divergence $d_\theta(\mathcal{Q}, \mathcal{Q}') := |\mathbb{E}_{\mathcal{Q}} g_\theta - \mathbb{E}_{\mathcal{Q}'} g_\theta|$ is smaller but does not bound $\sup_\theta |R_{\mathcal{Q}} - R_{\mathcal{Q}'}|$. Stronger divergences (total variation, Kullback-Leibler, Wasserstein on a fixed metric) would require realizability or mutual absolute continuity between $\mathcal{T}$ and $\mathcal{P}$, neither of which is available for arbitrary deployment distributions. We commit to the loss-class IPM because it gives a non-vacuous uniform bound under the minimum structural assumption compatible with the zero-shot claim.

Assumptions. We refer to the following assumptions by label at the point of use.

(A2) *Coordinate independence (*UNIFORM *only).* Conditional on a row x and the index set $\mathcal{I}$, under $s =$ UNIFORM the perturbed coordinates $\{\tilde{x}_k : k \in \mathcal{I}\}$ are mutually independent, holding by construction of the UNIFORM sampler. This assumption is *not* asserted for GAUSSIAN, MIXTURE, BOUNDARY, or EMPIRICAL (the EMPIRICAL strategy resamples a single row $x^{(r)}$ and copies its coordinates into all $k \in \mathcal{I}$, inducing non-trivial column dependence); no result in this appendix applies (A2) outside $\{$UNIFORM$\}$.

(A3) *Per-feature exit probability (*UNIFORM *only, data-dependent).* Let $\widehat{S} = \widehat{S}(X_{\mathrm{ref}}) = \prod_j [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ and let the deployed UNIFORM sampler resample each $\tilde{x}_k \sim \mathcal{U}[L_k, U_k]$ with $L_k = x_k^{\min} - \alpha|x_k^{\min}|$, $U_k = x_k^{\max} + \alpha|x_k^{\max}|$ (as defined in Appendix B.1). The per-feature exit probability is then given explicitly by

$$\delta_k(\alpha; X_{\mathrm{ref}}) \;=\; \left[1 \;-\; \frac{\hat{q}_k^{0.95} - \hat{q}_k^{0.05}}{U_k - L_k}\right]_+ \;\in\; [0,1],$$

and we write $\delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}}) := \min_k \delta_k(\alpha; X_{\mathrm{ref}})$ for the worst-case coordinate, where the minimum ranges over *non-degenerate* coordinates (those with $U_k > L_k$); constant columns with $U_k = L_k$ are excluded from the minimum and do not contribute to the exit event. The positive-part bracket ensures $\delta_k \geq 0$ even when numerical or symbolic roundoff would otherwise give a negative value. The expression depends on the reference data through both the quantile interval and the min/max-expanded range. Monotonicity: δ_k is non-decreasing in α for every k, so $\delta_{\min}$ is non-decreasing in α. *Symmetric-box sanity check.* If $x_k^{\min} = -x_k^{\max} = -c_k$ and $\hat{q}_k^{0.05} = -c_k$, $\hat{q}_k^{0.95} = c_k$, the expression reduces to $\delta_k = \alpha/(1+\alpha)$, the symmetric-expanded-interval constant. We do not use the symmetric form in the theorem; the general form is the data-dependent $\delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})$. (A3) is *not* asserted for GAUSSIAN, MIXTURE, or EMPIRICAL; the BOUNDARY strategy is handled separately as an observation (not a theorem) because it exits the empirical quantile box by construction for every perturbed coordinate.

(A5) *Bounded loss (clip-before-BCE).* Let $\tilde{p} := \mathrm{clip}(p,\, \epsilon,\, 1-\epsilon)$ for $\epsilon \in (0, 1/2)$. The per-row loss is the BCE applied to the already-clipped probability,

$$\ell_\epsilon(p, y) \;=\; -y \log \tilde{p} \;-\; (1-y)\log(1-\tilde{p}), \qquad \tilde{p} = \mathrm{clip}(p, \epsilon, 1-\epsilon),$$

so $\ell_\epsilon \in [0, B]$ with $B = \log(1/\epsilon)$ and $p \mapsto \ell_\epsilon(p, y)$ is convex on $[\epsilon, 1-\epsilon]$. The deployed implementation uses numerically stable logit-level BCE; the theory applies to the clipped analogue of the implemented loss and, at $\epsilon \sim 10^{-6}$, does not materially change the operating point. All bounds below quote the constant B whenever a boundedness argument is invoked.

(A2), (A3), and (A5) are all that the surviving results use, together with Architectural condition G.4 for the row-exchangeability theorem.

G.2. Exit behavior of minimal-subset perturbation

We formalize one property of the sampler: under the UNIFORM *strategy, the probability that a synthetic contrast with perturbation cardinality $k = \lceil \mu F \rceil$ leaves the empirical operational region $\widehat{S}(X_{\mathrm{ref}})$ grows geometrically in k (Theorem G.1, conditional on the source row lying in $\widehat{S}$). The* BOUNDARY *strategy exits by construction for every perturbed coordinate and is recorded as a separate observation, not a theorem. The surrounding design intuition (that a useful training signal concentrates at intermediate μ, rather than at $\mu \to 0$, indistinguishable contrasts, or $\mu \to 1$, trivially separable contrasts), is empirical (Fig. 1(b)) and is discussed informally in Section G.4(i) rather than proved here.*

G.2.1. Exit probability under a variable perturbation fraction

The first question is probabilistic: under the UNIFORM *sampler, what is the probability that a synthetic row leaves the empirical operational normal region $\widehat{S}(X_{\mathrm{ref}})$, conditional on the source row x lying inside $\widehat{S}$? The following theorem gives a geometric lower bound in the perturbation cardinality $k = \lceil \mu F \rceil$, data-dependent through $\delta_{\min}$. The bound is monotone in k but discrete, not continuous, in μ: at $\mu = 0$ the exit probability is zero; for $\mu \in (0, 1/F]$ the cardinality is $k = 1$ and the bound equals $\delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})$; it approaches 1 as $\mu F \to \infty$. The* BOUNDARY *strategy exits by construction for every perturbed coordinate and is handled as a separate observation below, not as a theorem.*

Theorem G.1 (Exit probability under variable μ, UNIFORM strategy)**.** *Fix X_{ref} and write $\widehat{S} := \widehat{S}(X_{\mathrm{ref}})$. Under (A2) and (A3), for $s =$ UNIFORM, any $x \in \widehat{S}$, and any $\mu \in (0, 1]$ with $k = \lceil \mu F \rceil$,*

$$\Pr_{\tilde{x} \sim \mathsf{Sampler}_{\text{UNIFORM},\alpha,\mu}(x)} \big(\tilde{x} \notin \widehat{S} \;\big|\; x \in \widehat{S}\big) \;\geq\; 1 - \big(1 - \delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})\big)^k.$$

The bound is monotone in k but discrete in μ: at $\mu = 0$ the exit probability is 0, and for $\mu \in (0, 1/F]$ it equals $\delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})$.

Proof. Fix $x \in \widehat{S}$, α, μ, and write $k = \lceil \mu F \rceil$. Let $\mathcal{I} \subseteq \{1, \ldots, F\}$ denote the random perturbation index set of size k. For $j \notin \mathcal{I}$, the sampler sets $\tilde{x}_j = x_j \in \mathcal{P}_j(\widehat{S})$. Since $\widehat{S} = \prod_j \mathcal{P}_j(\widehat{S})$ is an axis-aligned product, the event $\{\tilde{x} \in \widehat{S}\}$ is *equivalent* to $\{\forall j \in \mathcal{I} : \tilde{x}_j \in \mathcal{P}_j(\widehat{S})\}$ conditional on $(x, \mathcal{I})$. Assumption (A2) gives conditional independence across the coordinates of $\mathcal{I}$ under UNIFORM, so

$$\Pr(\tilde{x} \in \widehat{S} \mid x, \mathcal{I}) \;=\; \prod_{j \in \mathcal{I}} \Pr\big(\tilde{x}_j \in \mathcal{P}_j(\widehat{S}) \mid x\big) \;\leq\; \big(1 - \delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})\big)^k,$$

where the inequality uses (A3). Taking complements and averaging over $\mathcal{I}$ (the right-hand side does not depend on which subset of cardinality k is chosen) gives the claim. The piecewise statement in μ follows because $k = 0$ iff $\mu = 0$ and $k = 1$ for $\mu \in (0, 1/F]$. □

Remark G.2 (Tightness for UNIFORM on the symmetric box). In the symmetric special case $x_j^{\min} = -x_j^{\max} = -c_j$ and $\hat{q}_j^{0.05} = -c_j$, $\hat{q}_j^{0.95} = c_j$ for every coordinate (so the quantile box and the range-expansion are both symmetric), the per-coordinate bound is attained with equality: each perturbed coordinate re-enters $[-c_j, c_j]$ with probability exactly $1/(1+\alpha)$, independently across the k perturbed coordinates, and the bound becomes $1 - (1/(1+\alpha))^k$. The general data-dependent $\delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}})$ of (A3) replaces this symmetric closed form and can differ from $\alpha/(1+\alpha)$ in either direction depending on the reference data's skew and tail.

Per-coordinate view. The sharper primitive behind Theorem G.1 is coordinate-wise: under the UNIFORM strategy, for each perturbed coordinate j,

$$\Pr\big(\tilde{x}_j \notin \mathcal{P}_j(\widehat{S}) \,\big|\, X_{\mathrm{ref}}\big) \;=\; \delta_j(\alpha; X_{\mathrm{ref}}),$$

exactly (not just a lower bound), because UNIFORM puts mass uniformly on $[L_j, U_j]$ and the quantile interval $[\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ is a sub-interval. The joint theorem is the product-set consequence obtained when the k selected perturbed coordinates are considered together, invoking (A2)'s coordinate independence. Because $\widehat{S}$ is a joint 5 to 95% quantile box, the event $\{x \in \widehat{S}\}$ can be rare in high dimension (the per-coordinate mass is ≈ 0.9, so $\Pr(x \in \widehat{S}) \approx 0.9^F$ under weak coordinate independence, essentially zero for $F = 500$). Theorem G.1 should therefore be read as a conditional exit statement for rows that happen to lie in $\widehat{S}$, not as a marginal statement over all training rows. Rows outside $\widehat{S}$ already satisfy $L_S(x; X_{\mathrm{ref}}) = 1$ and contribute trivially to the deployment-side exit event.

Remark G.3 (Worst-case α under a random range expansion). If $\alpha \sim \mathcal{U}(\alpha_{\min}, \alpha_{\max})$ with $0 < \alpha_{\min} < \alpha_{\max}$, independently of the other sampler randomness, then by monotonicity of $\delta_{\min}$ in α,

$$\Pr\big(\tilde{x} \notin \widehat{S} \mid x \in \widehat{S}\big) \;\geq\; \mathbb{E}_\alpha\big[1 - (1 - \delta_{\min}(\text{UNIFORM}, \alpha; X_{\mathrm{ref}}))^k\big] \;\geq\; 1 - (1 - \delta_{\min}(\text{UNIFORM}, \alpha_{\min}; X_{\mathrm{ref}}))^k.$$

With the deployed $\alpha \sim \mathcal{U}(0.05, 0.5)$, the worst case $\alpha_{\min} = 0.05$ gives the symmetric-box reference value ≈ 0.0476 and $(0.952)^k$ decay.

Boundary observation (not a theorem). Under the empirical quantile-box operational region $\widehat{S}$, the BOUNDARY sampler draws each perturbed coordinate from $[L_j, \hat{q}_j^{0.05}] \cup [\hat{q}_j^{0.95}, U_j]$, i.e. from the complement of $\mathcal{P}_j(\widehat{S}) = [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ in the expanded range by construction. For any $k \geq 1$ (equivalently, $\mu > 0$), $\tilde{x}_{k_0} \notin \mathcal{P}_{k_0}(\widehat{S})$ for every perturbed index $k_0 \in \mathcal{I}$ almost surely, so $\Pr(\tilde{x} \notin \widehat{S} \mid x \in \widehat{S}) = 1$ up to measure-zero endpoint events. This case is trivial relative to $\widehat{S}$ and is not the content of Theorem G.1. The BOUNDARY strategy is included in the deployed 216-configuration inference grid as an *empirical configuration*; Appendix E records its construction-level exit property for completeness and does not claim it contributes independent theoretical information beyond the ablation-level evidence supporting its inclusion in the grid.

G.3. Row exchangeability and deployment-risk accounting

Querying a frozen in-context predictor with unordered reference sets requires a consistency property: the output should depend on the reference set only through its empirical distribution, not through the order in which rows arrive. This is a necessary architectural property for using the predictor on set-valued inputs; it is not *by itself sufficient for zero-shot generalization, which further requires the transfer-decomposition inequality of Section G.3.2. We isolate two structural ingredients: (i) row-exchangeability (Architectural condition G.4, Theorem G.5); and (ii) a deployment-risk decomposition naming the sources of error as pre-training risk, task-prior shift, and a sampler-label residual (Proposition G.6). A sharp Rademacher bound on the pre-training estimation gap would sit between these two ingredients; we do not state it, because the variable-length dual-axis-attention complexity class is open in the literature and nothing in the surviving analysis would be made quantitative by an abstract complexity placeholder.*

G.3.1. ROW EXCHANGEABILITY BY ARCHITECTURAL CONSTRUCTION

A necessary condition for the zero-shot claim is that the predictor's output on a query point does not depend on the order in which reference rows are presented. This is a structural property: no positional signal enters the predictor, and the training distribution is symmetric under row permutation. We first state the property as an architectural condition on the model class $\mathcal{F}_\theta$, then derive row-permutation invariance of the output from it.

Assumption G.4 (Architectural condition: row-permutation equivariance)**.** The predictor class $\mathcal{F}_\theta$ contains only architectures satisfying the following structural conditions on every layer of every $f_\theta \in \mathcal{F}_\theta$: (i) all normalization layers are token-wise (LayerNorm across the feature dimension of a single token; no BatchNorm, GroupNorm, or other cross-row aggregation); (ii) no learned row-index embedding, positional encoding, rotary/relative-position bias, or asymmetric row-aware mask is applied along the row axis; (iii) the predictor is constructed from alternating row-attention and column-attention layers separated by per-token operations, with support-row attention content-only and each per-token operation acting identically on all rows; (iv) the output head reads each query row through a shared per-token map independent of the row index; (v) the pre-training distribution is exchangeable under joint permutations of support rows. Each layer is therefore equivariant under joint permutations of support rows, and the per-query output is invariant under such permutations.

Verification. The deployed architecture class satisfies Architectural condition G.4 by direct inspection of the released Tab2D checkpoint implementation (Zhang et al., 2025): attention modules use standard scaled dot-product without relative-position bias, no sinusoidal or learned positional embedding is added to row tokens, and padding masks are symmetric across rows. A future variant that introduces any such component (e.g. a learned row-index embedding, a rotary or relative-position bias applied along the row axis, or an asymmetric row-aware mask) falls outside this class, and Architectural condition G.4 and Theorem G.5 must be re-verified against the new class.

Theorem G.5 (Row exchangeability)**.** *Under Architectural condition G.4, for any permutation $\sigma \in S_N$ acting jointly on the support rows (X_S, y_S) and any query X_Q,*

$$f_\theta(\sigma \cdot X_S,\, \sigma \cdot y_S,\, X_Q) \;=\; f_\theta(X_S, y_S, X_Q).$$

Equivalently, $f_\theta(\cdot \mid X_S, y_S)$ depends on (X_S, y_S) only through their empirical distribution.

Proof. Each layer of the predictor is a composition of per-token maps (layer normalization, per-row feed-forward) and attention operations. Per-token maps act identically on each token and therefore commute with any joint row permutation trivially: if T is per-token, then $T(\sigma \cdot Z) = \sigma \cdot T(Z)$ for any row-permutation $\sigma \in S_N$ applied to the stack of tokens Z.

Row-attention is the map $(Q_S, K_S, V_S) \mapsto \mathrm{softmax}(Q_S K_S^\top/\sqrt{d})\, V_S$, with no positional signal added (Architectural condition G.4). Under a joint permutation σ acting on support rows, $(\sigma \cdot Q_S, \sigma \cdot K_S, \sigma \cdot V_S)$ yields attention weights $\mathrm{softmax}((\sigma Q_S)(\sigma K_S)^\top/\sqrt{d}) = \mathrm{softmax}(\sigma Q_S K_S^\top \sigma^\top/\sqrt{d})$, whose output is $\sigma \cdot \mathrm{softmax}(Q_S K_S^\top/\sqrt{d})\, V_S = \sigma \cdot \mathrm{Attn}(Q_S, K_S, V_S)$. Hence row-attention is equivariant under joint row permutation.

Column-attention acts within each row across features and therefore commutes with any row permutation trivially. By induction on the stack of layers, the backbone's output on the support is row-equivariant, and therefore its output on the query (which reads each query row through a shared per-token map that does not depend on the row index; Architectural condition G.4) is row-invariant in the support. Equivalently, the output depends on (X_S, y_S) only through their empirical distribution.

For variable-length supports, the equivariance argument requires the padding mask to be a content-independent, symmetric indicator of the "real row" / "pad row" partition: any permutation that fixes this partition commutes with the mask and the equivariance argument goes through unchanged. A permutation that mixes real and pad rows is not a symmetry of the extended support and is not considered; Theorem G.5 is a statement about permutations acting on the real rows only.

Theorem G.5 is a *fixed-support-size* invariance statement: it asserts order-invariance within a support set of a given cardinality $|X_{\mathrm{ref}}|$, not stability of the output as $|X_{\mathrm{ref}}|$ itself varies. Variable-reference-set-size generalization (how the frozen predictor behaves when the deployment's X_{ref} is much smaller or larger than the pre-training support sizes) is empirical in this work and is not implied by row-exchangeability alone. □

G.3.2. Deployment-risk decomposition with sampler-label residual

The transfer ingredient is a decomposition identity that names the sources of deployment error, not a quantitative certification. The equality form below always holds; the inequality form that follows is informative only when the sampler-label residual $\epsilon_{\mathrm{Lab}}(\mathcal{P})$ is small compared to the loss constant B. On real deployments whose anomaly-generating process differs from the sampler (e.g. ODDS), $\epsilon_{\mathrm{Lab}}(\mathcal{P})$ is generically $\Theta(B)$ and the inequality is diagnostic rather than quantitative.

Purpose of the decomposition. The purpose of the decomposition is not to produce a small universal bound. It identifies *where* zero-shot transfer can fail: the pre-trained predictor may not fit its synthetic task distribution, the deployment feature distribution may fall outside the synthetic task family, or the real anomaly labels may disagree with the sampler-induced exit rule.

Proposition G.6 (Deployment-risk decomposition with sampler-label residual)**.** *Let $\mathcal{P}^{\mathrm{exit}}$ be the synthetic-rule deployment distribution of Section G.1, obtained from $\mathcal{P}$ by replacing the query labels y_Q with $L_S(X_Q; X_{\mathrm{ref}})$. Under (A5), for every $\theta \in \Theta$ the identity*

$$R_{\mathcal{P}}(\theta) \;=\; R_{\mathcal{T}}(\theta) \;+\; \underbrace{\big(R_{\mathcal{P}^{\mathrm{exit}}}(\theta) - R_{\mathcal{T}}(\theta)\big)}_{\textit{task-prior / feature shift}} \;+\; \underbrace{\big(R_{\mathcal{P}}(\theta) - R_{\mathcal{P}^{\mathrm{exit}}}(\theta)\big)}_{\textit{sampler-label residual}}$$

holds.

Proof. The identity is a telescoping sum: add and subtract $R_{\mathcal{P}^{\text{exit}}}(\theta)$ on the right-hand side of $R_{\mathcal{P}}(\theta) - R_{\mathcal{T}}(\theta)$, then regroup. No further hypothesis beyond (A5) is used. □

Remark G.7 (Optional IPM upper bound). The identity upper-bounds to an IPM-based inequality: with $\epsilon_{\text{Lab}}(\mathcal{P}) := \sup_{\theta\in\Theta} |R_{\mathcal{P}}(\theta) - R_{\mathcal{P}^{\text{exit}}}(\theta)|$ and $g_\theta = \ell_\epsilon \circ f_\theta \in \mathcal{G}_\theta$, the middle term is controlled by its loss-class IPM,

$$|R_{\mathcal{P}^{\text{exit}}}(\theta) - R_{\mathcal{T}}(\theta)| \;=\; |\mathbb{E}_{\mathcal{P}^{\text{exit}}} g_\theta - \mathbb{E}_{\mathcal{T}} g_\theta| \;\le\; d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta),$$

and summation gives

$$R_{\mathcal{P}}(\theta) \;\le\; R_{\mathcal{T}}(\theta) \;+\; d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta) \;+\; \epsilon_{\text{Lab}}(\mathcal{P}).$$

The inequality is informative only when both $d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta)$ and $\epsilon_{\text{Lab}}(\mathcal{P})$ can be independently controlled away from the loss scale B; on arbitrary real anomaly distributions it is a diagnostic upper bound, not a certification guarantee. The equality identity of Proposition G.6 is the main statement.

Remark G.8 (On the IPM argument vs. the naive domination step). An earlier natural-looking rewrite would replace $d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta)$ in the bound by $d_{\text{IPM}}(\mathcal{P}, \mathcal{T}; \mathcal{G}_\theta)$, on the grounds that "$\mathcal{P}$ and $\mathcal{P}^{\text{exit}}$ share their X_{ref} and X_Q marginals." This replacement is *not* generally valid for an IPM over functions of the 4-tuple $\tau = (X_S, y_S, X_Q, y_Q)$: for $g \in \mathcal{G}_\theta$ that depends non-trivially on y_Q, the expectations under $\mathcal{P}$ and $\mathcal{P}^{\text{exit}}$ can differ even though the feature-side marginals are shared, so there is no automatic inequality between $d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta)$ and $d_{\text{IPM}}(\mathcal{P}, \mathcal{T}; \mathcal{G}_\theta)$. We therefore keep the $\mathcal{P}^{\text{exit}}$-IPM form above and separate the labeling-rule gap into ϵ_{Lab} rather than absorbing it into the IPM.

Remark G.9 (On the ϵ_{Lab} term for real deployments). The residual $\epsilon_{\text{Lab}}(\mathcal{P})$ denotes the *deployment-side* real-vs-exit-rule label gap: the difference between $\mathcal{P}$'s real query labels y_Q and $L_S(X_Q; X_{\text{ref}})$ at the same feature point. The *training-side* generation-vs-exit-label gap (between the sampler's $y = 1$ label and $L_S(\tilde{X}; X_{\text{ref}})$, non-zero when a synthetic row's perturbed coordinates happen to remain inside $\widehat{S}$) is *not* inside ϵ_{Lab}; it lives inside the structural-shift term $d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta)$, which compares the exit-rule surrogate against the generation-label pre-training distribution. On real deployments whose anomaly-generating process differs structurally from the sampler (the norm on ODDS, where real anomalies are domain-specific outliers rather than minimal-subset perturbations of normal rows), $\epsilon_{\text{Lab}}(\mathcal{P})$ is generically $\Theta(B)$, of the same order as the loss upper bound. Proposition G.6's inequality is therefore informative only when the sampler's exit rule, pushed forward to $\mathcal{P}_X$'s support, agrees with real anomaly labels on most of the support. This is a structural limitation of the decomposition, not a bug: the transfer inequality precisely refuses to assume a label-level coupling between $\mathcal{T}$ and real deployments, and the ϵ_{Lab} term names that refusal.

Remark G.10 (Structural source of the task-prior-shift term). Under the pre-training task distribution $\mathcal{T}$, the anomaly marginal under the synthetic-labeling rule is a convex combination of sampler outputs across the 22 priors and 5 strategies: every positive-label $\tilde{X}$ is $\mathsf{Sampler}_{s,\alpha,\mu}(X)$ for some (π, s, α, μ) in the pool. The IPM term $d_{\text{IPM}}(\mathcal{P}^{\text{exit}}, \mathcal{T}; \mathcal{G}_\theta)$ is therefore large precisely when the deployment's feature-marginal anomaly density, evaluated under $L_S(\cdot; X_{\text{ref}})$, lies far from the convex hull of this reachable family in the loss-class-IPM pseudometric. We do not prove a quantitative lower bound on this shift: a natural candidate lower bound would require a witness function in $\mathcal{G}_\theta$ whose expectation gap between $\mathcal{P}^{\text{exit}}$ and $\mathcal{T}$ is directly controllable, which in turn needs realizability or a subset-of-loss-class argument we do not verify. Empirically, the synthetic-to-real gap manifests on the heavy-tail ODDS benchmarks (GLASS, PIMA, ARRHYTHMIA) where SPOT-FM underperforms tree-based and kernel baselines; this is the observable signature of a deployment whose anomaly marginal sits outside the reachable family, and we regard it as empirical evidence for the term's non-negligibility rather than as a formal certification.

G.4. Supporting results: inference, curriculum, auxiliary supervision

Three SPOT-FM *design choices (inference-time averaging, the μ-curriculum, and the auxiliary μ-prediction head) are not proved as theorems in this appendix. Inference-time averaging admits a one-paragraph Jensen-safety remark (below); the other two are discussed in the informal design-choice paragraphs that close this subsection.*

G.4.1. Multi-configuration ensembling: Jensen-safety remark

Jensen-safety remark. Under (A5), the clipped loss $p \mapsto \ell_\epsilon(p, y)$ is convex on $[\epsilon, 1-\epsilon]$. Jensen's inequality applied to this convex restriction gives, for any $\{\hat{p}_i\}_{i=1}^K \subset [\epsilon, 1-\epsilon]$ and any $y \in \{0, 1\}$,

$$\ell_\epsilon\left(\tfrac{1}{K}\sum_i \hat{p}_i,\; y\right) \;\le\; \tfrac{1}{K}\sum_{i=1}^K \ell_\epsilon(\hat{p}_i, y),$$

i.e. averaging clipped probabilities cannot increase the ℓ_ϵ loss. This is a log-loss safety statement only; the empirical $+1$ to 2 ROC/PR lift from the 216-configuration average (Appendix D.2) is not a consequence of Jensen, because ROC and PR are ranking metrics, not instances of ℓ_ϵ. A strict variance-reduction improvement would require a lower bound on ℓ'' (strong convexity); we do not formalize that here.

G.4.2. Informal justifications for design choices

Three SPOT-FM *design choices are not proved as theorems in this appendix, because each would require hypotheses that we cannot verify at the architectural or data-distribution level. We write one paragraph each and cite the empirical ablation that actually justifies the choice.*

(i) Proximity/reliability intuition for the perturbation fraction μ. Perturbing more coordinates raises the exit probability $\Pr(\tilde{x} \notin \widehat{S})$ geometrically in $\lceil \mu F \rceil$ (Theorem G.1), but negatives that leave the operational region by a large margin are easily classified and supply little training signal about the normal/anomalous boundary. The useful signal therefore lives at intermediate μ (Fig. 1(b)). We do not attempt a theoretical prediction of the best μ; the constant-μ inference grid of Appendix D.2 is justified empirically, and fixing $\mu \sim \mathcal{U}(0.05, 0.95)$ throughout training costs -1.3 ROC on average (Appendix B.2). Because only a μ-fraction of coordinates is changed and the range-expansion factor α is small, the generated negatives remain tied to the source row; we treat "boundary proximity" as a design intuition (consistent with the exit-probability result but not formally equivalent to a Hausdorff-localization statement), and validate it empirically (Fig. 1(b)).

(ii) μ-curriculum. The annealed schedule $\mathrm{Beta}(6, 2) \to \mathrm{Beta}(2, 5)$ over the first 40% of training (then fixed) is motivated by the standard easy-to-hard training heuristic. A formal curriculum-superiority theorem in the non-convex SGD setting would require a schedule-ordering hypothesis on gradient variance and a Polyak-Łojasiewicz-type condition to translate the Ghadimi-Lan stationarity-gap bound into a risk bound; neither is verifiable for a 76M-parameter transformer, and the first is not derivable from standard smoothness assumptions alone. We therefore do not claim a curriculum theorem. Empirically, the annealed schedule outperforms fixed $\mu \sim \mathcal{U}(0.05, 0.95)$ by $+1.3$ ROC on ODDS (Appendix B.2); this is the direct justification.

(iii) Auxiliary μ-prediction head. The auxiliary head outputs a Beta distribution over the task's μ, trained against the conditional-on-curriculum-time Beta that generated μ. Since μ is supplied to the model as an input token (μ_{tok}), the head is not inferring μ from the data; it is asked to reconstruct μ from the shared representation, and therefore is best interpreted as a representation-level reconstruction/regularization loss rather than latent task-difficulty inference or oracle-recovery estimation. A formal multi-task-bound theorem would require Maurer-style product-Rademacher hypotheses and a parameter-clipped Beta head; neither the clipping nor the shared-backbone product structure is verified here, so we do not state such a theorem. Empirically, ablating the auxiliary head costs -0.8 ROC on ODDS (Appendix D.3).

G.5. Limits of the analysis

All exit-probability statements are stated relative to the empirical quantile box $\widehat{S}(X_{\mathrm{ref}}) = \prod_j [\hat{q}_j^{0.05}, \hat{q}_j^{0.95}]$ induced by the reference set; other operational regions (Gaussian box, density sublevel set, Euclidean ball) would require their own per-coordinate exit constants and are not used. (A2) is asserted for the UNIFORM strategy only, since EMPIRICAL resamples a single row and induces column dependence via X's empirical covariance and GAUSSIAN/MIXTURE/BOUNDARY are not used in the theorem. (A3) is likewise UNIFORM-only, with data-dependent per-feature constant $\delta_k(\alpha; X_{\mathrm{ref}}) = 1 - (\hat{q}_k^{0.95} - \hat{q}_k^{0.05})/(U_k - L_k)$ (U_k, L_k the min/max-expanded range of Appendix B.1); the symmetric-box $\alpha/(1+\alpha)$ is a special case. The BOUNDARY strategy trivially satisfies $\Pr(\tilde{x} \notin \widehat{S} \mid x \in \widehat{S}) = 1$ by construction for every $k \geq 1$, which is recorded as an observation, not as a theorem. Theorem G.1 is conditional on $x \in \widehat{S}$: in high dimension the joint probability of lying inside the 5 to 95% box is small ($\approx 0.9^F$ under weak independence), so the theorem characterizes the behavior of contrasts generated from rows already inside $\widehat{S}$, while rows outside satisfy $L_S(x; X_{\mathrm{ref}}) = 1$ trivially. Theorem G.5 requires token-wise normalization, no learned row-index signal, and content-independent symmetric padding masks (Architectural condition G.4); a future variant violating these conditions falls outside the model class. The sampler-label residual $\epsilon_{\mathrm{Lab}}(\mathcal{P})$ in Proposition G.6 is generically $\Theta(B)$ on real deployments whose anomaly-generating process differs structurally from the sampler (the norm on ODDS); the inequality form of the proposition is informative only when $\epsilon_{\mathrm{Lab}}(\mathcal{P}) \ll B$, while the equality form always holds. $\mathcal{P}^{\mathrm{exit}}$ is the deployment distribution relabeled by the exit oracle L_S, not the generation-labeled

synthetic pre-training distribution; the training-side generation-vs-exit-label gap is absorbed into the task-prior / structural-shift term $d_{\mathrm{IPM}}(\mathcal{P}^{\mathrm{exit}}, \mathcal{T}; \mathcal{G}_\theta)$, while ϵ_{Lab} denotes only the deployment-side real-vs-exit gap; conflating the two would be technically incorrect. (A5) uses clip-before-BCE for a convex loss on $[\epsilon, 1-\epsilon]$ with $B = \log(1/\epsilon)$; the deployed logit-level BCE differs from this clipped analogue only on a measure-zero tail at $\epsilon \sim 10^{-6}$. The Jensen-safety remark of Section G.4.1 asserts only that averaging clipped probabilities cannot increase the ℓ_ϵ loss; the empirical $+1$ to $+2$ ROC/PR lift of the 216-configuration average is consistent with strong-convexity variance reduction but we do not formalize it, since ROC/PR are ranking metrics rather than instances of ℓ_ϵ. A sharp Rademacher bound for variable-length dual-axis attention is open in the statistical-learning literature (Edelman et al., 2022) and no result in this appendix uses a Rademacher placeholder. The theory therefore does *not* prove that SPOT-FM recovers the real anomaly manifold (Theorem G.1 is about leaving the operational region, not concentrating on a deployment's anomaly manifold; localization is evidenced by the per-dataset ODDS results of §5.2), that the 216-grid strictly improves AUC or PR (Jensen only rules out strict harm to the surrogate loss), that the μ-curriculum is optimal (the -1.3 ROC cost of fixed-μ training in Appendix B.2 is the empirical justification), or that transfer error is small on arbitrary real deployments (ϵ_{Lab} is generically $\Theta(B)$ on real ODDS deployments).