THE SURPRISING STRENGTH OF WEAK CLASSIFIERS FOR TWO-SAMPLE TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

The two-sample testing problem, a fundamental task in statistics and machine learning, seeks to determine whether two sets of samples, drawn from underlying distributions p and q, are in fact identically distributed (i.e. whether $p = q$). A popular and intuitive approach is the classifier two-sample test (C2ST), where a classifier is trained to distinguish between samples from p and q. Yet despite simplicity of the C2ST, its reliability hinges on access to a near-Bayes-optimal classifier, a requirement that is rarely met and difficult to verify. This raises a major open question: can a weak classifier still be useful for two-sample testing? We show that the answer is a definitive yes. Building on the work of Hu & Lei (2024), we analyze a conformal variant of the C2ST that converts the scores from any trained classifier—even if weak, biased, or overfit—into exact, finite-sample p-values. We establish two key theoretical properties of the conformal C2ST: (i) finite-sample Type-I error control, and (ii) non-trivial power that degrades gently in tandem with the error of the trained classifier. The upshot is that even poorly performing classifiers can yield powerful and reliable two-sample tests. This general framework finds a powerful application in Bayesian inference, particularly for validating Neural Posterior Estimation (NPE) models, where the task of comparing a learned posterior approximation $q(\theta \mid y)$ to the true posterior $p(\theta \mid y)$ can be framed as a two-sample test. Empirically, the Conformal C2ST outperforms classical discriminative tests across a wide range of benchmarks for this task. Our results establish the conformal C2ST as a practical, theoretically grounded diagnostic tool.

1 INTRODUCTION

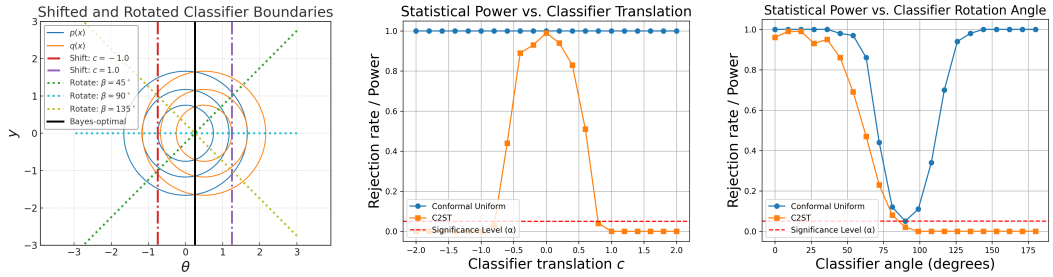
A fundamental problems in statistics and machine learning is to assess whether two sets of samples, drawn from distributions $p(x)$ and $q(x)$, are in fact distributed identically. To this end, two-sample tests (Lehmann & Romano, 2005) summarize the differences between the samples into a test statistic, which is then used to test the null hypothesis that $p = q$. A key application for modern two-sample tests is evaluating the sample quality of generative models, where p represents the true data-generating process and q is a neural approximation.

This evaluation challenge is particularly acute in Neural Posterior Estimation (NPE), an increasingly popular and practical tool for Bayesian inference. NPE methods use simulations to train a deep generative model $q(\theta \mid y)$ to approximate the true, typically intractable posterior $p(\theta \mid y)$ (Ho et al., 2020; Geffner et al., 2023; Chen et al., 2025; Gloeckler et al., 2024; Papamakarios et al., 2021; Wildberger et al., 2023; Kingma et al., 2013). While powerful, this approach creates a critical validation problem: how can we verify whether the learned posterior q faithfully approximates the true posterior p? Several diagnostics exist for this purpose, yet they all suffer from drawbacks. Simulation-based calibration (SBC) (Talts et al., 2018) is a widely used tool. Yet in its original form (rank-SBC), it operates only on one-dimensional marginals. This not only creates a potentially severe multiple-testing problem, but also leaves rank-SBC insensitive to inaccuracies that affect the joint distribution of all parameters without affecting their one-dimensional marginals. Proposed fixes that use the joint likelihood as a test statistic (Modrák et al., 2023) are inapplicable to many NPE problems, where likelihoods are intractable. Another recent method called Test of Accuracy with Random Points (TARP) (Lemos et al., 2023) is highly sensitive to the specification of a non-trainable proposal distribution, limiting its practical use.

Given the limitations of these specialized tools, a natural alternative is the classifier-based two-sample test (C2ST) (Lopez-Paz & Oquab, 2016; Linhart et al., 2023), a flexible, general-purpose approach. The C2ST reframes the problem as a classification task: if a classifier can reliably distinguish samples from p and q, the distributions must be different. Yet in practice, the C2ST depends critically on the quality of the classifier. If a model q "passes" the C2ST, that could mean either that $q = p$, or that $q \neq p$ but the classifier is too weak or poorly trained to tell the difference. This ambiguity persists even in more advanced methods classifier-based methods like Discriminative Calibration (DC) (Yao & Domke, 2023), which also relies on access to a near-optimal classifier.

Summary of contributions. We address these limitations by introducing the Conformal C2ST. Our approach builds on the conformal framework of (Hu & Lei, 2024) but departs on a crucial point. Their method fundamentally relies on a nearly-Bayes-optimal classifier to model the density ratio between distributions, an assumption that often fails in practice. This is especially problematic in challenging settings like neural posterior estimation, where classifiers are often trained on limited data and may be too weak to find an optimal decision boundary.

Our work is designed for this realistic, imperfect regime. We show how conformal calibration can transform a weak, misspecified, or overfit classifier into a powerful and trustworthy two-sample test. Specifically, we first show that conformal calibration effectively decouples type-I error control from the classifier's discriminative accuracy. More importantly, we also prove that conformal p-values maintain meaningful power when $p \neq q$, even when the classifier is weak or poorly trained. Intuitively, they do so by aggregating weak but informative ranking signals, a property that is crucial in scenarios where the traditional C2ST struggles, such as high-dimensional posteriors, small-sample regimes, and tasks with low signal-to-noise ratios. We support these theoretical developments with extensive empirical results, showing that the conformal C2ST exhibits state of the art performance for the two-sample testing problem across a wide range of benchmark problems motivated by NPE.



(a) Perturbed classifier boundaries (b) Comparison under shift (c) Comparison under rotation

Figure 1: Power of the C2ST and conformal C2ST under shift and rotate perturbations of the optimal decision boundary. The conformal test is much more robust to a weak or misspecified classifier.

A toy example Before detailing our results, we briefly focus on a toy example, to illustrate the fundamental advantage of conformal calibration. The setup mimics the NPE setting, where the marginal distribution of y is preserved under both the true and approximate joint distributions. Specifically, we consider two bivariate normal distributions: $p(\theta, y)$ is a standard bivariate normal, $\mathcal{N}(\mathbf{0}, I_2)$, while $q(\theta, y)$ is the same distribution with a mean shift of 0.5 in the θ coordinate, i.e., $\mathcal{N}([0.5, 0]^\top, I_2)$. The Bayes-optimal classifier for distinguishing between samples from p and q is a linear decision boundary at $\theta = 0.25$, bisecting $\mathbb{R}^2$ between the two means. Classifier scores are obtained by computing the signed distance from a sample point (θ, y) to the decision boundary.

We then degrade the classifier by manipulating the decision boundary in two systematic ways. First is *translation*, where the decision boundary is shifted horizontally by a parameter c, modifying the decision boundary to $\theta = 0.25 + c$. The parameter c can be positive, shifting the boundary toward q, or negative, shifting it toward p. As $|c|$ increases, the boundary moves further from its optimal position, increasing classification error. As we see in Figure 1b, this rapidly degrades the power of the C2ST for distinguishing p and q. Yet the conformal C2ST's performance remains remarkably stable, showing no degradation in power even for large shifts. This suggests a strong level of robustness to biased or poorly calibrated classifiers. (Although the classifier is trained to distinguish individual

draws from p and q, the goal of neural posterior testing is to assess whether two *collections* of samples arise from the same distribution, allowing aggregation across multiple draws and yielding greater power than would be expected from single-point classification alone.) The second form of degradation is *rotation*, where the decision boundary is rotated about the midpoint between the means of p and q. The boundary under rotation is described by $(\theta - 0.25)\cos\beta + y\sin\beta = 0$, where β controls the angle of rotation. A rotation of $\beta = 0$ corresponds to the optimal linear discriminant, while increasing $|\beta|$ introduces growing misalignment. Figure 1c shows that the conformal C2ST continues to perform significantly better under quite severe rotations, indicating resilience even when the classifier is badly misaligned. At $\beta = \pi/2$, the decision boundary becomes completely orthogonal to the separation axis of the two distributions, rendering the distribution of classifier scores identical under p and q. In this extreme case, the conformal C2ST has power 0.05, which is precisely the type-I error rate of the test.

These results illustrate the key advantage of the conformal C2ST: even when the classifier itself becomes progressively poorer or more miscalibrated, its ranking signals nonetheless remain useful.

2 Classifier-Based Posterior Testing and Conformal Calibration

2.1 Preliminaries

While the conformal C2ST is general two-sample test, we focus on the NPE setting as a motivating problem. Consider two distributions over joint parameter–observation pairs $(\theta, y) \in \Theta \times \mathcal{Y}$: the true posterior $p(\theta \mid y)$, defined via the joint density $p(\theta, y) = \pi(\theta)p(y \mid \theta)$, and an approximate posterior $q(\theta \mid y)$, learned using simulation-based inference (SBI). We are agnostic as to *how* q was learned; instead, our goal is to assess whether $q(\theta \mid y) = p(\theta \mid y)$ almost surely over $\Theta \times \mathcal{Y}$. Assuming a shared marginal $\pi(y)$, we define the approximate joint as $q(\theta, y) := \pi(y)q(\theta \mid y)$. Under this assumption, testing equality of posteriors reduces to testing the null hypothesis $H_0 : p(\theta, y) = q(\theta, y)$ almost surely over $\Theta \times \mathcal{Y}$. This naturally frames the problem as a *two-sample test* between the joint distributions $p(\theta, y)$ and $q(\theta, y)$.

In the NPE setting, i.i.d. samples from both p and q are readily available. To sample from p, we draw $\theta \sim \pi(\theta)$ and then $y \sim p(y \mid \theta)$. To sample from q, we use the y margin only of the true p, implicitly generating $y \sim \pi(y)$, and then we sample $\theta \sim q(\theta \mid y)$ from the learned model. For convenience, we write $x = (\theta, y)$, and use p and q as shorthand for the true and approximate joint densities over x. We denote samples as $\{X_i\}_{i=1}^n \sim p$ and $\{\tilde{X}_i\}_{i=1}^n \sim q$. Throughout, we assume that $p(\theta \mid y)$ and $q(\theta \mid y)$ are absolutely continuous with respect to a common base measure.

The classifier two-sample test (C2ST). The C2ST is a widely used approach for detecting differences between two distributions. In the context of posterior validation, suppose we are given two datasets $\{x_i\}_{i=1}^n \sim p$ and $\{\tilde{x}_i\}_{i=1}^n \sim q$. To test the null hypothesis $H_0 : p = q$, a classifier is trained to distinguish between the two samples. Specifically, each x_i is assigned label 1 and each $\tilde{x}_i$ is assigned label 0, producing a labeled dataset $\{(z_i, \ell_i)\}_{i=1}^{2n}$, where $z_i \in \mathcal{X}$ and $\ell_i \in \{0, 1\}$. The combined dataset is randomly split into disjoint training and testing subsets. A classifier $f : \mathcal{X} \to [0, 1]$ is trained on the training portion to estimate the conditional probability $\mathbb{P}[\ell = 1 \mid z]$, and evaluated on the test set $\mathcal{D}_{\text{te}}$. The C2ST test statistic is the classification accuracy on the test set:

$$\hat{t} := \frac{1}{n_{\text{te}}} \sum_{(z_i, \ell_i) \in \mathcal{D}_{\text{te}}} \mathbb{I}\left\{\mathbb{I}\{f(z_i) > 1/2\} = \ell_i\right\} ,$$

which is shown to have an asymptotic normal distribution in Lopez-Paz & Oquab (2016). The null hypothesis $H_0 : p = q$ is rejected if $\hat{t}$ is significantly greater than 0.5, indicating that the classifier has learned to distinguish between p and q.

While C2ST is easy to implement and effective in many cases, it has two key limitations. First, its power depends heavily on the classifier quality. Poorly trained or underfit classifiers can yield inconclusive results, even when the two distributions differ. Second, its test statistic relies on hard decisions via thresholding at 0.5, discarding potentially useful information about classifier confidence.

2.2 THE UNIFORM TEST: CONFORMAL CALIBRATION OF CLASSIFIER SCORES

To address these limitations, we adopt a conformal framework that calibrates classifier scores directly, rather than thresholding them. This yields a method we term the *Conformal C2ST*. The core idea is to treat each point as a test case, compute a score for that point (e.g., classifier log-odds), and use a conformal p-value to assess how extreme that score is relative to a calibration sample from the reference distribution (p). Each such p-value reflects the plausibility of a single test point from q under the null. To test whether $p = q$ globally, we repeat this procedure across many independent draws from q, aggregating the resulting p-values to assess overall deviation from uniformity. This allows the method to extract and accumulate weak signals for assessing distributional equality, even from underperforming classifiers, by calibrating based on ranks rather than raw accuracy.

Concretely, let $\{X_1, \ldots, X_m\} \sim p$ be a calibration sample and let $\tilde{X} \sim q$ be a test point. Let $s : \mathcal{X} \to \mathbb{R}$ be a deterministic scoring function, which in our case will be the output of a classifier trained to distinguish p from q, just like in C2ST. (Below we discuss the specific choice of s.) Define the nonconformity scores $S_i = s(X_i)$ for $i = 1, \ldots, m$, and $S_{m+1} = s(\tilde{X})$. The conformal p-value (Vovk et al., 2005; Lei & G'Sell, 2018) for $\tilde{X}$ is then given by:

$$U := \frac{1}{m+1}\left(\sum_{i=1}^{m+1} \mathbb{I}\{S_i < S_{m+1}\} + \xi \cdot \sum_{i=1}^{m+1} \mathbb{I}\{S_i = S_{m+1}\}\right), \tag{1}$$

where $\xi \sim \mathrm{Unif}[0, 1]$ is a tie-breaking random variable. The p-value U reflects the relative rank of the test point's score among the calibration scores.

2.3 RESULTS ON VALIDITY AND POWER

A key advantage of this approach is that it inherits the finite-sample validity guarantee from conformal prediction. Under H_0, the calibration points and the test point are exchangeable, implying the following marginal guarantee.

Lemma 1 (Uniformity under the null)**.** *Under the null hypothesis $H_0 : p = q$, the conformal p-value U defined in (1) satisfies:*

$$\mathbb{P}(U \le u \mid H_0) = u, \quad \forall u \in [0, 1],$$

where the probability is over the random draw of the calibration sample, the test point, and the tie-breaking variable.

This result holds for any deterministic scoring function s and for any finite calibration size m, making the conformal C2ST robust to classifier quality and sample size. This result is a standard property of conformal inference; see (Vovk et al., 2005; Lei & G'Sell, 2018).

From marginal p-values to a uniformity test. To turn the marginal validity established in Lemma 1 into a two-sample test of $H_0 : p = q$, we repeat the conformal p-value computation across multiple test points. Specifically, let $\{\tilde{X}_j\}_{j=1}^{n_q} \sim q$ be independent test samples. For each j, we draw an independent calibration set $\mathcal{C}_j = \{X_{j,i}\}_{i=1}^{m} \sim p$, and compute the corresponding conformal p-value

$$U_j := \frac{1}{m+1}\left(\sum_{i=1}^{m+1} \mathbb{I}\left\{s(X_{j,i}) < s(\tilde{X}_j)\right\} + \xi_j \cdot \sum_{i=1}^{m+1} \mathbb{I}\left\{s(X_{j,i}) = s(\tilde{X}_j)\right\}\right),$$

where $X_{j,m+1} = \tilde{X}_j$, and $\xi_j \sim \mathrm{Uniform}[0, 1]$ are independent tie-breaking variables. Under H_0, each $U_j \sim \mathrm{Unif}[0, 1]$, so we can aggregate the $\{U_j\}$ to form a global test statistic. We consider the empirical CDF $\hat{G}(u) = \frac{1}{n_q}\sum_{j=1}^{n_q} \mathbb{I}\{U_j \le u\}$ and apply the one-sample Kolmogorov–Smirnov (KS) test (Lehmann & Romano, 2005), using

$$T_{\mathrm{KS}} = \sup_{u \in [0,1]} \left|\hat{G}(u) - u\right|.$$

Because $H_0 : p = q$ implies exchangeability between the calibration and test samples, each U_j is uniformly distributed, and the KS test (or any similar test) controls Type-I error at level α. (We refer to this in the benchmarks as the "uniform" test.)

Power. While Lemma 1 ensures exact marginal validity of conformal p-values for *any* deterministic score function s, the power of the test under the alternative $H_1 : p \neq q$ crucially depends on how well s separates samples from p and q. A natural and theoretically grounded choice is the *oracle density ratio* between the joint distributions (or equivalently their conditionals for θ, since $p(y) = q(y)$). Let $r(x) = p(x)/q(x)$ denote this true density ratio. If $\eta(x) := \mathbb{P}(l = 1 \mid x)$ is the output of the Bayes classifier distinguishing between $x \sim p$ (label $l = 1$) and $x \sim q$ (label $l = 0$), then the density ratio is related to classifier scores via:

$$r(x) = \frac{\eta(x)}{1 - \eta(x)}. \tag{2}$$

We will soon consider what happens when r is estimated with error, but for now we suppose it is known. Because the transformation $t \mapsto t/(1-t)$ is strictly increasing on $(0, 1)$, the rankings induced by $r(x)$ and the probabilities $\eta(x)$ are identical.

These rankings are of central importance to the conformal method, which depends only on the orderings of scores, not their magnitudes. A natural way to quantify the quality of such a ranking is the area under the ROC curve (AUC) (Fawcett, 2006), which measures how well a scoring function separates the two distributions. Specifically, the AUC for $r(x)$ is given by:

$$\mathrm{AUC}(r) = \mathbb{P}_{X\sim p,\, \tilde{X}\sim q}\left[r(X) > r(\tilde{X})\right] + \frac{1}{2}\mathbb{P}\left[r(X) = r(\tilde{X})\right],$$

which reflects the probability that a randomly chosen p-sample is ranked above a q-sample by the scoring function $r(x)$. The next lemma formalizes that $r(x)$ (or any monotonic transformation of it) maximizes this quantity.

Lemma 2. *For any measurable scoring function $s : \mathcal{X} \to \mathbb{R}$, $\mathrm{AUC}(s) \leq \mathrm{AUC}(r)$, with equality if and only if there exists a strictly increasing function h such that $s(x) = h(r(x))$ for q-almost every x.*

This result justifies the use of the density ratio $r(x)$ as an optimal scoring function for discriminating between p and q; see Appendix A.1 for the full statement and proof.

Moreover, AUC serves as a useful proxy for the power of the conformal uniformity test, because it quantifies how well the scoring function ranks samples from p above those from q. Specifically, as the number of calibration samples $m \to \infty$, the conformal p-value for a test point $\tilde{X} \sim q$ converges almost surely to its population-level limit:

$$U \xrightarrow[m\to\infty]{\text{a.s.}} \mathop{\mathbb{P}}_{X\sim p}\left[r(X) < r(\tilde{X})\right] + \xi \mathop{\mathbb{P}}_{X\sim p}\left[r(X) = r(\tilde{X})\right], \quad \text{where } \xi \sim \mathrm{Unif}\,(0, 1)\,.$$

When $r(x)$ has good separation between the distributions, these p-values tend to concentrate below 0.5, making them detectably non-uniform.

As the number of test points $n_q \to \infty$, the empirical CDF $\hat{G}(u) = \frac{1}{n_q}\sum_{j=1}^{n_q} \mathbb{I}\{U_j \leq u\}$ converges to the population CDF $G(u) = \mathbb{P}(U \leq u)$, and the Kolmogorov–Smirnov test statistic converges to $\sup_{u\in[0,1]} |G(u) - u|$, which captures the deviation from uniformity. When AUC exceeds 0.5, i.e., $r(x)$ tends to rank p samples above q, conformal p-values become stochastically smaller than uniform. This deviation drives up the KS statistic and hence the test power. Moreover, the expected conformal p-value under the alternative is directly related to AUC, as the following lemma establishes.

Lemma 3. *Under the alternative $H_1 : p \neq q$, we have*

$$\mathbb{E}\,[U] = 1 - \mathrm{AUC}(r) \leq \frac{1}{2} - \frac{TV(p, q)}{2} < \frac{1}{2}$$

in the asymptotic limit of $m \to \infty$.

This result shows that, under the alternative, conformal p-values skew toward zero, with the extent of skewness proportional to the total variation distance between the two distributions. (Hu & Lei, 2024) also motivate the density ratio $r(x) = p(x)/q(x)$ from an information-theoretic perspective, showing that its variability under q controls the deviation from uniformity; see Lemma 4 in the Appendix.

Under degraded classifier scores. Our above analysis involves the oracle density ratio $r(x)$. But in practice, the density ratio is often approximated using a classifier trained to distinguish between

the true and approximate joint distributions. Accordingly, our analysis explicitly incorporates an error-prone plug-in estimate $\hat{r}$, derived from a potentially weak classifier $\hat{\eta}$ in (2).

Our main theoretical result shows that, under mild regularity conditions on the true and estimated density ratios, the uniform conformal test retains high power, even when the scoring function is imperfect. Our result specifically relates the error of the estimated density ratio to the performance of the conformal C2ST.

Assumption 1 (Bounded density near zero)**.** *The random variable $Z := r(\tilde{X}) - r(\tilde{X}')$, where $\tilde{X}, \tilde{X}' \overset{i.i.d.}{\sim} q$, has a density that is bounded in a neighborhood around zero.*

Assumption 2 (L^2 estimation error)**.** *The estimated density ratio $\hat{r} : \mathcal{X} \to (0, \infty)$ satisfies*

$$\mathbb{E}_{\tilde{X} \sim q}\left[(\hat{r}(\tilde{X}) - r(\tilde{X}))^2\right] \leq \varepsilon^2,$$

for some $\varepsilon > 0$, where $r(x) = p(x)/q(x)$ denotes the true density ratio.

Theorem 1 (Robustness to estimation error)**.** *Under Assumptions 1 and 2, let $\hat{U}$ denote the conformal p-value computed using the approximate score function $\hat{r}$. Then under the alternative $H_1 : p \neq q$, there exists $M > 0$, depending on ε, p, and q, such that*

$$\mathbb{E}[\hat{U}] - \mathbb{E}[U] \leq \mathcal{O}(\varepsilon^{2/3}) \quad \textit{for all } m > M.$$

This theorem shows that the uniform test enjoys a remarkable degree of robustness: even when the classifier is weak or undertrained, the resulting p-values still exhibit systematic deviation from uniformity under the alternative. In particular, the expected p-value computed using an approximate score $\hat{r}$ remains close to the ideal value obtained from the oracle score r, with the error controlled by the quality of the approximation. As long as the estimated score preserves a reasonable approximation to the true ranking, the test maintains power. This highlights a key advantage of the conformal approach: it leverages relative score orderings rather than relying on absolute classifier accuracy; We defer the proof to Appendix A.2. We further show in Lemma 5 in the Appendix, assuming zero tie probability for continuous classifier outputs, the variance of the conformal p-values scales as $\mathcal{O}(1/m)$. Consequently, the p-values converge quickly to their true non-uniform values under the alternative as one increases m, leading to a better test power.

We also emphasize that Assumption 1 is mild and typically satisfied in practice. When p and q have overlapping support and the true density ratio $r(x)$ is smooth, the difference $Z = r(\tilde{X}) - r(\tilde{X}')$ is a continuous random variable with a density that is finite around zero. This assumption rules out degenerate cases where $r(x)$ is constant or highly discontinuous, but permits a wide range of practical scenarios in which $r(x)$ is estimated via smooth classifier outputs. The condition is notably weaker than assuming global Lipschitz continuity of r.

2.4 The multiple test

A potential limitation of the conformal test described above is the need to generate a fresh calibration set for each test point. In many settings, this is not a big issue; drawing from the true joint distribution $p(\theta, y)$ will often be cheap, as it does not require forward simulation through the generative model. Although repeated calibration does incur multiple passes through the classifier, we find that even a small, low-capacity classifier, so long as it preserves good ranking behavior, can yield competitive results. As a result, the overall computational burden often remains modest.

Nonetheless, we include in our benchmark set an alternative method proposed by (Hu & Lei, 2024), which avoids repeated calibration by using a single shared calibration set. This approach is particularly appealing in applications where sampling from p is expensive, such as when $p(y \mid \theta)$ involves a costly forward simulation, since it requires only a single calibration set and avoids repeated draws from the true joint distribution. However, as our experiments show, this convenience comes at the cost of reduced statistical power in some settings.

To correct for the dependence introduced by using a shared calibration set, (Hu & Lei, 2024) analyze the average of the conformal p-values computed using a single calibration set as a two-sample U-statistic. They derive an asymptotically valid test statistic by normalizing the deviation of the average p-value from $1/2$, which is the expected value under the null. But since we assume the same

marginal distribution of y for both joint distributions, we can simplify their test statistic to obtain:

$$\hat{T} = \frac{\frac{1}{2} - \frac{1}{n_q}\sum_{j=1}^{n_q} \hat{U}_j}{\hat{\sigma}/\sqrt{n_p}} \tag{3}$$

where $\hat{U}_j := \frac{1}{n_p}\left(\sum_{i=1}^{n_p} \mathbb{I}\left\{\hat{r}(X_i) < \hat{r}(\tilde{X}_j)\right\} + \xi_j \cdot \sum_{i=1}^{n_p} \mathbb{I}\left\{\hat{r}(X_i) = \hat{r}(\tilde{X}_j)\right\}\right)$ is obtained from the entire calibration set (common to all test points) with $\hat{r}(\cdot)$ as the scoring function, and $\hat{\sigma}$ is the asymptotic estimated standard deviation of $\frac{\sqrt{n_p}}{n_q}\sum_{j=1}^{n_q} \hat{U}_j$; see Appendix A.4 for details. Under the null, $\hat{T} \sim \mathcal{N}(0,1)$, whereas under the alternative, $\hat{T} \to \infty$ under the assumptions mentioned in (Hu & Lei, 2024, Theorem 2). Although the original goal of this test was to enhance power under distribution shift, we find that in the NPE setting—where the marginals are matched and classifiers may be weak—the simpler uniform test often yields higher power in practice.

3 Experiments

We now empirically evaluate the performance of the conformal C2ST against baseline methods in a scenario with controlled perturbations of a known reference distribution. The experiments are designed to address two key questions. The first concerns *power under proper training:* when the classifier is well trained, how does the conformal C2ST compare to competing methods? The second concerns *robustness to classifier degradation:* as we progressively degrade the quality of the classifier, does the conformal C2ST retain power better than the the ordinary C2ST?

3.1 Controlled posterior perturbations.

Our first set of experiments involve a controlled simulation environment with a known ground-truth $p(\theta \mid y)$. From this ground truth, we generate a series of flawed approximations $q(\theta \mid y)$, by systematically applying a controlled perturbation of p. The magnitude of the perturbation is controlled by a scalar γ, which acts as a "difficulty dial." When $\gamma = 0$, the approximation is perfect $q = p$, allowing us to assess a method's Type-I error rate. As γ increases, the approximation becomes progressively worse, allowing us to measure a test's power.

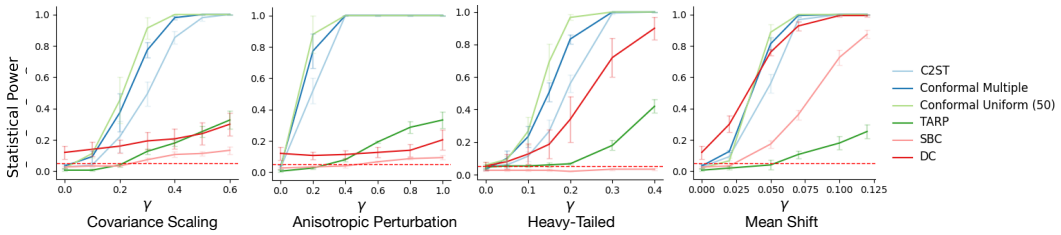
Our perturbations, described in detail in Appendix B, are designed to mimic common failure modes in NPE, such as biased means, overdispersion, miscalibrated covariance structure, or mode collapse. This framework allows us to directly assess whether a testing method can reliably detect meaningful discrepancies in a setting that mirrors real-world validation challenges. Specifically, we include four perturbation types from Chen et al. (2024) in the main text and defer the rest to Appendix B.

Testing classifier degradation. After training a classifier on a given benchmark problem at a fixed perturbation level γ, we generate a family of degraded classifiers by linearly interpolating the trained model parameters with those of a randomly initialized model:

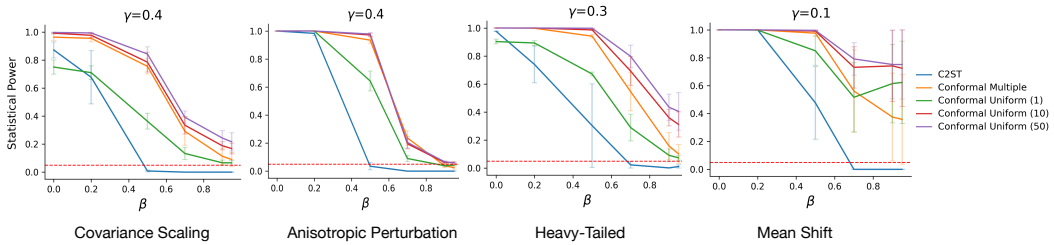
$$\psi_\beta := (1-\beta)\cdot\psi_{\hat{\eta}} + \beta \cdot \psi_{\text{rand}}, \quad \beta \in [0,1],$$

where $\psi_{\hat{\eta}}$ and ψ_{rand} are the parameter vectors of the trained and randomly initialized classifiers, respectively. This setup allows us to systematically degrade the classifier's quality by varying β, from a fully trained model ($\beta = 0$) to a random, uninformative one ($\beta = 1$). Class probabilities and nonconformity scores are calculated from the degraded classifier. We then evaluate the behavior of both the standard C2ST and the conformal C2ST across this interpolation path. This experiment provides a natural stress test for the conformal framework. If the conformal method maintains detection power even as the classifier degrades, this would offer compelling evidence of its practical robustness. We note that this experiment is specific to classifier-based testing methods; approaches like SBC and TARP, which do not involve classifiers, are not applicable in this setting.

Experiment settings and baselines. We benchmark the conformal C2ST against several well-established methods for assessing the quality of a neural posterior estimate. These include the *Classifier Two-Sample Test (C2ST)* (Lopez-Paz & Oquab, 2016), described in Section 2.1; *Simulation-Based Calibration (SBC)* (Talts et al., 2018), which computes rank statistics for each marginal of $q(\theta \mid y)$ and checks for uniformity under the true posterior; *Tests of Accuracy with Random*



(a) Power curves as a function of perturbation level γ at fixed classifier quality ($\beta = 0$).



(b) Power curves as a function of classifier degradation level β at fixed perturbation strength γ.

Figure 2: Statistical power of C2ST and conformal variants across benchmark perturbations. Panel (a) evaluates sensitivity to posterior mismatch; panel (b) evaluates robustness to classifier degradation.

Points (TARP) (Lemos et al., 2023), which uses randomly sampled reference points and distance-based statistics to construct a test that is both necessary and sufficient for posterior validity; and *Discriminative Calibration (DC)* which uses a multiclass classifier to get the strongest log-predictive-density (LPD) statistic (Yao & Domke, 2023, Algorithm 1).

We adapted each baseline to the validation task. For SBC, DC, and TARP, we drew m posterior samples from the approximate posterior for each observation y. For C2ST-based methods, we drew a single approximate posterior sample to create balanced datasets. For SBC, we performed a KS test on the rank statistics for each parameter dimension, using a Bonferroni correction for multiple testing. For TARP, we followed the recommended procedure to construct a KS test against a uniform distribution based on the TARP statistic under random reference points. For DC, we trained a multi-class classifier to compute the LPD statistic. Notably, this method is *far more* computationally expensive as the classifier's input dimension scales with m. For Conformal C2ST, we evaluated both our proposed "uniform" and "multiple" variants.

All tests used a significance level of $\alpha = 0.05$. Each experimental run used 1,000 independent samples from the true joint distribution p, and we evaluated test performance (power or Type I error) on 200 new data batched. All experiments were repeated with three random seeds, and we report the average results. Further details regarding training procedures and hyperparameters are provided in Appendix B.2.

In our experiments with the conformal-uniform test, we take advantage of the fact that samples from the true joint distribution $p(\theta, y)$ are typically cheap to generate in the NPE setting. For each test point drawn from the approximate posterior q, we sample m calibration points from p, resulting in a total of $m \cdot 1000$ labeled samples from p (with label 1), and 1000 labeled samples from q (with label 0). We refer to this configuration as **Conformal Uniform**(m), where $m \in \{1, 10, 50, 200\}$ controls the number of calibration points per test point used to calculate conformal p-values. Increasing m generally improves the power of the test, since more calibration samples yield more accurate p-values and finer resolution for detecting deviations from uniformity. For an empirical study of how test performance varies with m, see the ablation experiments in Appendix B.1.

Results. As shown in Figure 2a, our conformal methods demonstrate a clear advantage over. They consistently achieve higher power than the standard C2ST and match or outperform SBC, DC and TARP across all types of model error. Most notably, DC fails to control Type I error, rendering it an invalid test in this setting. The practical benefit of our approach is its ability to detect subtle

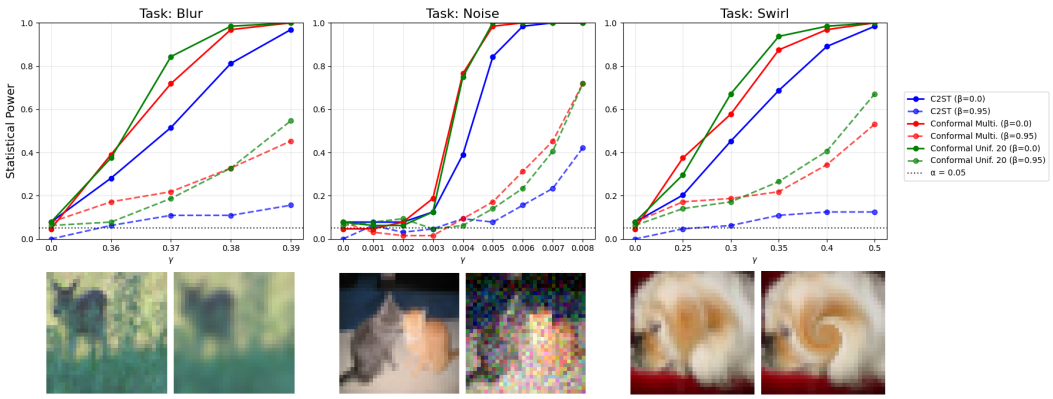


Figure 3: Conformal C2ST consistently outperforms standard C2ST across different tasks. Solid lines correspond to well-trained classifiers ($\beta = 0.0$), while dashed lines indicate weaker classifiers ($\beta = 0.95$). Representative qualitative examples for each task are shown below.

deviations from the true p. The Conformal Uniform and Conformal Multiple tests reliably identify misspecifications at low values of γ where other methods fail. While most tests can spot large errors, our methods provide a much lower detection threshold, making them more useful for identifying the small but significant imperfections common in generative models.

In our second experiment, we tested each method's robustness to a weak classifier. For a fixed model error γ, we systematically degraded the classifier's performance from well-trained ($\beta = 0$) to random ($\beta = 1$), as described previously. Figure 2b reveals that the standard C2ST is quite brittle; its power collapses as soon as the classifier's quality degrades. In contrast, our conformal methods are highly robust, maintaining high power even when the classifier is far from optimal. This resilience stems from a fundamental advantage. While the C2ST requires an accurate classifier to draw a sharp decision boundary, our conformal methods only need the classifier's scores to provide a weak but informative ranking. This makes them far more reliable in practice, where perfectly trained classifiers are rarely available.

3.2 High-Dimensional Image Example

To demonstrate the broad utility of the Conformal C2ST, we also evaluated it for on the general-purpose task of detecting distributional shift in high-dimensional images. We used the CIFAR-10 dataset (Krizhevsky et al., 2009) to create a two-sample problem: distinguishing original, clean images from versions altered by one of three corruptions: Gaussian blur, swirl distortion, or additive Gaussian noise. A parameter γ controlled the severity of the corruption. We compare the *clean* and *corrupted* distributions using C2ST, C2ST-MULTIPLE, and C2ST-UNIFORM (k=20). Power is reported (i) as a function of corruption strength γ with a well-trained classifier, and (ii) as a function of classifier quality β at fixed γ (larger β denotes a weaker classifier). As summarized in Figure 3, the Conformal C2ST achieved the highest power across all corruption types and severity levels. Crucially, it maintained its superior performance even when the underlying classifier was weak, confirming its robustness and practical value for general-purpose generative model validation.

Limitations. Our work has several limitations. First, in the NPE setting, our approach is designed to assess whether the learned posterior q is a faithful approximation of the model's true posterior p. We therefore do not address the separate, crucial problem of *statistical* model validation, where p may not accurately reflect the real-world data-generating process.

Second, our benchmarking framework benefits from cheap samples from $p(y \mid \theta)$, which may not be available in many NPE settings where the data arises from a black-box simulation model. While this enables rigorous evaluation in controlled experiments, it limits applicability to certain types of NPE problems. Finally, our theoretical analysis relies on bounding arguments that may be conservative; tighter techniques could potentially yield sharper error bounds and a more precise characterization of the test's behavior.

Ethics Statement

This work develops a new method for simulation based inference and hypothesis testing. Our framework does not involve human subjects, personal data, or harmful content, and thus poses minimal ethical risks.

Reproducibility Statement

We provide complete algorithmic and training details in the Appendix. To ensure reproducibility, we will release all code and model checkpoints upon acceptance of this manuscript.

References

Tianyu Chen, Vansh Bansal, and James Scott. NPTBench: A benchmark suite for neural posterior testing. Technical report, 2024. URL https://github.com/TianyuCodings/NPTBench.

Tianyu Chen, Vansh Bansal, and James G Scott. Conditional diffusions for amortized neural posterior estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 2377–2385. PMLR, 2025.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, pp. 11098–11116. PMLR, 2023.

Manuel Gloeckler, Michael Deistler, Christian Weilbach, Frank Wood, and Jakob H Macke. All-in-one simulation-based inference. *arXiv preprint arXiv:2404.09636*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154, 2024. doi: 10.1080/01621459.2023.2177165. URL https://doi.org/10.1080/01621459.2023.2177165.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, 3rd edition, 2005. ISBN 0-387-98864-5.

Jing Lei and Max G'Sell. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning*, pp. 19256–19273. PMLR, 2023.

Julia Linhart, Alexandre Gramfort, and Pedro L. C. Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference, 2023. URL https://arxiv.org/abs/2306.03580.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Martin Modrák, Angie H. Moon, Shinyoung Kim, Paul Bürkner, Niko Huurre, Kateřina Faltejsková, Andrew Gelman, and Aki Vehtari. Simulation-based calibration checking for bayesian computation: The choice of test quantities shapes sensitivity. *Bayesian Analysis*, 1(1), January 2023. ISSN 1936-0975. doi: 10.1214/23-ba1404. URL http://dx.doi.org/10.1214/23-BA1404.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Jonas Wildberger, Maximilian Dax, Simon Buchholz, Stephen Green, Jakob H Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference. *Advances in Neural Information Processing Systems*, 36:16837–16864, 2023.

Yuling Yao and Justin Domke. Discriminative calibration: Check bayesian computation from simulations and flexible classifier, 2023. URL https://arxiv.org/abs/2305.14593.

A THEORETICAL RESULTS

In this section, we restate our key lemmas in full and provide rigorous proofs to support our theoretical claims.

A.1 PROOFS OF LEMMA 2 AND 3

Lemma 2 (Full statement). *Let p and q be two probability densities on a measurable space $\mathcal{X}$, such that $p(x) > 0$ and $q(x) > 0$ for all $x \in \mathcal{X}$. Let $r(x) := \frac{p(x)}{q(x)}$ denote the oracle density ratio. For any measurable scoring function $s : \mathcal{X} \to \mathbb{R}$, define the area under the ROC curve (AUC) as*

$$\mathrm{AUC}(s) := \mathbb{P}[s(X) > s(\tilde{X})] + \frac{1}{2}\mathbb{P}[s(X) = s(\tilde{X})],$$

where $X \sim p$ and $\tilde{X} \sim q$ are independent.

Then $\mathrm{AUC}(s) \leq \mathrm{AUC}(r)$, with equality if and only if there exists a strictly increasing function h such that $s(x) = h(r(x))$ for $p \times q$-almost every x.

Proof. Define

$$\phi_s(x, \tilde{x}) := \mathbb{I}[s(x) > s(\tilde{x})] + \frac{1}{2}\mathbb{I}[s(x) = s(\tilde{x})],$$

so that

$$\mathrm{AUC}(s) = \iint \phi_s(x, \tilde{x})\, p(x)q(\tilde{x})\, dxd\tilde{x}.$$

Define the antisymmetric part

$$a_s(x, \tilde{x}) := \phi_s(x, \tilde{x}) - \phi_s(\tilde{x}, x) \in \{-1, 0, 1\}.$$

Note that $\phi_s(x, \tilde{x}) + \phi_s(\tilde{x}, x) = 1$, hence

$$\phi_s(x, \tilde{x}) = \frac{1}{2} + \frac{1}{2}a_s(x, \tilde{x}),$$

and therefore

$$\mathrm{AUC}(s) = \frac{1}{2} + \frac{1}{2}\iint a_s(x, \tilde{x})\, p(x)q(\tilde{x})\, dxd\tilde{x}.$$

Define the antisymmetric function

$$W(x, \tilde{x}) := p(x)q(\tilde{x}) - p(\tilde{x})q(x),$$

so that

$$\iint a_s(x, \tilde{x})\, p(x)q(\tilde{x})\, dxd\tilde{x} = \frac{1}{2}\iint a_s(x, \tilde{x})W(x, \tilde{x})\, dxd\tilde{x},$$

and thus

$$\mathrm{AUC}(s) = \frac{1}{2} + \frac{1}{4}\iint a_s(x, \tilde{x})W(x, \tilde{x})\, dxd\tilde{x}.$$

Now for each $(x, \tilde{x})$, the value $a_s(x, \tilde{x}) \in \{-1, 0, 1\}$ that maximizes the product $a_s(x, \tilde{x})W(x, \tilde{x})$ is $\mathrm{sign}(W(x, \tilde{x}))$. Therefore, the function

$$a^*(x, \tilde{x}) := \mathrm{sign}(W(x, \tilde{x}))$$

maximizes the integral.

Next, define the likelihood ratio $r(x) := \frac{p(x)}{q(x)}$. Then

$$r(x) > r(\tilde{x}) \iff \frac{p(x)}{q(x)} > \frac{p(\tilde{x})}{q(\tilde{x})} \iff p(x)q(\tilde{x}) > p(\tilde{x})q(x) \iff W(x, \tilde{x}) > 0,$$

so

$$a_r(x, \tilde{x}) := \mathrm{sign}(r(x) - r(\tilde{x})) = \mathrm{sign}(W(x, \tilde{x})) = a^*(x, \tilde{x}).$$

Thus, a_r maximizes the integral, and

$$\mathrm{AUC}(s) \leq \frac{1}{2} + \frac{1}{4} \iint |W(x, \tilde{x})| \, dx d\tilde{x} = \mathrm{AUC}(r).$$

Finally, equality occurs if and only if $a_s(x, \tilde{x}) = a_r(x, \tilde{x})$ for almost all $(x, \tilde{x})$, which implies that $s(x) > s(\tilde{x}) \iff r(x) > r(\tilde{x})$, i.e., $s = h(r)$ for some strictly increasing function h almost everywhere. □

Corollary 1 (Lower Bound on AUC via Total Variation)**.** *Under the setup of the previous lemma, let* $\mathrm{TV}(p, q) := \frac{1}{2} \int |p(x) - q(x)| dx$ *denote the total variation distance between* p *and* q*. Then*

$$\mathrm{AUC}(r) \geq \frac{1 + \mathrm{TV}(p, q)}{2}.$$

Proof. Consider the binary Bayes classifier

$$\phi^*(x) := \mathbb{I}[r(x) > 1] = \mathbb{I}\left[\frac{p(x)}{q(x)} > 1\right],$$

which is known to be the most powerful test at level $\alpha = \frac{1}{2}$. Its classification accuracy is:

$$\mathbb{P}[\phi^*(X) = 1] \cdot \frac{1}{2} + \mathbb{P}[\phi^*(\tilde{X}) = 0] \cdot \frac{1}{2} = \frac{1}{2} + \frac{1}{2}\mathrm{TV}(p, q).$$

Now note that if we treat $\phi^* \in \{0, 1\}$ as a scoring function, the AUC of this classifier is:

$$\mathrm{AUC}(\phi^*) = \mathbb{P}[\phi^*(X) > \phi^*(\tilde{X})] + \frac{1}{2}\mathbb{P}[\phi^*(X) = \phi^*(\tilde{X})] \leq \mathrm{AUC}(r),$$

since ϕ^* is a thresholding of r, and AUC is maximized by ranking with r.

But:

$$\mathrm{AUC}(\phi^*) = \frac{1}{2} + \frac{1}{2}\mathrm{TV}(p, q),$$

so we conclude:

$$\mathrm{AUC}(r) \geq \frac{1}{2} + \frac{1}{2}\mathrm{TV}(p, q) = \frac{1 + \mathrm{TV}(p, q)}{2}.$$

□

Lemma 3 (Expected conformal p-value under the alternative)**.** *Let* p *and* q *be as defined above, and let* U *denote the conformal p-value computed using the oracle density ratio* $r(x) = p(x)/q(x)$ *as the score function, as defined in (1), with* m *calibration samples drawn from* p *for each test point drawn from* q*. Then, under the alternative hypothesis* $H_1 : p \neq q$*, we have*

$$\mathbb{E}[U] = 1 - \mathrm{AUC}(r) \leq \frac{1}{2} - \frac{1}{2}\mathrm{TV}(p, q) < \frac{1}{2},$$

in the limit as $m \to \infty$*.*

Proof. We formalize our discussion in Section 2.3. First, by the strong law of large numbers, we have

$$U \xrightarrow[m\to\infty]{\text{a.s.}} \mathop{\mathbb{P}}_{X\sim p}\left[r(X) < r(\tilde{X})\right] + \xi \mathop{\mathbb{P}}_{X\sim p}\left[r(X) = r(\tilde{X})\right], \quad \text{where } \xi \sim \mathrm{Unif}\,(0, 1)\,.$$

Next, we take expectation with respect to $\tilde{X} \sim q$ and $\xi \sim \mathrm{Unif}\,(0, 1)$. However, since $U \leq 1$, Dominated Convergence Theorem gives that

$$\begin{aligned}\mathbb{E}\,[U] \xrightarrow[m\to\infty]{\text{a.s.}} & \mathbb{P}\left[r(X) < r(\tilde{X})\right] + \frac{1}{2}\,\mathbb{P}\left[r(X) = r(\tilde{X})\right] \\ =& 1 - \mathrm{AUC}(r)\end{aligned}$$

The result follows immediately from Corollary 1, since $0 < \mathrm{TV}(p, q) \leq 1$ under $H_1 : p \neq q$. □

We note that Hu & Lei (2024) present a related result from an information-theoretic perspective, showing that the variability of the density ratio $r(x)$ under q controls the deviation of conformal p-values from uniformity. In contrast, our focus is on quantifying the relationship between a classifier's discriminative ability and the statistical power of the resulting two-sample test. While the underlying intuition is similar, our formulation offers a more direct and operational perspective, grounded in the ranking statistic captured by the AUC score. For completeness, we restate their relevant lemma below.

Lemma 4 (Restated from Hu & Lei (2024)). *Under the alternative $H_1 : p \neq q$, we have*

$$\mathbb{E}[U] = \frac{1}{2} - \frac{1}{4}\,\mathbb{E}_{X,X'\sim q}\left[|r(X) - r(X')|\right] < \frac{1}{2} \quad \text{as } m \to \infty,$$

where X, X' are i.i.d. draws from q.

A.2 Proof of Theorem 1

Proof. We start by taking the expectation of our conformal p-value wrt the tie breaking uniform variable ξ:

$$2\,\mathbb{E}_\xi[\hat{U}] = \frac{1}{m+1}\left(\sum_{i=1}^{m}\mathbb{I}(\hat{r}(X_i) < \hat{r}(\tilde{X})) + \sum_{i=1}^{m}\mathbb{I}(\hat{r}(X_i) \leq \hat{r}(\tilde{X})) + 1\right).$$

By the Strong Law of Large Numbers (SLLN), as $m \to \infty$, we have:

$$2\,\mathbb{E}_\xi[\hat{U}] \to \mathbb{E}[\mathbb{I}(\hat{r}(X) < \hat{r}(\tilde{X})) \mid \tilde{X}] + \mathbb{E}[\mathbb{I}(\hat{r}(X) \leq \hat{r}(\tilde{X})) \mid \tilde{X}].$$

Taking expectation over $\tilde{X}$ and applying the Dominated Convergence Theorem (since $\hat{U} \leq 1$):

$$2\,\mathbb{E}[\hat{U}] \to \mathbb{E}[\mathbb{I}(\hat{r}(X) < \hat{r}(\tilde{X}))] + \mathbb{E}[\mathbb{I}(\hat{r}(X) \leq \hat{r}(\tilde{X}))].$$

Now define $\tilde{X}, \tilde{X}' \overset{\text{iid}}{\sim} q$. Using importance reweighting, we write:

$$2\,\mathbb{E}[\hat{U}] = \mathbb{E}\left[r(\tilde{X}') \cdot \mathbb{I}(\hat{r}(\tilde{X}') < \hat{r}(\tilde{X}))\right] + \mathbb{E}\left[r(\tilde{X}') \cdot \mathbb{I}(\hat{r}(\tilde{X}') \leq \hat{r}(\tilde{X}))\right].$$

Since $\mathbb{E}[r(\tilde{X}')] = 1$, this becomes:

$$2\,\mathbb{E}[\hat{U}] = 1 - \mathbb{E}[(r(\tilde{X}') - r(\tilde{X})) \cdot \mathbb{I}(\hat{r}(\tilde{X}') > \hat{r}(\tilde{X}))].$$

Define:

$$Z := r(\tilde{X}') - r(\tilde{X}), \quad \hat{Z} := \hat{r}(\tilde{X}') - \hat{r}(\tilde{X}), \quad \Delta := \hat{Z} - Z.$$

Then:

$$\mathbb{E}[\hat{U}] = \mathbb{E}\,[U] + \frac{1}{2}\delta, \quad \text{where} \quad \delta := \frac{1}{2}\mathbb{E}[|Z|] - \mathbb{E}[Z \cdot \mathbb{I}(\hat{Z} > 0)].$$

By symmetry of Z,

$$\delta = \frac{1}{2}\mathbb{E}[|Z|] - \mathbb{E}[Z \cdot \mathbb{I}(\hat{Z} > 0)] = \mathbb{E}[Z(\mathbb{I}(Z > 0) - \mathbb{I}(\hat{Z} > 0))].$$

Define events:

$$A := \{Z > 0, \hat{Z} \leq 0\}, \quad B := \{Z \leq 0, \hat{Z} > 0\}.$$

Then:

$$\delta = \mathbb{E}[Z \cdot \mathbb{I}_A] - \mathbb{E}[Z \cdot \mathbb{I}_B] \leq \mathbb{E}[|Z| \cdot \mathbb{I}_{|\Delta| \geq |Z|}].$$

For any threshold $t > 0$, we decompose:

$$\delta \leq \mathbb{E}[|Z| \cdot \mathbb{I}(|Z| \leq t)] + \mathbb{E}[|Z| \cdot \mathbb{I}(|\Delta| > t)].$$

The second term is bounded by Cauchy–Schwarz and Markov:

$$\mathbb{E}[|Z| \cdot \mathbb{I}(|\Delta| > t)] \leq \sqrt{\mathbb{E}[Z^2]} \cdot \sqrt{\mathbb{P}(|\Delta| > t)} \leq \sqrt{\mathbb{E}[Z^2]} \cdot \frac{2\varepsilon}{t}.$$

Assuming the density f_Z of Z is bounded near zero by C, we have:

$$\mathbb{E}[|Z| \cdot \mathbb{I}(|Z| \leq t)] \leq 2C \int_0^t z\,dz = Ct^2.$$

Combining,

$$\delta \leq Ct^2 + \frac{2\sqrt{\mathbb{E}[Z^2]}\,\varepsilon}{t}.$$

Minimizing the RHS by choosing $t = \left(\frac{2\sqrt{\mathbb{E}[Z^2]}\,\varepsilon}{C}\right)^{1/3}$, we obtain:

$$\delta = O(\varepsilon^{2/3}).$$

□

A.3 Variance scaling of conformal p-value

In this section we assume that the score function s gives continuous outputs, so that the probability of ties is zero. Note that the assumption is ubiquitously satisfied in practical settings when classifiers are trained using neural networks.

Lemma 5 (Variance of Conformal p-value)**.** *Let $\hat{U}$ be the conformal p-value for a test point $\tilde{X} \sim q$. The variance of $\hat{U}$, conditioned on the test point, is:*

$$\mathit{Var}(\hat{U}|\tilde{X}) = \frac{mU(1-U)}{(m+1)^2} = \mathcal{O}(1/m)$$

where $U = P(s(X) \leq s(\tilde{X})|X \sim p)$ is the true, population-level p-value (assuming no ties).

Proof. By definition, the conformal p-value (assuming no ties for simplicity) is given by:

$$\hat{U} = \frac{1}{m+1}\left(1 + \sum_{i=1}^{m} \mathbb{I}\{s(X_i) \leq s(\tilde{X})\}\right)$$

where $\{X_i\}_{i=1}^m$ are i.i.d. calibration samples from the distribution p.

When we condition on the test point $\tilde{X}$, the score $s(\tilde{X})$ becomes a fixed value. Let's define a set of indicator random variables B_i for $i = 1, \ldots, m$:

$$B_i = \mathbb{I}\{s(X_i) \leq s(\tilde{X})\}$$

Since the calibration samples X_i are drawn i.i.d. from p, the variables B_i are i.i.d. Bernoulli random variables. The probability of success for each B_i is:

$$P(B_i = 1) = P(s(X_i) \leq s(\tilde{X})) = U$$

where U is the population-level p-value as defined in the lemma statement. Thus, $B_i \sim \text{Bernoulli}(U)$.

We can now express $\hat{U}$ in terms of these Bernoulli variables:

$$\hat{U} = \frac{1}{m+1}\left(1 + \sum_{i=1}^{m} B_i\right)$$

The variance of $\hat{U}$ conditioned on $\tilde{X}$ is:

$$\begin{aligned}
\text{Var}(\hat{U}|\tilde{X}) &= \text{Var}\left(\frac{1}{m+1}\left(1 + \sum_{i=1}^{m} B_i\right)\right) \\
&= \left(\frac{1}{m+1}\right)^2 \text{Var}\left(1 + \sum_{i=1}^{m} B_i\right) \\
&= \frac{1}{(m+1)^2}\text{Var}\left(\sum_{i=1}^{m} B_i\right)
\end{aligned}$$

Since the B_i are i.i.d., the variance of their sum is the sum of their variances:

$$\text{Var}\left(\sum_{i=1}^{m} B_i\right) = \sum_{i=1}^{m} \text{Var}(B_i)$$

The variance of a Bernoulli(U) random variable is $U(1-U)$. Therefore:

$$\sum_{i=1}^{m} \text{Var}(B_i) = \sum_{i=1}^{m} U(1-U) = mU(1-U)$$

Substituting this back, we get the final result:

$$\text{Var}(\hat{U}|\tilde{X}) = \frac{mU(1-U)}{(m+1)^2}$$

As the number of calibration samples $m \to \infty$, the variance behaves as:

$$\frac{mU(1-U)}{(m+1)^2} \approx \frac{mU(1-U)}{m^2} = \frac{U(1-U)}{m} = \mathcal{O}(1/m)$$

This completes the proof. □

Hence, the resulting $\mathcal{O}(1/m)$ scaling of the variance of conformal p-value ensures that the test power quickly rises as the p-values converge quickly to their true non-uniform values under the alternative when one increases the number of calibration points m.

A.4 Multiple test details

In this section, we summarize the multiple conformal testing procedure proposed by Hu & Lei (2024), which accounts for the dependence among p-values arising from the use of a shared calibration set. Under the NPE setting, where the marginal distribution of y is assumed to be the same under both the true and approximate joint distributions, we derive a simplified form of their test statistic:

$$\hat{T} = \frac{\frac{1}{2} - \frac{1}{n_q}\sum_{j=1}^{n_q} \hat{U}_j}{\hat{\sigma}/\sqrt{n_p}}$$

where $\hat{U}_j := \frac{1}{n_p}\left(\sum_{i=1}^{n_p} \mathbb{I}\left\{\hat{r}(X_i) < \hat{r}(\tilde{X}_j)\right\} + \xi_j \cdot \sum_{i=1}^{n_p} \mathbb{I}\left\{\hat{r}(X_i) = \hat{r}(\tilde{X}_j)\right\}\right)$ is obtained from the entire calibration set (common to all test points) with $\hat{r}(\cdot)$ as the scoring function, and $\hat{\sigma}$ is the asymptotic estimated standard deviation of $\frac{\sqrt{n_p}}{n_q}\sum_{j=1}^{n_q} \hat{U}_j$. We also adapt their expression for the variance estimate to the same-marginal setting. Let $\hat{F}$ be the empirical CDF of $\left\{\hat{r}(\tilde{X}_j) : j \in [n_q]\right\}$ and $\hat{F}_-$ be its left limit. Then the variance estimate is given by:

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 + \frac{n_p}{12 \cdot n_q}$$

where $\hat{\sigma}_1^2$ is the empirical variance of $\left\{\hat{F}_{1/2}(\hat{r}(X_i)) : i \in [n_p]\right\}$ and $\hat{F}_{1/2}(t) = \left(\hat{F}(t) + \hat{F}_-(t)\right)/2$.

B Experiment Details

B.1 Benchmarking with perturbed Gaussians

Below, we include perturbations from Chen et al. (2024) which we use as our benchmarking suite. Section 3.1 of the main text shows results for the first four items in the following list.

- **Mean Shift.** To simulate systematic location bias, we perturb the posterior mean while keeping the covariance fixed: $q(\theta \mid y) = \mathcal{N}((1+\gamma)\mu_y, \Sigma_y)$. This mimics scenarios where the NPE model consistently misses the location of the true posterior mode.
- **Covariance Scaling.** To model over- or under-confidence in uncertainty quantification, we uniformly inflate (or deflate) the covariance matrix: $q(\theta \mid y) = \mathcal{N}(\mu_y, (1+\gamma)\Sigma_y)$. This captures calibration failures where the posterior has the correct shape and center but misrepresents its overall spread.

- **Anisotropic Covariance Perturbation.** We introduce structured distortion in the posterior shape by injecting uncertainty along the direction of least variance: $q(\theta \mid y) = \mathcal{N}(\mu_y, \Sigma_y + \gamma\Delta)$, where $\Delta = \mathbf{v}_{\min}\mathbf{v}_{\min}^\top$ and $\mathbf{v}_{\min}$ is the eigenvector of Σ_y corresponding to its smallest eigenvalue. This subtly alters the posterior geometry.
- **Heavy-Tailed Perturbation.** To explore deviations in tail behavior, we replace the Gaussian with a multivariate t-distribution: $q(\theta \mid y) = t_\nu(\mu_y, \Sigma_y)$ where $\nu = 1/(\gamma + \epsilon)$. As $\gamma \to 0$, the distribution approaches Gaussian; increasing γ yields heavier tails, modeling posterior approximations that spuriously introduce more extreme values.
- **Additional Mode.** We introduce a symmetric mode in the approximate posterior with weight γ, so that $q(\theta \mid y) = \gamma\mathcal{N}(-\mu_y, \Sigma_y) + (1-\gamma)\mathcal{N}(\mu_y, \Sigma_y)$ while $p(\theta \mid y) = \mathcal{N}(\mu_y, \Sigma_y)$. As $\gamma \to 0$, the two match, while increasing γ increases the mass of the spurious mode.
- **Mode Collapse.** We introduce a symmetric mode in the true posterior with weight γ, so that $p(\theta \mid y) = \gamma\mathcal{N}(-\mu_y, \Sigma_y) + (1-\gamma)\mathcal{N}(\mu_y, \Sigma_y)$ while $q(\theta \mid y) = \mathcal{N}(\mu_y, \Sigma_y)$. As $\gamma \to 0$, the two match, while increasing γ increases the mass of the missed mode.

Figures 6 to 11 present our experimental results, comparing our proposed methods against the baselines and including ablations on the number of calibration samples used in the conformal uniform test. Across all perturbation types, we observe consistent improvements over the classical C2ST under classifier degradation, highlighting the robustness of our method.

B.2 Training Details

We use a three-layer neural network classifier with skip connections and a hidden dimension of 256. The network is initialized using PyTorch's default parameter initialization. Optimization is performed using the Adam optimizer with a cosine annealing learning rate schedule.

The training set consists of 2,000 samples: 1,000 labeled examples of the form $\{y_i, \theta_i, 1\}$, drawn from the joint distribution $p(\theta \mid y)p(y)$, and 1,000 negative examples from $q(\theta \mid y)p(y)$. Each pair (θ, y) lies in $\mathbb{R}^{3\times 3}$. We train the model for 2,000 epochs using a fixed learning rate of 1×10^{-5}, which is sufficient to ensure convergence in our experiments.

For training the DC classifier, we set $m = 10$, consistent with the configuration in its `official repository`. Increasing m significantly prolongs training and makes the classifier harder to optimize effectively; we found $m = 10$ to be a practical sweet spot. The DC classifier is trained for 2000 epochs with a fixed learning rate of 1×10^{-5}

B.3 High-Dimensional Image Experiments

We investigate the behavior of conformal two-sample testing (C2ST) methods under controlled image corruptions applied to CIFAR-10 data. The experimental pipeline consists of: (i) sampling "true" (uncorrupted) data from the empirical distribution, (ii) generating "fake" data via structured corruption operators parameterized by a strength parameter γ, (iii) training a discriminative classifier between the two sources under a conditional or unconditional formulation, (iv) constructing a family of interpolated "weak" classifiers via a parameter β for robustness analysis, and (v) evaluating multiple conformal calibration strategies and a baseline C2ST p-value.

Data Sampling and Corruption. Let (θ, y) denote (image, class label) pairs from the empirical dataset D (CIFAR-10). In the conditional setting, y is a class label and θ the corresponding image; in the unconditional setting only θ is used. We define:

$$(\theta, y) \sim p, \qquad (\tilde{\theta}_\gamma, y) \sim q_\gamma,$$

where the corrupted image $\tilde{\theta}_\gamma$ is obtain as $\tilde{\theta}_\gamma = \mathcal{C}_\gamma(\theta)$ where $\theta \sim p(\cdot \mid y)$ and $\mathcal{C}_\gamma$ is a corruption operator with strength $\gamma \geq 0$. The following corruption families are considered:

1. **Gaussian Blur** ("blur"): per-channel convolution with a Gaussian kernel of standard deviation $\sigma = \gamma$ (implemented via `scipy.ndimage.gaussian_filter`).
2. **Swirl Transformation** ("swirl"): a geometric warp (`skimage.transform.swirl`) with angular distortion parameter (strength) set to γ and radius proportional to the image spatial extent.

3. **Additive Gaussian Noise** ("noise"): $\tilde{\theta}_\gamma = \theta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \gamma^2 I)$, followed by clipping to $[-3, 3]$ in normalized pixel space.

When $\gamma = 0$ the corruption reduces to the identity map. For each experiment we sample $N_{\text{true}} = 1024$ uncorrupted and $N_{\text{fake}} = 1024$ corrupted images to form the training pool. Additional independently sampled batches are used for evaluation replicates.

Preprocessing. CIFAR-10 images are normalized channelwise to zero mean and unit (scaled) variance via the standard transform with mean $(0.5, 0.5, 0.5)$ and std $(0.5, 0.5, 0.5)$. All downstream embedding extraction resizes inputs to 299×299 (bilinear) and converts grayscale to 3-channels when necessary.

Network Architecture. We employ a frozen Inception V3 backbone (pretrained on ImageNet) as a feature extractor producing a 2048-dimensional embedding $f(y) \in \mathbb{R}^{2048}$. In the conditional setting, a learnable label embedding $e(\theta) \in \mathbb{R}^{64}$ is concatenated to yield $[f(y); e(\theta)] \in \mathbb{R}^{2112}$. A feed-forward discriminator g consists of:

$$\text{Linear}(d_{\text{in}}, H) \to \text{ReLU} \to \text{Dropout}(0.5) \to \text{Linear}(H, H/2) \to \text{ReLU} \to \text{Dropout}(0.5) \to \text{Linear}(H/2, 1),$$

with $H = 256$ (so hidden layers of sizes 256 and 128). The model outputs a logit $\ell = g(\cdot)$, trained with binary cross-entropy against labels 1 (true) and 0 (fake). Only the classifier head and (if used) label embeddings are trainable; the Inception backbone remains frozen.

Weak Classifier Interpolation. To probe robustness and create a spectrum of discriminator strengths, we store (i) the randomly initialized classifier parameters ψ_{rand} and (ii) the fully trained parameters ψ_{trained}. For any $\beta \in [0, 1]$, we define an interpolated ("weak") classifier:

$$\psi_\beta \;=\; (1 - \beta)\, \psi_{\text{trained}} \;+\; \beta\, \psi_{\text{rand}}.$$

Thus $\beta = 0$ recovers the trained discriminator and $\beta \to 1$ approaches a near-random classifier. We evaluate each β independently in the downstream statistical tests.

Training Procedure. Training is conducted for 200 epochs with Adam (learning rate 10^{-4}) and cosine annealing LR scheduling. We employ Distributed Data Parallel (DDP) across all available GPUs. Instead of relying on a standard `DataLoader` with samplers, we materialize the full training set in CPU memory, deterministically shuffle each epoch (synchronized seeds across ranks), and partition indices among GPUs. Batch size per GPU is 256. This design avoids pinned-memory bottlenecks and enables explicit control of memory usage (with periodic cache clearing).

Evaluation and Test Statistics. Let the (potentially interpolated) classifier yield logit $\ell(y)$ (conditional case notationally suppressed). We estimate:

- **Baseline C2ST p-value**: using the held-out evaluation samples, comparing score distributions between true and fake.
- **Conformal Multiple**: a conformal calibration procedure applied to embeddings (concatenated with label embeddings if conditional), producing $p_{\text{conf, mult}}$.
- **Conformal Uniform Tests**: scalability check by enlarging the reference (true) sample size by multiplicative factors $m \in \mathcal{M}$ (e.g., $\{2, 5, 20\}$), yielding $p^{(m)}_{\text{conf, uni}}$.

For each (γ, β) configuration we perform n_{eval} independent replicates (distinct random seeds), each resampling true/fake evaluation sets (default 64 runs unless otherwise specified). Success rates are reported as the empirical frequency of $p < \alpha$ with $\alpha = 0.05$ for each test variant.

Hyperparameter Sweeps. We explore γ ranges tailored to the perceptual sensitivity of each corruption type. Table 1 summarizes the grids and the shared β values.

The narrow interval for blur emphasizes the phase transition region of detectability, while noise employs a single small variance ensuring subtle corruption. Swirl spans a broader geometric distortion spectrum.

Table 1: Corruption strength grids Γ per task and interpolation coefficients $\mathcal{B}$.

Task	γ (tested values)
Blur	$\{0.00, 0.36, 0.37, 0.38, 0.39\}$
Swirl	$\{0.00, 0.25, 0.30, 0.35, 0.40, 0.50\}$
Noise	$\{0.0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008\}$

$\mathcal{B} = \{0.0, 0.95\}$ (strong vs. weak discriminator regimes).

B.4 Evaluation Details

For evaluation, using the trained classifier, we sample $1000 \cdot m$ data points $\{y_i, \theta_i\}$ from $p(\theta \mid y)p(y)$, and another 1,000 samples from $q(\theta \mid y)p(y)$ to compute the rejection rate for C2ST and its conformal variants.

For classical C2ST and the Conformal Multiple testing variant, we set $m = 1$. For the Conformal Uniform test, we vary $m \in \{1, 2, 5, 10, 20, 50, 200\}$. We denote this setting as Conformal Uniform(m), where m specifies the evaluation sample budget.

For SBC and TARP, which require multiple posterior samples per observation y, we draw 200 posterior samples θ for each y. The overall evaluation budget remains consistent, using 1,000 pairs (y, θ) from $p(\theta \mid y)p(y)$. We denote these methods as SBC(200) and TARP(200) to indicate the number of posterior samples used. For DC, we also fix $m = 10$ and bootstrap 100 times to make sure it has sufficiently good power.

To ensure a fair comparison, we report results for C2ST, Conformal Multiple testing, Conformal Uniform(200), SBC(200), and TARP(200) when evaluating the statistical power of our method.

C Experiments on high dimensional posteriors with manifold structure

We further evaluate robustness and calibration of the proposed conformal tests in settings where the true posterior lies on a low-dimensional manifold embedded in a high-dimensional space. Specifically, we construct synthetic posteriors using a conditional normalizing flow model trained on data generated from a nonlinear spherical manifold, enabling controlled perturbations and precise comparisons across methods.

Problem setup. We first sample n points uniformly on the surface of the unit sphere in $\mathbb{R}^3$ using spherical coordinates $\phi \sim \mathcal{U}[0, \pi]$, $\tau \sim \mathcal{U}[0, 2\pi]$, and convert them to Cartesian coordinates:

$$a = \sin(\phi)\cos(\tau), \quad b = \sin(\phi)\sin(\tau), \quad c = \cos(\phi).$$

These coordinates are mapped into a higher-dimensional ambient space $\mathbb{R}^d$ (with $d = 100$) via a random projection matrix $W \in \mathbb{R}^{3\times d}$ sampled from a standard Gaussian, yielding

$$\theta = W^\top (a, b, c)^\top + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $\theta \in \mathbb{R}^d$ and $\sigma > 0$ controls the noise level. The corresponding conditioning variable is defined as $y = (\phi/\pi, \tau/2\pi)$, so that $y \in [0, 1]^2$.

Training details. We simulate posterior distributions using a conditional normalizing flow (CNF) model based on RealNVP Dinh et al. (2016). Our architecture consists of 8 RealNVP layers, each parameterized by a neural network with one hidden layer of width 512. The model is trained for 500 epochs to ensure convergence. The true posterior is obtained by applying the inverse flow to a base sample $z \sim \mathcal{N}(\gamma, I)$, where $\gamma = 0$.

To simulate an approximate posterior, we perturb the base distribution by introducing a mean shift $\gamma \in \mathbb{R}^d$, such that $z \sim \mathcal{N}(\gamma, I)$ instead of the true $z \sim \mathcal{N}(0, I)$. The resulting approximate posterior $q(\theta \mid y)$ is thus generated by applying the same flow model to this mis-specified base distribution.

Figure 4 illustrates samples from the true and approximate posteriors for various values of γ, projected onto the first two principal components using Principal Component Analysis (PCA).

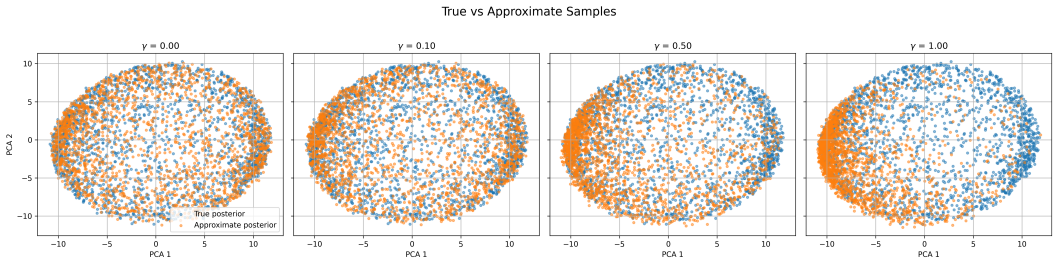
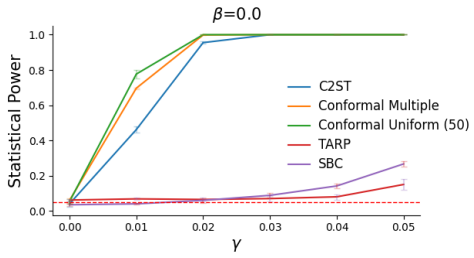
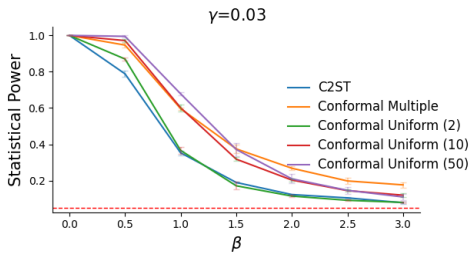


Figure 4: Approximate posterior samples (as a function of the base mean perturbation γ) projected on the first two principal axes

To distinguish between the true and approximate posteriors, we compute log density ratios by subtracting the log-probabilities assigned by the two CNF models. To emulate degradation in the classifier's discriminative ability, we add Gaussian noise $\delta \sim \mathcal{N}(0, \beta^2)$ to the log-density ratios. When $\beta = 0$, the scores are exact; as β increases, the scores become progressively noisier, reflecting reduced discriminative power. Figure 5 shows that conformal variants of C2ST outperform all baselines under increasing posterior perturbation and retain a better test power under classifier degradation. Note that even though DC achieves good power in some settings, it failed to control the Type-I error rate, which is always the first priority when we conduct hypothesis testing.

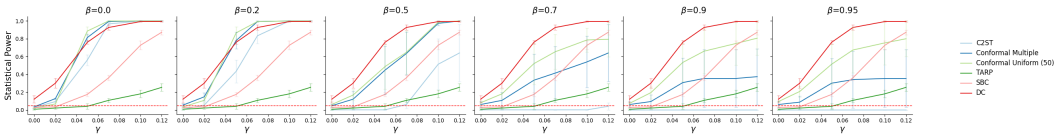


(a) Statistical power as a function of base mean perturbation γ

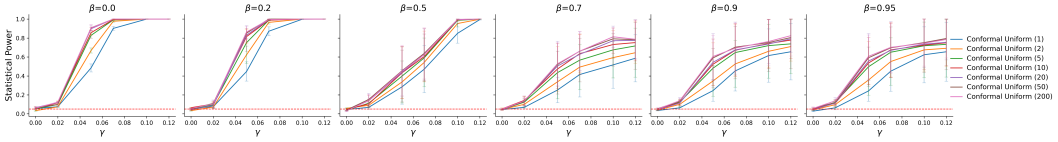


(b) Statistical power as a function of classifier noise β

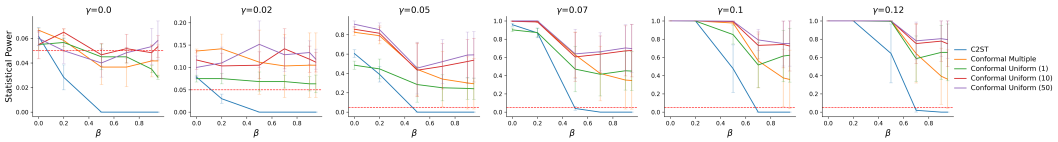
Figure 5: Power analysis under mean perturbation of the base distribution of the normalizing flow model.



(a) Statistical power as a function of γ

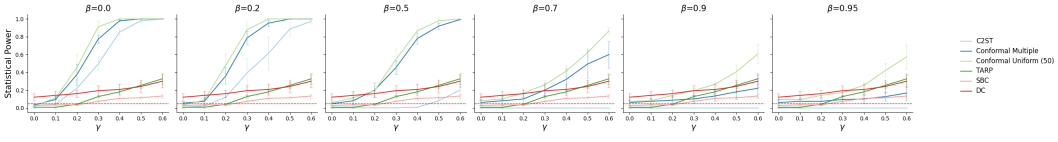


(b) Statistical power under Conformal Uniform with varying m

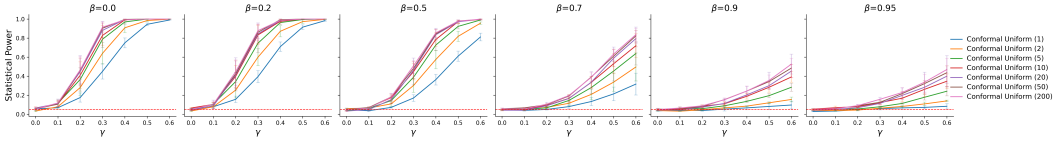


(c) Statistical Power degradation under classifier degeneration β

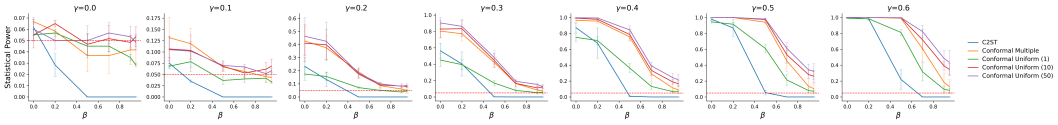
Figure 6: Power analysis under Mean Shift



(a) Statistical power as a function of γ

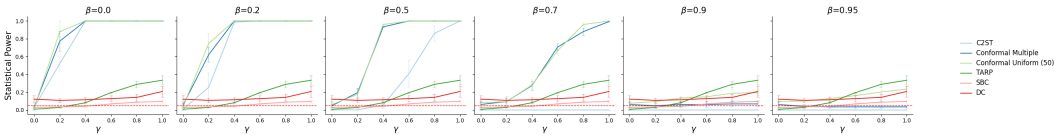


(b) Statistical power under Conformal Uniform with varying m

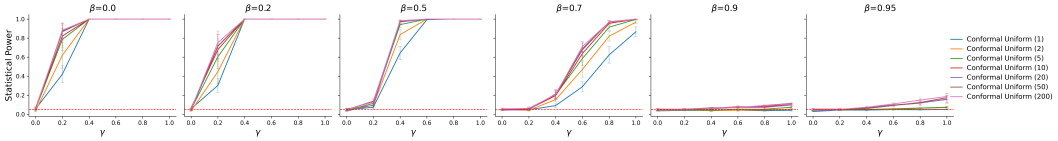


(c) Statistical Power degradation under classifier degeneration

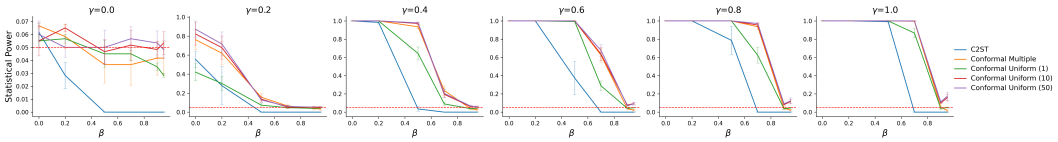
Figure 7: Power analysis under Covariance Scaling.



(a) Statistical power as a function of γ

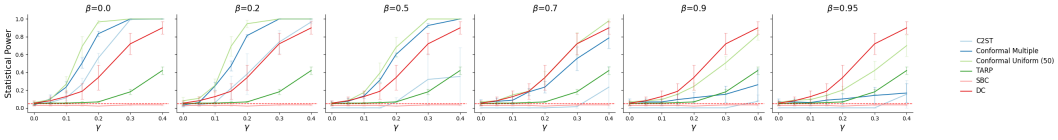


(b) Statistical power under Conformal Uniform with varying m.

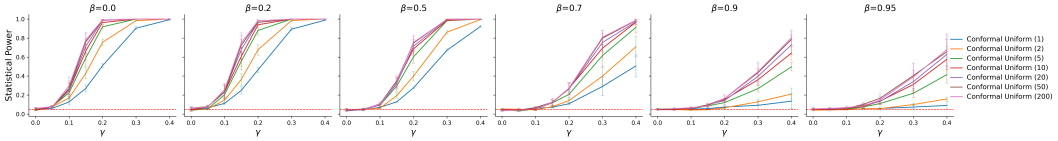


(c) Statistical Power degradation under classifier degeneration

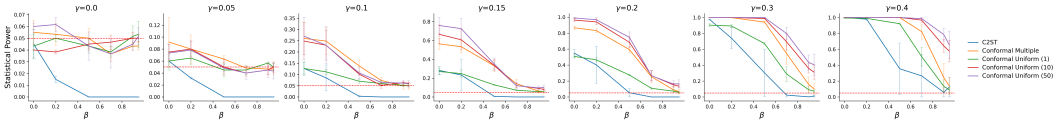
Figure 8: Power analysis under Anisotropic Covariance Perturbation.



(a) Statistical power as a function of γ

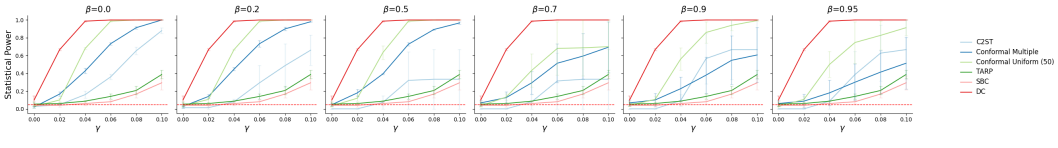


(b) Statistical power under Conformal Uniform with varying m

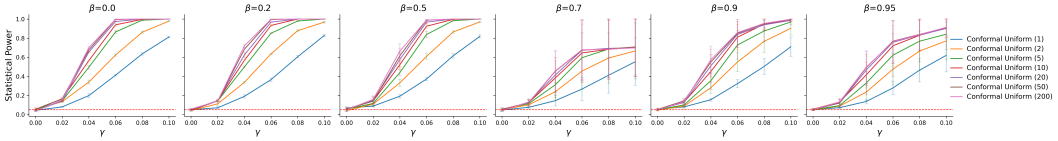


(c) Statistical Power degradation under classifier degeneration

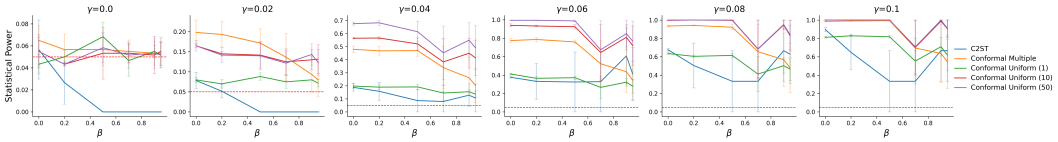
Figure 9: Power analysis under Heavy-Tailed Perturbation.



(a) Statistical power as a function of γ

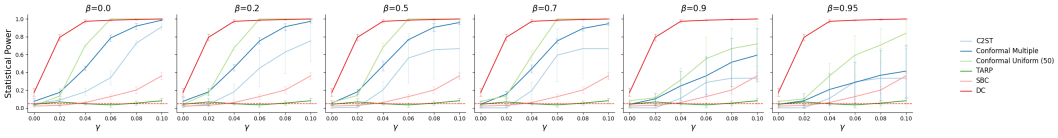


(b) Statistical power under Conformal Uniform with varying m

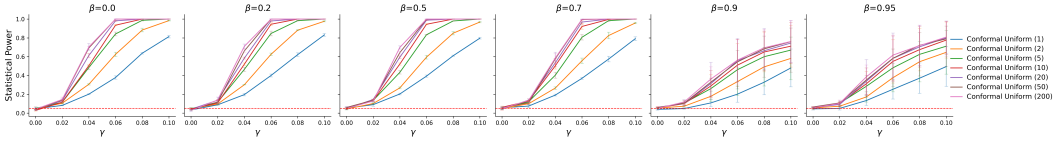


(c) Statistical Power degradation under classifier degeneration

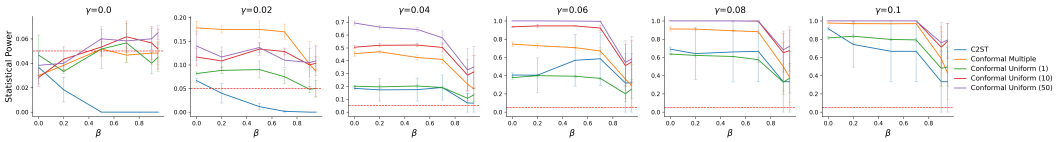
Figure 10: Power analysis under Additional Mode.



(a) Statistical power as a function of γ



(b) Statistical power under Conformal Uniform with varying k



(c) Statistical Power degradation under classifier degeneration

Figure 11: Power analysis under Mode Collapse.