ZERO-FLOW TWO-SAMPLE TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the success of modern flow-based generative models in modeling complex data, we study two-sample testing through the lens of flow-based methods. We propose the Zero-Flow Two-Sample Test (ZF2ST), built on the zero-flow criterion, which characterizes distributional equality through a time-reversal antisymmetry of a learnable velocity field. We extend this criterion and further develop the Zero-Flow Discrepancy, an identifying discrepancy that controls the Wasserstein distance, and derive a variational representation in terms of a witness function. This representation naturally leads to a witness-based test whose power is governed by the signal-to-noise ratio (SNR), allowing direct power maximization for witness learning. ZF2ST learns the witness on one data split and performs testing on held-out samples, thereby maintaining Type-I error control and admitting a simple asymptotic null distribution. Experimentally, ZF2ST performs competitively across various synthetic and real-world benchmarks, while showing particularly strong performance in distinguishing image distributions from different sources.

1 INTRODUCTION

Two-sample testing is a fundamental statistical problem for determining whether two sets of observations arise from the same distribution (Smirnov et al., 1939; Friedman & Rafsky, 1979; Fasano & Franceschini, 1987). A broad line of work has extended it beyond simple, low-dimensional settings to high-dimensional and structured data (Borgwardt et al., 2006; Sugiyama et al., 2011; Székely & Rizzo, 2013; Chwialkowski et al., 2015; Chen & Friedman, 2017; Ghoshdastidar & Von Luxburg, 2018). These advances have made two-sample testing a versatile tool in modern machine learning, with applications including dataset comparison, generative model evaluation, distribution shift detection, and privacy auditing (Gretton et al., 2012a; Lopez-Paz & Oquab, 2016; Sutherland et al., 2016; Rabanser et al., 2019; Ribero et al., 2026).

A dominant framework for nonparametric two-sample testing is the maximum mean discrepancy (MMD; Gretton et al., 2012a), which compares kernel mean embeddings of the two distributions (Muandet et al., 2017). However, its power depends critically on the kernel, with generic choices often degrading rapidly in high dimensions (Ramdas et al., 2015). Early work addressed this by optimizing kernels and test features for power (Gretton et al., 2012b; Jitkrittum et al., 2016; Sutherland et al., 2016). A major advance along this direction was the deep-kernel test of Liu et al. (2020), which parameterizes the kernel with a neural network and trains it to directly maximize testing power, enabling kernel tests to detect complex, high-dimensional alternatives such as differences between natural image distributions. Subsequent work has learned witness functions explicitly (Kübler et al., 2022a;b) and combined multiple kernels to avoid committing to a single choice (Schrab et al., 2023; Biggs et al., 2023; Zhou et al., 2025). These developments highlight that a central challenge in modern two-sample testing is to learn the structure of distributional differences, which may be subtle, embedded in high-dimensional data, and difficult to characterize a priori.

Recent advances in generative modeling offer a promising route to this challenge. Flow-based generative models have achieved remarkable success in capturing data distributions that are multimodal, highly structured, and rich in fine-grained detail (Lipman et al., 2023; Liu et al., 2022; Albergo et al., 2025). Beyond sample generation, they have also proven effective for statistical tasks such as missing-data imputation, mutual-information estimation, and simulation-based inference (Yu et al., 2025; Butakov et al., 2025; Wildberger et al., 2023). Motivated by these advances, we exploit the modeling capabilities of flow models to discern structured differences between distributions for two-sample testing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our approach builds on the zero-flow criterion (Wang et al., 2026), which characterizes distributional equality through the flow-matching velocity field. The original criterion requires non-vanishing characteristic functions **(A0)** and was developed primarily for representation learning, rather than as a statistical discrepancy or two-sample test. Recently, Shlenskii et al. (2026) study a corresponding divergence for generative modeling under bounded-support assumptions. In this work, we extend the zero-flow criterion to all $p, q \in \mathcal{P}_1(\mathbb{R}^d)$ [1] and introduce the Zero-Flow Discrepancy (ZFD), whose validity is established through Wasserstein control. We then derive a variational form of ZFD, yielding a simple two-sample test with asymptotic normality and an explicit power criterion. We call the resulting procedure the Zero-Flow Two-Sample Test (ZF2ST). It supports either standard flow-matching regression or direct optimization of testing power. We further introduce variance reduction to improve test power and a wild-bootstrap calibration with finite-sample validity. Finally, ZF2ST performs strongly across a range of hypothesis testing problems, with particular strength on high-dimensional image data comparisons. We summarize our contributions as follows.

- We introduce the *Zero-Flow Discrepancy* (ZFD), an identifying discrepancy with Wasserstein control that can be readily estimated through standard flow-matching regression.
- We propose ZF2ST, a ZFD-based test with asymptotic normality and an explicit power criterion, offering two training objectives based on either flow-matching regression or direct power maximization.
- We demonstrate that ZF2ST effectively detects subtle and complex distributional alternatives across synthetic and real-world hypothesis testing problems, with particular strength on challenging image data comparisons.

2 BACKGROUND

Two-sample testing. The goal of two-sample testing is to determine whether the two samples are drawn from the same distribution. Let p and q be Borel probability measures on $\mathbb{R}^d$. Given two sets of samples $\mathcal{S}_p = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{S}_q = \{\mathbf{y}_j\}_{j=1}^m$ drawn i.i.d. from p and q respectively, we test the null hypothesis $H_0 : p = q$ against the alternative $H_1 : p \neq q$. Throughout the paper we will assume $n = m$. A test is established on a test statistic $\zeta = \zeta(\mathcal{S}_p, \mathcal{S}_q)$ typically based on statistical discrepancies, where larger values indicate stronger evidence against H_0. Let ζ_0 denote the test statistic under H_0, whose distribution can be calibrated by bootstrap, permutation, or asymptotic theory (Efron & Tibshirani, 1985; Good, 2013; Gretton et al., 2012a). The right-tail p-value is $p_{\text{val}} = \Pr_{H_0}(\zeta_0 \geq \zeta)$ and H_0 is rejected at significance level $\alpha \in (0, 1)$ whenever $p_{\text{val}} \leq \alpha$. A desirable test controls the resulting Type-I error at level α while achieving high power, i.e., a high probability of rejecting H_0 under H_1.

Flow matching. Flow Matching (FM; Lipman et al., 2023; Liu et al., 2022) learns a vector field that transports samples from a prior distribution p to a target distribution q. Let $X \sim p$ and $Y \sim q$ be independently drawn. We define the linear interpolation

$$Z_t = (1 - t)X + tY, \ t \in [0, 1], \tag{1}$$

with $Z_0 = X$ and $Z_1 = Y$. This interpolation induces a path of marginal distributions $p_t := \text{Law}(Z_t)$ connecting $p = p_0$ and $q = p_1$. When the second moments are finite, FM learns the associated time-dependent velocity field by minimizing the regression loss $\mathcal{R}(\mathbf{u}) := \mathbb{E}_{t\sim\text{Unif}[0,1],X,Y}\|Y - X - \mathbf{u}(Z_t, t)\|^2$, whose optimal solution is the conditional expectation

$$\mathbf{v}_t(\mathbf{z}) = \mathbb{E}[Y - X | Z_t = \mathbf{z}]. \tag{2}$$

Under standard regularity conditions (Lipman et al., 2023), the same marginals p_t are realized by the ODE $\mathrm{d}Z'_t = \mathbf{v}_t(Z'_t)\,\mathrm{d}t$, initialized from $Z'_0 \sim p$. Numerically integrating this ODE therefore transports samples from p to q. Thus, FM converts generative modeling into the problem of learning the velocity field.

Zero-flow criterion. Recently, a key symmetry of the flow-matching velocity field was studied by Wang et al. (2026) and Shlenskii et al. (2026).

Proposition 2.1 (Zero-flow criterion; Wang et al. (2026)). *Suppose $p, q \in \mathcal{P}_1(\mathbb{R}^d)$ and satisfy*

[1] $\mathcal{P}_1(\mathbb{R}^d) := \left\{p \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|\,\mathrm{d}p(x) < \infty\right\}$, $\mathcal{P}(\mathbb{R}^d)$ is the set of all probability measures on $\mathbb{R}^d$.

(A0) Non-vanishing characteristic functions: $\varphi_p(\omega) \neq 0, \varphi_q(\omega) \neq 0$. [2]

Let $X \sim p$ and $Y \sim q$ be drawn independently. Then $\mathbf{v}_{1/2}(\mathbf{z}) = \mathbf{0}, \forall \mathbf{z}$ if and only if $p = q$.

Since $\mathbf{v}_{1/2}(\mathbf{z}) = \mathbb{E}[Y - X|(X + Y)/2 = \mathbf{z}]$, the exchangeability of X and Y under $p = q$ immediately implies $\mathbf{v}_{1/2}(\mathbf{z}) = \mathbf{0}$, while the converse holds under **(A0)**. Wang et al. (2026) further developed the criterion for conditional independence and representation learning, whereas Shlenskii et al. (2026) used it for generative modeling.

3 From Zero-Flow to an Identifying Discrepancy

A basic requirement for a general two-sample discrepancy is that it separates $p = q$ from distinct alternatives, as achieved by MMD with characteristic kernels (Sriperumbudur et al., 2011). The zero-flow criterion provides a natural starting point: under the null, the midpoint velocity satisfies $\mathbf{v}_{1/2}(\mathbf{z}) = \mathbf{0}$. However, its converse identification relies on **(A0)**. We therefore seek a quantity that retains the *full-time* structure of zero flow while identifying $p = q$ over a broader distribution class.

3.1 Extending the Zero-Flow Criterion

Indeed, we can extend zero-flow using the notion of probability flux, which combines velocity with probability mass. Wang et al. (2026) point to the fact that when $p = q$ the population velocity exhibits the time-reversal antisymmetry $\mathbf{v}_t(\mathbf{z}) = \mathbb{E}[Y - X|Z_t = \mathbf{z}] = \mathbb{E}[X - Y|Z_{1-t} = \mathbf{z}] = -\mathbf{v}_{1-t}(\mathbf{z})$. This follows from the exchangeability of X and Y: swapping the two endpoints reverses the displacement while mapping the interpolation state at time t to that at time $1 - t$,

$$Z_t = (1 - t)X + tY \quad \longleftrightarrow \quad Z_{1-t} = (1 - t)Y + tX. \tag{3}$$

The same exchangeability further implies $Z_t \overset{d}{=} Z_{1-t}$ for all $t \in [0, 1]$, and hence the marginal symmetry $p_t = p_{1-t}$. The continuity equation (Ambrosio et al., 2005),

$$\partial_t p_t + \nabla \cdot \big(p_t \mathbf{v}_t\big) = 0. \tag{4}$$

couples the marginal and velocity through the *probability flux* $\mathbf{j}_t := p_t \mathbf{v}_t$. [3] Under $p = q$, the marginal symmetry and velocity antisymmetry therefore combine into the flux cancellation $\mathbf{j}_t + \mathbf{j}_{1-t} = 0$.

Theorem 3.1 (Extended Zero-Flow criterion)**.** *Suppose $p, q \in \mathcal{P}_1(\mathbb{R}^d)$, and let $X \sim p$ and $Y \sim q$ be drawn independently. Then $\mathbf{j}_t + \mathbf{j}_{1-t} = \mathbf{0}, \forall t \in [0, 1]$ if and only if $p = q$.*

A proof can be found in Appendix A.1. Unlike Proposition 2.1, the extended zero-flow criterion does not require **(A0)**. To see the reverse characterization, applying eq. (4) to p_t and p_{1-t} gives

$$\partial_t(p_t - p_{1-t}) = -\nabla \cdot (\mathbf{j}_t + \mathbf{j}_{1-t}). \tag{5}$$

Thus, if $\mathbf{j}_t + \mathbf{j}_{1-t} = 0$ for all $t \in [0, 1]$, then $p_t - p_{1-t}$ is constant in t; evaluating at $t = 1/2$ gives $p_t = p_{1-t}$, and hence $p = q$ at the endpoints.

Remark 1 (Relationship with Zero-Flow criterion). At $t = 1/2$, $\mathbf{j}_{1/2} = p_{1/2}\mathbf{v}_{1/2} = \mathbf{0}$ implies $\mathbf{v}_{1/2} = \mathbf{0}$ $p_{1/2}$-a.s. Thus, Theorem 3.1 recovers the condition in Proposition 2.1, while the converse implication from $\mathbf{v}_{1/2} = \mathbf{0}$ to $p = q$ still requires **(A0)**.

3.2 Zero-Flow Discrepancy

In order to turn Theorem 3.1 into a quantitative statistical discrepancy, we compare the flux at t and $1 - t$. Let $\widetilde{p}_t = (p_t + p_{1-t})/2$ be the symmetrized marginal and define the normalized time-varying flux residual $\mathbf{g}(\cdot, \cdot) : \mathbb{R}^d \times [0, 1/2] \to \mathbb{R}^d$ as [4]

$$\mathbf{g}(\mathbf{z}, t) := \frac{\mathbf{j}_t(\mathbf{z}) + \mathbf{j}_{1-t}(\mathbf{z})}{p_t(\mathbf{z}) + p_{1-t}(\mathbf{z})}. \tag{6}$$

[2] $\varphi_p(\omega) := \mathbb{E}_{X \sim p}[e^{i\omega^\top X}]$.

[3] For ease of exposition, we use density notation; the corresponding measure-theoretic formulation, which does not require the existence of densities, is provided in Appendix A.1.

[4] The same density convention is used here; see Appendix A.2 for the measure-theoretic definition.

At each interpolation state $\mathbf{z}$, $\mathbf{g}(\mathbf{z}, t)$ measures the violation of flux cancellation on t. We thus aggregate its squared magnitude over $\widetilde{p}_t$ to define the *Zero-Flow discrepancy*.

Definition 3.1 (Zero-Flow Discrepancy)**.** The Zero-Flow Discrepancy $\mathrm{ZFD} : \mathcal{P}_1(\mathbb{R}^d) \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\mathrm{ZFD}(p, q) := \mathbb{E}_{t\sim \mathrm{Unif}[0,\frac{1}{2}]} \mathbb{E}_{\mathbf{z}\sim\widetilde{p}_t} \left[\|\mathbf{g}(\mathbf{z}, t)\|^2 \right]. \tag{7}$$

Integrating eq. (5) over $t \in [0, 1/2]$ shows that the accumulated flux residual recovers the endpoint distribution difference:

$$p - q = \nabla \cdot \int_0^{1/2} (\mathbf{j}_t + \mathbf{j}_{1-t}) \, \mathrm{d}t. \tag{8}$$

This identity leads to the following Wasserstein control. [5]

Proposition 3.1 (Wasserstein control)**.** *For any $p, q \in \mathcal{P}_1(\mathbb{R}^d)$, $W_1^2(p, q) \le \mathrm{ZFD}(p, q)$.*

A proof can be found in Appendix A.2. Integrating eq. (8) against any 1-Lipschitz function and applying Cauchy-Schwarz yields a uniform bound, whose supremum is identified with W_1 by the Kantorovich-Rubinstein duality (Villani et al., 2009, Ch. 5). Together with the symmetry and nonnegativity of the construction, Proposition 3.1 immediately gives the definiteness of ZFD.

Corollary 3.1 (Symmetry and definiteness of ZFD)**.** *For any $p, q \in \mathcal{P}_1(\mathbb{R}^d)$,*

$$\mathrm{ZFD}(p, q) = \mathrm{ZFD}(q, p) \ge 0, \qquad \mathrm{ZFD}(p, q) = 0 \iff p = q.$$

A proof can be found in Appendix A.3. Furthermore, the normalized flux residual $\mathbf{g}$ admits the following conditional expectation form.

Lemma 3.1. *Define $\widetilde{Z}_t = (1 - B)Z_t + BZ_{1-t}$, where $B \sim \mathrm{Ber}(1/2)$. Then $\mathrm{Law}(\widetilde{Z}_t) = \widetilde{p}_t$ and*

$$\mathbf{g}(\mathbf{z}, t) = \mathbb{E}[Y - X \mid \widetilde{Z}_t = \mathbf{z}], \quad \mathbf{g}(\cdot, \cdot) = \arg\min_{\mathbf{u}} \mathbb{E}_{t\sim\mathrm{Unif}[0,1/2],X,Y,B} \|Y - X - \mathbf{u}(\widetilde{Z}_t, t)\|^2. \tag{9}$$

A proof can be found in Appendix A.4. Lemma 3.1 shows that $\mathbf{g}$ can be learned using the standard flow-matching regression objective. Consequently, ZFD can be estimated by first learning $\mathbf{g}$ and then squaring the resulting estimator. We note that our residual field $\mathbf{g}$ coincides with the flipped conditional field introduced by Shlenskii et al. (2026). Accordingly, up to time normalization, Definition 3.1 recovers their time-integrated Midpoint Divergence. Their work establishes identification under bounded-support assumptions and uses the resulting divergence as a critic for training generative models. In contrast, our focus is on its role in two-sample testing, where directly squaring a learned residual raises a distinct statistical challenge that motivates the construction developed next.

Remark 2 (Finite state spaces). Definition 3.1 does not directly apply to abstract finite or categorical state spaces $\mathcal{X}$, where the linear interpolation (1) and displacement $Y - X$ are not intrinsically defined. It therefore requires an embedding $\phi : \mathcal{X} \to \mathbb{R}^D$, such as one-hot encoding, and the resulting discrepancy depends on the induced Euclidean representation rather than the native discrete state space.

4 From ZFD to ZF2ST

Although ZFD admits a naïve plug-in estimator, under $H_0 : p = q$, the population field satisfies $\mathbf{g}(\mathbf{z}, t) \equiv \mathbf{0}$, while its finite-sample estimate $\widehat{\mathbf{g}}(\mathbf{z}, t)$ generally does not. Squaring the estimated field turns the null statistic into a quadratic functional of the estimation error. Such plug-in quadratic statistics generally do not admit a universal null distribution: their asymptotic behavior depends on the empirical-process limit of the underlying estimator and its estimation regime (Hall, 1984; Anderson et al., 1994). Consequently, even under H_0, calibrating the squared plug-in ZFD may require an estimator-specific asymptotic analysis. We circumvent this difficulty through two decoupling steps. First, we avoid squaring the estimated field by reformulating ZFD in terms of linear directional alignment. Second, we separate witness learning from testing through data splitting.

[5] $W_1(p, q) := \inf_{\pi\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\| \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$, where $\Pi(p, q)$ is the set of couplings of p and q.

4.1 A SCALE-INVARIANT FORM OF ZFD

Inspired by MMD, which admits both an RKHS norm representation in terms of the difference between kernel mean embeddings and an equivalent variational representation as the maximal alignment with a *witness function* over the RKHS unit ball (Gretton et al., 2012a), we derive an analogous scale-invariant variational representation of ZFD.

Proposition 4.1 (Scale-invariant ZFD)**.** *Let $t \sim \mathrm{Unif}[0, 1/2]$ and set $\mu := \mathrm{Law}(\widetilde{Z}_t, t)$. Denote by $L^2(\mu; \mathbb{R}^d)$ the Hilbert space of measurable time-dependent vector fields $\mathbf{h} : \mathbb{R}^d \times [0, 1/2] \to \mathbb{R}^d$ satisfying $\mathbb{E}_\mu[\|\mathbf{h}(\widetilde{Z}_t, t)\|^2] < \infty$, equipped with the canonical inner product $\langle \cdot, \cdot \rangle_\mu$ and norm $\| \cdot \|_\mu$. Suppose $\mathrm{ZFD}(p, q) < \infty$. Then*

$$\mathrm{ZFD}(p, q) = \sup_{\mathbf{h} \in L^2(\mu; \mathbb{R}^d);\ \|\mathbf{h}\|_\mu > 0} \frac{\langle \mathbf{h}, \mathbf{g} \rangle_\mu^2}{\|\mathbf{h}\|_\mu^2}. \tag{10}$$

The supremum is attained by any nonzero rescaling $\mathbf{h} = c\mathbf{g}$ with $c \neq 0$ whenever $\mathrm{ZFD}(p, q) > 0$. If $\mathrm{ZFD}(p, q) = 0$, every admissible $\mathbf{h}$ attains value zero.

A proof can be found in Appendix A.6. Similarly, $\mathbf{h}$ can be viewed as the witness function for ZFD, as it probes the discrepancy between p and q through the alignment $\langle \mathbf{h}, \mathbf{g} \rangle_\mu$. Crucially, this alignment equals $\mathrm{ZFD}(p, q)$ when $\mathbf{h} = \mathbf{g}$, while under H_0 it vanishes for *any fixed* $\mathbf{h}$ since $\mathbf{g} \equiv \mathbf{0}$. This linear alignment naturally leads to a two-stage *learn-then-test* procedure based on sample splitting (Kübler et al., 2022a), thereby avoiding the squaring problem: under H_0, the held-out alignment remains centered at zero for any learned $\mathbf{h}$, so estimation error in the witness does not induce a positive quadratic bias. We also note that Shlenskii et al. (2026) derive a quadratic variational form of the corresponding Midpoint Divergence, whereas our formulation takes a scale-invariant Rayleigh-quotient form that depends only on the witness direction and is naturally suited to the witness learning construction.

4.2 ZERO-FLOW TWO-SAMPLE TEST (ZF2ST)

Test statistic. Given paired samples $\mathcal{S} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ formed from $\mathcal{S}_p$ and $\mathcal{S}_q$, we partition $\mathcal{S}$ into a training split $\mathcal{S}^{\mathrm{tr}}$ of size n_{tr} and an independent held-out split $\mathcal{S}^{\mathrm{te}}$ of size n_{te}, where $n_{\mathrm{tr}} + n_{\mathrm{te}} = n$. *Stage-I* learns a flexible witness $\mathbf{h}$ on $\mathcal{S}^{\mathrm{tr}}$, while *Stage-II* evaluates the fixed witness on $\mathcal{S}^{\mathrm{te}}$. Thus, witness learning determines testing power through its alignment with $\mathbf{g}$, while sample splitting *preserves* type-I error control. To construct the test statistic, consider the single score of the alignment

$$\tau(\mathbf{h}) := \mathbb{E}[\mathfrak{s}_{\mathbf{h}}(X, Y, t, B)] \overset{(*)}{=} \langle \mathbf{h}, \mathbf{g} \rangle_\mu, \quad \mathfrak{s}_{\mathbf{h}}(X, Y, t, B) := \langle \mathbf{h}(\widetilde{Z}_t, t), Y - X \rangle, \tag{11}$$

where $t \sim \mathrm{Unif}[0, 1/2]$ and $B \sim \mathrm{Ber}(1/2)$ are sampled independently of (X, Y) and $(*)$ holds by the tower property (Lemma A.1). Importantly, estimating $\tau(\mathbf{h})$ does not require explicitly estimating $\mathbf{g}$. Moreover, when $\mathbf{h}$ is fixed, its empirical mean over i.i.d. paired observations provides a simple unbiased estimator with an asymptotically normal distribution by the central limit theorem (Serfling, 2009).

Proposition 4.2 (Asymptotic normality)**.** *Fix a witness $\mathbf{h} \in L^2(\mu; \mathbb{R}^d)$ and let $\sigma^2(\mathbf{h}) := \mathrm{Var}[\mathfrak{s}_{\mathbf{h}}(X, Y, t, B)]$. Let $(X_i, Y_i, t_i, B_i)_{i=1}^n$ be i.i.d., with $X_i \sim p$ and $Y_i \sim q$ drawn independently, $t_i \sim \mathrm{Unif}[0, 1/2]$, and $B_i \sim \mathrm{Ber}(1/2)$. Denote by $\hat{\tau}(\mathbf{h})$ and $\hat{\sigma}^2(\mathbf{h})$ the sample mean and sample variance of $\{\mathfrak{s}_{\mathbf{h}}(X_i, Y_i, t_i, B_i)\}_{i=1}^n$. If $0 < \sigma^2(\mathbf{h}) < \infty$, then, as $n \to \infty$,*

$$\frac{\sqrt{n}\big(\hat{\tau}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\hat{\sigma}(\mathbf{h})} \xrightarrow{d} \mathcal{N}(0, 1).$$

A proof can be found in Appendix A.7. Since the witness $\mathbf{h}$ is learned exclusively from the training split $\mathcal{S}^{\mathrm{tr}}$ in Stage-I, it is *fixed* when evaluated on the independent held-out split $\mathcal{S}^{\mathrm{te}}$. Applying Proposition 4.2 to the n_{te} held-out pairs, we denote the corresponding sample mean and standard deviation by $\hat{\tau}_{\mathrm{te}}(\mathbf{h})$ and $\hat{\sigma}_{\mathrm{te}}(\mathbf{h})$, respectively. The final ZF2ST statistic is the studentized empirical alignment

$$\zeta_{\mathrm{ZF}} := \frac{\sqrt{n_{\mathrm{te}}}\hat{\tau}_{\mathrm{te}}(\mathbf{h})}{\hat{\sigma}_{\mathrm{te}}(\mathbf{h})}. \tag{12}$$

Testing Power and Consistency. For a fixed witness $\mathbf{h}$, ζ_{ZF} is asymptotically $\mathcal{N}(0,1)$ under $H_0 : p = q$. Following Kübler et al. (2022a), we use a one-sided test because Stage-I is designed to orient $\mathbf{h}$ toward positive alignment with $\mathbf{g}$. A two-sided test discards this directional information and therefore uses a more conservative rejection threshold for positively aligned alternatives. We thus reject H_0 whenever $\zeta_{\mathrm{ZF}} \geq \Phi^{-1}(1-\alpha)$.

Proposition 4.3 (Power Approximation)**.** *Fix witness* $\mathbf{h}$*, and suppose* $0 < \sigma^2(\mathbf{h}) < \infty$*. For sufficiently large* n_{te}*, the power of the one-sided level-*α *test admits the Gaussian approximation*

$$\Pr\left(\zeta_{\mathrm{ZF}} \geq \Phi^{-1}(1-\alpha)\right) \approx 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{n_{\mathrm{te}}}\,\mathrm{SNR}(\mathbf{h})\right),\ \mathrm{SNR}(\mathbf{h}) := \frac{\tau(\mathbf{h})}{\sigma(\mathbf{h})}. \tag{13}$$

Under any fixed alternative satisfying $\tau(\mathbf{h}) > 0$*, the power converges to one as* $n_{\mathrm{te}} \to \infty$*.*

A proof can be found in Appendix A.8. Here, $\mathrm{SNR}(\mathbf{h})$ denotes the signal-to-noise ratio, measuring the strength of the population alignment relative to the variability of the corresponding score. Equation (13) shows that the power depends on $\sqrt{n_{\mathrm{te}}}\mathrm{SNR}(\mathbf{h})$, implying consistency for any positively aligned witness with $\tau(\mathbf{h}) > 0$, provided the score variance $\sigma^2(\mathbf{h})$ is finite and nonzero.

4.3 Practical Implementation

Stage I - Witness learning. We consider two objectives for learning the witness $\mathbf{h}$ from the training split $\mathcal{S}^{\mathrm{tr}}$. The first follows directly from Lemma 3.1: we learn the population field $\mathbf{g}$ by minimizing the empirical flow-matching regression risk,

$$\widehat{\mathbf{h}}_{\mathrm{reg}} = \arg\min_{\mathbf{h}\in\mathcal{F}} \widehat{\mathcal{R}}_{\mathrm{reg}}(\mathbf{h}), \quad \widehat{\mathcal{R}}_{\mathrm{reg}}(\mathbf{h}) := \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \|Y_i - X_i - \mathbf{h}(\widetilde{Z}_{t_i}, t_i)\|^2, \tag{14}$$

where $\mathcal{F}$ denotes a neural-network function class, t_i, and B_i are sampled on the fly for each pair, with $\widetilde{Z}_{t_i} = (1-B_i)Z_{t_i} + B_i Z_{1-t_i}$. This is implemented in the same manner as standard flow-matching training. We refer to this training objective as **ZF-Reg**. Alternatively, following Liu et al. (2020), we directly optimize the empirical regularized SNR:

$$\widehat{\mathbf{h}}_{\mathrm{snr}} = \arg\min_{\mathbf{h}\in\mathcal{F}} \widehat{\mathcal{R}}_{\mathrm{snr}}(\mathbf{h}), \quad \widehat{\mathcal{R}}_{\mathrm{snr}}(\mathbf{h}) = -\frac{\hat{\tau}_{\mathrm{tr}}(\mathbf{h})}{\hat{\sigma}^{\lambda}_{\mathrm{tr}}(\mathbf{h})}. \tag{15}$$

Here, $\hat{\tau}(\mathbf{h})$ is the empirical mean of i.i.d. scores defined in Proposition 4.2 and $\hat{\sigma}^{\lambda}_{\mathrm{tr}}(\mathbf{h})$ denotes the empirical counterpart of the regularized standard deviation $\sigma^{\lambda}(\mathbf{h}) := \sqrt{\sigma^2(\mathbf{h}) + \lambda\|\mathbf{h}\|^2_{\mu}}$ on $\mathcal{S}^{\mathrm{tr}}$ with $\lambda > 0$ providing ridge stabilization of the SNR denominator. We refer to this training objective as **ZF-SNR**. For both ZF-Reg and ZF-SNR, we use dynamic pairing during training to better approximate the empirical product measure $\hat{p} \otimes \hat{q}$. Implementation details are provided in B.2.

Stage II - Wild bootstrap calibration. For finite-sample calibration, we apply the wild bootstrap to the empirical held-out alignment $\hat{\tau}_{\mathrm{te}}(\mathbf{h})$ with Rademacher multipliers (Shao, 2010; Chwialkowski et al., 2014). Given a fixed witness $\mathbf{h}$, for the b-th bootstrap replicate, we draw a Rademacher vector $\boldsymbol{\epsilon}^{(b)} = [\epsilon_1^{(b)}, \ldots, \epsilon_{n_{\mathrm{te}}}^{(b)}]^{\top}$, where the entries are independent and take values in $\{+1, -1\}$ with equal probability, and compute

$$\hat{\tau}^{(b)}_{\mathrm{te}}(\mathbf{h}) := \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \epsilon_i^{(b)} \mathfrak{s}_{\mathbf{h}}(X_i, Y_i, t_i, B_i), \quad \hat{p}_{\mathrm{val}} := \frac{1 + \sum_{b=1}^{n_{\mathrm{perm}}} \mathbf{1}\left\{\hat{\tau}^{(b)}_{\mathrm{te}}(\mathbf{h}) \geq \hat{\tau}_{\mathrm{te}}(\mathbf{h})\right\}}{n_{\mathrm{perm}} + 1}. \tag{16}$$

We reject H_0 whenever $\hat{p}_{\mathrm{val}} \leq \alpha$. The resulting calibration controls the finite-sample Type-I error and is consistent under fixed alternatives with positive alignment, as formalized in Appendix A.9. In our paired setting, the wild bootstrap reduces to an exact sign-flip randomization (Hemerik et al., 2020): under H_0, the transformation $(X, Y, B) \mapsto (Y, X, 1-B)$ preserves the joint distribution while reversing the score $\mathfrak{s}_{\mathbf{h}}(Y, X, t, 1-B) = -\mathfrak{s}_{\mathbf{h}}(X, Y, t, B)$. Therefore, the null distribution can be generated entirely from the pre-computed held-out scores. As in witness-based tests (Kübler et al., 2022a;b), these scores are computed in a *single* evaluation pass over the held-out data and reused for all subsequent randomization replicates, enabling fast calibration with only sign flips and no additional witness evaluations.

Variance reduction. To further improve testing power, we introduce a variance-reduced ZF-SNR objective that preserves the population alignment $\tau(\mathbf{h})$ while reducing variance induced by auxiliary variables. As shown in Proposition 4.3, holding the alignment unchanged, a smaller score variance increases the SNR and consequently the asymptotic testing power. It can be shown that by the law of total variance,

$$\sigma^2(\mathbf{h}) = \underbrace{\sigma^2_{X,Y}}_{\text{pair } X,\,Y} + \underbrace{\mathbb{E}_{X,Y}[\sigma^2_{t|X,Y}]}_{\text{time } t} + \underbrace{\mathbb{E}_{X,Y,t}[\sigma^2_{B|X,Y,t}]}_{\text{branch } B}. \tag{17}$$

where $\sigma^2_{X,Y} := \mathrm{Var}_{X,Y}(\mathbb{E}_{t,B}[\mathfrak{s}_\mathbf{h}|X,Y])$, $\sigma^2_{t|X,Y} := \mathrm{Var}_t(\mathbb{E}_B[\mathfrak{s}_\mathbf{h}|X,Y,t] \mid X,Y)$ and $\sigma^2_{B|X,Y,t} := \mathrm{Var}_B(\mathfrak{s}_\mathbf{h}|X,Y,t)$. The last two terms arise solely from sampling t and B. We average out B exactly and reduce the remaining time-sampling variance using K i.i.d. samples $t_{1:K} \sim \mathrm{Unif}[0,1/2]$:

$$\mathfrak{s}_{\mathbf{h},K}(X,Y,t_{1:K}) := \frac{1}{K}\sum_{k=1}^{K} \bar{\mathfrak{s}}_\mathbf{h}(t_k). \quad \bar{\mathfrak{s}}_\mathbf{h}(t) := \frac{1}{2}\left\langle \mathbf{h}(Z_t,t) + \mathbf{h}(Z_{1-t},t), Y - X \right\rangle. \tag{18}$$

Here, $\bar{\mathfrak{s}}_\mathbf{h}(t) = \mathbb{E}_B[\mathfrak{s}_\mathbf{h} \mid X,Y,t]$, with the dependence on (X,Y) suppressed for brevity. Letting $\sigma^2_K(\mathbf{h}) := \mathrm{Var}(\mathfrak{s}_{\mathbf{h},K})$, we have $\sigma^2_K(\mathbf{h}) = \sigma^2_{X,Y} + \frac{1}{K}\mathbb{E}_{X,Y}[\sigma^2_{t|X,Y}] \le \sigma^2(\mathbf{h})$. Moreover, $\mathbb{E}[\mathfrak{s}_{\mathbf{h},K}] = \tau(\mathbf{h})$ for any $K \ge 1$, so the population alignment is preserved while the auxiliary variance is reduced. Accordingly, during training, we replace the ZF-SNR objective $\mathcal{R}_{\mathrm{snr}}(\mathbf{h}) := -\tau(\mathbf{h})/\sigma^\lambda(\mathbf{h})$ with its variance-reduced counterpart $\mathcal{R}_{\mathrm{snr},K}(\mathbf{h}) := -\tau(\mathbf{h})/\sigma^\lambda_K(\mathbf{h})$, and optimize the corresponding empirical objective $\widehat{\mathcal{R}}_{\mathrm{snr},K}$ on $\mathcal{S}^{\mathrm{tr}}$. An explicit form is provided in Appendix B.3.

At test time, ZF-SNR uses the same K as in training to align the training objective with the test statistic. Since $\sigma^2_K(\mathbf{h}) \le \sigma^2(\mathbf{h})$ while the alignment is unchanged, any fixed witness with positive alignment attains no lower asymptotic power when evaluated using $\mathfrak{s}_{\mathbf{h},K}$. Moreover, $\mathfrak{s}_{\mathbf{h},K}$ preserves swap antisymmetry, so wild-bootstrap validity remains unchanged. We therefore also use $\mathfrak{s}_{\mathbf{h},K}$ for ZF-Reg at test time. This requires $2K$ witness evaluations during testing for both methods. We summarize the full ZF2ST procedure in Algorithm 1, with a detailed complexity analysis provided in Appendix B.4.

5 Experiments

We compare ZF2ST against a broad range of kernel-based and witness-learning two-sample tests, spanning both sample-splitting and non-splitting approaches. Among sample-splitting methods, we consider deep-kernel MMD (MMD-D; Liu et al., 2020), classifier two-sample tests based on sign (C2ST-S; Lopez-Paz & Oquab, 2016) and linear scores (C2ST-L; Cheng & Cloninger, 2022), the KFDA witness test (KFDA-WiTS; Kübler et al., 2022a) and its more flexible AutoML-based extension (AutoML; Kübler et al., 2022b). We further include strong non-splitting kernel tests, namely MMDagg (Schrab et al., 2023) and MMD-FUSE (Biggs et al., 2023).

For sample-splitting methods, we run 10 independent trainings, each followed by 100 independent held-out test trials. We use $n = n_{\mathrm{tr}} + n_{\mathrm{te}}$ with $n_{\mathrm{tr}} = n_{\mathrm{te}}$. Sample-splitting methods train on n_{tr} and test on n_{te}, while non-splitting methods use all n samples over 100 independent trials. For non-image experiments, ZF2ST *matches* the MLP architecture of MMD-D; for image experiments, ZF2ST uses a *substantially smaller* network, with architecture and capacity comparisons provided in Appendix D.2. For ZF-SNR, we use its variance-reduced version and set $(K,\lambda) = (3, 10^{-3})$ for non-image experiments and $(2, 10^{-2})$ for image experiments. ZF2ST uses wild bootstrap calibration in all experiments. Further experiment details are provided in Appendix C.

Synthetic benchmarks. Following Liu et al. (2020), we consider the `Blobs` dataset with multiple Gaussian components of different covariances and the high-dimensional Gaussian mixture (`HDGM`) with differences only in the first two dimensions. For `HDGM`, we vary either the sample size with $d = 10$ fixed or the ambient dimension with $n = 2000$ fixed. As shown in the top row of Figure 1, ZF-SNR rapidly gains power on `Blobs`, while ZF-Reg shows a more modest improvement as the sample size increases. On `HDGM`, both ZF-SNR and ZF-Reg perform strongly across two settings. ZF-Reg is particularly robust as the ambient dimension increases and also performs best in the larger-sample regime, suggesting that regression-based witness learning may be especially well suited to this type of sparse, high-dimensional alternative.

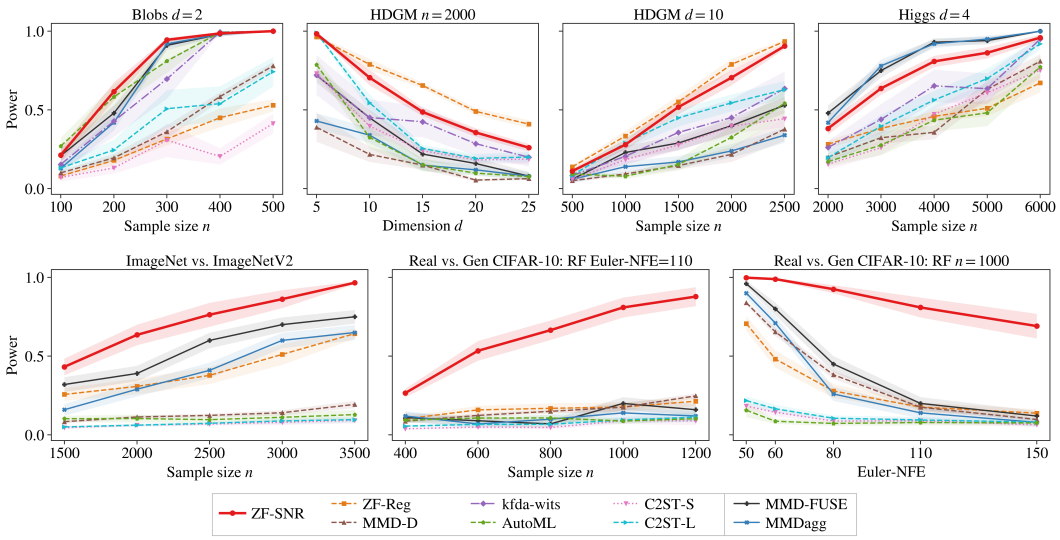

Figure 1: Power experiments at level $\alpha = 0.05$. The top row reports `Blobs`, `HDGM`, and `Higgs`, varying either sample size n or dimension d as indicated. The bottom row reports ImageNet vs. ImageNetV2 and real vs. RF-generated CIFAR-10, varying either the sample size or the NFE of the Euler solver. All images are resized to $3 \times 64 \times 64$. KFDA-WiTS is evaluated only on the non-image data. Lines show mean test power, with shaded regions indicating ± 1 standard error. Corresponding Type-I error results are provided in Figure 5.

Higgs dataset (Baldi et al., 2014). Following Liu et al. (2020), we consider the `Higgs` dataset, where the signal is known to be relatively weak (Chwialkowski et al., 2015). As shown in the top-right panel of Figure 1, ZF-SNR consistently outperforms the other sample-splitting methods. It also remains competitive with the non-splitting kernel tests, despite using only held-out samples for testing, while its improvement over ZF-Reg indicates an additional benefit from the SNR-based training objective.

Real-data distribution shift. We first consider CIFAR-10 (Krizhevsky et al., 2009) vs. CIFAR-10.1 (Recht et al., 2019) where CIFAR-10.1 was independently recollected to reproduce the original CIFAR-10 image classification task while using a separate data-collection process. This provides a natural real-data distribution-shift benchmark and has been studied in two-sample testing by Liu et al. (2020) and Biggs et al. (2023). We further consider ImageNet (Deng et al., 2009) vs. ImageNetV2 (Recht et al., 2019), which applies the same replication principle to the more challenging ImageNet classification task using newly collected test images. Liu et al. (2020) identified this comparison as a more challenging extension.

On CIFAR-10 vs. CIFAR-10.1, Table 1a shows that both ZF variants achieve power close to one and outperform all the competing methods. On ImageNet vs. ImageNetV2, ZF-SNR consistently achieves the highest power as the sample size increases, as shown in the bottom-left panel of Figure 1. Together, these results provide strong statistical evidence of distribution shifts in both benchmarks and demonstrate that ZF2ST remains effective on high-dimensional natural images.

CIFAR-10 with modern generative models. Comparing generated samples with observed data has long been a natural application of two-sample testing for evaluating generative models (Sutherland et al., 2016; Lopez-Paz & Oquab, 2016). Since generative models are explicitly trained to make their output distribution q closely approximate the data distribution p, distinguishing the two distributions can be particularly challenging when the remaining differences are subtle and concentrated in complex high-dimensional image statistics. Recent advances in generative modeling have dramatically improved the quality of synthesized images (Dhariwal & Nichol, 2021; Rombach et al., 2022), further narrowing the gap between generated and real data. We therefore evaluate ZF2ST across several recent generative models to assess its ability to detect these residual distributional differences.

We first compare real CIFAR-10 with samples generated by a 1-Rectified Flow (RF, Liu et al., 2022) model using an Euler solver. The Number of Function Evaluations (NFE) determines the number of Euler steps used to solve the learned ODE when generating each sample. As noted by Liu et al.

(a) CIFAR-10 vs. CIFAR-10.1.

Tests	Power (%)
ZF-SNR	99.5 ± 0.27
ZF-Reg	**99.8** ± 0.13
MMD-FUSE	94.0 ± 2.39
MMDagg	86.0 ± 3.49
MMD-D	83.3 ± 3.65
AutoML	78.4 ± 7.39
C2ST-L	43.3 ± 8.17
C2ST-S	40.1 ± 8.37

(b) Real vs. generated CIFAR-10 across different generative models.

Tests ($n = 1000$)	ScoreSDE (cFID_{10k}=6.48)	RF RK45 (cFID_{10k}=7.38)	MeanFlow (cFID_{10k}=7.48)	SGXL (cFID_{10k}=6.29)
ZF-SNR	**32.3** ± 5.51	**40.3** ± 4.78	59.6 ± 3.71	**36.3** ± 5.53
ZF-Reg	10.0 ± 3.32	16.6 ± 3.54	35.0 ± 5.37	23.2 ± 3.32
MMD-FUSE	17.0 ± 3.78	27.0 ± 4.46	**79.0** ± 4.09	18.0 ± 3.86
MMDagg	15.0 ± 3.59	15.0 ± 3.59	75.0 ± 4.35	22.0 ± 4.16
MMD-D	9.1 ± 1.26	8.2 ± 0.59	28.1 ± 2.90	14.7 ± 1.34
AutoML	11.4 ± 1.71	9.1 ± 1.85	16.1 ± 5.17	26.8 ± 2.66
C2ST-L	8.0 ± 1.65	6.3 ± 1.15	10.0 ± 0.92	14.2 ± 2.01
C2ST-S	6.9 ± 1.43	5.7 ± 0.99	8.1 ± 0.91	12.6 ± 1.86

Table 1: Power experiments at level $\alpha = 0.05$ on **(a)** CIFAR-10 vs. CIFAR-10.1 **(b)** real CIFAR-10 vs. samples generated by four different models, with $n = 1000$ samples per distribution. cFID_{10k} denotes clean FID computed using 10,000 generated samples. All images are resized to $3 \times 64 \times 64$. Test power (%) is reported as mean ± standard error, with the best result shown in bold.

(2022), increasing the Euler NFE reduces discretization error and improves sampling quality. We therefore consider two settings: varying the sample size n with NFE $= 110$ fixed, and varying the Euler NFE with $n = 1000$ fixed. As shown in the bottom-middle and bottom-right panels of Figure 1, ZF-SNR exhibits a striking advantage over all baselines, with power increasing rapidly with sample size and remaining substantially higher as the Euler NFE increases.

We further evaluate ZF2ST across several representative generative model families to examine whether its strong performance extends beyond a particular generation mechanism. The benchmark includes multi-step generative models from both diffusion and flow paradigms, represented by ScoreSDE (Song et al., 2020) and 1-Rectified Flow with a higher-accuracy RK45 solver (Liu et al., 2022), respectively. We also consider the adversarially trained StyleGAN-XL (SGXL, Sauer et al., 2022) and MeanFlow (Geng et al., 2025), a recent high-quality one-step generative model. For each model, we use the official pretrained checkpoint to generate 10,000 samples, matching the size of the CIFAR-10 test set, and perform two-sample testing with $n = 1000$ samples from each distribution.

As shown in Table 1b, ZF-SNR achieves the highest power on three of the four generators. On MeanFlow, non-splitting MMD methods perform better, while ZF-SNR still achieves substantially higher power than the remaining sample-splitting baselines. Overall, these results suggest that the strong performance of ZF-SNR generalizes across diverse generative mechanisms rather than being tied to a particular model family or numerical solver. We also report clean FID (cFID, Parmar et al., 2022) computed using 10,000 generated samples, providing a more standardized and fair comparison, with lower values indicating better generative quality. Test power in Table 1b generally increases with cFID, consistent with lower-quality generated samples being easier to distinguish from real data. However, SGXL is a notable exception: despite having the lowest cFID, all tests achieve higher power than on ScoreSDE, indicating that cFID does not fully capture the distributional differences detected by two-sample tests.

6 Conclusion

In this work, we proposed ZF2ST, a flow-based framework for two-sample testing built on the zero-flow characterization of distributional equality. We developed a principled discrepancy and testing procedure and demonstrated competitive performance across synthetic, tabular, and image benchmarks, with ZF-Reg and ZF-SNR showing complementary strengths and particularly strong image performance for SNR optimization. A current limitation is the reliance on sample splitting, which reduces data efficiency, while a complete finite-sample theory for data-dependent learned witnesses remains open.

Future work could develop more data-efficient testing procedures that relax or avoid sample splitting, establish finite-sample guarantees that account for witness estimation, and extend ZF2ST to other hypothesis testing problems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting high-level section structures, polishing writing, retrieving and discovering related work, and checking mathematical arguments and presentation. All relevant literature was read and verified by the authors to ensure accurate representation. The core research ideas, theorem statements, and proofs were developed and written by the authors; generative AI tools were not used to generate the main research ideas or to prove mathematical claims.

Our **CIFAR-10 with modern generative models** experiments in Section 5 use synthetic image samples produced by the pretrained generative models being evaluated. These samples were not generated by an LLM or AI assistant; the corresponding models, checkpoints, and sampling procedures are fully disclosed in Appendix C.2. All AI-assisted content was reviewed and verified by the authors.

The authors take full responsibility for the final content of the paper, including all text, claims, proofs, code, and experimental results produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a general statistical methodology for two-sample testing and does not involve the collection of new human-subject data or personally identifiable information. All empirical evaluations use publicly available benchmark datasets and publicly released pretrained generative models, with data sources and model checkpoints documented in Appendix C. We do not release any new dataset. As with general-purpose statistical testing tools, ZF2ST may be applied to domains involving sensitive data; such applications should follow the relevant requirements regarding data access, privacy, consent, and responsible interpretation of statistical results. We are not aware of other direct ethical concerns specific to the methodology introduced in this work.

REPRODUCIBILITY STATEMENT

We provide detailed material to support reproducibility of both the theoretical and empirical results. Complete proofs of the main theoretical results, including the measure-theoretic formulation and the validity of the wild bootstrap calibration, are given in Appendix A. Appendix B provides the full ZF2ST algorithm together with implementation details for dynamic pairing, the variance-reduced objective, and computational complexity. Appendix C specifies the experimental protocols, datasets and preprocessing, model architectures, baseline implementations, hyperparameters, Type-I error results, and runtime measurements. Additional ablation studies are provided in Appendix D, and visualizations of the datasets and generated samples are included in Appendix E. For experiments involving pretrained generative models, Appendix C also provides the corresponding official code repositories, checkpoints, and sampling configurations.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014.

Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36: 75151–75188, 2023.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Ivan Butakov, Alexander Semenenko, Valeriya Kirova, Alexey Frolov, and Ivan Oseledets. Fmmi: Flow matching mutual information estimation. *arXiv preprint arXiv:2511.08552*, 2025.

Hao Chen and Jerome H Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409, 2017.

Xiuyuan Cheng and Alexander Cloninger. Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, 68(10): 6631–6662, 2022.

Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. *Advances in neural information processing systems*, 27, 2014.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Bradley Efron and Robert Tibshirani. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17):1–35, 1985.

Giovanni Fasano and Alberto Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987.

Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of statistics*, pp. 697–717, 1979.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2025.

Debarghya Ghoshdastidar and Ulrike Von Luxburg. Practical methods for graph two-sample testing. *Advances in Neural Information Processing Systems*, 31, 2018.

Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012a.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012b.

Peter Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, 14(1):1–16, 1984.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *Test*, 27(4):811–825, 2018.

Jesse Hemerik, Jelle J Goeman, and Livio Finos. Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):841–864, 2020.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pp. 1403–1419. PMLR, 2022a.

Jonas M Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf. Automl two-sample test. *Advances in Neural Information Processing Systems*, 35:15929–15941, 2022b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and TrendsÂ® in Machine Learning*, 10(1-2):1–141, 2017.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Mónica Ribero, Antonin Schrab, and Arthur Gretton. Regularized f-divergence kernel tests. *arXiv preprint arXiv:2601.19755*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Xiaofeng Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105 (489):218–235, 2010.

Daniil Shlenskii, Nikita Gushchin, Lev Novitskiy, Dmitry V Dylov, and Alexander Korotin. Midpoint generative models. *arXiv preprint arXiv:2605.29920*, 2026.

Nikolai V Smirnov et al. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks*, 24(7):735–751, 2011.

Danica J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Yakun Wang, Leyang Wang, Song Liu, and Taiji Suzuki. Zero-flow encoders. In *International Conference on Machine Learning*, 2026.

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.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. *Advances in Neural Information Processing Systems*, 38:80324–80352, 2025.

Zhijian Zhou, Xunye Tian, Liuhua Peng, Chao Lei, Antonin Schrab, Danica J Sutherland, and Feng Liu. Dual: Learning diverse kernels for aggregated two-sample and independence testing. *Advances in Neural Information Processing Systems*, 38:130661–130695, 2025.

APPENDIX

A PROOFS

A.1 PROOF OF THEOREM 3.1

Proof. We first give a measure-theoretic formulation of the probability flux, which does not require the interpolating distributions to admit densities. Recall that

$$Z_t = (1-t)X + tY, \qquad p_t := \mathrm{Law}(Z_t).$$

Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ-algebra on $\mathbb{R}^d$. We define the vector-valued measure

$$\mathbf{J}_t(A) := \mathbb{E}\big[(Y-X)\mathbf{1}_{\{Z_t\in A\}}\big], \qquad A \in \mathcal{B}(\mathbb{R}^d).$$

Let $\|\mathbf{J}_t\|_{\mathrm{TV}} := |\mathbf{J}_t|(\mathbb{R}^d)$ denote the total variation norm of $\mathbf{J}_t$. Since $p, q \in \mathcal{P}_1(\mathbb{R}^d)$,

$$\|\mathbf{J}_t\|_{\mathrm{TV}} \le \mathbb{E}\|Y-X\| \le \mathbb{E}\|X\| + \mathbb{E}\|Y\| < \infty,$$

so $\mathbf{J}_t$ is a finite vector-valued measure. Moreover, by the tower property, for every $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mathbf{J}_t(A) &= \mathbb{E}\big[(Y-X)\mathbf{1}_{\{Z_t\in A\}}\big] \\ &= \mathbb{E}\big[\mathbb{E}[Y-X \mid Z_t]\,\mathbf{1}_{\{Z_t\in A\}}\big] \\ &= \int_A \mathbb{E}[Y-X \mid Z_t = \mathbf{z}]\,\mathrm{d}p_t(\mathbf{z}).\end{aligned}$$

Hence $\mathbf{J}_t \ll p_t$, with Radon–Nikodym derivative

$$\frac{\mathrm{d}\mathbf{J}_t}{\mathrm{d}p_t}(\mathbf{z}) = \mathbb{E}[Y-X \mid Z_t = \mathbf{z}] = \mathbf{v}_t(\mathbf{z}), \qquad p_t\text{-a.e.}$$

Whenever p_t admits a density with respect to the Lebesgue measure, which we also denote by $p_t(z)$, this reduces to

$$\mathrm{d}\mathbf{J}_t(\mathbf{z}) = p_t(\mathbf{z})\mathbf{v}_t(\mathbf{z})\,\mathrm{d}\mathbf{z} = \mathbf{j}_t(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Thus $\mathbf{J}_t$ is the measure-theoretic counterpart of the probability flux $\mathbf{j}_t$ used in the main text.

We first prove the forward direction. Suppose $p = q$. Since X and Y are independent and identically distributed, (X, Y) is exchangeable. Therefore, for every $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mathbf{J}_{1-t}(A) &= \mathbb{E}\big[(Y-X)\mathbf{1}_{\{tX+(1-t)Y\in A\}}\big] \\ &= \mathbb{E}\big[(X-Y)\mathbf{1}_{\{tY+(1-t)X\in A\}}\big] \\ &= -\mathbb{E}\big[(Y-X)\mathbf{1}_{\{(1-t)X+tY\in A\}}\big] \\ &= -\mathbf{J}_t(A).\end{aligned}$$

Hence

$$\mathbf{J}_t + \mathbf{J}_{1-t} = 0, \qquad \forall t \in [0,1].$$

When the fluxes admit densities, this is exactly

$$\mathbf{j}_t + \mathbf{j}_{1-t} = \mathbf{0}.$$

Conversely, suppose

$$\mathbf{J}_t + \mathbf{J}_{1-t} = 0, \qquad \forall t \in [0,1].$$

For any test function $\varphi \in C_c^1(\mathbb{R}^d)$, where $C_c^1(\mathbb{R}^d)$ denotes the space of continuously differentiable functions with compact support, define

$$F_\varphi(t) := \int_{\mathbb{R}^d} \varphi(\mathbf{z})\,\mathrm{d}p_t(\mathbf{z}) = \mathbb{E}[\varphi(Z_t)].$$

Since $\nabla\varphi$ is bounded and $\mathbb{E}\|Y-X\| < \infty$, differentiation under the expectation is justified by dominated convergence, and

$$F_\varphi'(t) = \mathbb{E}\big[\nabla\varphi(Z_t)^\top (Y-X)\big] = \int_{\mathbb{R}^d} \nabla\varphi(\mathbf{z})^\top\,\mathrm{d}\mathbf{J}_t(\mathbf{z}).$$

This is the weak form of the continuity equation (Ambrosio et al., 2005).

Now consider the difference between the marginals at times t and $1-t$. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(\mathbf{z})\,\mathrm{d}\big(p_t-p_{1-t}\big)(\mathbf{z}) &= F'_\varphi(t)+F'_\varphi(1-t)\\ &= \int_{\mathbb{R}^d}\nabla\varphi(\mathbf{z})^\top\,\mathrm{d}\big(\mathbf{J}_t+\mathbf{J}_{1-t}\big)(\mathbf{z})\\ &= 0.\end{aligned}$$

Therefore,

$$\int_{\mathbb{R}^d}\varphi(z)\,\mathrm{d}\big(p_t-p_{1-t}\big)(z)$$

is constant in t. At $t=1/2$, this quantity is zero, since $p_{1/2}-p_{1/2}=0$. Hence,

$$\int_{\mathbb{R}^d}\varphi(\mathbf{z})\,\mathrm{d}p_t(\mathbf{z})=\int_{\mathbb{R}^d}\varphi(\mathbf{z})\,\mathrm{d}p_{1-t}(\mathbf{z}),\qquad\forall\varphi\in C_c^1(\mathbb{R}^d),\quad\forall t\in[0,1].$$

Since $C_c^1(\mathbb{R}^d)$ is measure determining, it follows that

$$p_t=p_{1-t},\qquad\forall t\in[0,1].$$

Taking $t=0$ gives

$$p=p_0=p_1=q.$$

This completes the proof. □

A.2 Proof of Proposition 3.1

Proof. We first give a measure-theoretic definition of the normalized flux residual in Equation (6). Let $\mathbf{J}_t$ be the vector-valued flux measure defined in Appendix A.1, and define

$$\mathbf{R}_t:=\mathbf{J}_t+\mathbf{J}_{1-t}.$$

Since $\mathbf{J}_t\ll p_t$ and $\mathbf{J}_{1-t}\ll p_{1-t}$, we have

$$\mathbf{R}_t\ll p_t+p_{1-t}.$$

We therefore define

$$\mathbf{g}(\cdot,t):=\frac{\mathrm{d}\mathbf{R}_t}{\mathrm{d}(p_t+p_{1-t})}=\frac{\mathrm{d}(\mathbf{J}_t+\mathbf{J}_{1-t})}{\mathrm{d}(p_t+p_{1-t})}.$$

Equivalently, since

$$\widetilde{p}_t=\frac{p_t+p_{1-t}}{2},$$

we have

$$\mathrm{d}\mathbf{R}_t(\mathbf{z})=2\mathbf{g}(\mathbf{z},t)\,\mathrm{d}\widetilde{p}_t(\mathbf{z}).\tag{19}$$

Whenever p_t and p_{1-t} admit densities with respect to the Lebesgue measure, this definition reduces to

$$\mathbf{g}(\mathbf{z},t)=\frac{\mathbf{j}_t(\mathbf{z})+\mathbf{j}_{1-t}(\mathbf{z})}{p_t(\mathbf{z})+p_{1-t}(\mathbf{z})},$$

which is precisely Equation (6).

If $\mathrm{ZFD}(p,q)=+\infty$, the result is immediate. Hence, suppose $\mathrm{ZFD}(p,q)<\infty$. For any continuously differentiable 1-Lipschitz function $\varphi:\mathbb{R}^d\to\mathbb{R}$, we have

$$\int_{\mathbb{R}^d}\varphi(\mathbf{z})\,\mathrm{d}p_t(\mathbf{z})=\mathbb{E}[\varphi(Z_t)].$$

Since $\|\nabla\varphi(\mathbf{z})\|\le 1$ and $\mathbb{E}\|Y-X\|<\infty$, the difference quotients are dominated by the integrable random variable $\|Y-X\|$. Hence, differentiation under the expectation is justified by dominated convergence, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(\mathbf{z})\,\mathrm{d}p_t(\mathbf{z})=\mathbb{E}\big[\nabla\varphi(Z_t)^\top(Y-X)\big]=\int_{\mathbb{R}^d}\nabla\varphi(\mathbf{z})^\top\,\mathrm{d}\mathbf{J}_t(\mathbf{z}).$$

Applying this identity to $p_t - p_{1-t}$ and integrating over $t \in [0, 1/2]$ yields

$$\int_{\mathbb{R}^d} \varphi(\mathbf{z})\, \mathrm{d}(p-q)(\mathbf{z}) = -\int_0^{1/2} \int_{\mathbb{R}^d} \nabla\varphi(\mathbf{z})^\top \, \mathrm{d}\mathbf{R}_t(\mathbf{z})\, \mathrm{d}t.$$

Using Equation (19) and $\|\nabla\varphi(\mathbf{z})\| \le 1$, we obtain

$$\begin{aligned}\left|\int_{\mathbb{R}^d} \varphi(\mathbf{z})\, \mathrm{d}(p-q)(\mathbf{z})\right| &\le 2\int_0^{1/2} \int_{\mathbb{R}^d} \|\mathbf{g}(\mathbf{z},t)\|\, \mathrm{d}\widetilde{p}_t(\mathbf{z})\, \mathrm{d}t \\ &= \mathbb{E}_{t\sim\mathrm{Unif}[0,\frac{1}{2}]}\mathbb{E}_{\mathbf{z}\sim\widetilde{p}_t}\left[\|\mathbf{g}(\mathbf{z},t)\|\right].\end{aligned} \tag{20}$$

By the Cauchy-Schwarz inequality,

$$\mathbb{E}_{t\sim\mathrm{Unif}[0,\frac{1}{2}]}\mathbb{E}_{\mathbf{z}\sim\widetilde{p}_t}\left[\|\mathbf{g}(\mathbf{z},t)\|\right] \le \left(\mathbb{E}_{t\sim\mathrm{Unif}[0,\frac{1}{2}]}\mathbb{E}_{\mathbf{z}\sim\widetilde{p}_t}\left[\|\mathbf{g}(\mathbf{z},t)\|^2\right]\right)^{1/2} = \sqrt{\mathrm{ZFD}(p,q)}.$$

Combining the above inequalities gives

$$\left|\int_{\mathbb{R}^d} \varphi(\mathbf{z})\, \mathrm{d}(p-q)(\mathbf{z})\right| \le \sqrt{\mathrm{ZFD}(p,q)} \tag{21}$$

for every continuously differentiable 1-Lipschitz function φ. By a standard smooth approximation argument, Equation (21) extends to every 1-Lipschitz function.

Recall that the 1-Wasserstein distance is defined by

$$W_1(p,q) := \inf_{\pi\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|\, \mathrm{d}\pi(\mathbf{x},\mathbf{y}),$$

where $\Pi(p,q)$ denotes the set of couplings of p and q. By the Kantorovich-Rubinstein duality (Villani et al., 2009, Ch. 5),

$$W_1(p,q) = \sup_{\|\varphi\|_{\mathrm{Lip}}\le 1} \left|\int_{\mathbb{R}^d} \varphi(\mathbf{z})\, \mathrm{d}(p-q)(\mathbf{z})\right|.$$

Since Equation (21) holds for every 1-Lipschitz φ,

$$W_1^2(p,q) = \left(\sup_{\|\varphi\|_{\mathrm{Lip}}\le 1} \left|\int_{\mathbb{R}^d} \varphi(\mathbf{z})\, \mathrm{d}(p-q)(\mathbf{z})\right|\right)^2 \le \mathrm{ZFD}(p,q),$$

which proves the claim. □

A.3 Proof of Corollary 3.1

Proof. We first show symmetry. Let $\mathbf{J}_t$ and

$$\mathbf{R}_t := \mathbf{J}_t + \mathbf{J}_{1-t}$$

be the flux measure and residual flux associated with (p,q), as defined at the beginning of Appendix A.1 and A.2, respectively. For the reversed pair (q,p), denote the corresponding quantities by $\mathbf{J}_t^{\mathrm{rev}}$ and $\mathbf{R}_t^{\mathrm{rev}}$.

Swapping the roles of X and Y maps

$$Z_t = (1-t)X + tY \quad \text{to} \quad Z_{1-t} = (1-t)Y + tX$$

and reverses the displacement $Y - X$. Hence,

$$\mathbf{J}_t^{\mathrm{rev}} = -\mathbf{J}_{1-t},$$

and therefore

$$\mathbf{R}_t^{\mathrm{rev}} = \mathbf{J}_t^{\mathrm{rev}} + \mathbf{J}_{1-t}^{\mathrm{rev}} = -\mathbf{R}_t.$$

Meanwhile, the symmetrized marginal is unchanged under the swap,

$$\widetilde{p}_t^{\mathrm{rev}} = \frac{p_{1-t}+p_t}{2} = \widetilde{p}_t.$$

It follows from the definition of the normalized residual that

$$\mathbf{g}^{\mathrm{rev}}(\mathbf{z}, t) = -\mathbf{g}(\mathbf{z}, t), \qquad \widetilde{p}_t\text{-a.e.}$$

Hence, by Definition 3.1

$$\mathrm{ZFD}(q, p) = \mathrm{ZFD}(p, q).$$

Nonnegativity follows immediately from Definition 3.1

$$\mathrm{ZFD}(p, q) = \mathbb{E}_{t\sim\mathrm{Unif}[0,1/2]}\mathbb{E}_{\mathbf{z}\sim\widetilde{p}_t}\left[\|\mathbf{g}(\mathbf{z}, t)\|^2\right] \geq 0.$$

It remains to establish definiteness. If $p = q$, then by Theorem 3.1,

$$\mathbf{J}_t + \mathbf{J}_{1-t} = 0, \qquad \forall t \in [0, 1],$$

and hence $\mathbf{g}(\mathbf{z}, t) = 0$ almost everywhere, which gives

$$\mathrm{ZFD}(p, q) = 0.$$

Conversely, if $\mathrm{ZFD}(p, q) = 0$, then Proposition 3.1 implies

$$W_1^2(p, q) \leq \mathrm{ZFD}(p, q) = 0.$$

Since W_1 is a metric on $\mathcal{P}_1(\mathbb{R}^d)$, we obtain $p = q$. Therefore,

$$\mathrm{ZFD}(p, q) = 0 \quad \Longleftrightarrow \quad p = q.$$

□

A.4 Proof of Lemma 3.1

Proof. We first show that $\widetilde{Z}_t$ follows the symmetrized marginal. Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ-algebra on $\mathbb{R}^d$. For any $A \in \mathcal{B}(\mathbb{R}^d)$, since $B \sim \mathrm{Ber}(1/2)$ is independent of (X, Y),

$$\begin{aligned}\mathbb{P}(\widetilde{Z}_t \in A) &= \frac{1}{2}\mathbb{P}(Z_t \in A) + \frac{1}{2}\mathbb{P}(Z_{1-t} \in A) \\ &= \frac{1}{2}p_t(A) + \frac{1}{2}p_{1-t}(A) = \widetilde{p}_t(A).\end{aligned}$$

Hence,

$$\mathrm{Law}(\widetilde{Z}_t) = \widetilde{p}_t.$$

Next, recall from Appendix A.2 that

$$\mathbf{R}_t := \mathbf{J}_t + \mathbf{J}_{1-t}, \qquad \mathrm{d}\mathbf{R}_t(\mathbf{z}) = 2\mathbf{g}(\mathbf{z}, t)\,\mathrm{d}\widetilde{p}_t(\mathbf{z}).$$

For any $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mathbb{E}\Big[(Y - X)\mathbf{1}_{\{\widetilde{Z}_t\in A\}}\Big] &= \frac{1}{2}\mathbb{E}\big[(Y - X)\mathbf{1}_{\{Z_t\in A\}}\big] + \frac{1}{2}\mathbb{E}\big[(Y - X)\mathbf{1}_{\{Z_{1-t}\in A\}}\big] \\ &= \frac{1}{2}\mathbf{R}_t(A) \\ &= \int_A \mathbf{g}(\mathbf{z}, t)\,\mathrm{d}\widetilde{p}_t(\mathbf{z}).\end{aligned}$$

Since $\mathrm{Law}(\widetilde{Z}_t) = \widetilde{p}_t$, this identifies $\mathbf{g}(\cdot, t)$ as the conditional expectation,

$$\mathbf{g}(\mathbf{z}, t) = \mathbb{E}[Y - X \mid \widetilde{Z}_t = \mathbf{z}], \qquad \widetilde{p}_t\text{-a.e.}$$

For the regression characterization, we additionally assume $p, q \in \mathcal{P}_2(\mathbb{R}^d)$. Let $t \sim \mathrm{Unif}[0, 1/2]$ be independent of (X, Y, B). Then $\mathbb{E}\|Y - X\|^2 < \infty$, and by the standard L^2 projection property of conditional expectation, for any measurable $\mathbf{u}$ with finite second moment,

$$\begin{aligned}&\mathbb{E}\left[\|Y - X - \mathbf{u}(\widetilde{Z}_t, t)\|^2\right] \\ &\quad = \mathbb{E}\left[\|Y - X - \mathbf{g}(\widetilde{Z}_t, t)\|^2\right] + \mathbb{E}\left[\|\mathbf{u}(\widetilde{Z}_t, t) - \mathbf{g}(\widetilde{Z}_t, t)\|^2\right].\end{aligned}$$

Therefore, the unique minimizer up to almost-everywhere equivalence is

$$\mathbf{g}(\cdot, \cdot) = \arg\min_{\mathbf{u}} \mathbb{E}_{t\sim\mathrm{Unif}[0,1/2],\, X,Y,B}\left[\|Y - X - \mathbf{u}(\widetilde{Z}_t, t)\|^2\right].$$

This proves Equation (9). □

A.5 Tower Property for the Alignment

Lemma A.1 (Tower property). *Let $t \sim \mathrm{Unif}[0, 1/2]$ and $\mu = \mathrm{Law}(\widetilde{Z}_t, t)$. For any $\mathbf{h} \in L^2(\mu; \mathbb{R}^d)$ such that $\mathbb{E}[\|\mathbf{h}(\widetilde{Z}_t, t)\| \, \|Y - X\|] < \infty$, we have*

$$\langle \mathbf{h}, \mathbf{g} \rangle_\mu = \mathbb{E}\Big[\Big\langle \mathbf{h}(\widetilde{Z}_t, t), Y - X \Big\rangle\Big].$$

Proof. By Lemma 3.1, we have $\mathbf{g}(\mathbf{z}, t) = \mathbb{E}[Y - X | \widetilde{Z}_t = \mathbf{z}]$, hence

$$\mathbf{g}(\widetilde{Z}_t, t) = \mathbb{E}[Y - X \mid \widetilde{Z}_t, t].$$

Therefore, by the tower property,

$$\begin{aligned}
\langle \mathbf{h}, \mathbf{g} \rangle_\mu &= \mathbb{E}\Big[\Big\langle \mathbf{h}(\widetilde{Z}_t, t), \mathbf{g}(\widetilde{Z}_t, t) \Big\rangle\Big] \\
&= \mathbb{E}\Big[\Big\langle \mathbf{h}(\widetilde{Z}_t, t), \mathbb{E}[Y - X \mid \widetilde{Z}_t, t] \Big\rangle\Big] \\
&= \mathbb{E}\Big[\Big\langle \mathbf{h}(\widetilde{Z}_t, t), Y - X \Big\rangle\Big].
\end{aligned}$$

□

A.6 Proof of Proposition 4.1

Proof. Since $\mu = \mathrm{Law}(\widetilde{Z}_t, t)$ with $t \sim \mathrm{Unif}[0, 1/2]$, the definition of ZFD gives

$$\mathrm{ZFD}(p, q) = \mathbb{E}_\mu\Big[\|\mathbf{g}(\widetilde{Z}_t, t)\|^2\Big] = \|\mathbf{g}\|_\mu^2.$$

Hence, the assumption $\mathrm{ZFD}(p, q) < \infty$ implies $\mathbf{g} \in L^2(\mu; \mathbb{R}^d)$.

For any $\mathbf{h} \in L^2(\mu; \mathbb{R}^d)$ with $\|\mathbf{h}\|_\mu > 0$, the Cauchy-Schwarz inequality yields

$$\langle \mathbf{h}, \mathbf{g} \rangle_\mu^2 \le \|\mathbf{h}\|_\mu^2 \|\mathbf{g}\|_\mu^2.$$

Therefore,

$$\frac{\langle \mathbf{h}, \mathbf{g} \rangle_\mu^2}{\|\mathbf{h}\|_\mu^2} \le \|\mathbf{g}\|_\mu^2 = \mathrm{ZFD}(p, q),$$

and hence

$$\sup_{\substack{\mathbf{h} \in L^2(\mu; \mathbb{R}^d) \\ \|\mathbf{h}\|_\mu > 0}} \frac{\langle \mathbf{h}, \mathbf{g} \rangle_\mu^2}{\|\mathbf{h}\|_\mu^2} \le \mathrm{ZFD}(p, q).$$

If $\mathrm{ZFD}(p, q) > 0$, then $\|\mathbf{g}\|_\mu > 0$. Taking $\mathbf{h} = c\mathbf{g}$ for any $c \neq 0$ gives

$$\frac{\langle c\mathbf{g}, \mathbf{g} \rangle_\mu^2}{\|c\mathbf{g}\|_\mu^2} = \frac{c^2 \|\mathbf{g}\|_\mu^4}{c^2 \|\mathbf{g}\|_\mu^2} = \|\mathbf{g}\|_\mu^2 = \mathrm{ZFD}(p, q).$$

Thus the upper bound is attained by any nonzero rescaling of $\mathbf{g}$, and

$$\mathrm{ZFD}(p, q) = \sup_{\substack{\mathbf{h} \in L^2(\mu; \mathbb{R}^d) \\ \|\mathbf{h}\|_\mu > 0}} \frac{\langle \mathbf{h}, \mathbf{g} \rangle_\mu^2}{\|\mathbf{h}\|_\mu^2}.$$

If $\mathrm{ZFD}(p, q) = 0$, then $\|\mathbf{g}\|_\mu = 0$, so $\mathbf{g} = 0$ μ-almost everywhere. Consequently,

$$\langle \mathbf{h}, \mathbf{g} \rangle_\mu = 0$$

for every admissible $\mathbf{h}$, and the variational objective is identically zero. This proves the claim. □

A.7 Proof of Proposition 4.2

Proof. For brevity, write

$$\mathfrak{s}_i := \mathfrak{s}_{\mathbf{h}}(X_i, Y_i, t_i, B_i), \qquad i = 1, \dots, n.$$

Since $\mathbf{h}$ is fixed and $\{(X_i, Y_i, t_i, B_i)\}_{i=1}^n$ are i.i.d., the random variables $\{\mathfrak{s}_i\}_{i=1}^n$ are also i.i.d. By Lemma A.1,

$$\mathbb{E}[\mathfrak{s}_i] = \tau(\mathbf{h}), \qquad \mathrm{Var}(\mathfrak{s}_i) = \sigma^2(\mathbf{h}).$$

Since $0 < \sigma^2(\mathbf{h}) < \infty$, the classical central limit theorem gives

$$\frac{\sqrt{n}\big(\hat{\tau}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\sigma(\mathbf{h})} \xrightarrow{d} \mathcal{N}(0,1).$$

Moreover, finite variance implies $\mathbb{E}[\mathfrak{s}_i^2] < \infty$. Hence, by the law of large numbers, the sample variance is consistent:

$$\hat{\sigma}^2(\mathbf{h}) \xrightarrow{p} \sigma^2(\mathbf{h}).$$

Since $\sigma^2(\mathbf{h}) > 0$,

$$\frac{\hat{\sigma}(\mathbf{h})}{\sigma(\mathbf{h})} \xrightarrow{p} 1.$$

Applying Slutsky's theorem therefore yields

$$\frac{\sqrt{n}\big(\hat{\tau}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\hat{\sigma}(\mathbf{h})} = \frac{\sqrt{n}\big(\hat{\tau}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\sigma(\mathbf{h})} \frac{\sigma(\mathbf{h})}{\hat{\sigma}(\mathbf{h})} \xrightarrow{d} \mathcal{N}(0,1),$$

which proves the claim. □

A.8 Proof of Proposition 4.3

Proof. Recall that $\hat{\tau}_{\mathrm{te}}(\mathbf{h})$ and $\hat{\sigma}_{\mathrm{te}}(\mathbf{h})$ are the empirical mean and standard deviation computed from the held-out test split $\mathcal{S}_{\mathrm{te}}$. We decompose

$$\zeta_{\mathrm{ZF}} = \frac{\sqrt{n_{\mathrm{te}}}\big(\hat{\tau}_{\mathrm{te}}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\hat{\sigma}_{\mathrm{te}}(\mathbf{h})} + \sqrt{n_{\mathrm{te}}}\,\frac{\tau(\mathbf{h})}{\hat{\sigma}_{\mathrm{te}}(\mathbf{h})}.$$

By Proposition 4.2,

$$\frac{\sqrt{n_{\mathrm{te}}}\big(\hat{\tau}_{\mathrm{te}}(\mathbf{h}) - \tau(\mathbf{h})\big)}{\hat{\sigma}_{\mathrm{te}}(\mathbf{h})} \xrightarrow{d} \mathcal{N}(0,1).$$

Following the power analysis in Kübler et al. (2022b, Eq. (4)), we additionally assume that the population standard deviation $\sigma(\mathbf{h})$ is known for the Gaussian power approximation. Under this approximation,

$$\zeta_{\mathrm{ZF}} \approx Z + \sqrt{n_{\mathrm{te}}}\,\frac{\tau(\mathbf{h})}{\sigma(\mathbf{h})} = Z + \sqrt{n_{\mathrm{te}}}\,\mathrm{SNR}(\mathbf{h}), \qquad Z \sim \mathcal{N}(0,1).$$

Therefore,

$$\Pr\left(\zeta_{\mathrm{ZF}} \geq \Phi^{-1}(1-\alpha)\right) \approx 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \sqrt{n_{\mathrm{te}}}\,\mathrm{SNR}(\mathbf{h})\right),$$

which gives Equation (13).

We next establish consistency directly, without relying on the Gaussian power approximation. Under any fixed alternative satisfying $\tau(\mathbf{h}) > 0$, the law of large numbers gives

$$\hat{\tau}_{\mathrm{te}}(\mathbf{h}) \xrightarrow{p} \tau(\mathbf{h}), \qquad \hat{\sigma}_{\mathrm{te}}(\mathbf{h}) \xrightarrow{p} \sigma(\mathbf{h}).$$

Hence,

$$\frac{\zeta_{\mathrm{ZF}}}{\sqrt{n_{\mathrm{te}}}} = \frac{\hat{\tau}_{\mathrm{te}}(\mathbf{h})}{\hat{\sigma}_{\mathrm{te}}(\mathbf{h})} \xrightarrow{p} \frac{\tau(\mathbf{h})}{\sigma(\mathbf{h})} = \mathrm{SNR}(\mathbf{h}) > 0.$$

It follows that

$$\zeta_{\mathrm{ZF}} \xrightarrow{p} +\infty,$$

and therefore

$$\Pr\left(\zeta_{\mathrm{ZF}} \geq \Phi^{-1}(1-\alpha)\right) \longrightarrow 1.$$

□

A.9 FINITE-SAMPLE VALIDITY AND CONSISTENCY OF THE WILD BOOTSTRAP

We here establish the finite-sample validity and consistency of the wild-bootstrap calibration used in **Stage-II** of Section 4.3.

Proposition A.1 (Validity and consistency of wild-bootstrap calibration)**.** *Suppose the witness* $\mathbf{h}$ *is fixed and independent of the held-out test split* $\mathcal{S}_{\mathrm{te}}$*. Under* $H_0 : p = q$*, the exhaustive wild-bootstrap calibration based on* $\hat{\tau}_{\mathrm{te}}(\mathbf{h})$ *controls the finite-sample Type-I error at level* α*. The same guarantee holds for the corrected Monte Carlo* p*-value* $\hat{p}_{\mathrm{val}}$ *in Equation* (16)*.*

Under any fixed alternative satisfying the positive alignment $\tau(\mathbf{h}) > 0$ *with* $\sigma^2(\mathbf{h}) < \infty$*, the exhaustive wild-bootstrap test is consistent, i.e., its power converges to one as* $n_{\mathrm{te}} \to \infty$*. The corrected Monte Carlo test is also consistent provided* $\frac{1}{n_{\mathrm{perm}}+1} \le \alpha$*.*

Proof. Condition on the fixed witness $\mathbf{h}$, which is independent of the held-out test split $\mathcal{S}_{\mathrm{te}}$. For brevity, write

$$\mathfrak{s}_i := \mathfrak{s}_{\mathbf{h}}(X_i, Y_i, t_i, B_i), \qquad i = 1, \dots, n_{\mathrm{te}}.$$

We first establish finite-sample validity under H_0. For $\epsilon \in \{-1, +1\}$, define the pairwise transformation

$$\mathrm{swap}_\epsilon(X, Y, t, B) := \begin{cases} (X, Y, t, B), & \epsilon = +1, \\ (Y, X, t, 1 - B), & \epsilon = -1. \end{cases}$$

For a sign vector

$$\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_{n_{\mathrm{te}}}) \in \{-1, +1\}^{n_{\mathrm{te}}},$$

let $\mathrm{swap}_{\boldsymbol{\epsilon}}$ apply $\mathrm{swap}_{\epsilon_i}$ independently to the i-th held-out tuple (X_i, Y_i, t_i, B_i).

Under $H_0 : p = q$, X_i and Y_i are i.i.d., and $B_i \sim \mathrm{Ber}(1/2)$. Hence, for every i,

$$(X_i, Y_i, t_i, B_i) \overset{d}{=} (Y_i, X_i, t_i, 1 - B_i).$$

Since the held-out tuples are independent, for every $\boldsymbol{\epsilon} \in \{-1, +1\}^{n_{\mathrm{te}}}$,

$$\mathrm{swap}_{\boldsymbol{\epsilon}}\left(\{(X_i, Y_i, t_i, B_i)\}_{i=1}^{n_{\mathrm{te}}}\right) \overset{d}{=} \{(X_i, Y_i, t_i, B_i)\}_{i=1}^{n_{\mathrm{te}}}.$$

For each transformed tuple, the randomized interpolation state is unchanged while the displacement is reversed, and therefore

$$\mathfrak{s}_{\mathbf{h}}(Y_i, X_i, t_i, 1 - B_i) = -\mathfrak{s}_{\mathbf{h}}(X_i, Y_i, t_i, B_i).$$

Consequently,

$$\mathrm{swap}_{\boldsymbol{\epsilon}} : (\mathfrak{s}_1, \dots, \mathfrak{s}_{n_{\mathrm{te}}}) \longmapsto (\epsilon_1 \mathfrak{s}_1, \dots, \epsilon_{n_{\mathrm{te}}} \mathfrak{s}_{n_{\mathrm{te}}}).$$

Thus, applying $\mathrm{swap}_{\boldsymbol{\epsilon}}$ to the held-out data is exactly equivalent to multiplying the pre-computed score vector by Rademacher signs, which is precisely the wild-bootstrap replicate in Equation (16). Since the transformed held-out sample has the same distribution as the original sample under H_0, the exhaustive wild-bootstrap calibration is an exact finite randomization test and controls the Type-I error at level α.

When only n_{perm} Rademacher sign vectors are sampled, the corrected Monte Carlo p-value in Equation (16) preserves finite-sample level control by the standard randomization-test argument (Hemerik & Goeman, 2018).

We next establish consistency under a fixed alternative. Recall that $\mathbb{E}[\mathfrak{s}_i] = \tau(\mathbf{h}) > 0$. Hence, by the law of large numbers,

$$\hat{\tau}_{\mathrm{te}}(\mathbf{h}) = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \mathfrak{s}_i \xrightarrow{p} \tau(\mathbf{h}) > 0.$$

For a wild-bootstrap replicate, let $\epsilon_1, \dots, \epsilon_{n_{\mathrm{te}}}$ be i.i.d. Rademacher variables and define

$$\hat{\tau}^*_{\mathrm{te}}(\mathbf{h}) := \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \epsilon_i \mathfrak{s}_i.$$

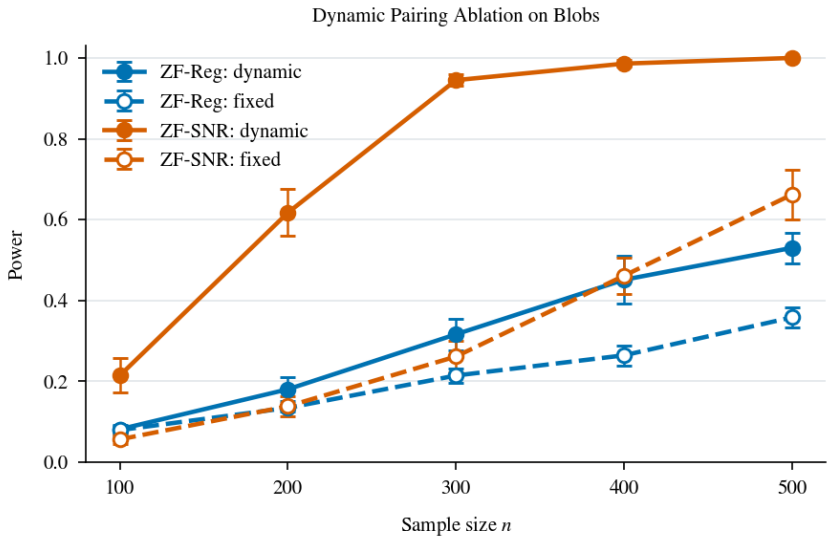

Figure 2: Dynamic pairing ablation for both tests on the `Blobs` dataset.

Conditionally on the held-out scores, we have

$$\mathbb{E}_\epsilon\big[\hat{\tau}^*_{\mathrm{te}}(\mathbf{h})\big] = 0, \quad \mathrm{Var}_\epsilon\big[\hat{\tau}^*_{\mathrm{te}}(\mathbf{h})\big] = \frac{1}{n_{\mathrm{te}}^2}\sum_{i=1}^{n_{\mathrm{te}}} \mathfrak{s}_i^2.$$

Since $\mathbb{E}[\mathfrak{s}_i^2] = \sigma^2(\mathbf{h}) + \tau^2(\mathbf{h}) < \infty$, the law of large numbers gives

$$\frac{1}{n_{\mathrm{te}}}\sum_{i=1}^{n_{\mathrm{te}}} \mathfrak{s}_i^2 = O_p(1),$$

and hence

$$\mathrm{Var}_\epsilon\big[\hat{\tau}^*_{\mathrm{te}}(\mathbf{h})\big] = O_p(n_{\mathrm{te}}^{-1}).$$

Therefore, the wild-bootstrap distribution concentrates around zero, whereas the observed statistic $\hat{\tau}_{\mathrm{te}}(\mathbf{h})$ converges to the strictly positive population alignment $\tau(\mathbf{h})$. It follows that the power of the exhaustive wild-bootstrap test converges to one.

Finally, consider the corrected Monte Carlo calibration with a fixed n_{perm}. Each sampled wild-bootstrap replicate also concentrates around zero, while $\hat{\tau}_{\mathrm{te}}(\mathbf{h}) \xrightarrow{p} \tau(\mathbf{h}) > 0$. Consequently,

$$\hat{p}_{\mathrm{val}} \xrightarrow{p} \frac{1}{n_{\mathrm{perm}} + 1}.$$

Therefore, whenever

$$\frac{1}{n_{\mathrm{perm}} + 1} \le \alpha,$$

the corrected Monte Carlo test is also consistent. □

B Implementation Details in ZF2ST

B.1 Full Algorithm

We summarize the full ZF2ST procedure in Algorithm 1.

Algorithm 1 Zero-Flow Two-Sample Test (ZF2ST).

Stage I: ZF-Reg

Input: $\mathcal{S}_p, \mathcal{S}_q$, level α
Split: $\mathcal{S}_p = \mathcal{S}_p^{\mathrm{tr}} \cup \mathcal{S}_p^{\mathrm{te}}, \mathcal{S}_q = \mathcal{S}_q^{\mathrm{tr}} \cup \mathcal{S}_q^{\mathrm{te}}$.
Initialize witness $\mathbf{h}_\theta$
for $r = 1, \ldots, N_{\mathrm{iter}}$ **do**
  $\{(X_i, Y_i)\}_{i=1}^m \leftarrow \mathrm{DP}(\mathcal{S}_p^{\mathrm{tr}}, \mathcal{S}_q^{\mathrm{tr}}, m)$ ▷ Alg. 3
  Sample $t_{1:m} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, 1/2]$
  Sample $B_{1:m} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2)$
  Form $\{(X_i, Y_i, t_i, B_i)\}_{i=1}^m$
  Compute $\widehat{\mathcal{R}}_{\mathrm{reg}}(\mathbf{h}_\theta)$ using (14)
  Update $\mathbf{h}_\theta$: $\theta \leftarrow \theta - \eta \nabla_{\mathrm{Adam}} \widehat{\mathcal{R}}_{\mathrm{reg}}(\mathbf{h}_\theta)$
end for
$\mathbf{h} \leftarrow \mathbf{h}_\theta$ ▷ Fixed witness

Stage I: ZF-SNR

Input: $\mathcal{S}_p, \mathcal{S}_q$, level α
Split: $\mathcal{S}_p = \mathcal{S}_p^{\mathrm{tr}} \cup \mathcal{S}_p^{\mathrm{te}}, \mathcal{S}_q = \mathcal{S}_q^{\mathrm{tr}} \cup \mathcal{S}_q^{\mathrm{te}}$.
Initialize witness $\mathbf{h}_\theta$
for $r = 1, \ldots, N_{\mathrm{iter}}$ **do**
  $\{(X_i, Y_i)\}_{i=1}^m \leftarrow \mathrm{DP}(\mathcal{S}_p^{\mathrm{tr}}, \mathcal{S}_q^{\mathrm{tr}}, m)$ ▷ Alg. 3
  Sample $t_{1:m,1:K} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, 1/2]$
  # Branch B is averaged out
  Form $\{(X_i, Y_i, t_{i,1:K})\}_{i=1}^m$
  Compute $\widehat{\mathcal{R}}_{\mathrm{snr},K}(\mathbf{h}_\theta)$ using (23)
  Update $\mathbf{h}_\theta$: $\theta \leftarrow \theta - \eta \nabla_{\mathrm{Adam}} \widehat{\mathcal{R}}_{\mathrm{snr},K}(\mathbf{h}_\theta)$
end for
$\mathbf{h} \leftarrow \mathbf{h}_\theta$ ▷ Fixed witness

Stage II: Wild bootstrap calibration

$\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}} \leftarrow (\mathcal{S}_p^{\mathrm{te}}, \mathcal{S}_q^{\mathrm{te}})$
Sample $t_{1:n_{\mathrm{te}},1:K} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, 1/2]$
Form $\{(X_i, Y_i, t_{i,1:K})\}_{i=1}^{n_{\mathrm{te}}}$
$\mathfrak{s}_i \leftarrow \mathfrak{s}_{\mathbf{h},K}(X_i, Y_i, t_{i,1:K}), \quad i = 1, \ldots, n_{\mathrm{te}}$
$\hat{\tau}_{\mathrm{te}}(\mathbf{h}) \leftarrow \dfrac{1}{n_{\mathrm{te}}} \displaystyle\sum_{i=1}^{n_{\mathrm{te}}} \mathfrak{s}_i$
for $b = 1, \ldots, n_{\mathrm{perm}}$ **do**
  Draw $\epsilon_{1:n_{\mathrm{te}}}^{(b)} \overset{\text{i.i.d.}}{\sim} \{-1, +1\}$ with equal prob.
  $\hat{\tau}_{\mathrm{te}}^{(b)}(\mathbf{h}) \leftarrow \dfrac{1}{n_{\mathrm{te}}} \displaystyle\sum_{i=1}^{n_{\mathrm{te}}} \epsilon_i^{(b)} \mathfrak{s}_i$
end for
$\hat{p}_{\mathrm{val}} \leftarrow \dfrac{1 + \sum_{b=1}^{n_{\mathrm{perm}}} \mathbf{1}\left\{\hat{\tau}_{\mathrm{te}}^{(b)}(\mathbf{h}) \geq \hat{\tau}_{\mathrm{te}}(\mathbf{h})\right\}}{n_{\mathrm{perm}} + 1}$
return 1 **if** $\hat{p}_{\mathrm{val}} \leq \alpha$ **else** 0

Algorithm 2 Fixed pairing (FP).

```
def fixed_pairing(
    x, y, batch_size
):
    # shared indices for x and
y
    idx = randint(
        len(x), batch_size
    )

    xb = x[idx]
    yb = y[idx]

    return xb, yb
```

Algorithm 3 Dynamic pairing (DP).

```
def dynamic_pairing(
    x, y, batch_size
):
    # independently sample x
and y
    ix = randint(
        len(x), batch_size
    )
    iy = randint(
        len(y), batch_size
    )

    xb = x[ix]
    yb = y[iy]

    return xb, yb
```

Figure 3: Comparison between Fixed pairing and Dynamic Pairing

B.2 Training Details: Dynamic Pairing

During witness training, we *dynamically* construct cross-pairs by sampling the indices of the two training sets independently with replacement, as illustrated in Algorithm 3. The pairing of endpoint samples can be viewed as a choice of coupling, with standard flow-matching constructions typically relying on independent sampling from the two marginals (Pooladian et al., 2023). Let

$$\hat{p}_{\mathrm{tr}} := \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \delta_{X_i}, \qquad \hat{q}_{\mathrm{tr}} := \frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \delta_{Y_j}.$$

Independent index sampling then draws each training pair directly from the empirical product coupling

$$\hat{p}_{\mathrm{tr}} \otimes \hat{q}_{\mathrm{tr}} = \frac{1}{n_{\mathrm{tr}}^2} \sum_{i=1}^{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \delta_{(X_i, Y_j)},$$

which directly mirrors the independent population coupling $p \otimes q$ used in ZF2ST. Re-sampling the indices at every optimization iteration explores different cross-pair combinations without explicitly constructing all n_{tr}^2 possible pairs. Since the indices are sampled *with replacement* from the empirical product coupling, the minibatch size is not restricted by n_{tr} and may exceed the number of training samples, allowing more cross-pair evaluations within each optimization step.

We further examine the effect of dynamic pairing on the `Blobs` dataset. As shown in Figure 2, dynamically resampling cross-pairs consistently improves the test power of both ZF-Reg and ZF-SNR compared with a fixed one-to-one pairing, with a more pronounced gain for ZF-SNR.

Importantly, dynamic pairing is used *only* during training. At test time, the held-out pairs are kept fixed so that the resulting scores remain i.i.d., as assumed in Proposition 4.2. Averaging over cross-pairs at test time would instead lead to a two-sample U-statistic, requiring a different asymptotic analysis (Hoeffding, 1992; Serfling, 2009).

B.3 An Explicit Form of the Empirical Risk for the Variance-Reduced Objective

We explicitly write out the empirical risk used for the variance-reduced ZF-SNR objective. Recall that

$$\mathfrak{s}_{\mathbf{h},K}(X, Y, t_{1:K}) := \frac{1}{2K} \sum_{k=1}^{K} \langle \mathbf{h}(Z_{t_k}, t_k) + \mathbf{h}(Z_{1-t_k}, t_k), Y - X \rangle .$$

The empirical alignment is

$$\hat{\tau}(\mathbf{h}) := \frac{1}{n} \sum_{i=1}^{n} \mathfrak{s}_{\mathbf{h},K}(X_i, Y_i, t_{i,1:K}). \tag{22}$$

The empirical variance is

$$\hat{\sigma}_K^2(\mathbf{h}) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathfrak{s}_{\mathbf{h},K}(X_i, Y_i, t_{i,1:K}) - \hat{\tau}(\mathbf{h})\right)^2 .$$

Define the K-sample witness energy

$$\mathcal{E}_{\mathbf{h},K}(X, Y, t_{1:K}) := \frac{1}{2K} \sum_{k=1}^{K} \left[\|\mathbf{h}(Z_{t_k}, t_k)\|^2 + \|\mathbf{h}(Z_{1-t_k}, t_k)\|^2 \right],$$

where

$$Z_{t_k} = (1 - t_k)X + t_k Y, \qquad Z_{1-t_k} = t_k X + (1 - t_k)Y.$$

The regularized denominator is

$$\hat{\sigma}_K^{\lambda}(\mathbf{h}) := \left[\hat{\sigma}_K^2(\mathbf{h}) + \frac{\lambda}{n} \sum_{i=1}^{n} \mathcal{E}_{\mathbf{h},K}(X_i, Y_i, t_{i,1:K}) \right]^{1/2}.$$

Thus, on the training split $\mathcal{S}^{\mathrm{tr}}$, the corresponding empirical risk is

$$\widehat{\mathcal{R}}_{\mathrm{snr},K}(\mathbf{h}) := -\frac{\hat{\tau}_{\mathrm{tr}}(\mathbf{h})}{\hat{\sigma}_{K,\mathrm{tr}}^{\lambda}(\mathbf{h})}. \tag{23}$$

Method	Training per update	Testing
ZF-Reg	$\mathcal{O}(mC_{\mathbf{h}})$	$\mathcal{O}(2Kn_{\mathrm{te}}C_{\mathbf{h}} + Bn_{\mathrm{te}})$
ZF-SNR	$\mathcal{O}(2KmC_{\mathbf{h}})$	$\mathcal{O}(2Kn_{\mathrm{te}}C_{\mathbf{h}} + Bn_{\mathrm{te}})$
MMD-D (Liu et al., 2020)	$\mathcal{O}(mC_\phi + m^2C_k)$	$\mathcal{O}(n_{\mathrm{te}}C_\phi + n_{\mathrm{te}}^2C_k + Bn_{\mathrm{te}}^2)$
C2ST (Lopez-Paz & Oquab, 2016)	$\mathcal{O}(mC_f)$	$\mathcal{O}(n_{\mathrm{te}}C_f + Bn_{\mathrm{te}})$
MMD-FUSE (Biggs et al., 2023)	–	$\mathcal{O}(cBn^2)$
MMDagg (Schrab et al., 2023)	–	$\mathcal{O}\big(c(B+B')n^2\big)$

Table 2: Leading-order computational complexity. Here m is the mini-batch size, n_{tr} and n_{te} are the training and testing sample sizes with $n = n_{\mathrm{tr}} + n_{\mathrm{te}}$, K is the number of time samples used by ZF2ST, B is the number of randomization/calibration replicates, and c is the number of candidate kernels. For MMDagg, B' denotes the additional randomization replicates used for aggregation-level correction. We denote by $C_{\mathbf{h}}$ the cost of one witness-network evaluation, C_ϕ the cost of one feature-network evaluation, C_k the cost of one kernel evaluation, and C_f the cost of one classifier evaluation.

B.4 Complexity of ZF2ST

Let m denote the mini-batch size, n_{te} the number of held-out pairs, K the number of time samples used in variance reduction, and B the number of wild-bootstrap replicates. Let $C_{\mathbf{h}}$ denote the cost of evaluating the witness network at one interpolation state; we suppress the fixed data dimension and constant factors.

Training complexity. For ZF-Reg, each training pair requires one witness evaluation for the randomly sampled interpolation state, giving a per-update cost of $\mathcal{O}(mC_{\mathbf{h}})$. For ZF-SNR, the variance-reduced objective evaluates both Z_{t_k} and Z_{1-t_k} for each of the K time samples, resulting in a per-update cost of $\mathcal{O}(2KmC_{\mathbf{h}})$. The empirical alignment, variance, and ridge terms require only linear-time reductions over these pre-computed witness outputs. Dynamic pairing independently samples the two sets of indices at every update and incurs an additional $\mathcal{O}(m)$ cost, without explicitly constructing the n_{tr}^2 empirical cross-pairs.

Testing complexity. Both ZF-Reg and ZF-SNR use the same variance-reduced score at test time. Pre-computing the held-out scores requires $\mathcal{O}(2Kn_{\mathrm{te}}C_{\mathbf{h}})$. Once these scores have been computed, each wild-bootstrap replicate requires only Rademacher sign flips and a reduction over the n_{te} scores, with no additional witness evaluations. The total testing and calibration cost is therefore $\mathcal{O}(2Kn_{\mathrm{te}}C_{\mathbf{h}} + Bn_{\mathrm{te}})$. In particular, the witness-evaluation cost does not scale with the number of wild-bootstrap replicates.

C Experimental Details

C.1 Protocols

All tests are conducted at significance level $\alpha = 0.05$. For sample-splitting methods, we perform 10 independent training runs, each followed by 100 independent held-out test trials with the trained model fixed. Unless otherwise specified, we use $n = n_{\mathrm{tr}} + n_{\mathrm{te}}$ with $n_{\mathrm{tr}} = n_{\mathrm{te}}$; sample-splitting methods train on n_{tr} samples and evaluate on disjoint n_{te} held-out samples, whereas non-splitting methods use all n samples in each of 100 independent trials. For CIFAR-10 vs. CIFAR-10.1, following Liu et al. (2020), we instead use $n_{\mathrm{tr}} = 1000$ and $n_{\mathrm{te}} = 1021$. Standard errors for sample-splitting methods are computed across the 10 independent training runs, while those for non-splitting methods are computed across the 100 independent trials. ZF2ST, MMD-D, C2ST, and KFDA-WiTS use 200 calibration randomizations, except AutoML, which uses its default 10,000 permutations. MMD-FUSE and MMDagg use the default calibration settings of their official implementations.

KFDA-WiTS (Kübler et al., 2022a) is evaluated only on non-image datasets, as neither the experiments in the paper nor the released code include an image-specific pipeline.

Table 3: Summary of the real-world datasets used in our experiments.

Dataset	Samples	Pixels	Data type	URL
Higgs	11000000	–	Tabular	https://archive.ics.uci.edu/dataset/280/higgs
CIFAR-10 (test)	10000	$3 \times 32 \times 32$	Image	https://www.cs.toronto.edu/~kriz/cifar.html
CIFAR-10.1 (v4_data)	2021	$3 \times 32 \times 32$	Image	https://github.com/modestyachts/CIFAR-10.1/tree/master/datasets
ImageNet (validation)	50000	Variable	Image	https://www.image-net.org/challenges/LSVRC/2012/2012-downloads
ImageNetV2 (MatchedFrequency)	10000	Variable	Image	https://huggingface.co/datasets/vaishaal/ImageNetV2/tree/main

C.2 Dataset Metadata and Preprocessing

For the synthetic Blobs and HDGM datasets, we follow the distribution specifications given in Liu et al. (2020, Table 6).

Table 3 briefly summarizes the real-world datasets used in our experiments. We additionally provide the code repositories and pretrained checkpoints used to generate CIFAR-10 samples from the generative models considered in our experiments.

- **ScoreSDE** (Song et al., 2020): the code is available at https://github.com/yang-song/score_sde_pytorch. We use the official pretrained checkpoint available at https://drive.google.com/drive/folders/11s6A_xM7qiztdj8AHQWqaIAUSC3I7uX2.
- **1-Rectified Flow** (Liu et al., 2022): the code is available at https://github.com/gnobitab/RectifiedFlow. We use the official pretrained checkpoint available at https://drive.google.com/file/d/10aPF5KC30SjVwr6rOnNosStpSGXnELXn/view.
- **MeanFlow** (Geng et al., 2025): the code is available at https://github.com/Gsunshine/py-meanflow. We use the official pretrained checkpoint available at https://drive.google.com/uc?id=1Ev9aZdt7hNVkOfaTQ7rrLAqs3jA_vnvj.
- **StyleGAN-XL** (Sauer et al., 2022): the code is available at https://github.com/autonomousvision/stylegan-xl. We use the official pretrained checkpoint available at https://s3.eu-central-1.amazonaws.com/avg-projects/stylegan_xl/models/cifar10.pkl.

We note that StyleGAN-XL is trained conditionally on CIFAR-10 class labels. Following the official generation procedure, we generate samples without specifying a class label, so the labels are sampled uniformly by the released code. For the Rectified Flow experiments that vary the sample size and number of function evaluations (NFE), we use the official Euler solver. For all remaining experiments, we keep the official sampling configurations unchanged.

Visualizations of all image datasets used in our experiments, including generated samples, are provided in Appendix E.

Image preprocessing. Following Liu et al. (2020), all CIFAR-10, CIFAR-10.1, and generated CIFAR-10 images are kept in RGB format, resized from 32×32 to 64×64, and normalized channel-wise from $[0, 1]$ to $[-1, 1]$ using mean and standard deviation $(0.5, 0.5, 0.5)$, yielding $3 \times 64 \times 64$ inputs.

For ImageNet and ImageNetV2, we preserve the RGB channels, resize each image so that its shorter side is 256 pixels, and apply a 224×224 center crop. The images are then normalized using the ImageNet channel-wise mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$, and finally resized to $3 \times 64 \times 64$.

C.3 Configurations

All experiments are conducted on an NVIDIA RTX 5090 GPU, except KFDA-WiTS and AutoML, which are run on an AMD Ryzen 7 9700X 8-Core CPU. We run MMD-D, C2ST-S, and C2ST-L

using the official code provided by Liu et al. (2020). For C2ST, we train a single shared classifier and compute both C2ST-S and C2ST-L statistics from its outputs. For KFDA-WiTS, we use the official implementation (Kübler et al., 2022a) but replace the FALKON-based solver with a full Gram-matrix implementation for compatibility. AutoML is run using the official Python package `autotst` (Kübler et al., 2022b). We use the official Python packages `mmdagg` and `mmdfuse` released by Schrab et al. (2023) and Biggs et al. (2023), respectively.

For the non-image experiments, we keep the original network configurations of MMD-D, C2ST-S, and C2ST-L from Liu et al. (2020) unchanged whenever applicable. For both ZF-Reg and ZF-SNR, we *match* the five-layer fully connected architecture used by MMD-D on `Blobs`, `HDGM`, and `Higgs`, using the same hidden widths of 50, $3 \times d$, and 20, respectively, where d denotes the data dimension. The only architectural changes are to concatenate the scalar time t to the input, replace the Softplus activation with SiLU, and map the final hidden representation to the d-dimensional witness through a linear output layer.

For all image experiments, we preprocess the images to $3 \times 64 \times 64$ inputs (see Appendix C.2) and keep the original CIFAR-10 network configurations of MMD-D, C2ST-S, and C2ST-L from Liu et al. (2020, Figures 8 and 9) unchanged. Both ZF-Reg and ZF-SNR use the same convolutional witness architecture across all image experiments. As shown in Figure 4, the network consists of an initial convolutional layer, a residual block with Group Normalization (GN; Wu & He, 2018) and SiLU activations, and a final convolutional layer that outputs the witness.

For MMD-D, C2ST-S, C2ST-L, ZF-Reg, and ZF-SNR, we use the Adam optimizer (Kingma & Ba, 2014) for training in all experiments.

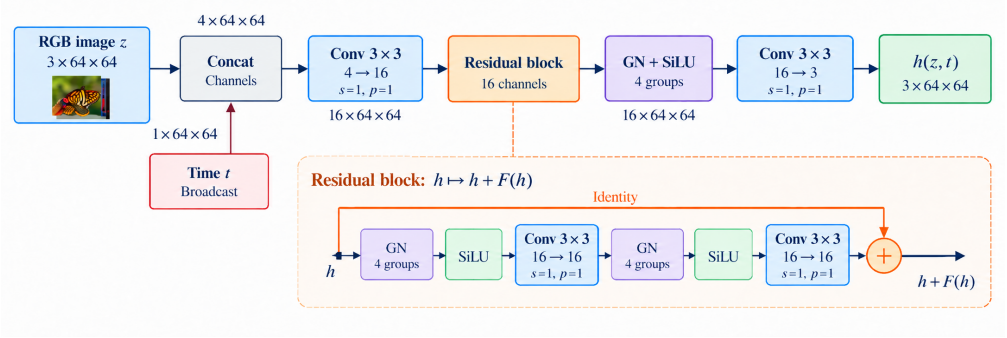

Figure 4: CNN of witness $\mathbf{h}$ used in *all* image experiments. All convolutional layers use a 3×3 kernel with stride $s = 1$ and padding $p = 1$.

C.4 Hyperparameters of test methods

Following Liu et al. (2020), we use the default Adam hyperparameters except for the learning rate. Both ZF-Reg and ZF-SNR are trained for $2{,}000$ iterations in all experiments. For `Blobs`, `HDGM`, and `Higgs`, we use learning rates of 5×10^{-3}, 10^{-3}, and 3×10^{-3}, respectively, with batch size $\min(2048, 6 \times n_{\mathrm{tr}})$. For all image experiments, we use a learning rate of 5×10^{-4} and batch size 100. We use one validation set of the same size as the training set to select the above hyperparameters and λ.

For ZF-SNR, we use the variance-reduced objective throughout, with $(K, \lambda) = (3, 10^{-3})$ for non-image experiments and $(2, 10^{-2})$ for image experiments. The same values of K are used for the variance-reduced test score of both ZF-Reg and ZF-SNR.

For MMD-D, C2ST-S, and C2ST-L, we keep all hyperparameters and their selection rules unchanged from the original implementations in Liu et al. (2020).

For AutoML, we set the training time parameter `time_limit` to 60 seconds.

C.5 Type-I Errors

The corresponding Type-I error results are reported in Figure 5 and Table 4.

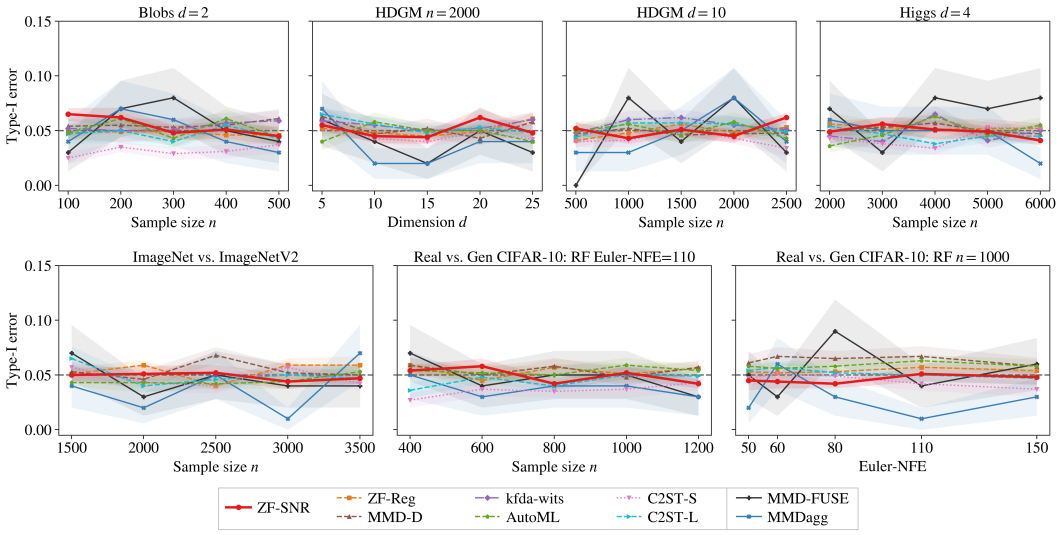

Figure 5: Type-I error results at significance level $\alpha = 0.05$ for the corresponding experimental settings. See Figure 1 for the associated power curves.

(a) CIFAR-10 vs. CIFAR-10.1

Method	Type-I
ZF-Reg	0.052 ± 0.008
ZF-SNR	0.052 ± 0.006
MMD-D	0.046 ± 0.007
C2ST-S	0.052 ± 0.006
C2ST-L	0.048 ± 0.005
MMD-FUSE	0.060 ± 0.024
MMDagg	0.050 ± 0.022
AutoML	0.063 ± 0.007

(b) Generative model experiments

Method	Type-I
ZF-Reg	0.051 ± 0.009
ZF-SNR	0.052 ± 0.006
MMD-D	0.048 ± 0.008
C2ST-S	0.037 ± 0.008
C2ST-L	0.050 ± 0.007
MMD-FUSE	0.060 ± 0.024
MMDagg	0.050 ± 0.022
AutoML	0.057 ± 0.006

Table 4: Type-I error at significance level $\alpha = 0.05$. Results are reported as mean $\pm$ one standard error. For the generative-model experiments in Table 1b, we report a single Type-I error result since all experiments share the same null CIFAR-10 distribution.

C.6 RUNTIMES

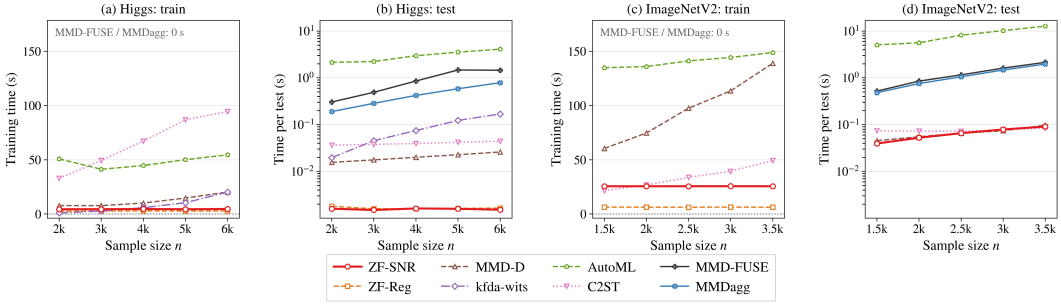

Figure 6: Training and testing runtimes on `Higgs` and ImageNet vs. ImageNetV2 as the sample size varies. Training time is averaged over 10 independent training runs, while testing time is averaged over 100 independent test trials. MMD-FUSE and MMDagg require no separate training stage and are therefore assigned zero training time. Testing times are shown on logarithmic scales.

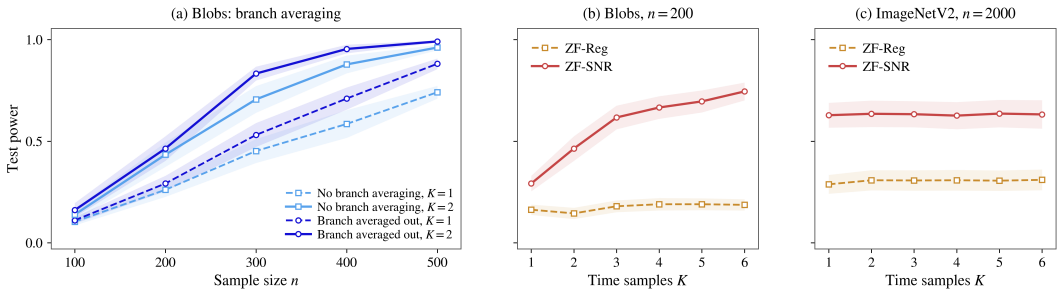

Figure 7: Ablation of the variance-reduction strategy. (a) ZF-SNR test power on `Blobs` with and without averaging out the branch variable B, for $K = 1$ and $K = 2$. (b) Effect of the number of time samples K on `Blobs` with $n = 200$. (c) Effect of K on ImageNet vs. ImageNetV2 with $n = 2000$. Lines show mean test power and shaded regions indicate ± 1 standard error.

We report the training and testing runtimes in Figure 6. ZF-Reg and ZF-SNR evaluate the learned witness on the held-out samples in a single pass, after which the wild-bootstrap randomizations reuse the precomputed scores without additional witness evaluations, resulting in relatively low testing overhead. C2ST-S and C2ST-L share the same trained classifier, so we report their runtime jointly as C2ST.

Among the non-splitting methods, MMDagg is faster than MMD-FUSE in our experiments. In the equal-sample-size setting, the default MMDagg implementation uses wild-bootstrap calibration, whereas MMD-FUSE uses label permutations; differences in bandwidth selection and first-call overhead, including `JAX` compilation where applicable, may also affect the observed runtimes.

C2ST uses epoch-based training with small minibatches, so its number of parameter updates increases with sample size, whereas ZF-Reg and ZF-SNR use a fixed budget of 2,000 updates. AutoML is configured with `time_limit=60`, but we report the elapsed time of the complete training call, which may exceed 60 seconds. Overall, the reported wall-clock times reflect both computational costs and the optimization and implementation choices used by each method.

D Ablation Study

D.1 Variance Reduction

We ablate the two components of the variance-reduction strategy introduced in Section 4.3: averaging out the branch variable B and averaging over multiple time samples K. As shown in Figure 7(a), branch averaging consistently improves test power on `Blobs` for both $K = 1$ and $K = 2$, while increasing K provides an additional improvement. Figure 7(b) further shows that increasing K substantially improves ZF-SNR on `Blobs`, with diminishing gains beyond a small number of time samples, whereas ZF-Reg is less sensitive to K. On ImageNet vs. ImageNetV2, both methods are relatively stable across the tested values of K (Figure 7(c)). Overall, these results support using a small number of time samples: we set $K = 3$ for non-image experiments to capture most of the observed gain, and $K = 2$ for image experiments where larger K provides little additional benefit.

D.2 Network Architecture

For the non-image experiments, ZF-Reg and ZF-SNR use fully connected networks with hidden-layer widths matched to those of MMD-D and C2ST, with adaptations to the input, activation functions, and output layer for the time-conditioned witness (see Appendix C.3). To further examine whether the strong image performance of ZF2ST can be attributed to its network architecture, we replace the original CNN backbones of MMD-D and C2ST with an adapted version of the ZF2ST CNN in Figure 4, and evaluate all three baseline tests on ImageNet vs. ImageNetV2 with $n = 2000$.

The adapted backbone retains the 16-channel input convolution, one residual block, and the GroupNorm and SiLU layers. We remove the time-conditioning channel so that the network takes only RGB images as input, and replace the vector-field output layer with adaptive average pooling to a

Table 5: Backbone ablation on ImageNet vs. ImageNetV2 ($n = 2000$, $\alpha = 0.05$). Test power (%) is reported as mean $\pm$ SE over 10 training runs, each with 100 test trials. MMD-D and C2ST use either their original CNN backbones (Liu et al., 2020, Figures 8 and 9) or the adapted ZF2ST backbone described above. ZF-Reg and ZF-SNR use their native CNN shown in Figure 4.

Method	Official backbone	ZF2ST backbone
MMD-D	11.4 ± 1.0	7.0 ± 0.6
C2ST-S	6.2 ± 0.8	10.0 ± 2.3
C2ST-L	6.2 ± 1.0	12.5 ± 2.5
ZF-Reg	–	30.8 ± 5.0
ZF-SNR	–	$\mathbf{63.5 \pm 6.5}$

4×4 spatial grid. The resulting 256-dimensional representation is passed to each baseline's original fully connected head, with the input dimension adjusted accordingly. For MMD-D, the head outputs 300-dimensional features; for C2ST, it consists of a 300-unit hidden layer with ReLU activation followed by a two-class softmax output so that both match the output dimension Liu et al. (2020, Fig. 8 & 9). For each baseline, the training objective, optimization settings, and testing procedure are held fixed across the two backbone variants. The total numbers of parameters, including the output heads, are $712{,}588/82{,}284$ for MMD-D and $713{,}190/82{,}886$ for C2ST using the official/adapted backbones, respectively. ZF-Reg and ZF-SNR each use $5{,}763$ parameters.

As shown in Table 5, the adapted ZF2ST backbone does not improve MMD-D, while both C2ST-S and C2ST-L achieve moderate gains in test power. Together with the substantially smaller parameter count of ZF2ST, these results suggest that its performance advantage is *not* due to using a larger network.

E DATASETS VISUALIZATION

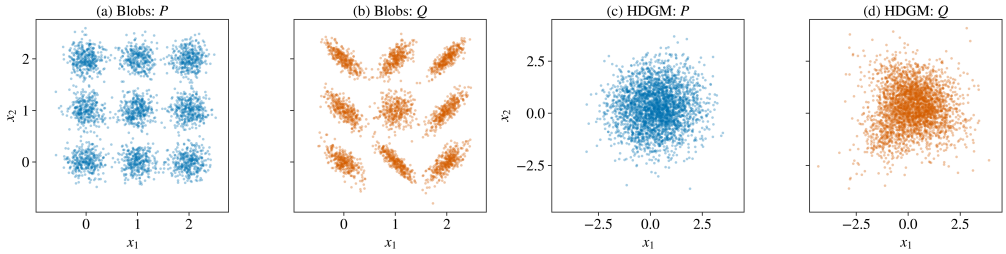

Figure 8: Illustration of the synthetic `Blobs` and `HDGM` distributions. For `Blobs`, p and q differ through local covariance structures across a 3×3 grid of Gaussian components. For `HDGM`, the distributional difference is confined to the first two dimensions.

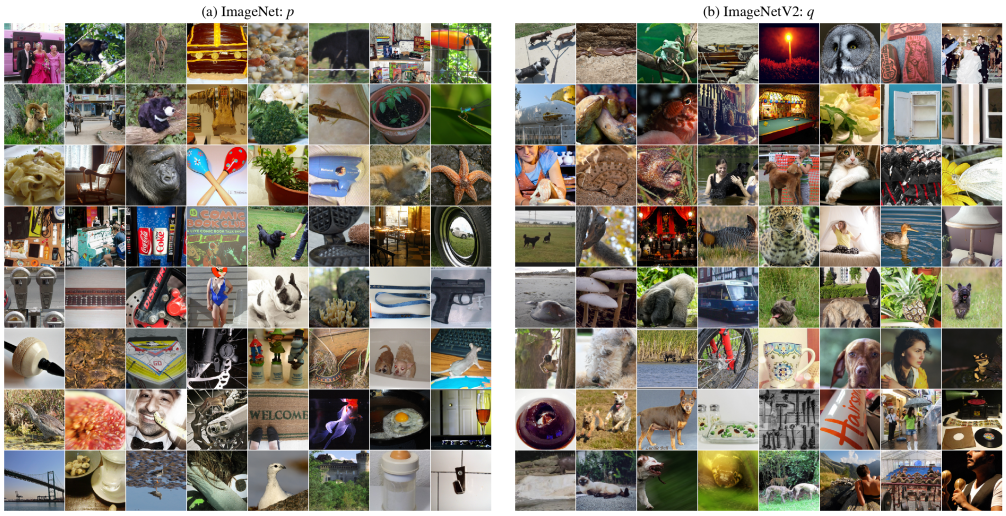

Figure 9: **Left:** ImageNet; **Right:** ImageNetV2

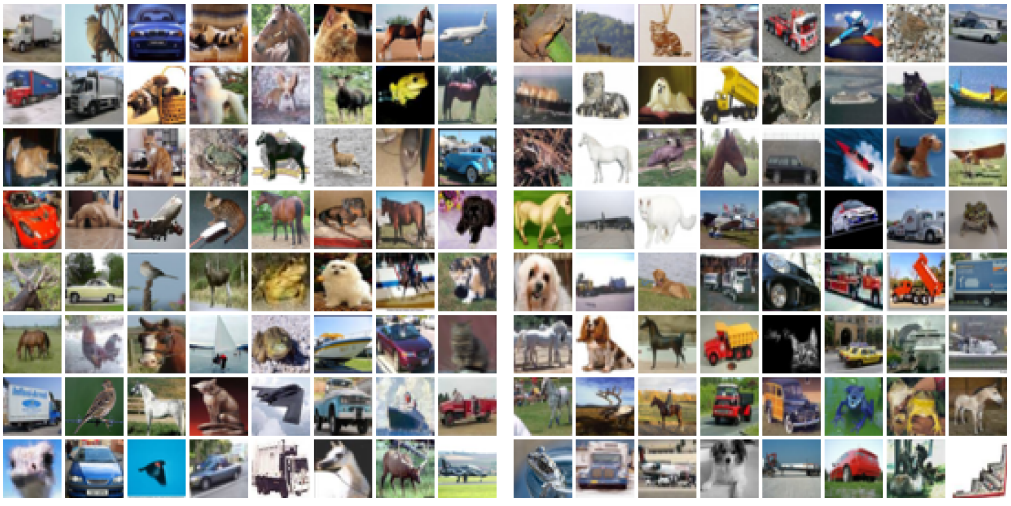

Figure 10: **Left:** CIFAR-10; **Right:** CIFAR-10.1

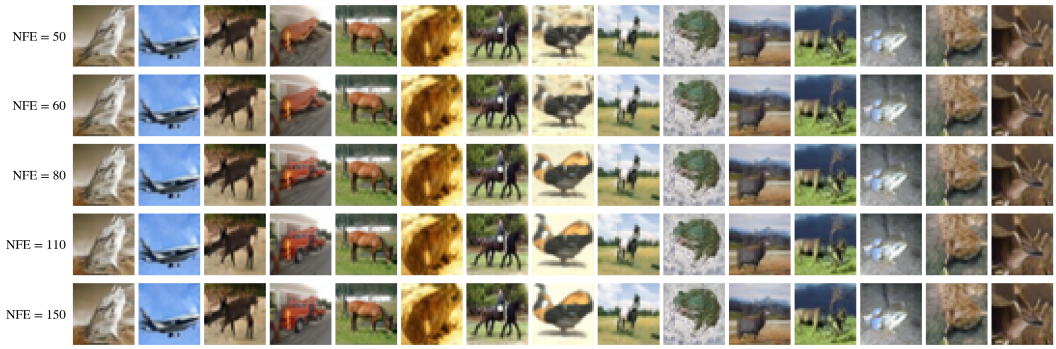

Figure 11: Samples generated by 1-Rectified Flow using the Euler solver with different NFE. Each column uses the same initial noise across NFE values, illustrating how the generated samples evolve as the number of Euler steps increases.



Figure 12: CIFAR-10 samples generated by (top left) ScoreSDE, (top right) RF with RK45, (bottom left) MeanFlow, and (bottom right) SGXL.