A Semi-Supervised Kernel Two-Sample Test

Gyumin Lee
Dept. of Statistics
Pennsylvania State University

Shubhanshu Shekhar
Dept. of EECS
University of Michigan

Ilmun Kim
Dept. of Mathematical Sciences
KAIST

Abstract

We consider the problem of two-sample testing in a semi-supervised setting with abundant unlabeled covariate data. Standard two-sample tests neglect covariate information, which has the potential to significantly boost performance. However, incorporating covariates potentially breaks the exchangeability assumption under the null, which further complicates a calibration procedure. To address these issues, we propose a semi-supervised method that produces a test statistic with asymptotic normality, while effectively integrating additional information from covariates. Our test is straightforward to calibrate due to the asymptotic normality under the null and achieves asymptotic power that is often much higher than existing kernel tests without covariates. Furthermore, we formally show that the proposed method is consistent in power against fixed and local alternatives. Simulations confirm the practical and theoretical strengths of our approach.

1 INTRODUCTION

In recent years, the realm of statistics and machine learning has seen notable progress in the development of semi-supervised methodologies that exploit both labeled and unlabeled data. These techniques present promising options for addressing numerous issues where labeled data are scarce or costly to gather, while large quantities of unlabeled data are typically available. The integration of both data types in semi-supervised learning has drawn considerable interest for its efficacy in enhancing predictive modeling for various tasks, including classification and regression (e.g., Chapelle et al., 2006; Zhu, 2008; Van Engelen and Hoos, 2020). These methods are widely adopted across various domains. In healthcare, for example, obtaining sufficient labeled medical data is challenging due to privacy concerns and the rarity of certain diseases (e.g., Han et al., 2024; Jiao et al., 2024; Chebli et al., 2018). Medical image annotation also demands substantial time and effort. Similar challenges occur in cyber-security (e.g., Mvula et al., 2024; Watkins et al., 2017), drug discovery (e.g., Zhang et al., 2018), and part-of-speech tagging (e.g., Wang et al., 2007). In such fields, semi-supervised learning has allowed practitioners to leverage unlabeled data for more accurate predictions.

Traditionally, semi-supervised learning research has focused on improving classification performance. However, more recent work has expanded its scope to address a wider range of problems, including tasks in statistical estimation and inference. A central goal of this body of work is to use unlabeled data to improve statistical methods when labeled data are scarce but unlabeled data are abundant. In line with this direction, our objective in this paper is to adapt and extend traditional two-sample testing to a semi-supervised setting to effectively incorporate unlabeled data.

The goal of two-sample testing is to determine whether two samples originate from the same underlying distribution. Numerous methods have been proposed to address this problem (see Stolte et al., 2024, for a recent review). Among them, a popular technique leverages the kernel-based framework introduced by Gretton et al. (2012a). This method employs an estimator of the kernel maximum mean discrepancy (MMD) as the test statistic where it measures the maximum difference in expectations over functions within the unit ball of a reproducing kernel Hilbert space (RKHS). The MMD has gained widespread adoption due to its nonparametric nature and the strong theoretical guarantees provided by RKHS theory. Despite these advantages, its application is often constrained by the intractability of its null distribution, which complicates the implementation of direct inference.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

To circumvent this issue, permutation-based methods are commonly used to determine the threshold, setting $\tau := \tau(\alpha)$ as the $(1-\alpha)$-quantile of the statistic over B permuted datasets. While these methods ensure finite-sample validity, they are computationally intensive, requiring $B+1$ evaluations of the test statistic, with B typically exceeding 100. Moreover, incorporating additional covariates (say V and W), which are possibly correlated with the primary samples (X and Y), introduces further complications: under the null hypothesis, the distributions of V and W do not need to match, which violates the *exchangeability* assumption. This lack of exchangeability undermines the validity of permutation-based methods, which poses a critical challenge in the semi-supervised setting.

Various permutation-free methods for determining the threshold τ have been proposed specifically to address the computational challenges associated with the permutation test. However, these methods entail notable limitations that warrant further attention. Some are overly conservative, resulting in a type-I error rate much smaller than the target size α (e.g., Gretton et al., 2006; Kim, 2021). Others rigorously control the size only under restrictive conditions, such as when the kernel remains fixed as the sample size increases (e.g., Gretton et al., 2009; Chwialkowski et al., 2015; Jitkrittum et al., 2016). Additionally, some methods lack theoretical guarantees of size control, as they are heuristic in nature (e.g., Gretton et al., 2006, 2009). In contrast to these approaches, Shekhar et al. (2022) proposed a permutation-free kernel two-sample test that leverages the dimension-agnostic framework introduced by Kim and Ramdas (2024). It offers rigorous theoretical guarantees, achieving consistency and minimax-rate optimality against local alternatives, but it only utilizes labeled data. To our knowledge, there exists no method for two-sample testing in a semi-supervised setting, and in this paper we develop a general framework that addresses this gap in the literature.

1.1 Contributions

With the preceding background in place, the main contributions of this work are summarized below.

General framework. We propose a general framework for semi-supervised two-sample testing that effectively leverages unlabeled data. This framework does not require permutation-based inference and is provably more powerful than the corresponding supervised methods in various scenarios. Additionally, we introduce a cross-fitting procedure to broaden the applicability of our method.

Semi-supervised kernel two-sample test. As a specific implementation of our general framework, we propose a semi-supervised kernel two-sample test. This test can be seen as a natural extension of the method introduced by Shekhar et al. (2022), adapted to effectively leverage the additional information available from unlabeled data in a semi-supervised setting.

Power analysis. We provide both theoretical and empirical evidence showing that our method retains the desirable properties of existing approaches while achieving higher power across diverse scenarios. A key element of our analysis is the asymptotic normality of a studentized test statistic. Unlike prior work, we establish the asymptotic normality of the test statistic under both the null and alternative hypotheses, which is critical to our power analysis. Furthermore, we demonstrate that our test statistic maintains consistency in power against both fixed and local alternatives.

1.2 Related Work

Semi-Supervised Inference. Semi-supervised inference has emerged as an important area in statistics, with numerous studies exploring how unlabeled data can enhance estimation and testing. For instance, Zhang et al. (2019) introduced methods for semi-supervised mean estimation, demonstrating the potential of unlabeled data to improve inference accuracy. Chakrabortty et al. (2019) examined the use of semi-supervised techniques in high-dimensional settings, while Cai and Guo (2020) focused on variance estimation. In the context of linear regression, Chakrabortty and Cai (2018); Azriel et al. (2022) improved standard estimators by incorporating unlabeled data, and Chakrabortty et al. (2022) extended this idea to quantile estimation. Further advancements were made by Angelopoulos et al. (2023); Zrnic and Candès (2024), who introduced the concept of *prediction-powered inference*, providing a unified framework for constructing predictive models that leverage both labeled and unlabeled data. More recently, Kim et al. (2024) analyzed semi-supervised U-statistics, which offers a comprehensive framework for integrating unlabeled data into nonparametric inference.

Kernel Two-Sample Tests. Kernel-based two-sample testing has gained widespread attention for handling complex and high-dimensional data. Since its introduction by Gretton et al. (2012a), numerous advancements have followed: developing optimal kernels to enhance test power (Gretton et al., 2012b; Sutherland et al., 2017; Liu et al., 2020), extending applicability to manifold data (Cheng and Xie, 2024), and devising alternative methods to reduce computational overhead (Zaremba et al., 2013; Song and Chen, 2021; Schrab et al., 2022; Choi and Kim, 2024). A more recent line of work has focused on boosting test power by aggregating MMD estimates over multiple kernels (Schrab et al., 2023; Biggs et al., 2024; Hagrass et al., 2024; Chatterjee and Bhattacharya, 2025). Taking a different approach, Tian et al. (2024) proposed a unified representation learning framework that utilizes the entire dataset to learn discriminative features. While their approach focuses on embedding learning via self-supervised learning, our method explicitly leverages the functional relationship between the target variable and abundant unlabeled covariates to reduce the variance of the test statistic.

Permutation-Free Approaches. Permutation-free methods have been actively studied to address the computational challenge of traditional permutation-based methods for large-scale analyses. Among these, Kim and Ramdas (2024) proposed a dimension-agnostic framework that uses sample-splitting to construct a studentized test statistic, asymptotically Gaussian under the null. This approach proves especially valuable when the null distribution is intractable or computationally expensive to estimate. Building on this, Shekhar et al. (2022) extended these ideas to the kernel-MMD setting, introducing the cross-MMD statistic. This innovative approach overcomes the degeneracy issues of the classical kernel-MMD statistic under the null. A similar framework has been applied in several studies to develop kernel-based independence testing (Shekhar et al., 2023), conditional independence testing (Lundborg et al., 2024), and kernel-based treatment effect testing (Martinez Taboada et al., 2023).

The above works collectively form the foundation of our approach, integrating semi-supervised inference, kernel-based testing, and advanced estimation techniques to address challenges in two-sample testing with additional covariates.

2 GENERAL SEMI-SUPERVISED TWO-SAMPLE TEST

In this section, we start by describing the problem setup and presenting the key idea underlying our approach. We then formally introduce a general semi-supervised two-sample test, which serves as the cornerstone for the semi-supervised kernel two-sample test detailed in Section 3.

We first clarify the terminology of labeled and unlabeled data as used in this paper. While the term *label* often refers to class variables in supervised learning tasks such as classification, we adopt a broader usage that is standard in recent literature on semi-supervised inference (e.g., Zhang et al., 2019; Angelopoulos et al., 2023; Kim et al., 2024). In our setting, labeled data refers to observations for which the primary response variable is available, whereas unlabeled data consists of covariates without associated responses. Although these covariates are not directly analyzed, they are typically easier to obtain and exhibit meaningful associations with the primary variables of interest.

In particular, we refer to the paired samples (X, V) and (Y, W) as labeled data, where X and Y are the primary variables of interest, and V and W denote covariates associated with the responses X and Y, respectively. While V and W come from the same feature space, we distinguish them to reflect the two different groups. In this context, supervised approaches rely solely on the labeled pairs, while semi-supervised methods additionally exploit the unlabeled covariates to improve statistical power.

Problem Setting. Let us formalize the setting of semi-supervised two-sample testing where we observe mutually independent labeled and unlabeled datasets as follows:

- *Labeled data*: $\mathcal{L}_{XV} \coloneqq \{(X_i, V_i)\}_{i=1}^{n_1} \overset{\text{i.i.d.}}{\sim} P_{XV}$ and $\mathcal{L}_{YW} \coloneqq \{(Y_i, W_i)\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} P_{YW}$
- *Unlabeled data*: $\mathcal{U}_V \coloneqq \{V_i\}_{i=n_1+1}^{n_1+m_1} \overset{\text{i.i.d.}}{\sim} P_V$ and $\mathcal{U}_W \coloneqq \{W_i\}_{i=n_2+1}^{n_2+m_2} \overset{\text{i.i.d.}}{\sim} P_W$

Using these observations, we would like to test the null hypothesis that the marginal distributions of X and Y are equal, that is, $H_0 : P_X = P_Y$ against the alternative $H_1 : P_X \neq P_Y$. Unlike the classical two-sample testing, covariates V and W are available, and our goal is to create a testing procedure that boosts statistical power by incorporating these covariates, while ensuring robustness when they are independent of X and Y.

2.1 Oracle Test

Before presenting a practical version, we first build intuition by considering an oracle test, assuming that we know the true conditional expectation.

Key idea. To clarify the key idea behind our approach, let us revisit semi-supervised mean estimation (e.g., Zhang et al., 2019; Zhang and Bradic, 2022). Specifically, consider the problem of estimating the population mean of some real-valued function $f(X)$. A natural idea is to use the sample mean, $n_1^{-1}\sum_{i=1}^{n_1} f(X_i)$, which has the minimum variance among all possible unbiased estimators. However, the situation changes when additional unlabeled datasets become available. For simplicity of our discussion, we assume the conditional expectation $\mathbb{E}[f(X_i) \mid V_i]$ is known, and address the unknown case in Section 2.2. Under this setup, one can construct another estimator

$$\widehat{\mu}_{X,f} \coloneqq \frac{1}{n_1}\sum_{i=1}^{n_1}\{f(X_i) - \mathbb{E}[f(X_i)\,|\,V_i]\} + \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\mathbb{E}[f(X_i)\,|\,V_i],$$

which is also an unbiased estimator. Importantly, the variance of $\widehat{\mu}_{X,f}$ is never greater than that of the ordinary sample mean. This property arises from the observation that the two summations in $\widehat{\mu}_{X,f}$ are uncorrelated. As a result, the variance of $\widehat{\mu}_{X,f}$ can be expressed as $\sigma^2_{X,f} \coloneqq n_1^{-1}\sigma^2_{1,X,f} + (n_1 + m_1)^{-1}\sigma^2_{2,X,f}$ where $\sigma^2_{1,X,f} \coloneqq \mathbb{E}[\mathrm{Var}\{f(X)\,|\,V\}]$ and $\sigma^2_{2,X,f} \coloneqq \mathrm{Var}[\mathbb{E}\{f(X)\,|\,V\}]$. Moreover, the ordinary sample mean is equivalent to $\widehat{\mu}_{X,f}$ with $m_1 = 0$, whose variance equals $n_1^{-1}\mathrm{Var}[f(X)] = n_1^{-1}\sigma^2_{1,X,f} + n_1^{-1}\sigma^2_{2,X,f}$ by the law of total variance. This directly confirms the variance reduction achieved by incorporating the additional unlabeled data. Furthermore, $\widehat{\mu}_{X,f}$ is a linear statistic that is expected to converge to a normal distribution under regularity conditions. As a result, statistical inference based on $\widehat{\mu}_{X,f}$ would be more efficient than that based on the ordinary sample mean, which lies at the heart of recent advancements in semi-supervised inference.

Oracle Test Construction. Building on the idea that incorporating unlabeled data can lead to variance reduction, we now introduce a general semi-supervised two-sample test. To delineate the procedure, define $\widehat{\mu}_{Y,f}$ analogously to $\widehat{\mu}_{X,f}$ using $\mathcal{L}_{YW}$ and $\mathcal{U}_W$. Here, f is treated as a certain feature map, mapping inputs to $\mathbb{R}$, chosen to effectively distinguish P_X and P_Y under the alternative. For instance, f can be a certain basis function (Zhou et al., 2017), an estimated witness function of an integral probability metric (Kim and Ramdas, 2024) or a deep kernel feature map (Liu et al., 2020). Importantly, we assume that f is independent of $\mathcal{L}_{XV}$, $\mathcal{L}_{YW}$, $\mathcal{U}_V$, and $\mathcal{U}_W$.

Remark 2.1. When f is a random function, all expectations and variances below are *implicitly conditional on* f unless stated otherwise. For example, $\mathbb{E}[f(X)]$ then denotes the conditional expectation of $f(X)$ given the σ-algebra generated by the randomness of f.

Our general procedure compares the studentized difference between $\widehat{\mu}_{X,f}$ and $\widehat{\mu}_{Y,f}$. To this end, we estimate the variance of $\widehat{\mu}_{X,f}$ as $\widehat{\sigma}^2_{X,f} = n_1^{-1}\widehat{\sigma}^2_{1,X,f} + (n_1+m_1)^{-1}\widehat{\sigma}^2_{2,X,f}$ by combining two components defined in (13) of Appendix F.1. We similarly define $\widehat{\sigma}^2_{Y,f} = n_2^{-1}\widehat{\sigma}^2_{1,Y,f} + (n_2+m_2)^{-1}\widehat{\sigma}^2_{2,Y,f}$ as an estimator of $\sigma^2_{Y,f} \coloneqq n_2^{-1}\sigma^2_{1,Y,f} + (n_2+m_2)^{-1}\sigma^2_{2,Y,f}$, which is the variance of $\widehat{\mu}_{Y,f}$. Using these estimates, we define an oracle test statistic as

$$T_{\mathrm{oracle}} = \frac{\widehat{\mu}_{X,f} - \widehat{\mu}_{Y,f}}{\sqrt{\widehat{\sigma}^2_{X,f} + \widehat{\sigma}^2_{Y,f}}}. \tag{1}$$

Given $\alpha \in (0,1)$, the resulting oracle test rejects the null when $T_{\mathrm{oracle}} > z_{1-\alpha}$ (or $|T_{\mathrm{oracle}}| > z_{1-\alpha/2}$ for a two-sided test) without requiring permutations. Here $z_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$. To analyze the oracle test, we make the following moment assumption, using $n \coloneqq n_1 \wedge n_2$ to denote the minimum sample size throughout. Additional notational conventions are provided in Appendix B.

Assumption 2.2. Suppose there exists $\delta > 0$ such that

$$\frac{\mathbb{E}[|f(X) - \mathbb{E}[f(X)]|^{2+\delta}]}{\sigma^{2+\delta}_{1,X,f} \wedge \sigma^{2+\delta}_{2,X,f}} = o_P\big(n_1^{\delta/2}\big) \quad \text{and}$$
$$\frac{\mathbb{E}[|f(Y) - \mathbb{E}[f(Y)]|^{2+\delta}]}{\sigma^{2+\delta}_{1,Y,f} \wedge \sigma^{2+\delta}_{2,Y,f}} = o_P\big(n_2^{\delta/2}\big) \quad \text{as } n \to \infty.$$

The o_P notation above is used to accommodate the random case of f discussed in Remark 2.1. The next theorem derives an asymptotic power expression of the oracle test and highlights the power gain obtained through the unlabeled dataset.

Theorem 2.3. *Under Assumption 2.2, the power function of the oracle test defined in* (1) *approximates, unconditionally on f, that*

$$\Phi\left(z_\alpha + \frac{\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]}{\sqrt{\sigma^2_{X,f} + \sigma^2_{Y,f}}}\right) \quad \text{as } n \to \infty.$$

Under the null, Theorem 2.3 indicates that the oracle test asymptotically maintains the correct size α. Regarding power, the oracle test achieves an asymptotic power *never less* than that of the standard two-sample t-test. This is evident from the fact that the standard t-test is a special case of the oracle test when $m_1 = m_2 = 0$, and that $\sigma^2_{X,f} + \sigma^2_{Y,f}$ is non-decreasing in m_1 and m_2. In other words, incorporating unlabeled data reduces variance while keeping the mean unchanged, ultimately resulting in increased power. Theorem 2.3 allows f to change with the sample sizes. This flexibility requires conditions stronger than the finite second moment of $f(X)$ and $f(Y)$ as specified in Assumption 2.2. It is also worth highlighting that Theorem 2.3 puts no restrictions on m_1 and m_2, which can grow much faster than $n = n_1 \wedge n_2$.

2.2 Procedure with Cross-Fitting

In the previous subsection, we constructed the oracle test under the assumption that both the conditional expectations $\mathbb{E}[f(X)\,|\,V]$ and $\mathbb{E}[f(Y)\,|\,W]$ are known. We now eliminate this assumption and propose a practical procedure using the estimated conditional expectations $\widehat{\mathbb{E}}[f(X) \mid V]$ and $\widehat{\mathbb{E}}[f(Y) \mid W]$. For this purpose, we employ cross-fitting, a commonly used technique in semi-parametric statistics. Cross-fitting is a practical, efficient method of data splitting, typically applied to correct for bias arising from nuisance estimation, ease stringent conditions on the parameter space, and regain full efficiency (e.g., Zheng and Van Der Laan, 2010; Chernozhukov et al., 2018; Newey and Robins, 2018; Wasserman et al., 2020; Kennedy, 2023; Kim et al., 2024). This method involves partitioning the dataset into two segments: one is used to estimate nuisance parameters, while the other is employed to form an initial estimator. The roles of these partitions are then alternated, and the procedure is repeated. Finally, the two resulting statistics are aggregated to yield the final estimator. For simplicity, we assume that n_1, n_2, m_1, m_2 are even numbers, which allows us to avoid asymmetry in the cross-fitting procedure and thus simplifies the analysis.

Cross-Fit Test Construction. To describe the idea, we split the dataset $\mathcal{L}_{XV}$ into two parts: $\mathcal{L}_{XV,a} := \{(X_i, V_i) : i \in [n_1], i \text{ is odd}\}$ and $\mathcal{L}_{XV,b} := \{(X_i, V_i) : i \in [n_1], i \text{ is even}\}$. Write $\widehat{\mathbb{E}}[f(X_i) \mid V_i]$ as an estimator of $\mathbb{E}[f(X_i) \mid V_i]$ trained on $\mathcal{L}_{XV,a}$ if the index i is even and on $\mathcal{L}_{XV,b}$ if i is odd. This estimator can be obtained using methods such as neural nets or random forests by regressing $f(X)$ on V. We similarly construct $\widehat{\mathbb{E}}[f(Y_i) \mid W_i]$ as an estimator of $\mathbb{E}[f(Y_i) \mid W_i]$. The test statistic T_{cross} is then computed in the same way as T_{oracle}, replacing $\mathbb{E}[f(X_i) \mid V_i]$ and $\mathbb{E}[f(Y_i) \mid W_i]$ with their estimators. We finally reject the null if T_{cross} exceeds $z_{1-\alpha}$. This cross-fit test retains the same asymptotic properties as the oracle test, provided that the estimated conditional expectations satisfy the required convergence conditions.

Corollary 2.4. *Suppose Assumption 2.2 holds and, additionally, the following condition is satisfied:*

$$\frac{\mathbb{E}\big[\{\widehat{\mathbb{E}}[f(X) \mid V] - \mathbb{E}[f(X) \mid V]\}^2\big]}{\sigma^2_{1,X,f} \wedge \sigma^2_{2,X,f}} = o_P(1), \qquad (2)$$

as $n \to \infty$, and the analogous condition holds for (Y, W). Then the power function of the cross-fit test approximates that of the oracle test as in Theorem 2.3.

Similarly to Assumption 2.2, the o_P notation accounts for the randomness of f. The validity of Corollary 2.4 primarily depends on accurately estimating the conditional expectation associated with f, a problem well-studied in the statistical literature (e.g., Györfi et al., 2006; Wainwright, 2019).

Up to this point, we have developed a general semi-supervised two-sample test with a generic function f and demonstrated its power gain through the incorporation of unlabeled data. We next focus on a specific instantiation of f constructed as the difference between empirical kernel mean embeddings.

3 SEMI-SUPERVISED KERNEL TEST

In this section, we introduce a semi-supervised kernel two-sample test, regarded as a semi-supervised extension of the xMMD test (Shekhar et al., 2022). In Section 3.1, we first provide a brief overview of the xMMD test and then describe our proposed semi-supervised extension. Section 3.2 presents the theoretical analysis of the proposed test.

3.1 Testing Procedure

As mentioned earlier, one notable method for addressing the two-sample testing problem involves using an empirical version of the kernel-MMD (Gretton et al., 2012a). For a positive definite kernel k and its associated RKHS $\mathcal{H}_k$, the kernel-MMD quantifies the distance between distributions P and Q by computing the supremum of the difference in expectations $\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]$ over all functions f in the unit ball of $\mathcal{H}_k$. The empirical MMD statistic, based on U- or V-statistics, has an intractable limiting distribution under the null, which is often addressed using the permutation method. However, this resampling method is computationally expensive due to repeated evaluation of a test statistic. Beyond the computational issue, the permutation method may not be valid in the semi-supervised setting where V and W do not necessarily share the same distribution under the null. This violates the exchangeability assumption, which is crucial for the validity of the permutation test.

To address the computational issue of permutation-based MMD tests, Shekhar et al. (2022) introduced the xMMD test, which is essentially the two-sample t-test applied to data projected onto the optimal witness function. To provide a brief overview, let $\{\widetilde{X}_i\}_{i=1}^{n_1}$ and $\{\widetilde{Y}_i\}_{i=1}^{n_2}$ be i.i.d. copies of $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$, respectively, which can be obtained through sample splitting. The xMMD test is then implemented through the following two steps:

1. Optimal Witness Function Estimation. Estimate the optimal witness function that achieves the supremum in the definition of MMD based on $\{\widetilde{X}_i\}_{i=1}^{n_1}$ and $\{\widetilde{Y}_i\}_{i=1}^{n_2}$:

$$\widehat{f}(\cdot) := \frac{1}{n_1}\sum_{i=1}^{n_1} k(\widetilde{X}_i, \cdot) - \frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i, \cdot).$$

2. Projection and t-Test. Project $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$ onto the direction $\widehat{f}$, which results in (conditionally independent) univariate two samples:

$$\widehat{f}(X_1), \ldots, \widehat{f}(X_{n_1}) \quad \text{and} \quad \widehat{f}(Y_1), \ldots, \widehat{f}(Y_{n_2}). \tag{3}$$

The xMMD test rejects the null when the corresponding two-sample t-statistic exceeds $z_{1-\alpha}$.

Shekhar et al. (2022) showed that the xMMD test is asymptotically level α under a certain moment condition, consistent in power and minimax-rate optimal against local L_2 alternatives. Moreover, the xMMD test offers a notable computational advantage over the permutation-based MMD test as it avoids the need for repeated resampling to determine a critical value.

xssMMD Test. Building on the work of Shekhar et al. (2022), we propose a new method called the xssMMD test, which extends the xMMD test to a semi-supervised setting. The main idea is to apply the general semi-supervised two-sample test introduced in Section 2 to the two projected samples based on the estimated witness function $\widehat{f}$. Specifically, we define the cross-fit statistic T_{cross} based on the projected samples in (3) as

$$\mathrm{xss}\widehat{\mathrm{MMD}}^2 = \frac{\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \widehat{\mu}^{\dagger}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}, \tag{4}$$

where $\dagger$ denotes the use of cross-fitting. The exact mathematical formulas for the cross-fitted components are detailed in Appendix F.3. The xssMMD test then rejects the null when $\mathrm{xss}\widehat{\mathrm{MMD}}^2 > z_{1-\alpha}$. A schematic illustration is provided in Figure 3 of A.

3.2 Theoretical Analysis

We now shift our focus to the theoretical analysis of the xssMMD test. It is already clear from the results of Section 2 that the xssMMD test is asymptotically level α and provably more powerful than the xMMD test under certain conditions. Our goal is to present more concrete conditions for these properties tailored to the kernel-MMD setting.

Comparison with xMMD. To establish the asymptotic properties of the tests, we require certain regularity and moment conditions. First, we assume that the witness function $\widehat{f}$ is measurable and that the Bochner integral $\int_{\mathcal{X}} \|k(x,\cdot)\|_{\mathcal{H}} dP(x)$ is finite to ensure the existence of the mean embedding. Note that for bounded kernels, such as the Gaussian kernel, this condition is automatically satisfied since $\|k(x,\cdot)\|_{\mathcal{H}} = \sqrt{k(x,x)}$ is bounded.

Next, we state the assumptions required for our theoretical guarantees. These involve key quantities defined through the centered kernel $\overline{k}_X$ and its expected product $\overline{g}_X$. Formal definitions are deferred to Appendix C.

Assumption 3.1 (Null Condition)**.** Assume $X_1, X_2, X_3 \overset{\text{i.i.d.}}{\sim} P_{X,n}$ where $P_X := P_{X,n}$ and the kernel $k := k_n$ potentially changing with n satisfy

$$\frac{\mathbb{E}[\overline{k}(X_1,X_2)^4] + n_1 \mathbb{E}[\overline{k}(X_1,X_2)^2 \overline{k}(X_1,X_3)^2]}{n_1^2 \{\mathbb{E}[\overline{g}_X(X,X)]\}^2} = o(1).$$

Assumption 3.2 (Consistency of Conditional Expectation)**.** Assume that the estimated conditional expectation of $\widehat{f}(X)$ given $V, \widehat{f}$ and that of $\widehat{f}(Y)$ given $W, \widehat{f}$ satisfy

$$\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(X) \mid V, \widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]\}^2 \mid \widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(X) \mid \widehat{f}\}} = o_P(1), \text{ and}$$
$$\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(Y) \mid W, \widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]\}^2 \mid \widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(Y) \mid \widehat{f}\}} = o_P(1). \tag{5}$$

Assumption 3.3 (Alternative Condition)**.** Assume that $P_X := P_{X,n}$ and $P_Y := P_{Y,n}$ have Lebesgue density functions p_X and p_Y, respectively, satisfying $\|p_X/p_Y\|_{L_\infty} \vee \|p_Y/p_X\|_{L_\infty} < C$ for some constant $C > 0$, with $\|f\|_{L_\infty}$ denoting $\inf\{M \geq 0 : \mathrm{Leb.}(\{x : |f(x)| > M\}) = 0\}$. Furthermore, assume that

$$\frac{\mathrm{MMD}(P_X,P_Y)^4 \times \mathbb{E}[\overline{k}_X(X,X)^2]}{\{n_1 \mathbb{E}[\overline{g}_X(X,X)] + n_1^2 \mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} = o(1), \tag{6}$$

where $X \sim P_X$ and $Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} P_Y$.

Assumption 3.1 is essentially a Lyapunov-type condition required for the Central Limit Theorem. At a high level, it ensures that the "tails" of the test statistic's distribution are not too heavy relative to its variance, guaranteeing that no single data point dominates the statistic, thus allowing it to converge to a normal distribution. Assumption 3.2 guarantees that the estimation error from the cross-fitting procedure decays sufficiently fast relative to the variance, preventing it from dominating the asymptotic behavior of the test statistic. Assumption 3.3 provides sufficient conditions to establish asymptotic normality under the alternative. The bounded density ratio simplifies the mathematical derivations, while the moment condition (6) ensures the Lyapunov central limit theorem holds for a broad range of alternatives. We discuss further implications of these assumptions in Appendix C.

The next theorem compares the xssMMD test and the xMMD test based on their asymptotic properties under specific conditions. For brevity, these conditions are presented and discussed in Appendix C. In the following, we let $\Psi_{\mathrm{x}} := \mathbb{1}(\mathrm{x}\widehat{\mathrm{MMD}}^2 > z_{1-\alpha})$ and $\Psi_{\mathrm{xss}} := \mathbb{1}(\mathrm{xss}\widehat{\mathrm{MMD}}^2 > z_{1-\alpha})$ denote the xMMD and xssMMD tests, respectively.

Theorem 3.4. *The tests Ψ_{x} and Ψ_{xss} satisfy the following asymptotic guarantees:*

Level. *Suppose Assumption 3.1 and Assumption 3.2 hold with $n_1 \asymp m_1$ and $n_2 \asymp m_2$. Then both tests control the size α under H_0 such that $\lim_{n\to\infty} \mathbb{E}_{H_0}[\Psi_{\mathrm{x}}] = \lim_{n\to\infty} \mathbb{E}_{H_0}[\Psi_{\mathrm{xss}}] = \alpha$.*

Power. *Suppose Assumption 3.3 also holds under H_1. Then the asymptotic power of Ψ_{xss} is at least as that of Ψ_{x}, satisfying $\lim_{n\to\infty}\{\mathbb{E}_{H_1}[\Psi_{\mathrm{xss}}] - \mathbb{E}_{H_1}[\Psi_{\mathrm{x}}]\} \geq 0$.*

The above theorem confirms that Ψ_{xss} is asymptotically level α under the null and achieves at least the same power as Ψ_{x} under the alternative. As for the general test, the key insight behind the power gain of Ψ_{xss} lies in the effective use of unlabeled data, which reduces the variance of the test statistic while preserving the same mean. This insight, together with the asymptotic normality, enables a direct comparison of the power of the xssMMD and xMMD tests.

The novelty of Theorem 3.4 is in extending the conditions for asymptotic normality to the alternative, whereas prior work has primarily focused on the null. This extension is crucial for power comparisons and requires substantial effort to establish. Unlike Theorem 2.3, Theorem 3.4 additionally assumes $n_1 \asymp m_1$ and $n_2 \asymp m_2$. These conditions are imposed to facilitate a comparison of Ψ_{x} and Ψ_{xss} under common and concrete moment assumptions, which could be relaxed under more abstract conditions. Alternatively, when $m_1 \geq n_1$ and $m_2 \geq n_2$, one could discard a portion of the unlabeled samples to ensure the asymptotic balance condition.

Consistency in Power. The power property of the xssMMD test, as stated in Theorem 3.4, is established under the assumptions that the centered test statistic converges to a normal distribution under the alternative. Here, we present independent conditions under which the xssMMD test remains consistent in power (i.e., the power approaches one), without relying on the asymptotic normality. Below, a subscript n is added to indicate that the corresponding sequence may vary with $n = n_1 \wedge n_2$.

Lemma 3.5. *Let $\{\delta_n : n \geq 2\}$ be any positive sequence such that $\delta_n \to 0$, and $\gamma_n := \mathrm{MMD}(P_{X,n}, P_{Y,n})$. If*

$$\sup_{(P_{X,n},P_{Y,n})\in\mathcal{P}_n} \left\{ \frac{\mathbb{E}_{P_{X,n},P_{Y,n}}\big[\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}\big]}{\delta_n \gamma_n^4} + \frac{\mathrm{Var}_{P_{X,n},P_{Y,n}}\big[\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \widehat{\mu}^{\dagger}_{Y,\widehat{f}}\big]}{\gamma_n^4} \right\} = o(1),$$

then Ψ_{xss} is consistent in power uniformly over $\mathcal{P}_n$ as $\inf_{(P_{X,n},P_{Y,n})\in\mathcal{P}_n} \mathbb{E}_{P_{X,n},P_{Y,n}}[\Psi_{\mathrm{xss}}] = 1$.

The lemma above corresponds to Shekhar et al. (2022, Theorem 8), which forms the primary foundation for their other results, including minimax-rate optimality. In Appendix D.1, we show that the condition in Shekhar et al. (2022, Theorem 8) is stronger than that in Lemma 3.5, provided that $\widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]$ and $\widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]$ exhibit "well-behaved" properties. This implies that Ψ_{xss} is consistent in power whenever Ψ_{x} is. Importantly, this result does not rely on the consistency of $\widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]$ and $\widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]$ with the true conditional expectations. Instead, it requires that the residuals $\widehat{f}(X) - \widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]$ and $\widehat{f}(Y) - \widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]$ have second moments comparable to the variances of $\widehat{f}(X)$ and $\widehat{f}(Y)$, respectively—a much weaker condition than the full consistency of the conditional expectations. We discuss further implications of Lemma 3.5 in Appendix D.1.

4 EXPERIMENTS

We now experimentally validate the theoretical results stated in the previous sections. In particular, our experiments show that (i) the limiting null distribution of the proposed test statistic in (4) follows a $N(0, 1)$ distribution across a wide range of dimensions d, sample sizes n_1, n_2, m_1, m_2, and kernel k, and (ii) the power of the xssMMD test is comparable to and often much higher than that of the xMMD test and the kernel-MMD permutation (MMD-perm) test. Moreover, we examine its performance on several real-world datasets. Additional experimental findings can be found in Appendix E.

Limiting Null Distribution. We demonstrated in Theorem 3.4 that the xssMMD test is asymptotically level α under the null, given some assumptions. We empirically validate this result by considering the case where $P_{XV} = P_{YW} = N(\mathbf{0}_{2d}, I_{2d})$. We study the effects of dimensionality, sample skewness, labeled-unlabeled sample size ratio, methods for estimating conditional expectation, and choice of kernel on the null distribution of the test statistic. Specifically, we consider two scenarios:

- **Scenario 1 (Null).** $d = 10$, $n_1/n_2 = 1$, $n_1/m_1 = n_2/m_2 = 1$, Gaussian kernel with the median heuristic.
- **Scenario 2 (Null).** $d = 100$, $n_1/n_2 = 0.1$, $n_1/m_1 = n_2/m_2 = 0.5$, bilinear kernel.

Note that we applied a bandwidth determined by the median heuristic when using a Gaussian kernel. Each scenario considers different methods for estimating the conditional expectation, including k-nearest neighbors (knn), kernel regression (kernel), and random forest (rf). As shown in Figure 1, the null distribution of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ is robust to all these factors and closely approximates $N(0, 1)$. This confirms that our test, calibrated using

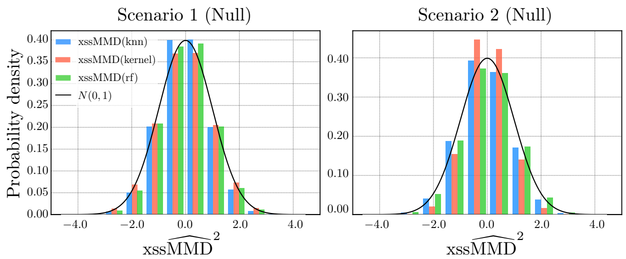

Figure 1: Experimental results for the distribution of $\text{xss}\widehat{\text{MMD}}^2$ under the null hypothesis. The plots demonstrate that the test statistic closely follows a $N(0, 1)$ distribution across various parameter settings, confirming the validity of the xssMMD test.

the normal quantile, successfully controls the level α in the scenarios considered. Additional results and implementation details are provided in Appendix E.

Power Analysis. Theorem 3.4 shows that the xss-MMD test achieves power at least as high as the xMMD test under certain conditions. We empirically validate this by comparing our test (Ψ_{xss}) with the xMMD test (Ψ_{x}) using a Gaussian kernel and the MMD-perm test ($B = 200$ permutations). We perform simulations at $\alpha = 0.05$ and present results from 1,000 trials in Figure 2.

We consider the case of $P_V = N(\mathbf{0}_d, \Sigma_V)$ and $P_W = N(\mathbf{a}_{\epsilon,j}, \Sigma_W)$, where $\mathbf{a}_{\epsilon,j} \in \mathbb{R}^d$ has its first j entries equal to ϵ and the rest zero. We let $\Sigma_V = \Sigma_W = \rho \mathbb{1}_d \mathbb{1}_d^\top + (1-\rho) I_d$ and obtain $\{V_i\}_{i=1}^{n_1+m_1}$ by sampling $n_1 + m_1$ independent samples from P_V. We then construct $\mathbb{V} = (V_1^\top, \ldots, V_{n_1}^\top)^\top \in \mathbb{R}^{n_1 \times d}$ and obtain a set of n_1 labeled samples, $\mathbb{X} = \mathbb{V} \cdot \mathbf{b}$, where $\mathbf{b} = (b_i)_{i=1}^d \in \mathbb{R}^d$ with $b_i = 1$ if i belongs to an index set $\mathcal{I}$ and $b_i = 0$ otherwise. A similar construction is applied to $\mathbb{Y}$ and $\mathbb{W}$. Each scenario differs based on how we construct X and V, which determines the dependence between the labeled and unlabeled data. In particular, we consider four scenarios:

- **Scenario 1 (Alt).** $\rho = 0.95$, $\mathcal{I} = \{1, d-1, d\}$
- **Scenario 2 (Alt).** $\rho = 0.95$, $\mathcal{I} = [d]$
- **Scenario 3 (Alt).** $\rho = 0.1$, $\mathcal{I} = \{1, d-1, d\}$
- **Scenario 4 (Alt).** $\rho = 0.1$, $\mathcal{I} = [d]$

In all scenarios in Figure 2, we fix parameters at $\epsilon = 0.3$, $j = 3$, $d = 10$, $n_1/n_2 = 1$, and $n_1/m_1 = n_2/m_2 = 0.1$, using a Gaussian kernel with the median heuristic. The main factors controlling the dependence between X and V are ρ and $\mathbf{b}$. For example, in Scenarios 1 and 3 (Alt), X and Y are sums of the first and last two entries of V and W, so the covariance vector has $1 + 2\rho$ for the first and last two entries and 3ρ otherwise. In contrast, in Scenarios 2 and 4 (Alt), X and Y are sums of all entries, yielding uniform covariance of $1+(d-1)\rho$. This leads to stronger dependence when $d \geq 3$, with larger ρ further enhancing it and improving the performance of the xssMMD test.

As shown in Figure 2, the xssMMD test significantly outperforms other methods when additional covariates strongly correlate with the labeled data (Scenarios 1 and 2). Even when the correlation is weaker (Scenarios 3 and 4), the xssMMD still demonstrates consistently better or comparable performance. These results highlight the advantage of leveraging auxiliary covariates, particularly when dependencies are strong. Further implementation details are provided in Appendix E.

Experiment on HTRU2 dataset. We next evaluate the performance of xssMMD using the HTRU2 dataset (Lyon, 2015), which involves the classification of pulsars versus non-pulsars based on radio signal features of the integrated pulse profile (IP) and DM-SNR curve (DM). We examine several scenarios of labeled data with various levels of Gaussian noise added. A detailed description of the experimental setup and results are provided in Appendix E.5. As shown in Table 9, xssMMD consistently outperforms baseline methods across most of the settings and noise levels. Even when Gaussian noise degrades the labeled data, xssMMD maintains a high power, leveraging auxiliary covariates effectively. These results highlight the strength of the method in extracting signals from complementary, unlabeled information under semi-supervised conditions.

Experiment on Caltech-UCSD Bird dataset. We also examine the performance of our proposed methods on the Caltech-UCSD Bird dataset (Wah et al., 2011), which contains 11,788 images of 200 bird species. Each image has 10 detailed single-sentence descriptions collected through the Amazon Mechanical Turk (AMT) platform (Reed et al., 2016). In this experiment, we conduct two-sample tests to detect differences among groups of birds categorized by their diet and habitat. A detailed description of the experimental setup is provided in Appendix E.6. The MMD-perm and xMMD tests rely solely on textual descriptions, while the xssMMD test additionally incorporates image data as auxiliary covariates. As shown in Table 1, the xssMMD test consistently outperforms the other tests in all cases. This confirms that the use of additional covariates improves the power of the test. This superior performance is likely due to strong dependency between the textual descriptions (labeled data) and images (unlabeled data), allowing xssMMD to extract informative representations from the additional covariates.

Experiment on MNIST dataset. We further evaluate the performance of our proposed methods on the MNIST dataset (LeCun et al., 2010). We construct

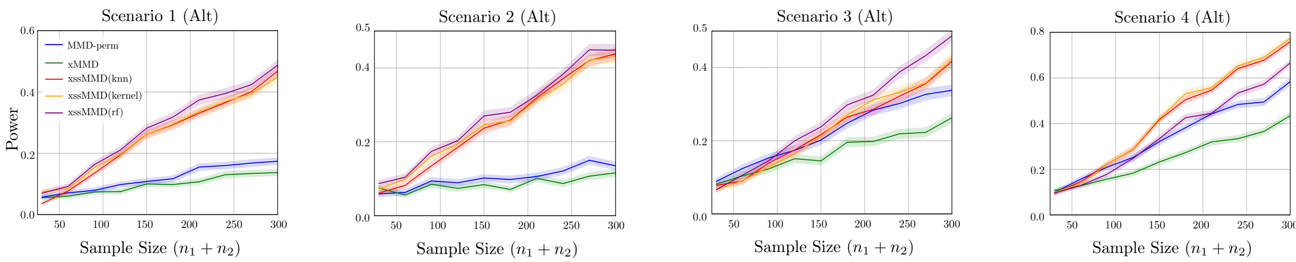

Figure 2: Power comparisons across different dependence scenarios. The xssMMD tests, employing various regression methods, outperform existing approaches in the considered scenarios, particularly when X and Y exhibit strong dependence on V and W.

Table 1: Estimated test power for detecting the difference between two bird groups with test level $\alpha = 0.05$.

Group 1	Group 2	Test	Power
		MMD-perm	0.957
Insect	Forest	xMMD	0.837
		xssMMD	**0.989**
		MMD-perm	0.626
Fish	Wetland	xMMD	0.471
		xssMMD	**0.808**
		MMD-perm	0.992
Seed	Scrub	xMMD	0.920
		xssMMD	**0.998**

a testing problem to detect distributional differences between two distinct groups of handwritten digits. We use clean images as labeled data and images with Gaussian noise as unlabeled data. A detailed description of the experimental setup and results are provided in Appendix E.7. As shown in Table 12, the xssMMD test outperforms the MMD-perm and xMMD tests across most of the tested conditions, especially when the noise level is low. Despite the increase of the noise level, the power of the xssMMD test is still higher than that of the xMMD test. This result demonstrates that xssMMD effectively utilizes information from the additional covariates, even when those covariates are corrupted by Gaussian noise. The successful integration of noisy auxiliary data underscores the strength of the method, boosting the power in semi-supervised settings.

5 DISCUSSION

In this paper, we present a semi-supervised framework for two-sample testing that incorporates both labeled and unlabeled covariate data to improve power while maintaining asymptotic level control. Leveraging sample-splitting and cross-fitting, the proposed method integrates covariate information and achieves asymptotic properties such as power consistency. Our analysis highlights the benefits of utilizing unlabeled data and provides conditions ensuring the validity of our tests. Along with numerical experiments, these results emphasize the potential of the framework as a theoretically sound tool for semi-supervised inference.

Several promising directions remain for exploration. First, extending the framework to broader contexts, such as k-sample testing and independence testing, would expand its applicability to complex semi-supervised problems. Exploring witness functions beyond MMD offers another avenue for future research. Moreover, studying methods for estimating conditional mean embeddings and exploring alternative variance reduction techniques, such as control covariates, may further refine the proposed framework.

Acknowledgements

Ilmun Kim gratefully acknowledges support from the Korean government (RS-2023-00211073) and KAIST startup funding (KAIST-G04250059).

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.

Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., and Zhao, L. (2022). Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136.

Biggs, F., Schrab, A., and Gretton, A. (2024). MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36.

Cai, T. and Guo, Z. (2020). Semi-supervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal*

Statistical Society Series B: Statistical Methodology, 82(2):391–419.

Chakrabortty, A. and Cai, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572.

Chakrabortty, A., Dai, G., and Carroll, R. J. (2022). Semi-Supervised Quantile Estimation: Robust and Efficient Inference in High Dimensional Settings. *arXiv preprint arXiv:2201.10208*.

Chakrabortty, A., Lu, J., Cai, T. T., and Li, H. (2019). High dimensional M-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press.

Chatterjee, A. and Bhattacharya, B. B. (2025). Boosting the power of kernel two-sample tests. *Biometrika*, 112(1):0–48.

Chebli, A., Djebbar, A., and Marouani, H. F. (2018). Semi-supervised learning for medical application: A survey. In *2018 international conference on applied smart systems (ICASS)*, pages 1–9. IEEE.

Cheng, X. and Xie, Y. (2024). Kernel two-sample tests for manifold data. *Bernoulli*, 30(4):2572–2597.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning foratment and structural parameters. *The Econometrics Journal*, 21(1).

Choi, I. and Kim, I. (2024). Computational-Statistical Trade-off in Kernel Two-Sample Testing with Random Fourier Features. *arXiv preprint arXiv:2407.08976*.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, 22.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hagrass, O., Sriperumbudur, B., and Li, B. (2024). Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101.

Han, K., Sheng, V. S., Song, Y., Liu, Y., Qiu, C., Ma, S., and Liu, Z. (2024). Deep semi-supervised learning for medical image segmentation: A review. *Expert Systems with Applications*, page 123052.

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., and Jin, C. (2024). Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840.

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29.

Kennedy, E. H. (2023). Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects. *Electronic Journal of Statistics*, 17(2).

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli*, 27(1):419–441.

Kim, I. and Ramdas, A. (2024). Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683–711.

Kim, I., Wasserman, L., Balakrishnan, S., and Neykov, M. (2024). Semi-Supervised U-statistics. *arXiv preprint arXiv:2402.18921*.

Kübler, J. M., Jitkrittum, W., Schölkopf, B., and Muandet, K. (2022). A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR.

LeCun, Y., Cortes, C., Burges, C., et al. (2010). Mnist handwritten digit database.

Li, T. and Yuan, M. (2024). On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *Journal of Machine Learning Research*, 25(334):1–62.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR.

Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2024). The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878.

Lyon, R. (2015). HTRU2. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DK6R.

Martinez Taboada, D., Ramdas, A., and Kennedy, E. (2023). An efficient doubly-robust test for the kernel treatment effect. *Advances in Neural Information Processing Systems*, 36:59924–59952.

Mvula, P. K., Branco, P., Jourdan, G.-V., and Viktor, H. L. (2024). A survey on the applications of semi-supervised learning to cyber-security. *ACM Computing Surveys*, 56(10):1–41.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Reed, S., Akata, Z., Lee, H., and Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete U-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807.

Shekhar, S., Kim, I., and Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180.

Shekhar, S., Kim, I., and Ramdas, A. (2023). A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68.

Song, H. and Chen, H. (2021). A fast and effective large-scale two-sample test based on kernels. *arXiv preprint arXiv:2110.03118*.

Stolte, M., Kappenberg, F., Rahnenführer, J., and Bommert, A. (2024). Methods for quantifying dataset similarity: a review, taxonomy and comparison. *Statistic Surveys*, 18:163–298.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. In *International Conference on Learning Representations*.

Tian, X., Peng, L., Zhou, Z., Gong, M., Gretton, A., and Liu, F. (2024). A unified data representation learning for non-parametric two-sample testing. *arXiv preprint arXiv:2412.00613*.

Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD birds-200-2011 dataset.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, W., Huang, Z., and Harper, M. (2007). Semi-supervised learning for part-of-speech tagging of Mandarin transcribed speech. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–137. IEEE.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Watkins, L., Beck, S., Zook, J., Buczak, A., Chavis, J., Robinson, W. H., Morales, J. A., and Mishra, S. (2017). Using semi-supervised machine learning to address the big data problem in DNS networks. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–6. IEEE.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26.

Zhang, A., Brown, L. D., and Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566.

Zhang, X., Wang, S., Zhu, F., Xu, Z., Wang, Y., and Huang, J. (2018). Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 404–413.

Zhang, Y. and Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403.

Zheng, W. and Van Der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation.

Zhou, W.-X., Zheng, C., and Zhang, Z. (2017). Two-sample smooth tests for the equality of distributions. *Bernoulli*, 23(2):951–989.

Zhu, X. J. (2008). Semi-Supervised Learning Literature Survey. *Technical Report*.

Zrnic, T. and Candès, E. J. (2024). Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The paper clearly states all assumptions for each theorem. Our general framework is outlined in Section 2. We subsequently detail our method in Section 3 and discuss assumptions in Appendix C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] In Section 3.2, we prove several asymptotic properties of our model. Further theoretical results and all proofs are provided in Appendix D and Appendix F.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The supplementary materials include all code and instructions, with dependencies clearly specified.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] We provide Assumption 2.2 for Theorem 2.3 and state additional assumptions for theoretical analysis in Section 3.2.
   (b) Complete proofs of all theoretical results. [Yes] All proofs for our theoretical results are provided in Appendix F.
   (c) Clear explanations of any assumptions. [Yes] All assumptions are provided with intuitive sketches and explanations.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] In the supplementary material, we provide code and detailed instructions to reproduce the experiments.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Section 4 details our experiments, with further settings and additional results provided in Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] In Section 4 and Appendix E, we specify the experimental settings and describe the evaluation measures used in our experiments.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] In Appendix E, we describe the computing environments used in the experiments.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] In Section 4, several datasets used in our experiments are publicly available and cited properly.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] In Section 4, we demonstrate our method using several datasets. The supplementary materials provide the necessary code.
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Semi-Supervised Kernel Two-Sample Test: Supplementary Materials

A Overview of the xssMMD Framework and Theoretical Contributions

A.1 Visual Overview of the xssMMD Framework

To complement the main text, we include here an illustration of the xssMMD construction.

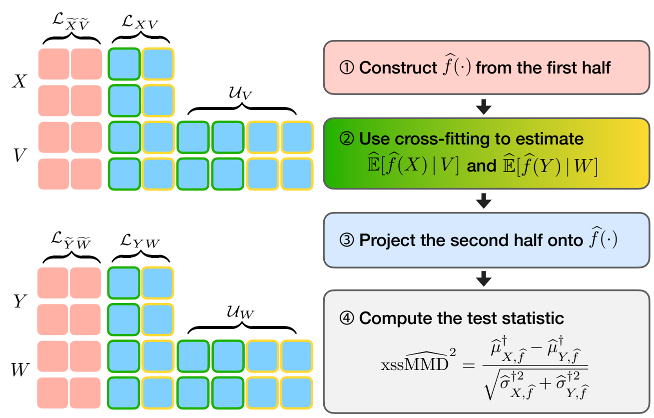

Figure 3: An illustration of the construction of the xssMMD statistic based on the same principles as the general framework and using an empirical estimate of the MMD witness function.

This figure provides a visual breakdown of how the test statistic $\widehat{\mathrm{xssMMD}}^2$ is derived in practice. Specifically, it illustrates the key steps involved in the cross-fitting procedure: splitting the data, estimating the witness function f from the first half, and computing the statistic by projecting the second half onto the learned function $\widehat{f}$. By leveraging auxiliary covariates (such as V and W), the method estimates conditional expectations $\mathbb{E}[\widehat{f}(X)|V]$ and $\mathbb{E}[\widehat{f}(Y)|W]$, thereby integrating semi-supervised information into the testing framework. This enables the test to maintain nonparametric flexibility while enhancing power.

A.2 Roadmap of Theoretical Results

This section provides a structured overview of the theoretical guarantees established in this work. To facilitate navigation through the various lemmas, propositions, and theorems, we summarize our key theoretical contributions and their exact locations in Table 2.

B Notation

For a sequence of random variables $(X_n)_{n\geq 1}$ and another random variable X, we write $X_n \xrightarrow{D} X$ when X_n converges in distribution to X. Likewise, we write $X_n \xrightarrow{P} X$ when X_n converges in probability to X. For a sequence of positive numbers $(a_n)_{n\geq 1}$, we denote $a_n \lesssim b_n$ if there exists some constant $C > 0$, which may depend on some fixed parameters, such that $a_n \leq Cb_n$ for all $n \geq 1$. Also, we write $a_n \asymp b_n$ if there exist some positive constants C_1, C_2 such that $C_1 \leq |a_n/b_n| \leq C_2$ for all $n \geq 1$. We say $X_n = o_P(a_n)$ when $X_n/a_n \xrightarrow{P} 0$, and $a_n = o(1)$ when $a_n \to 0$ as $n \to \infty$. Also, we write $a_n = O(1)$ when $|a_n| \leq C$ for some constant $C > 0$ for all large n. The symbol Φ represents the cumulative distribution function of the standard normal random variable

Table 2: Summary of Theoretical Results and Guarantees.

Result	Brief Description	Location
Theorem 2.3	Asymptotic power and size control of the oracle semi-supervised test	Section 2.1
Corollary 2.4	Asymptotic power approximation of the cross-fit test to the oracle test	Section 2.2
Theorem 3.4	Asymptotic level and power guarantees of the xssMMD test compared to xMMD	Section 3.2
Lemma 3.5	Power consistency of the xssMMD test against fixed and local alternatives	Section 3.2
Theorem D.1	Power consistency of xssMMD under weaker conditions on estimated conditional expectations	Appendix D.1
Corollary D.2	Power consistency of the xssMMD test against fixed alternatives	Appendix D.1
Corollary D.3	Power consistency of the xssMMD test against smooth local alternatives	Appendix D.1
Corollary D.4	Sufficient conditions for Assumption C.2 using linear smoothers	Appendix D.2
Theorem D.6	Explicit asymptotic power expression of xssMMD under Gaussianity and a linear kernel	Appendix D.3

$N(0,1)$ and the α quantile of $N(0,1)$ is denoted as $z_\alpha = \Phi^{-1}(\alpha)$. For two real numbers a and b, we use $a \wedge b$ and $a \vee b$ to denote $\min(a,b)$ and $\max(a,b)$, respectively. In numerical studies, we denote $\mathbf{0}_d$ as the all-zeros vector in $\mathbb{R}^d$, $\mathbb{1}_d$ as the all-ones vector in $\mathbb{R}^d$, and I_d as the $d \times d$ identity matrix.

C Detailed Discussion on Assumptions for Theorem 3.4

In this section, we formally define the centered kernel quantities and provide a more in-depth discussion of the assumptions introduced in Section 3.2. These assumptions are crucial for establishing the asymptotic properties of the xssMMD test and provide a framework for understanding the conditions under which the xssMMD test achieves improved performance. The assumptions involve key quantities defined through the centered kernel $\overline{k}_X$, which captures pairwise relationships while removing marginal effects. Specifically, $\overline{k} \coloneqq \overline{k}_X$ with respect to $X_1, X_2 \overset{\text{i.i.d.}}{\sim} P_X$ is defined as:

$$\overline{k}(x_1, x_2) \coloneqq k(x_1, x_2) - \mathbb{E}[k(X_1, X_2) \mid X_1 = x_1] - \mathbb{E}[k(X_1, X_2) \mid X_2 = x_2] + \mathbb{E}[k(X_1, X_2)]. \tag{7}$$

Based on this, we further define $\overline{g}_X(x_1, x_2) \coloneqq \mathbb{E}[\overline{k}(x_1, X_1)\overline{k}(x_2, X_1)]$, which quantifies the dependence structure of X and encodes covariance-like properties.

Discussion on Assumption 3.1: We note that similar conditions have been considered in the literature. For example, Kim and Ramdas (2024, page 23, condition (29b)) introduced a related assumption in the context of high-dimensional testing with U-statistics. Likewise, Shekhar et al. (2022, Theorem 5) proposed a similar but stronger condition to establish the asymptotic normality of the test statistic in the xMMD test. This moment condition is also used in the proof of Li and Yuan (2024, Theorem 1), where it is shown to hold for the Gaussian kernel with bandwidths that grow at specific rates relative to the sample size.

As highlighted in the main text, Assumption 3.1 is a Lyapunov-type condition. A set of bounded kernels (such as the Gaussian kernel used in our experiments) serves as a primary example of a class satisfying this condition, provided that the variance is lower-bounded by a constant. Because bounded kernels imply that the centered kernel functions are uniformly bounded, all higher-order moments in the numerator of Assumption 3.1 are bounded by finite constants. Meanwhile, the denominator, with the variance term not vanishing too rapidly, grows with the sample size n. Consequently, the ratio vanishes asymptotically ($o(1)$), implying that the assumption is satisfied for this practical class of kernels.

Discussion on Assumption 3.2: Assumption 3.2 ensures that the conditional expectations of $\widehat{f}(X)$ and $\widehat{f}(Y)$ are estimated with sufficient accuracy relative to their variances. Under these conditions, we rigorously show that the test statistic, after applying cross-fitting, is asymptotically equivalent to the oracle test statistic using the true conditional expectations.

Importantly, the boundedness of the kernel ensures that the witness function is uniformly bounded. In non-parametric regression theory (Györfi et al., 2006), the boundedness of the target function is a sufficient condition to establish the consistency of standard estimators like k-NN and kernel regression. Therefore, the use of a bounded kernel ensures that the numerator in Assumption 3.2 converges to zero, satisfying the condition for any consistent regression method. In Appendix D.2, we provide a more concrete condition for the consistency of the conditional expectations in the context of linear smoothers.

Discussion on Assumption 3.3: The condition for bounded ratios between density functions is not necessary to obtain the asymptotic normality of the test statistic, but it greatly simplifies our conditions. In particular, this allows the expectations associated with X and Y to be comparable up to a constant factor, which results in a more concise expression of the condition. For instance, under the condition, we can write $\mathbb{E}[\bar{g}_X(Y,Y)] \asymp \mathbb{E}[\bar{g}_X(X,X)]$ which helps simplify the necessary conditions. The additional condition in Assumption 3.3 expected to be satisfied under a broad range of alternatives with small $\mathrm{MMD}(P_X, P_Y)$. At a high-level, this condition is derived while verifying the Lyapunov central limit theorem for the test statistic. For further details, refer to the proof of Theorem 3.4.small $\mathrm{MMD}(P_X, P_Y)$.

It is important to clarify the implications if Assumption 3.3 fails. Technically, this assumption is a sufficient condition for establishing asymptotic normality under the alternative, rather than a strict condition for variance reduction. Consequently, if it fails, we cannot use the analytic formula derived in Theorem 3.4 to directly compare the power. However, this does not automatically imply that the asymptotic power of Ψ_{xss} is lower than that of Ψ_x.

That being said, there are specific scenarios where Ψ_{xss} could underperform relative to Ψ_x. This typically occurs in finite-sample regimes when the conditional expectation is estimated poorly, leading to increased variance. For instance, if one uses a complex regression model (e.g., a deep neural network) on a small sample size where covariates V are completely independent of X, the model may overfit to the noise in V. In this case, the estimated residuals will have higher variance than the original data, which can lead to a power loss compared to the supervised baseline Ψ_x. This highlights the important distinction between the condition required for asymptotic normality and the finite-sample estimation risks.

D Additional Theoretical Results

In this section, we extend our theoretical findings in several directions. First, we compare the conditions under which the xssMMD test achieves power consistency with those of the xMMD test. We demonstrate that the xssMMD test attains consistency under weaker assumptions than the xMMD test, thereby highlighting its broader applicability. Moreover, we consider a linear smoothing approach for estimating conditional expectations, which enables us to derive a simplified version of Theorem 3.4.

D.1 Consistency in Power

In this subsection, we further examine the consistency in power of the xssMMD test and its relationship with the xMMD test. Notably, the xMMD statistic can be regarded as a special case of the xssMMD statistic with $m_1 = m_2 = 0$. To explicitly define the xMMD statistic, we set $\widetilde{\mu}_{X,\widehat{f}} = n_1^{-1}\sum_{i=1}^{n_1}\widehat{f}(X_i)$, $\widetilde{\mu}_{Y,\widehat{f}} = n_2^{-1}\sum_{i=1}^{n_2}\widehat{f}(Y_i)$, $\widetilde{\sigma}^2_{X,\widehat{f}} = n_1^{-2}\sum_{i=1}^{n_1}\{\widehat{f}(X_i) - \widetilde{\mu}_{X,\widehat{f}}\}^2$, and $\widetilde{\sigma}^2_{Y,\widehat{f}} = n_2^{-2}\sum_{i=1}^{n_2}\{\widehat{f}(Y_i) - \widetilde{\mu}_{Y,\widehat{f}}\}^2$. The xMMD test statistic is then defined as

$$\mathrm{x}\widehat{\mathrm{MMD}}^2 = \frac{\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}}{\sqrt{\widetilde{\sigma}^2_{X,\widehat{f}} + \widetilde{\sigma}^2_{Y,\widehat{f}}}}.$$

Similarly to Lemma 3.5, Shekhar et al. (2022, Theorem 8) shows that the xMMD test is consistent in power under the following condition:

$$\sup_{(P_{X,n},P_{Y,n})\in\mathcal{P}_n}\left\{\frac{\mathbb{E}_{P_{X,n},P_{Y,n}}\big[\widetilde{\sigma}^2_{X,\widehat{f}} + \widetilde{\sigma}^2_{Y,\widehat{f}}\big]}{\delta_n\gamma_n^4} + \frac{\mathrm{Var}_{P_{X,n},P_{Y,n}}\big[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\big]}{\gamma_n^4}\right\} = o(1), \tag{8}$$

where δ_n is any positive sequence converging to zero and $\gamma_n = \mathrm{MMD}(P_{X,n}, P_{Y,n})$. We now show that the condition in (8) is stronger than that in Lemma 3.5 whenever the second moments of the residuals $\widehat{f}(X) - \widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]$ and $\widehat{f}(Y) - \widehat{\mathbb{E}}[\widehat{f}(Y)\,|\,W,\widehat{f}]$ are comparable to the variances of $\widehat{f}(X)$ and $\widehat{f}(Y)$, respectively.

Theorem D.1. *Suppose the consistency condition for the xMMD test in* (8) *holds. Moreover, suppose that* $\widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}] := \widehat{u}_X(V)$ *and* $\widehat{\mathbb{E}}[\widehat{f}(Y)\,|\,W,\widehat{f}] := \widehat{u}_Y(W)$ *satisfy the following conditions:*

$$\frac{\mathbb{E}\big[\big\{\widehat{f}(X) - \widehat{u}_X(V)\big\}^2\big]}{\mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big]} \lesssim 1 \quad \text{and} \quad \frac{\mathbb{E}\big[\big\{\widehat{f}(Y) - \widehat{u}_Y(W)\big\}^2\big]}{\mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(Y)\,|\,\widehat{f}\big\}\big]} \lesssim 1. \tag{9}$$

Then Lemma 3.5 remains valid.

Proof. The proof can be found in Appendix F.5. □

As emphasized in the main text, the conditions in (9) are much weaker than the full consistency of the conditional expectations $\widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]$ and $\widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]$. In particular, when $\widehat{\mathbb{E}}[\widehat{f}(X) \mid V, \widehat{f}]$ and $\widehat{\mathbb{E}}[\widehat{f}(Y) \mid W, \widehat{f}]$ are the true conditional expectations, the conditions in (9) are automatically satisfied by the law of total variance. This finding demonstrates that the xssMMD test is consistent in power whenever the xMMD test is consistent in power under weak conditions on the residuals of the estimated conditional expectations. Furthermore, this result can be applied to establish other consistency results from Shekhar et al. (2022) such as consistency against fixed alternatives and against L_2 local alternatives.

Fixed alternatives. We begin by applying Theorem D.1 to the setting where P_X and P_Y are fixed distributions, and show that the xssMMD test equipped with a characteristic kernel achieves asymptotic power of one in distinguishing P_X and P_Y.

Corollary D.2. *Suppose that distributions P_X, P_Y and a kernel k do not vary with n. If k is a characteristic kernel with $\mathbb{E}_{P_X}[k(X_1, X_1)] < \infty$ and $\mathbb{E}_{P_Y}[k(Y_1, Y_1)] < \infty$, and condition* (9) *holds, then the* xssMMD *test is consistent against the fixed alternative $H_1 : P_X \neq P_Y$.*

The proof of this statement is given in Appendix F.6.

Smooth local alternatives. We next demonstrate Theorem D.1 to the setting where the distributions $P_{X,n}$ and $P_{Y,n}$ admit Lebesgue densities $p_{X,n}$ and $p_{Y,n}$ which belong to a Sobolev ball of order β for some $\beta > 0$. Specifically, we consider the following class of smooth densities:

$$\mathcal{W}^{\beta,2}(M) := \left\{ f : \mathcal{X} \to \mathbb{R} \mid f \text{ is almost surely continuous and } \int \left(1 + \omega^2\right)^{\beta/2} \|\mathcal{F}(f)(\omega)\|^2 d\omega < M < \infty \right\},$$

where $\mathcal{F}(f)$ is the Fourier transform of f and $\|\cdot\|$ denotes the Euclidean norm. We then define a class of alternative distributions that is Δ_n-close to the null hypothesis in the L_2-norm:

$$\mathcal{P}_n^{(1)} = \left\{ (P_X, P_Y) \text{ with densities } p_X, p_Y \in \mathcal{W}^{\beta,2}(M) : \|p_X - p_Y\|_{L_2} \geq \Delta_n \right\}$$

for some sequence Δ_n decaying to zero. The next theorem, which is the corresponding result to Shekhar et al. (2022, Theorem 9), establishes the consistency in power of the xssMMD test against $\mathcal{P}_n^{(1)}$.

Corollary D.3. *Consider the* xssMMD *test Ψ_{xss} with the Gaussian kernel $k_{s_n}(x, y) = \exp(-s_n \|x - y\|^2)$ with the scale parameter $s_n \asymp n^{4/(d+4\beta)}$. If condition* (9) *holds and $\lim_{n\to\infty} \Delta_n n^{2\beta/(d+4\beta)} = \infty$ with $n_1 = n_2 = n$, then Ψ_{xss} is consistent against $\mathcal{P}_n^{(1)}$ as*

$$\lim_{n\to\infty} \inf_{(P_{X,n}, P_{Y,n}) \in \mathcal{P}_n^{(1)}} \mathbb{E}_{P_{X,n}, P_{Y,n}}[\Psi_{\mathrm{xss}}] = 1.$$

Proof. We omit the proof of Corollary D.3 since it is a direct consequence of Theorem D.1 and the proof of Shekhar et al. (2022, Theorem 9). □

The corollary shows that the xssMMD test is consistent against smooth local alternatives under the same conditions as the xMMD test. This result highlights that the xssMMD test achieves the same separation rate as the xMMD test, provided the estimated conditional expectations satisfy the requirements in (9).

D.2 Linear Smoother for Estimating Conditional Expectations

In this subsection, we consider a linear smoother (e.g., k-nearest neighbors and kernel regression) for estimating conditional expectations, which provides a more interpretable condition for the consistency of the conditional expectations in Assumption 3.2. The simplification of condition (5) is derived using the spectral decomposition of the centered kernel $\overline{k}$ in (7):

$$\overline{k}(x_1, x_2) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x_1) \phi_i(x_2),$$

where $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues, and $\{\phi_i\}_{i=1}^{\infty}$ are the orthonormal eigenfunctions under the marginal distribution of X. Given this decomposition, we express the conditional expectation and variance in terms of eigenfunctions and derive a clearer condition as follows:

Corollary D.4. *Suppose that the estimators for the conditional expectations satisfy* $\widehat{\mathbb{E}}[a\widehat{f}(X)+b\,|\,V,\widehat{f}] = a\widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]+b$ *and* $\widehat{\mathbb{E}}[a\widehat{f}(Y)+b\,|\,W,\widehat{f}] = a\widehat{\mathbb{E}}[\widehat{f}(Y)\,|\,W,\widehat{f}]+b$ *for all* $a,b\in\mathbb{R}$. *Suppose further that*

$$\sup_{i\geq 1}\mathbb{E}[\Delta_{X,i}^2] = o(1), \quad \textit{and} \quad \sup_{i\geq 1}\mathbb{E}[\Delta_{Y,i}^2] = o(1) \tag{10}$$

where $\Delta_{X,i} = \mathbb{E}[\phi_i(X)\,|\,V] - \widehat{\mathbb{E}}[\phi_i(X)\,|\,V]$ *and* $\Delta_{Y,i} = \mathbb{E}[\phi_i(Y)\,|\,W] - \widehat{\mathbb{E}}[\phi_i(Y)\,|\,W]$, *and Assumption 3.1 holds with* $n_1 \asymp m_1$ *and* $n_2 \asymp m_2$. *Then Assumption 3.2 is satisfied.*

Proof. The proof can be found in Appendix F.7. □

The condition in (10) essentially states that if the regression estimator is linear and consistent for estimating the conditional expectations of the eigenfunctions $\mathbb{E}[\phi_i(X)\,|\,V]$ and $\mathbb{E}[\phi_i(Y)\,|\,W]$, then Assumption 3.2 holds. This condition is particularly notable as it translates the relatively abstract stochastic requirement in (5) into a deterministic one that does not depend on $\widehat{f}$. Moreover, we note that the linearity of the regression estimator can be relaxed to asymptotic linearity with more technical efforts.

D.3 Asymptotic Power Expression using a Linear Kernel

In the main text, we showed that the xssMMD test achieves asymptotic power at least as high as that of the xMMD test while maintaining controlled type-I error. To further support this finding, we formalize the asymptotic expression of the power of our statistic using the linear kernel $k(x,y)=\langle x,y\rangle$, and compare it to the heuristic results of the kernel-MMD test and the xMMD test. Before proceeding with further discussion, let us assume the following conditions to ease our analysis.

Assumption D.5. Suppose that the following assumptions are satisfied.

(a) Gaussianity: We observe d-dimensional i.i.d. copies of random vectors $(X,V)^\top$ and $(V,W)^\top$ from a Gaussian distributions

$$P_{XV} = \begin{pmatrix} P_X \\ P_V \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_V \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \quad \text{and}$$
$$P_{YW} = \begin{pmatrix} P_Y \\ P_W \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_W \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

(b) Bounded eigenvalues: For $i=1,2$, there exist constants c and $C>0$ such that $c\leq\lambda_1(\Sigma_{ii})\leq\cdots\leq\lambda_d(\Sigma_{ii})\leq C$.

(c) Local alternative: $\mu_X^\top\mu_X = O(\sqrt{d}/n_1)$, $\mu_Y^\top\mu_Y = O(\sqrt{d}/n_2)$.

(d) Dimension-to-sample size ratio: $d/n_1\to\tau_1\in(0,\infty)$, $d/n_2\to\tau_2\in(0,\infty)$.

(e) Labeled-unlabeled sample size ratio: $m_1/(n_1+m_1)\to r_1\in(0,1)$, $m_2/(n_2+m_2)\to r_2\in(0,1)$.

We note that these conditions are only necessary for deriving the concrete, asymptotic power expression of the proposed test. These conditions are analogous to those given by Kim and Ramdas (2024, Assumption 2.5) and Assumption D.5 can be seen as its two-sample testing version extension in the semi-supervised setting.

For the xssMMD test defined as $\Psi_{\mathrm{xss}} := \mathbb{1}(\mathrm{xss}\widehat{\mathrm{MMD}}^2 > z_{1-\alpha})$, we analyze its power assuming the previous Assumption D.5 holds.

Theorem D.6. *Suppose that Assumption 3.2 and Assumption D.5 are fulfilled under the alternative. Assume that* $(X,V)^\top$ *and* $(V,W)^\top$ *have equal sample sizes and equal covariance matrices, i.e.,* $n_1=n_2=m_1=m_2=n$ *and* $\Sigma_{ij}=\tilde{\Sigma}_{ij}$ *for* $i,j\in\{1,2\}$. *Then, it holds that*

$$\mathbb{E}[\Psi_{\mathrm{xss}}] = \Phi\left(z_\alpha + \frac{n(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y)}{\sqrt{4\mathrm{tr}(\Sigma_{11}^2) - 2\mathrm{tr}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11})}}\right) + o_P(1).$$

Proof. The proof can be found in Appendix F.8 □

We note that the constant value of 2 comes from $4r$ where r denotes the labeled-unlabeled sample size ratio $r := r_1 = r_2$ and we assumed $r = 1/2$ in the above theorem. From this, we can also obtain the result when there is no unlabeled data, meaning $r = 0$, which is identical to the heuristic result of the permutation-free kernel two-sample test from Shekhar et al. (2022). This suggests that our finding can be viewed as an extension of the earlier results from the xMMD test, incorporating additional covariates, which may result in a reduction of variance calculated in the denominator and an increase in power.

Conversely, the asymptotic expressions for the power of the kernel two-sample test using permutation on MMD, as suggested by Gretton et al. (2012a), denoted as Ψ_{perm}, and the permutation-free kernel two-sample test using studentized MMD by Shekhar et al. (2022), represented as Ψ_{x}, can be expressed as follows:

$$\mathbb{E}[\Psi_{\text{perm}}] \approx \Phi\left(z_\alpha + \frac{n(\mu_X - \mu_Y)^\top(\mu_X - \mu_Y)}{\sqrt{2\text{tr}(\Sigma_{11}^2)}}\right), \quad \text{and}$$
$$\mathbb{E}[\Psi_{\text{x}}] \approx \Phi\left(z_\alpha + \frac{n(\mu_X - \mu_Y)^\top(\mu_X - \mu_Y)}{\sqrt{4\text{tr}(\Sigma_{11}^2)}}\right).$$

We note that the result for Ψ_{perm} is estimated in heuristic manner, while that of Ψ_{x} is derived from its asymptotic normality under the alternative, shown in Theorem 3.4. Observe that the power of Ψ_{x} is $\sqrt{2}$ times lower than that of Ψ_{perm}, a result stemming from sample-splitting. In this context, the value of our test statistic is straightforward, as we initially assumed there were $2n_1$ and $2n_2$ labeled samples at first, then utilized one half to construct the witness function and the other to compute the studentized statistic. These distinctions in power highlight the inherent trade-off introduced by sample-splitting, where the power of the test is reduced in exchange for computational efficiency. However, incorporating additional covariates into our xssMMD framework mitigates this drawback by utilizing the unlabeled data to reduce variance, thereby narrowing the gap in power performance while maintaining robustness. This is evident when comparing the powers of Ψ_{x} and Ψ_{xss}, where $\text{tr}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}) > 0$ indicates that Ψ_{xss} has greater power than Ψ_{x}. Furthermore, under Assumption D.5, we identified explicit conditions under which the power of Ψ_{xss} exceeds that of Ψ_{perm}. Specifically, when $2\text{tr}(\Sigma_{11}^2) < \text{tr}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11})$, Ψ_{xss} demonstrates superior power compared to Ψ_{perm}, which implies that leveraging highly informative additional covariates can enhance the power. This underscores the significant practical advantage of our proposed test, especially in scenarios where unlabeled data is abundant and can be effectively utilized.

E Additional Experiments

In this section, we present additional numerical results and provide detailed information about our experiments. In implementing our proposal, we incorporate various methods for conditional expectation estimation, including k-nearest neighbors (knn), kernel regression, and random forest (rf). In all tables throughout this section, boldface indicates the best performance: the highest test power under the alternative, and the lowest Type-I error rate under the null.

We have limited our scope to a standard kernel-based baseline (e.g., MMD with a Gaussian kernel using median heuristic) to clearly isolate and evaluate the contribution of unlabeled data to statistical power. Incorporating recent advanced methods (e.g., Biggs et al., 2024; Schrab et al., 2023; Kübler et al., 2022) could potentially yield stronger empirical performance. However, doing so at this stage may conflate gains attributable to semi-supervised information with those arising from more refined kernel choices. We believe that establishing the value of unlabeled data in a controlled setting is a necessary first step.

Experiments under the null and alternative are lightweight and can be conducted on a local machine without GPU acceleration, taking approximately an hour each. The HTRU2 experiments run efficiently on CPU and complete within minutes. The MNIST experiments also run on CPU but complete within an hour. In contrast, the CUB experiments require computing image embeddings using a pretrained model, for which GPU acceleration is beneficial. Once embeddings are obtained, the remaining computations are lightweight. We used an NVIDIA RTX A6000 GPU and it is done within minutes for each embedding setting. Reproducible code is available at https://github.com/gyumin-lee68/ssk2st under the MIT License.

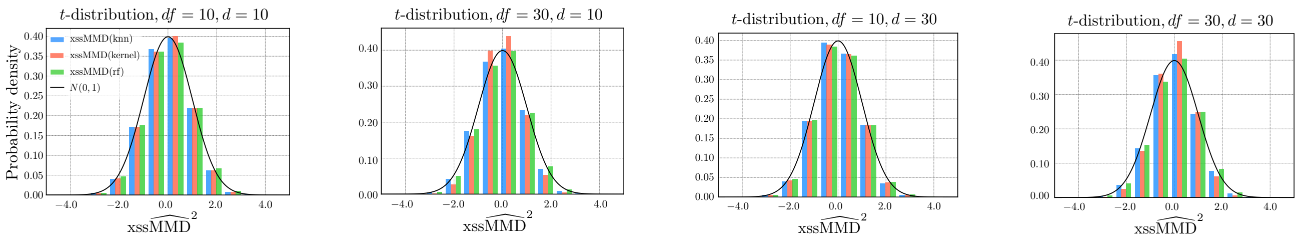

Figure 4: Experimental results for the distribution of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ under the null, using t-distributed data with varying dimension and degrees of freedom. The plots demonstrate that the test statistic $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ asymptotically follows a $N(0, 1)$ distribution under the null, even when the data deviates from Gaussianity.

E.1 Limiting Null Distribution with Different Settings

Along with Section 4, this subsection examines the asymptotic normality of the xssMMD test statistic under the null across various scenarios and parameter settings.

To validate the asymptotic normality of the xssMMD test under the null, we conduct experiments across various settings, including different data distributions, dimensions, and dependency structures. First, we examine the behavior of the test statistic when the data follows a t-distribution. The results are presented in Figure 4, with varying degrees of freedom df and dimension d: $df = 10$ & $d = 10$, $df = 30$ & $d = 10$, $df = 10$ & $d = 30$, and $df = 30$ & $d = 30$ from left to right. Note that we fixed the other parameters as $n_1 = n_2 = 100$ and $m_1 = m_2 = 200$ using a Gaussian kernel with the median heuristic. The results confirm that $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ follows $N(0, 1)$ under the null, even when the data deviate from Gaussianity. This demonstrates that the asymptotic normality of the proposed statistic remains robust across different distributional settings.

Next, we show that the xssMMD test consistently achieves asymptotic normality under the null, regardless of the specific factors outlined in Section 4. In detail, we demonstrate that its asymptotic normality is consistently achieved despite the effects of dimensionality, sample skewness, labeled-unlabeled sample size ratio, methods for estimating conditional expectation, and choice of kernel on the null distribution of the test statistic. The experimental results are summarized in Figure 5 whose each column represents the different cases of sample skewness $r_{sample} := n_1/n_2$ and labeled-unlabeled sample size ratio $r_{label} := n_1/m_1 = n_2/m_2$: $r_{sample} = 1$ & $r_{label} = 1$, $r_{sample} = 0.1$ & $r_{label} = 0.5$, $r_{sample} = 1$ & $r_{label} = 1$, $r_{sample} = 0.1$ & $r_{label} = 0.5$ from left to right with fixed $n_1 = 100$. Each row corresponds to the different cases based on dimension d and kernel choice: $d = 10$ & bilinear kernel, $d = 100$ & bilinear kernel, $d = 10$ & Gaussian kernel, $d = 100$ & Gaussian kernel from top to bottom. Note that we used the median heuristic as the bandwidth for the Gaussian kernel. These results confirm that $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ follows $N(0, 1)$ under the null, regardless of variations in dimensionality, sample skewness, the labeled-unlabeled sample size ratio, estimation methods for conditional expectations, and kernel choice. This robustness further supports the validity of our theoretical findings, demonstrating that the asymptotic normality of the xssMMD test consistently holds across diverse settings.

Lastly, we investigate the robustness of the xssMMD test under different structural dependencies. We conduct experiments where X and V (or Y and W) exhibit dependence which correspond to the settings used in the power analysis (Scenario 1 (Alt) to Scenario 4 (Alt) in Section 4). Figure 6 displays the empirical distribution of the xssMMD test statistic in these scenarios. The results demonstrate that even in the presence of dependencies, the standardized test statistic consistently follows $N(0, 1)$. This confirms that the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ holds even when the covariates are not independent, thereby supporting the validity of the xssMMD test across a wide range of dependency structures.

These findings confirm the robustness of the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ across a wide range of conditions. This further affirms the validity of our theoretical results that the xssMMD test is asymptotically level α under the null, as shown in Theorem 3.4.

E.2 Power Curve with Different Settings

In this subsection, we investigate the power curves of the xssMMD test statistic in different settings.

First, we evaluate the performance of the xssMMD test when additional covariates are independent of the labeled data. Unlike cases where unlabeled data provides valuable information to enhance power, this experiment demonstrates that the xssMMD test does not outperform other tests when the additional covariates contain no useful information. The first two subfigures in Figure 7 illustrate each case corresponding to Scenario 1 (Alt)

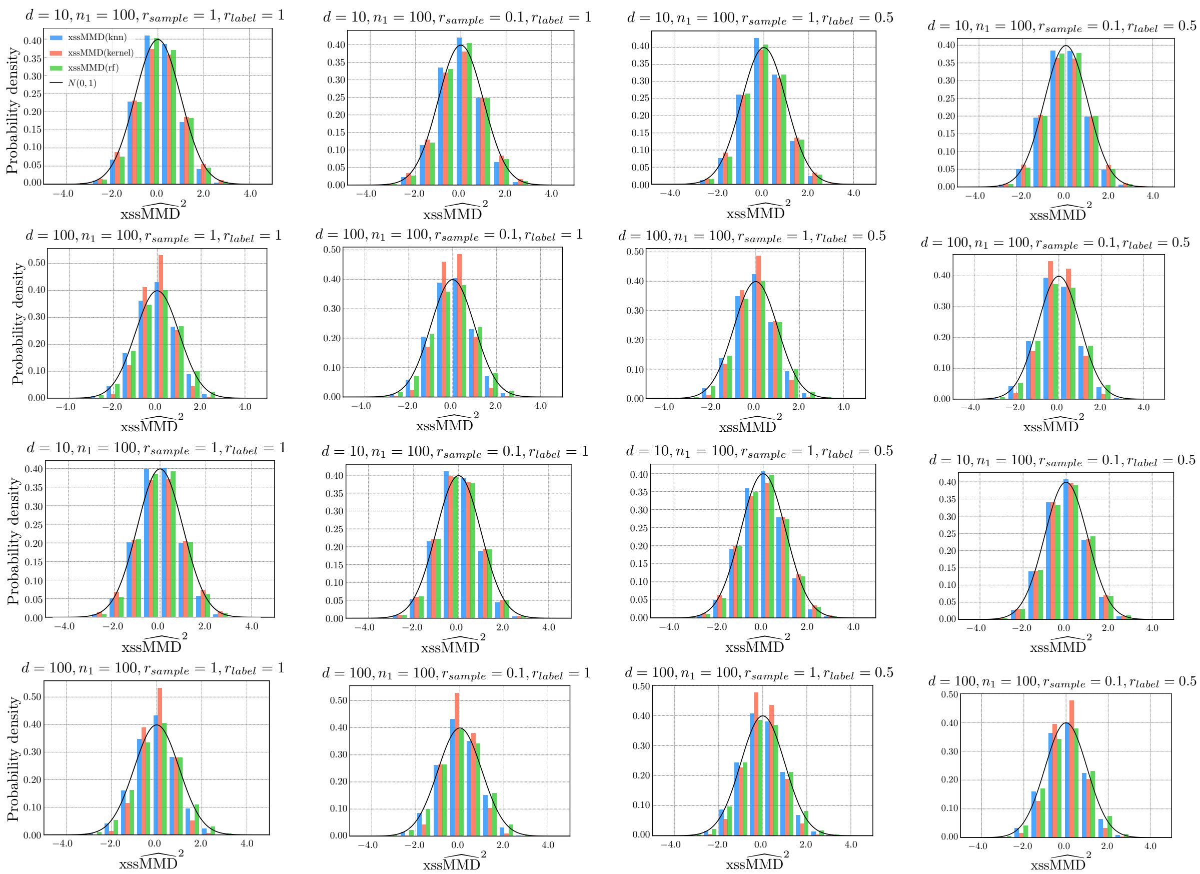

Figure 5: Experimental results for the distribution of $\text{xss}\widehat{\text{MMD}}^2$ under the null hypothesis across all scenarios explained in Section 4. The plots illustrate that the test statistic consistently adheres to a $N(0, 1)$ distribution under various parameter settings. These comprehensive results confirm the validity of the xssMMD test across a broad range of conditions.

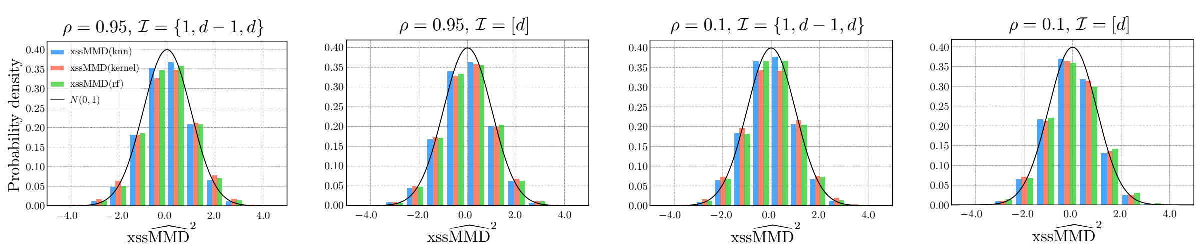

Figure 6: Experimental results for the distribution of $\text{xss}\widehat{\text{MMD}}^2$ under the null across the scenarios in Section 4, particularly from Scenario 1 (Alt) to Scenario 4 (Alt). These settings introduce dependencies between X and V or Y and W, deviating from the standard independence assumption. The plots illustrate that even in the presence of such dependencies, the test statistic continues to follow a $N(0, 1)$ under the null.

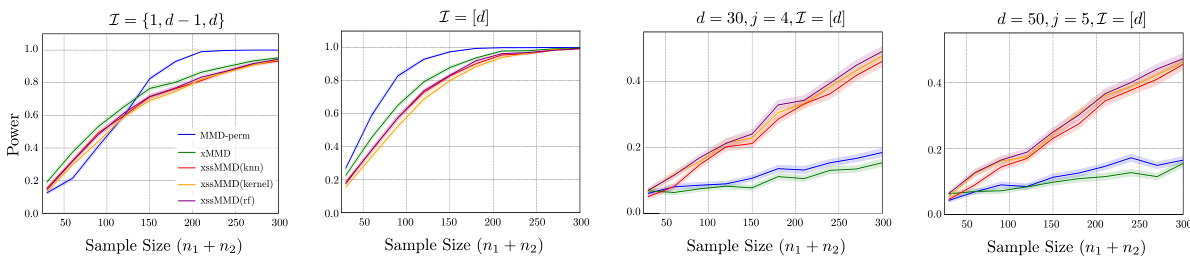

Figure 7: Power analysis of the xssMMD test in various settings. The first two subfigures depict scenarios in which additional covariates are independent of the labeled data. This result confirms that independent covariates do not enhance the performance of the xssMMD test. The last two subfigures illustrate the impact of varying the dimension d, demonstrating that the xssMMD test maintains superior power when X and Y show strong dependence on V and W, even as the dimension increases.

and Scenario 2 (Alt), respectively, but with X and Y sampled independently from P_X and P_Y, both following $N(\mathbf{0}_d, I_d)$. In such scenarios, the power curves of the xssMMD test closely align with those of the xMMD test and fall below those of the MMD-perm test, consistent with our conclusions about the impact of independent covariates.

Additionally, we conduct experiments to verify that the theoretical results on the power derived in Theorem 3.4 hold across various dimensions. Using the same construction method outlined in Section 4 for Scenario 1 (Alt) through Scenario 4 (Alt), we vary the dimension d and the parameter j, which represents the difference between the mean vectors of V and W. Recall that both are sampled from a Gaussian distribution with mean vectors $\mathbf{0}_d$ and $\mathbf{a}_{\epsilon,j}$ where the first j entries of $\mathbf{a}_{\epsilon,j}$ are equal to ϵ and the remaining entries are equal to 0. The last two subfigures in Figure 7 correspond to the cases where the dimension is set to $d = 30$ and $d = 50$, while the mean vector difference is set to $j = 4$ and $j = 5$, respectively. As illustrated in the last two subfigures in Figure 7, even as the dimension increases, the xssMMD test remains robust and consistently surpasses both the MMD-perm and xMMD tests, confirming its superior power in settings where the unlabeled data are informative.

Furthermore, we investigate how the amount of unlabeled data influences the test power. We conducted additional simulations adopting the settings of Scenario 1 (Alt) through Scenario 4 (Alt) from Section 4. Fixing the labeled sample sizes at $n_1 = n_2 = 100$, we varied the unlabeled sample sizes $m_1 = m_2$ from 0 to 2000. As shown in Table 3, the power of the xssMMD test increases monotonically as the size of the unlabeled data grows, confirming that our method effectively leverages additional information. In contrast, the power of MMD-perm and xMMD remains constant. Notably, when $m_1 = m_2 = 0$, xssMMD is mathematically identical to xMMD, yielding the same power.

E.3 Comparison with Joint Two-Sample Tests

In this section, we highlight the specific advantages of our semi-supervised test by comparing it against a standard two-sample test applied directly to the joint distributions, P_{XV} vs. P_{YW}. To systematically evaluate this, we generate 10-dimensional Gaussian vectors with a strong correlation of $\rho = 0.95$ and induce distributional differences using a mean shift of $\epsilon = 0.3$ and $j = 1$.

Under the null ($P_X = P_Y$), we deliberately construct a shift in the auxiliary covariates ($P_V \neq P_W$). Specifically, both X and Y contain the mean shift in their first coordinate. However, V is extracted from the shifted first two coordinates of X, while W is extracted from the unshifted last two coordinates of Y, creating a pure auxiliary shift. As shown in Table 4, the joint MMD test exhibits a Type-I error rate approaching 1.0, as it incorrectly flags the irrelevant shift as a discrepancy of interest. In contrast, xssMMD effectively ignores this shift in the auxiliary space and reliably maintains nominal control of $\alpha = 0.05$. To explicitly demonstrate this robustness, Table 5 shows the effect of varying the shift magnitude ϵ. As the auxiliary shift intensifies from 0.0 to 1.0, the joint tests rapidly collapse, whereas all variants of xssMMD consistently maintain the Type-I error near the nominal level, remaining unaffected by the auxiliary noise.

Under the alternative ($P_X \neq P_Y$), we configure the environment such that the auxiliary covariates are completely identical ($P_V = P_W$). Here, only X contains the mean shift, but both V and W are extracted strictly from the

Table 3: Estimated test power for Scenario 1 (Alt) through Scenario 4 (Alt) with fixed $n_1 = n_2 = 100$ and varying unlabeled sample sizes $m_1 = m_2$.

		$m_1 = m_2$									
Scenario	Test	0	222	444	666	888	1111	1333	1555	1777	2000
Scenario 1 (Alt)	MMD-perm	**0.136**	0.147	0.129	0.157	0.126	0.125	0.116	0.128	0.151	0.120
	xMMD	0.097	0.105	0.107	0.125	0.117	0.106	0.092	0.090	0.122	0.114
	xssMMD(knn)	0.097	0.179	0.242	0.279	0.319	0.340	0.350	0.366	0.403	0.388
	xssMMD(ker)	0.097	**0.215**	0.251	0.286	0.315	0.344	0.348	0.365	0.375	0.360
	xssMMD(rf)	0.097	0.193	**0.254**	**0.301**	**0.343**	**0.362**	**0.366**	**0.405**	**0.434**	**0.421**
Scenario 2 (Alt)	MMD-perm	**0.129**	0.099	0.155	0.129	0.126	0.112	0.121	0.113	0.107	0.108
	xMMD	0.088	0.088	0.116	0.109	0.092	0.092	0.091	0.088	0.093	0.089
	xssMMD(knn)	0.088	0.159	0.215	0.261	0.286	0.309	0.331	0.342	0.372	0.387
	xssMMD(ker)	0.088	**0.176**	**0.244**	0.261	0.312	0.316	0.320	0.325	0.351	0.374
	xssMMD(rf)	0.088	0.161	0.242	**0.279**	**0.317**	**0.326**	**0.358**	**0.361**	**0.399**	**0.424**
Scenario 3 (Alt)	MMD-perm	0.264	0.257	0.231	0.253	0.267	0.230	0.272	0.261	0.247	0.251
	xMMD	0.208	0.197	0.193	0.189	0.200	0.179	0.205	0.188	0.177	0.206
	xssMMD(knn)	0.208	0.249	0.260	0.265	0.290	0.256	0.292	0.279	0.275	0.266
	xssMMD(ker)	0.208	0.261	0.263	0.297	0.293	0.248	0.307	0.298	0.292	0.272
	xssMMD(rf)	0.208	**0.271**	**0.292**	**0.317**	**0.322**	**0.282**	**0.337**	**0.322**	**0.315**	**0.309**
Scenario 4 (Alt)	MMD-perm	**0.406**	0.394	0.389	0.394	0.388	0.435	0.406	0.405	0.416	0.414
	xMMD	0.286	0.284	0.285	0.283	0.281	0.302	0.278	0.295	0.289	0.278
	xssMMD(knn)	0.286	0.441	0.500	0.508	0.557	0.549	0.532	0.544	0.533	0.538
	xssMMD(ker)	0.286	**0.456**	**0.511**	**0.540**	**0.564**	**0.564**	**0.544**	**0.559**	**0.547**	**0.575**
	xssMMD(rf)	0.286	0.385	0.413	0.407	0.400	0.453	0.439	0.429	0.445	0.448

Table 4: Estimated Type-I error of the joint and marginal tests under the null when $P_X = P_Y$ and $P_V \neq P_W$. The joint MMD-perm test incorrectly rejects the null, while xssMMD maintains nominal control of $\alpha = 0.05$.

	$n_1 + n_2$									
Test	20	40	60	80	100	120	140	160	180	200
MMD-perm (Joint)	0.143	0.272	0.459	0.670	0.837	0.929	0.969	0.996	1.000	1.000
xMMD	0.170	0.246	0.407	0.546	0.851	0.769	0.711	0.916	0.953	0.973
xssMMD(knn)	**0.037**	**0.046**	0.052	**0.050**	0.051	0.049	0.044	0.049	**0.048**	**0.042**
xssMMD(ker)	0.046	0.048	**0.043**	0.051	**0.047**	**0.046**	0.053	**0.048**	0.051	0.050
xssMMD(rf)	0.046	0.055	0.051	0.055	0.056	**0.046**	**0.043**	0.049	0.053	0.054

Table 5: Estimated Type-I error of the joint and marginal tests under the null when $P_X = P_Y$ and $P_V \neq P_W$ with varying shift magnitude ϵ.

	ϵ					
Test	0.0	0.2	0.4	0.6	0.8	1.0
MMD-perm	0.079	0.225	0.800	1.000	1.000	1.000
xMMD	0.178	0.457	0.736	0.925	0.984	1.000
xssMMD(knn)	**0.047**	**0.038**	**0.042**	0.057	0.047	0.055
xssMMD(ker)	0.061	0.049	0.045	**0.056**	**0.044**	**0.054**
xssMMD(rf)	0.056	0.053	0.054	0.061	0.045	0.056

unshifted last two coordinates of X and Y, respectively. In this scenario, the joint test yields lower power as shown in Table 6 since identical auxiliary marginal distributions act as high-dimensional noise, diluting the overall signal. In contrast, xssMMD leverages the underlying correlation between X and V to reduce variance, substantially outperforming both MMD-perm and xMMD. To further validate the utility of our approach, we examine the test power across different correlation strengths ρ in Table 7. When the auxiliary data are uninformative ($\rho = 0.0$), xssMMD retains the baseline power of the standard xMMD test, avoiding any negative transfer. As the correlation increases to 0.95, xssMMD efficiently exploits the dependency structure to achieve higher power, whereas the power of the joint tests remains low.

Table 6: Estimated test power of the joint and marginal tests under the alternative when $P_X \neq P_Y$ and $P_V = P_W$. xssMMD achieves higher power, whereas joint tests struggle to capture the marginal difference.

	$n_1 + n_2$									
Test	20	40	60	80	100	120	140	160	180	200
MMD-perm (Joint)	0.078	0.089	0.121	0.180	0.244	0.303	0.385	0.495	0.633	0.737
xMMD	0.117	0.175	0.274	0.360	0.420	0.496	0.553	0.637	0.680	0.716
xssMMD(knn)	0.097	0.234	0.522	0.696	0.830	0.905	0.935	0.948	**0.975**	0.985
xssMMD(ker)	**0.152**	0.344	0.552	0.681	0.787	0.875	0.906	0.935	0.962	0.969
xssMMD(rf)	0.151	**0.380**	**0.604**	**0.763**	**0.858**	**0.919**	**0.946**	**0.958**	**0.975**	**0.987**

Table 7: Estimated test power of the joint and marginal tests under the alternative when $P_X \neq P_Y$ and $P_V = P_W$ across varying correlation strengths ρ between target and auxiliary covariates.

	ρ					
Test	0.00	0.20	0.40	0.60	0.80	0.95
MMD-perm	**0.185**	0.156	0.132	0.090	0.097	0.080
xMMD	0.146	0.150	0.176	0.140	0.165	0.154
xssMMD(knn)	0.159	**0.165**	0.210	0.195	**0.298**	0.414
xssMMD(ker)	0.165	0.159	**0.221**	0.202	0.297	0.369
xssMMD(rf)	0.155	0.162	0.201	**0.203**	0.290	**0.462**

E.4 Running-Time Comparison

While permutation-based MMD tests provide strong finite-sample guarantees, they are often computationally prohibitive for large datasets. To demonstrate the computational efficiency of our proposed framework, we compared the execution time (in seconds) of xssMMD against xMMD and MMD-perm. We adopted the setting from Scenario 1 (Alt) to Scenario 4 (Alt), simultaneously varying the labeled sample size $n_1 = n_2$ from 10 to 100 and the unlabeled sample size $m_1 = m_2$ from 100 to 1000.

The results, presented in Table 8, show a consistent efficiency ranking across most of the settings: xMMD requires the least amount of time, followed closely by xssMMD, while MMD-perm is the most computationally intensive. Notably, there is a substantial time difference between xssMMD and MMD-perm as the sample size increases. For instance, at $n_1 + n_2 = 200$, xssMMD(knn) is approximately 13 times faster than MMD-perm in every scenarios. These results indicate that xssMMD provides a computationally efficient semi-supervised approach. It mitigates the computational burden of standard permutation tests without sacrificing the statistical benefits of incorporating unlabeled data.

E.5 Details of the Real-World Experiment: HTRU2 Pulsar dataset

This subsection provides a detailed description of our experiments on real-world data using the HTRU2 pulsar dataset. The HTRU2 dataset consists of measurements from radio astronomy, with each sample characterized by eight continuous features derived from the integrated pulse profile (IP) and the DM-SNR curve (DM). The dataset contained 1639 pulsar and 16259 non-pulsar observations and we constructed a testing problem to determine whether the distribution of features differed between the two classes.

Table 8: Running-time comparison (in seconds) across varying sample sizes. The total labeled sample size $n_1 + n_2$ is shown, with the unlabeled sample size fixed at ten times the labeled size ($m_1 + m_2 = 10(n_1 + n_2)$).

		$n_1 + n_2$									
Scenario	Method	20	40	60	80	100	120	140	160	180	200
Scenario 1 (Alt)	MMD-perm	32.594	55.570	354.557	440.820	503.935	500.720	548.205	566.169	575.530	865.501
	xMMD	0.709	0.650	1.061	1.008	3.121	3.489	4.312	3.565	4.147	3.464
	xssMMD(knn)	9.130	13.794	16.985	21.670	28.756	35.236	45.231	49.880	57.319	64.845
	xssMMD(ker)	10.337	16.008	20.623	26.023	34.358	41.907	55.003	57.790	68.878	78.862
	xssMMD(rf)	278.681	270.142	254.706	256.005	268.618	281.739	297.154	308.513	333.594	321.440
Scenario 2 (Alt)	MMD-perm	29.597	58.656	366.912	448.648	500.716	503.202	547.274	570.824	552.232	838.050
	xMMD	0.656	0.721	1.019	1.029	3.224	3.448	4.259	3.798	3.375	3.594
	xssMMD(knn)	8.169	14.346	17.180	22.723	30.735	37.411	49.342	52.368	56.562	66.373
	xssMMD(ker)	9.164	16.864	20.907	27.751	37.016	44.399	60.935	59.687	68.209	81.675
	xssMMD(rf)	277.267	270.677	250.552	259.682	272.121	286.572	314.231	317.082	329.089	325.893
Scenario 3 (Alt)	MMD-perm	29.407	58.977	371.145	442.269	506.101	501.124	553.109	577.414	547.026	824.445
	xMMD	0.660	0.675	0.994	1.015	3.229	3.003	3.908	3.685	3.076	3.356
	xssMMD(knn)	7.972	14.525	16.849	22.682	30.519	36.164	47.830	51.671	55.769	65.830
	xssMMD(ker)	9.031	17.077	20.441	27.328	36.228	42.917	59.162	57.778	65.760	81.043
	xssMMD(rf)	279.430	273.422	250.447	261.256	272.815	287.031	315.216	317.947	332.128	331.039
Scenario 4 (Alt)	MMD-perm	30.948	59.501	352.966	434.063	486.332	480.628	565.353	573.585	540.560	856.321
	xMMD	0.622	0.646	1.066	1.052	3.176	3.600	4.570	3.366	3.637	3.894
	xssMMD(knn)	7.725	13.961	17.659	23.076	30.688	37.273	49.264	51.885	58.866	67.369
	xssMMD(ker)	9.842	16.055	21.594	27.440	36.015	44.364	60.609	59.432	69.667	80.912
	xssMMD(rf)	272.036	265.660	256.013	258.103	269.625	292.752	316.483	321.325	338.474	339.847

Table 9: Estimated test power under different scenarios and increasing noise levels (independent Gaussian noise with standard deviation $\sigma \in \{0, 0.1, 0.3, 0.5, 0.7, 1.0\}$). Each value corresponds to the average test power across 1000 trials. Our proposed method, xssMMD, was implemented using random forest for the conditional expectation estimation.

Labeled Data	Test	$\sigma = 0$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$	$\sigma = 1.0$
IP(Mean), DM(Mean)	MMD-perm	**1.000**	**1.000**	**0.999**	0.859	0.442	0.120
	xMMD	0.964	0.962	0.891	0.690	0.398	0.156
	xssMMD	0.998	0.997	0.983	**0.869**	**0.567**	**0.232**
IP(Mean, SD)	MMD-perm	0.106	0.095	0.052	0.015	0.090	0.013
	xMMD	0.205	0.185	0.123	0.062	0.038	0.030
	xssMMD	**0.560**	**0.531**	**0.342**	**0.181**	**0.105**	**0.061**
IP(Mean, SD, EK, Skew)	MMD-perm	0.262	0.250	0.173	0.084	0.054	0.014
	xMMD	**0.402**	**0.388**	**0.282**	**0.173**	**0.069**	0.053
	xssMMD	0.367	0.361	0.271	0.170	0.064	**0.064**
IP(Mean, SD), DM(Mean, SD)	MMD-perm	**1.000**	**0.999**	**0.985**	0.762	0.360	0.072
	xMMD	0.929	0.927	0.832	0.605	0.256	0.135
	xssMMD	0.999	0.998	0.982	**0.849**	**0.535**	**0.222**
IP(EK, Skew), DM(EK, Skew)	MMD-perm	0.758	0.726	0.552	0.284	0.109	0.043
	xMMD	0.694	0.682	0.608	0.390	0.218	0.085
	xssMMD	**0.941**	**0.944**	**0.880**	**0.674**	**0.405**	**0.183**
IP(Mean, SD, EK), DM(Mean, SD, EK)	MMD-perm	0.985	0.983	0.932	0.636	0.251	0.057
	xMMD	0.860	0.845	0.740	0.537	0.296	0.118
	xssMMD	**0.994**	**0.987**	**0.944**	**0.812**	**0.528**	**0.208**
IP(SD, EK, Skew), DM(SD, EK, Skew)	MMD-perm	0.811	0.786	0.612	0.307	0.106	0.022
	xMMD	0.719	0.713	0.621	0.432	0.232	0.089
	xssMMD	**0.934**	**0.932**	**0.865**	**0.681**	**0.417**	**0.165**

For each trial, we randomly selected $n_1 = 100$ pulsar and $n_2 = 100$ non-pulsar samples as labeled data and the remaining dataset formed the unlabeled data. In detail, X comprised some covariates of random pulsar samples,

Table 10: Estimated test power under different scenarios with different unlabeled data and increasing noise levels (independent Gaussian noise with standard deviation $\sigma \in \{0, 0.1, 0.3, 0.5, 0.7, 1.0\}$). Each value corresponds to the average test power across 1000 trials.

Data Features		Test	$\sigma = 0$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$	$\sigma = 1.0$
Labeled Data	Unlabeled Data							
IP(Mean, SD), DM(Mean, SD)	V: IP(EK, Skew) W: DM(EK, Skew)	MMD-perm	**1.000**	**1.000**	**0.985**	**0.757**	0.290	0.072
		xMMD	0.935	0.932	0.837	0.622	0.345	0.138
		xssMMD	0.968	0.961	0.889	0.687	**0.370**	**0.145**
	V: IP(EK) W: DM(EK)	MMD-perm	**1.000**	**1.000**	**0.985**	**0.757**	0.290	0.072
		xMMD	0.935	0.932	0.837	0.622	0.345	0.138
		xssMMD	0.962	0.956	0.875	0.663	**0.349**	**0.145**
IP(EK, Skew), DM(EK, Skew)	V: IP(Mean, SD) W: DM(Mean, SD)	MMD-perm	0.753	0.730	0.549	0.392	0.130	0.050
		xMMD	0.729	0.728	0.640	0.431	0.259	0.098
		xssMMD	**0.878**	**0.872**	**0.776**	**0.569**	**0.323**	**0.135**
	V: IP(Mean) W: DM(Mean)	MMD-perm	0.753	0.730	0.549	0.392	0.130	0.050
		xMMD	0.729	0.728	0.640	0.431	0.259	0.098
		xssMMD	**0.84**	**0.833**	**0.750**	**0.547**	**0.321**	**0.133**
IP(SD, EK, Skew), DM(SD, EK, Skew)	V: IP(Mean), DM(Mean) W: IP(Mean)	MMD-perm	0.811	0.786	0.612	0.307	0.106	0.022
		xMMD	0.719	0.713	0.621	0.432	0.232	0.089
		xssMMD	**0.851**	**0.836**	**0.760**	**0.544**	**0.306**	**0.122**
	V: IP(Mean) W: DM(Mean)	MMD-perm	0.811	0.786	0.612	0.307	0.106	0.022
		xMMD	0.719	0.713	0.621	0.432	0.232	0.089
		xssMMD	**0.833**	**0.819**	**0.728**	**0.516**	**0.274**	**0.100**

V comprised other covariates of the remaining pulsar observations, and the same applied to Y and W. Input features were standardized before testing.

To simulate meaningful and challenging scenarios, we chose feature subsets as the labeled data with varying information and added noise. We considered six experimental settings, each defined by combining IP and DM statistics: (1) the means of IP and DM; (2) the mean and standard deviation (SD) of the IP; (3) the mean, SD, EK, and Skew of IP; (4) the mean and SD of both IP and DM; (5) the EK and Skew of both IP and DM; and (6) the mean, SD, and EK from both IP and DM. We used the remaining covariates as the unlabeled data for each scenario. For each setting, we added independent Gaussian noise with a standard deviation from 0 to 1 to the labeled data, thereby gradually reducing their discriminative power. This process enabled an examination of how feature groups responded to increasing corruption and how effectively the auxiliary covariates could be leveraged when the primary covariates were severely degraded by noise. Finally, each method used 50 random splits with 20 repetitions per split, and we reported the average test power.

The result is summarized in Table 9. Our method, xssMMD, achieved performance comparable to or significantly exceeding that of both the standard kernel test (MMD-perm) and xMMD across most of the feature settings and noise levels, demonstrating higher power, particularly under higher noise or with weakly informative features. By using unlabeled data effectively, xssMMD is more sensitive in detecting distributional differences. Incorporating auxiliary covariates is especially beneficial when the main features are weak, as shown in this setup. Using additional information clearly enhanced test sensitivity and robustness, as our results demonstrate.

We further investigated the robustness of xssMMD in a more challenging scenario. In this case, the unlabeled data for the pulsar and non-pulsar classes contained different covariates so they could not be used to test the two-sample problem alone. In other words, V and W came from non-comparable feature spaces. The result is summarized in Table 10. Even in this case, xssMMD consistently maintained a test power comparable to or higher than MMD-perm and xMMD across most of the levels of Gaussian noise. This result highlights the key advantage of xssMMD. It successfully integrates information from the auxiliary sets V and W for each class, even though V and W come from different feature spaces and cannot be directly compared. This integration improves the effectiveness of the primary two-sample test on labeled data.

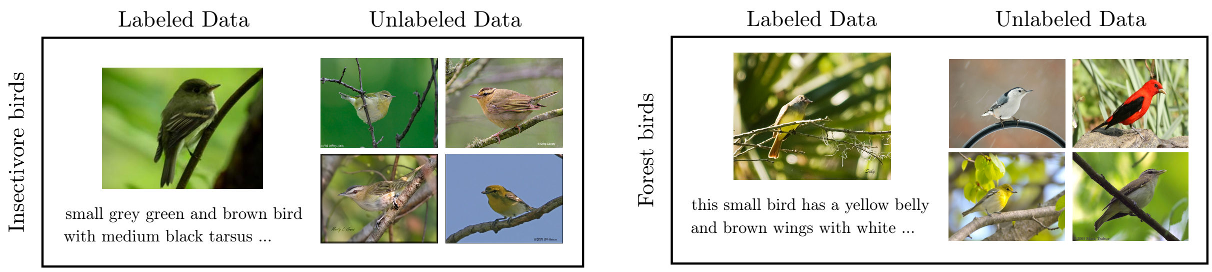

Figure 8: An example of data construction when testing coastal birds against grassland birds. Labeled data consists of both text and corresponding images, while unlabeled data consists only of images.

E.6 Details of the Real-World Experiment: Caltech-UCSD Bird dataset

This subsection provides a detailed description of our experiments on real-world data using the Caltech-UCSD Bird dataset (CUB-200-2011). The dataset includes both images and text descriptions for each bird species. This setup allows us to model a two-sample testing problem using multi-modal data. There are 11,788 images and 2,000 sentences of 200 bird species. We extracted text features (primary covariates) by obtaining embeddings with a pre-trained BERT base model ('bert-base-uncased'). The CLS token embedding from BERT's last hidden state was then passed through a single-layer MLP (hidden dimension 128, output dimension 4, dropout 0.2, ReLU activation) for dimensionality reduction. For image features (auxiliary covariates), a pre-trained ResNet-152 was used, with standard preprocessing (resize to 64×64 pixels, normalization). These features were further reduced with a separate MLP (hidden dimensions 1024 and 256, output 32, dropout 0.2, ReLU). Batch normalization followed each linear layer in the text and image MLPs. Following Biggs et al. (2024), these text and image embeddings were used within the two-sample testing framework.

We grouped bird species by diet (Insect, Seed, Fish) and habitat (Forest, Scrub, Wetland), forming three distinct comparison pairs. This grouping tested whether primary covariates (text) alone could distinguish species, while auxiliary covariates (image) provided complementary information. This classification was chosen specifically to create a scenario where textual descriptions, which often emphasize species-specific attributes, would be less sufficient for accurate group differentiation, whereas the image backgrounds would provide crucial contextual information about the habitat, making the auxiliary data more significant. The three pairs were Insect (12 species) vs. Forest (9), Seed (7) vs. Scrub (7), and Fish (8) vs. Wetland (7); in every comparison, 6 species overlapped, simulating realistic conditions with significant species overlap. To further investigate this phenomenon, we conducted additional experiments focusing on specific species comparisons: we tested 15 species of sparrow against 18 species of ground picker, compared 10 species of cuckoo with 13 species of foliage gleaner, and also examined 14 species of warbler against 12 species of canopy explorer. This data setup is illustrated in Figure 8. For each trial, we randomly selected $n_1 = 150$ Group 1 and $n_2 = 150$ Group 2 labeled samples, then chose 200 samples as unlabeled data. For example, X contained random text samples from insectivorous species, V contained images of corresponding species, including those not in X; the same approach applied to Y and W. We applied the standard kernel MMD (MMD-perm) and xMMD tests using text embeddings only, while our proposed test, xssMMD, incorporated both text and image embeddings. Each method used 50 random splits with 20 repetitions per split, and we reported average test power.

The estimated power for detecting distributional differences between bird groups is shown in Table 1 and Table 11. Across all testing scenarios, the permutation-based method MMD-perm generally achieved higher power than the permutation-free method xMMD. Importantly, our proposed method, xssMMD, consistently showed the highest test power in all comparisons. This underscores the advantage of our method, which leverages image embeddings to boost test power. The consistently strong performance of xssMMD across all pairs highlights the benefit of using complementary information from different modalities.

E.7 Details of the Real-World Experiment: MNIST dataset

This subsection provides a detailed description of our experiments on real-world data using the MNIST digit dataset, a widely used benchmark for visual recognition. The MNIST dataset consists of 28×28 grayscale images of handwritten digits (0 through 9). In this experiment, we constructed a two-sample testing problem

Table 11: Additional results of the estimated test power for detecting the difference between two bird groups with test level $\alpha = 0.05$. Our proposed method, xssMMD, was implemented using knn for the conditional expectation estimation.

Group 1	Group 2	Test	Power
Sparrow	Ground picker	MMD-perm	0.783
		xMMD	0.712
		xssMMD	**0.978**
Cuckoo	Foliage gleaner	MMD-perm	0.766
		xMMD	0.697
		xssMMD	**0.882**
Warbler	Canopy explorer	MMD-perm	0.883
		xMMD	0.650
		xssMMD	**0.991**

by partitioning the dataset into two distinct classes ($\mathcal{D}_1$ vs. $\mathcal{D}_2$) to detect distributional differences between the classes. The experimental design was motivated by the work of Schrab et al. (2023) and Chatterjee and Bhattacharya (2025), aiming to simulate a challenging scenario where the primary information is easily corrupted but supplemented by abundant auxiliary covariates.

For the main test, the labeled data X and Y consisted of pixel data from clear, original images. The unlabeled auxiliary covariates V and W consisted of the full pixel data of the images, into which we systematically injected independent Gaussian noise ϵ with increasing standard deviation σ. This process was motivated by examining the method's robustness against data corruption. Specifically, for each observation and each pixel entry i, j, the noise $\epsilon_{i,j}$ was sampled independently from a univariate normal distribution with zero mean and variance σ^2. The final noisy covariates were generated by adding this noise ϵ to the normalized original images, V and W. Subsequently, the resulting pixel values were constrained to remain within the valid range of $[0, 1]$. The operation for each entry of the noisy matrix is precisely defined as:

$$(V_{\text{noisy}})_{i,j} = \begin{cases} 0, & \text{if } V_{i,j} + \epsilon_{i,j} < 0 \\ 1, & \text{if } V_{i,j} + \epsilon_{i,j} > 1 \\ V_{i,j} + \epsilon_{i,j}, & \text{otherwise} \end{cases}$$

This matrix-wise operation ensures that the added noise ϵ is independent across all pixels and observations, and the resulting pixel values remain within the valid range. This setup allows us to examine how our method utilizes auxiliary information under increasing levels of corruption.

We considered three experimental settings for partitioning the data into two classes: (1) $\mathcal{D}_1 = \{0, 1, 2, 3, 4, 5, 9\}$ vs $\mathcal{D}_2 = \{0, 1, 2, 3, 4, 5, 8\}$; (2) $\mathcal{D}_1 = \{0, 1, 2, 3, 9\}$ vs $\mathcal{D}_2 = \{0, 1, 2, 3, 6\}$; and (3) $\mathcal{D}_1 = \{0, 1, 2, 3, 5, 8\}$ vs $\mathcal{D}_2 = \{0, 1, 2, 3, 5, 9\}$. For each setting, the dataset was divided by randomly sampling $n_1 = 200$ images from $\mathcal{D}_1$ and $n_2 = 200$ images from $\mathcal{D}_2$ to create the labeled data for each group. From the remaining images of each class in $\mathcal{D}_1$ and $\mathcal{D}_2$, 2000 samples per class were randomly selected and used as the unlabeled auxiliary covariates. Each method was then evaluated on 100 random partitions of the data, with 10 repetitions per partition, and average test power was reported.

The results are summarized in Table 12. Across all feature settings and noise levels, the xssMMD test consistently outperformed the standard kernel test (MMD-perm) and xMMD test. When the auxiliary data was clear ($\sigma = 0$), xssMMD showed superior power because the additional clear images effectively increased the sample size of the informative pixel features, confirming that leveraging the full, uncorrupted image data significantly enhances test power. More importantly, even when substantial Gaussian noise was present, xssMMD maintained a significant advantage over MMD-perm and xMMD. Notably, in the highest noise scenario ($\sigma = 1$), severely degrading the visual quality of the auxiliary covariates, xssMMD consistently demonstrated superior performance compared to xMMD. These results confirm the crucial utility of the auxiliary covariates and demonstrate the robustness of xssMMD in utilizing noisy information; our method successfully extracts the underlying structural differences between the two digit distributions from the corrupted auxiliary covariates, confirming that the utility of incorporating additional information persists, even when the discriminative signal in the primary labeled features is weak and the auxiliary data is corrupted.

We further investigated the scenario where the amount of unlabeled data was significantly reduced to 200 samples per class, matching the number of labeled samples ($n_1 = n_2 = 200$). With the same partitioning of (1), (2), and (3), we compared the performance of xssMMD against the baseline tests performed solely on the labeled data (MMD-perm(X,Y), xMMD) and the unlabeled data (MMD-perm(V,W)). The results are summarized in Table 13. When the noise level was minimal ($\sigma = 0$), testing on the unlabeled data yielded high performance since this setting is the same as testing on the labeled data. However, the performance of MMD-perm on the unlabeled data drastically decreased as the noise level σ increased. On the other hand, xssMMD demonstrated superior power compared to the MMD-perm and xMMD tests in most scenarios when the noise is not extremely large. As the noise increases, the power of xssMMD test decreases gradually, while the power of the MMD-perm test on the unlabeled data decreases drastically, falling below that of xssMMD in high-noise scenarios. These results indicate that xssMMD effectively leverages the structural information of the limited noisy auxiliary covariates to enhance test sensitivity, even when the auxiliary covariates themselves are not sufficiently distinct for successful permutation testing.

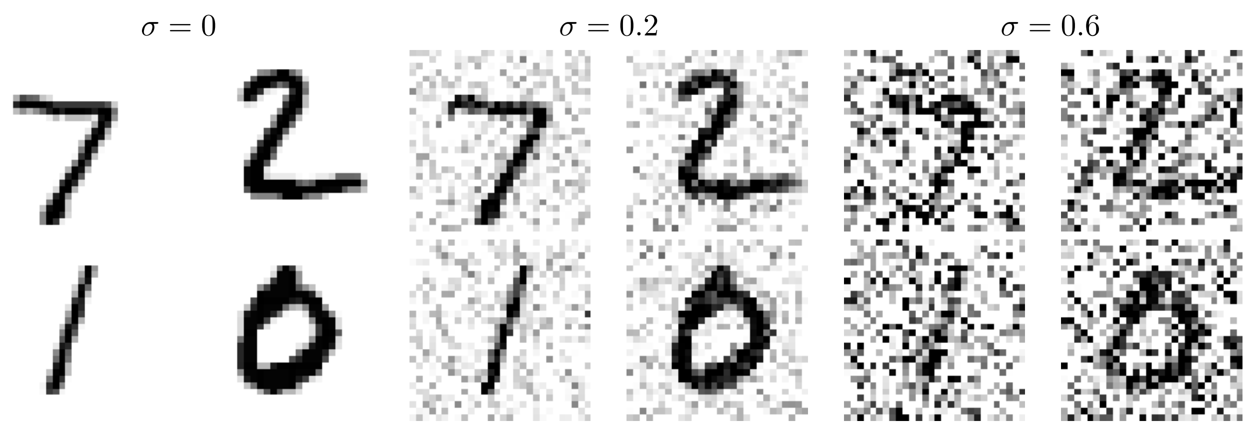

Figure 9: An example of data construction of images with Gaussian noise of $\sigma = 0,\ 0.2,\ 0.6$. Labeled data consists of both clear image without any noise and corresponding image with noise, while unlabeled data consists only of images with Gaussian noise.

Table 12: Estimated test power under different scenarios and increasing noise levels (independent Gaussian noise with standard deviation $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$). Each value corresponds to the average test power across 1000 trials. Our proposed method, xssMMD, was implemented using knn for the conditional expectation estimation.

Scenario	MMD-perm	xMMD	xssMMD					
			$\sigma = 0$	$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.6$	$\sigma = 0.8$	$\sigma = 1.0$
$\{0,1,2,3,5,8\}$ vs $\{0,1,2,3,5,9\}$	0.731	0.601	**0.85**	**0.852**	**0.862**	**0.837**	**0.776**	0.721
$\{0,1,2,3,9\}$ vs $\{0,1,2,3,6\}$	0.985	0.915	**0.999**	**0.999**	**0.999**	**0.999**	**0.998**	**0.997**
$\{0,1,2,3,4,5,9\}$ vs $\{0,1,2,3,5,8\}$	0.55	0.402	**0.666**	**0.679**	**0.694**	**0.659**	**0.577**	0.496

F Proof of Main Results

F.1 Proof of Theorem 2.3

Consider the centered oracle statistic with the population variance

$$\overline{T}_{\text{oracle}} = \frac{\widehat{\mu}_{X,f} - \widehat{\mu}_{Y,f} - \mathbb{E}[f(X)] + \mathbb{E}[f(Y)]}{\sqrt{\sigma_{X,f}^2 + \sigma_{Y,f}^2}}. \tag{11}$$

Table 13: Estimated test power under different scenarios including kernel test on the unlabeled data only and increasing noise levels (independent Gaussian noise with standard deviation $\sigma \in \{0, 0.5, 1.0, 1.5, 2.0\}$). Each value corresponds to the average test power across 1000 trials. Our proposed method, xssMMD, was implemented using knn for the conditional expectation estimation.

Labeled Data	Test	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.5$	$\sigma = 2.0$
$\{0, 1, 2, 3, 5, 8\}$ vs $\{0, 1, 2, 3, 5, 9\}$	MMD-perm(X,Y)	0.724	0.724	**0.724**	**0.724**	**0.724**
	MMD-perm(V,W)	0.721	0.662	0.330	0.179	0.118
	xMMD	0.581	0.581	0.581	0.581	0.581
	xssMMD	**0.769**	**0.779**	0.651	0.572	0.545
$\{0, 1, 2, 3, 9\}$ vs $\{0, 1, 2, 3, 6\}$	MMD-perm(X,Y)	0.988	0.988	0.988	**0.988**	**0.988**
	MMD-perm(V,W)	0.982	0.963	0.703	0.352	0.214
	xMMD	0.912	0.912	0.912	0.912	0.912
	xssMMD	**0.999**	**0.999**	**0.993**	0.958	0.937
$\{0, 1, 2, 3, 4, 5, 9\}$ vs $\{0, 1, 2, 3, 5, 8\}$	MMD-perm(X,Y)	0.550	0.550	**0.550**	**0.550**	**0.550**
	MMD-perm(V,W)	**0.836**	**0.698**	0.277	0.114	0.080
	xMMD	0.419	0.419	0.419	0.419	0.419
	xssMMD	0.580	0.611	0.450	0.390	0.369

To establish the desired result, it suffices to prove that $\overline{T}_{\mathrm{oracle}}$ is asymptotically $N(0, 1)$ as $n = n_1 \wedge n_2 \to \infty$, and that the empirical variance estimates are ratio-consistent. We will prove these statements in order. Throughout this proof, we write $\mathbb{E}[f(X)]$ as $\mathbb{E}[f(X) \,|\, f]$ (and similarly for other quantities) to highlight that we condition on the randomness inherent in f.

Step 1: Asymptotic Normality of $\overline{T}_{\mathrm{oracle}}$. Note that the numerator of $\overline{T}_{\mathrm{oracle}}$ is $\widehat{\mu}_{X,f} - \widehat{\mu}_{Y,f} - \mathbb{E}[f(X)] + \mathbb{E}[f(Y)] = \sum_{i=1}^{n_1+m_1} G_i - \sum_{i=1}^{n_2+m_2} H_i$ where

$$G_i = \begin{cases} \frac{1}{n_1}\big\{f(X_i) - \mathbb{E}[f(X_i) \,|\, f]\big\} - \frac{m_1}{n_1(n_1+m_1)}\big\{\mathbb{E}[f(X_i) \,|\, V_i, f] - \mathbb{E}[f(X_i) \,|\, f]\big\} & \text{if } 1 \le i \le n_1, \\ \frac{1}{n_1+m_1}\big\{\mathbb{E}[f(X_i) \,|\, V_i, f] - \mathbb{E}[f(X_i) \,|\, f]\big\} & \text{if } n_1 + 1 \le i \le n_1 + m_1, \end{cases}$$

which are centered and conditionally independent given f. Similarly, we have

$$H_i = \begin{cases} \frac{1}{n_2}\big\{f(Y_i) - \mathbb{E}[f(Y_i) \,|\, f]\big\} - \frac{m_2}{n_2(n_2+m_2)}\big\{\mathbb{E}[f(Y_i) \,|\, W_i, f] - \mathbb{E}[f(Y_i) \,|\, f]\big\} & \text{if } 1 \le i \le n_2, \\ \frac{1}{n_2+m_2}\big\{\mathbb{E}[f(Y_i) \,|\, W_i, f] - \mathbb{E}[f(Y_i) \,|\, f]\big\} & \text{if } n_2 + 1 \le i \le n_2 + m_2. \end{cases}$$

Since the distributions of G_i and H_i may vary with the sample sizes depending on the choice of f, we use Lyapunov central limit theorem (CLT) to prove the desired statement. To apply Lyapunov CLT, we further define the variance as

$$\begin{aligned} s_{n+m}^2 &:= \mathrm{Var}(\widehat{\mu}_{X,f} - \widehat{\mu}_{Y,f} \,|\, f) \\ &= \sum_{i=1}^{n_1+m_1} \mathrm{Var}(G_i \,|\, f) + \sum_{i=1}^{n_2+m_2} \mathrm{Var}(H_i \,|\, f) \\ &= \sigma_{X,f}^2 + \sigma_{Y,f}^2 \\ &= \frac{1}{n_1}\sigma_{1,X,f}^2 + \frac{1}{n_1+m_1}\sigma_{2,X,f}^2 + \frac{1}{n_2}\sigma_{1,Y,f}^2 + \frac{1}{n_2+m_2}\sigma_{2,Y,f}^2. \end{aligned}$$

For some $\delta > 0$, (conditional) Lyapunov's condition (Lundborg et al., 2024, Lemma S8) stated as

$$\sum_{i=1}^{n_1+m_1} \mathbb{E}\left[\left|\frac{G_i}{s_{n+m}}\right|^{2+\delta} \,\middle|\, f\right] + \sum_{i=1}^{n_2+m_2} \mathbb{E}\left[\left|\frac{H_i}{s_{n+m}}\right|^{2+\delta} \,\middle|\, f\right] = o_P(1) \tag{12}$$

ensures the asymptotic normality of $\overline{T}_{\mathrm{oracle}}$ unconditional on f.

Now, we show that the above condition (12) is satisfied. Letting $r_1 = \frac{m_1}{n_1+m_1}$ and $r_2 = \frac{m_2}{n_2+m_2}$, the proportions of the unlabeled data in the total dataset for X and Y, respectively, we have an upper bound for the first term with

G_i in (12) as

$$
\begin{aligned}
\sum_{i=1}^{n_1+m_1} \mathbb{E}\Big[\Big|\frac{G_i}{s_{n+m}}\Big|^{2+\delta} \Big| f\Big] \;\lesssim\; & \frac{1}{n_1^{1+\delta}}\mathbb{E}\Big[\frac{|f(X)-\mathbb{E}[f(X)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big] \\
& + \frac{r_1^{2+\delta}}{n_1^{1+\delta}}\mathbb{E}\Big[\frac{|\mathbb{E}[f(X)\,|\,V,f]-\mathbb{E}[f(X)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big] \\
& + \frac{r_1}{(n_1+m_1)^{1+\delta}}\mathbb{E}\Big[\frac{|\mathbb{E}[f(X)\,|\,V,f]-\mathbb{E}[f(X)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big].
\end{aligned}
$$

Similarly, for the second term of (12) with H_i, we have an upper bound as

$$
\begin{aligned}
\sum_{i=1}^{n_2+m_2} \mathbb{E}\Big[\Big|\frac{H_i}{s_{n+m}}\Big|^{2+\delta} \Big| f\Big] \;\lesssim\; & \frac{1}{n_2^{1+\delta}}\mathbb{E}\Big[\frac{|f(Y)-\mathbb{E}[f(Y)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big] \\
& + \frac{r_2^{2+\delta}}{n_2^{1+\delta}}\mathbb{E}\Big[\frac{|\mathbb{E}[f(Y)\,|\,W,f]-\mathbb{E}[f(Y)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big] \\
& + \frac{r_2}{(n_2+m_2)^{1+\delta}}\mathbb{E}\Big[\frac{|\mathbb{E}[f(Y)\,|\,W,f]-\mathbb{E}[f(Y)\,|\,f]|^{2+\delta}}{s_{n+m}^{2+\delta}}\Big| f\Big].
\end{aligned}
$$

By conditional Jensen's inequality, we have an upper bound

$$
\mathbb{E}\big[\big|\mathbb{E}[f(X)\,|\,V,f]-\mathbb{E}[f(X)\,|\,f]\big|^{2+\delta}\big] \;\le\; \mathbb{E}\big[\big|f(X)-\mathbb{E}[f(X)\,|\,f]\big|^{2+\delta}\big],
$$

and similarly

$$
\mathbb{E}\big[\big|\mathbb{E}[f(Y)\,|\,W,f]-\mathbb{E}[f(Y)\,|\,f]\big|^{2+\delta}\big] \le\; \mathbb{E}\big[\big|f(Y)-\mathbb{E}[f(Y)\,|\,f]\big|^{2+\delta}\big].
$$

Next, we observe that the term s_{n+m}^2, representing the combined sample variance, satisfies the lower bound:

$$
\begin{aligned}
s_{n+m}^2 \;\ge\; \sigma_{X,f}^2 &= \frac{1}{n_1}\sigma_{1,X,f}^2 + \frac{1}{n_1+m_1}\sigma_{2,X,f}^2 \\
&\ge \Big(\frac{1}{n_1}+\frac{1}{n_1+m_1}\Big)\times\big(\sigma_{1,X,f}^2\wedge\sigma_{2,X,f}^2\big).
\end{aligned}
$$

Similarly, the combined variance also satisfies:

$$
s_{n+m}^2 \ge \Big(\frac{1}{n_2}+\frac{1}{n_2+m_2}\Big)\times\big(\sigma_{1,Y,f}^2\wedge\sigma_{2,Y,f}^2\big).
$$

These bounds, combined with the earlier results, ensure that the terms in (12) decay appropriately as

$$
\sum_{i=1}^{n_1+m_1} \mathbb{E}\Big[\Big|\frac{G_i}{s_{n+m}}\Big|^{2+\delta} \Big| f\Big] \;\lesssim\; \frac{1}{n_1^{\delta/2}}\frac{\mathbb{E}\big[\big\{f(X)-\mathbb{E}[f(X)\,|\,f]\big\}^{2+\delta}\big]}{\sigma_{1,X,f}^{2+\delta}\wedge\sigma_{2,X,f}^{2+\delta}}.
$$

A similar upper bound can be obtained for

$$
\sum_{i=1}^{n_2+m_2} \mathbb{E}\Big[\Big|\frac{H_i}{s_{n+m}}\Big|^{2+\delta} \Big| f\Big] \;\lesssim\; \frac{1}{n_2^{\delta/2}}\frac{\mathbb{E}\big[\big\{f(Y)-\mathbb{E}[f(Y)\,|\,f]\big\}^{2+\delta}\big]}{\sigma_{1,Y,f}^{2+\delta}\wedge\sigma_{2,Y,f}^{2+\delta}}.
$$

Consequently, under the moment condition (2.2), Lyapunov's condition is fulfilled, which implies the asymptotic normality of $\overline{T}_{\text{oracle}}$.

Step 2: Asymptotic Normality with Sample Variance In this step, we aim to show that the ratio of the sample variance to the population variance converges to 1 in probability. This result in conjunction with Slutsky's theorem and continuous mapping theorem in turn confirms that $\overline{T}_{\text{oracle}}$ and T_{oracle} share the same asymptotic distribution.

We first formally define the estimated variance of $\widehat{\sigma}^2_{X,f} = n_1^{-1}\widehat{\sigma}^2_{1,X,f} + (n_1+m_1)^{-1}\widehat{\sigma}^2_{2,X,f}$ where each term is defined as

$$\begin{aligned}\widehat{\sigma}^2_{1,X,f} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\bigl\{f(X_i) - \mathbb{E}[f(X_i)\,|\,V_i]\bigr\}^2 \quad \text{and} \\ \widehat{\sigma}^2_{2,X,f} &= \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\Biggl\{\mathbb{E}[f(X_i)\,|\,V_i] - \frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\mathbb{E}[f(X_j)\,|\,V_j]\Biggr\}^2.\end{aligned} \tag{13}$$

We similarly define the variance of $\widehat{\mu}_{Y,f}$ as $\widehat{\sigma}^2_{Y,f} = n_2^{-1}\widehat{\sigma}^2_{1,Y,f} + (n_2+m_2)^{-1}\widehat{\sigma}^2_{2,Y,f}$.

Based on these definitions, we prove the following convergence:

$$\frac{\widehat{\sigma}^2_{X,f} + \widehat{\sigma}^2_{Y,f}}{\sigma^2_{X,f} + \sigma^2_{Y,f}} - 1 = o_P(1) \quad \Longleftrightarrow \quad \frac{\widehat{\sigma}^2_{X,f} - \sigma^2_{X,f} + \widehat{\sigma}^2_{Y,f} - \sigma^2_{Y,f}}{\sigma^2_{X,f} + \sigma^2_{Y,f}} = o_P(1).$$

By the triangle inequality, it suffices to show that

$$\frac{\widehat{\sigma}^2_{X,f} - \sigma^2_{X,f}}{\sigma^2_{X,f}} = o_P(1) \quad \text{and} \quad \frac{\widehat{\sigma}^2_{Y,f} - \sigma^2_{Y,f}}{\sigma^2_{Y,f}} = o_P(1). \tag{14}$$

Without loss of generality, we focus on the first convergence result. Using the lower bound for $\sigma^2_{X,f}$:

$$\sigma^2_{X,f} = \frac{1}{n_1}\sigma^2_{1,X,f} + \frac{1}{n_1+m_1}\sigma^2_{2,X,f} \geq \biggl(\frac{1}{n_1} + \frac{1}{n_1+m_1}\biggr) \times \bigl(\sigma^2_{1,X,f} \wedge \sigma^2_{2,X,f}\bigr), \tag{15}$$

as well as the definition of $\widehat{\sigma}^2_{X,f} = n_1^{-1}\widehat{\sigma}^2_{1,X,f} + (n_1+m_1)^{-1}\widehat{\sigma}^2_{2,X,f}$, we have

$$\frac{\bigl|\widehat{\sigma}^2_{X,f} - \sigma^2_{X,f}\bigr|}{\sigma^2_{X,f}} \leq \underbrace{\frac{\bigl|\widehat{\sigma}^2_{1,X,f} - \sigma^2_{1,X,f}\bigr|}{\sigma^2_{1,X,f} \wedge \sigma^2_{2,X,f}}}_{\text{(I)}} + \underbrace{\frac{\bigl|\widehat{\sigma}^2_{2,X,f} - \sigma^2_{2,X,f}\bigr|}{\sigma^2_{1,X,f} \wedge \sigma^2_{2,X,f}}}_{\text{(II)}}.$$

A conditional version of the weak law of large numbers (Lundborg et al., 2024, Lemma S9) under Assumption 2.2 guarantees that the first term (I) is $o_P(1)$. For the second term (II), we rewrite $\widehat{\sigma}^2_{2,X,f}$ as

$$\underbrace{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\bigl\{\mathbb{E}[f(X_i)\,|\,V_i,f] - \mathbb{E}[f(X_i)\,|\,f]\bigr\}^2}_{\text{(II)}_1} - \underbrace{\biggl(\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\bigl\{\mathbb{E}[f(X_i)\,|\,V_i,f] - \mathbb{E}[f(X_i)\,|\,f]\bigr\}\biggr)^2}_{\text{(II)}_2^2}.$$

By the weak law of large numbers, again, under Assumption 2.2, it can be seen that

$$\frac{|\text{(II)}_1 - \sigma^2_{2,X,f}|}{\sigma^2_{1,X,f} \wedge \sigma^2_{2,X,f}} = o_P(1).$$

For (II)_2, we first note by Jensen's inequality that

$$\begin{aligned}\mathbb{E}\Biggl[\biggl|\frac{\bigl\{\mathbb{E}[f(X_i)\,|\,V_i,f] - \mathbb{E}[f(X_i)\,|\,f]\bigr\}}{\sigma_{1,X,f} \wedge \sigma_{2,X,f}}\biggr|^{1+\delta}\,\Bigg|\,f\Biggr] &\leq \sqrt{\mathbb{E}\Biggl[\biggl|\frac{\bigl\{\mathbb{E}[f(X_i)\,|\,V_i,f] - \mathbb{E}[f(X_i)\,|\,f]\bigr\}}{\sigma_{1,X,f} \wedge \sigma_{2,X,f}}\biggr|^{2+2\delta}\,\Bigg|\,f\Biggr]} \\ &= o_P(n^{\delta}),\end{aligned}$$

where the last approximation holds under Assumption 2.2. Hence by Lundborg et al. (2024, Lemma S9), it holds that $\text{(II)}_2^2 = o_P(1)$, which in turn implies $\text{(II)} = o_P(1)$ as required. Combining the results for (I) and (II), we conclude

$$\frac{\widehat{\sigma}^2_{X,f} - \sigma^2_{X,f}}{\sigma^2_{X,f}} = o_P(1).$$

A similar argument applies to $\widehat{\sigma}^2_{Y,f}$, which verifies the ratio consistency of the sample variance.

F.2 Proof of Corollary 2.4

As mentioned in the main text, we assume that the sample sizes n_1, n_2, m_1, m_2 are even for simplicity. The other cases can be proven similarly by minor modifications. Similarly to Theorem 2.3, we consider the centered cross-fitted statistic with the empirical variance

$$\overline{T}_{\text{cross}} = \frac{\widehat{\mu}^{\dagger}_{X,f} - \widehat{\mu}^{\dagger}_{Y,f} - \mathbb{E}[f(X)\,|\,f] + \mathbb{E}[f(Y)\,|\,f]}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,f} + \widehat{\sigma}^{\dagger 2}_{Y,f}}}.$$

Above, we define $\widehat{\mu}^{\dagger}_{X,f}$ and $\widehat{\mu}^{\dagger}_{Y,f}$ as the counterparts of $\widehat{\mu}_{X,f}$ and $\widehat{\mu}_{Y,f}$, replacing $\mathbb{E}[f(X_i)\,|\,V_i, f]$ and $\mathbb{E}[f(Y_i)\,|\,W_i, f]$ with their estimates using $\widehat{\mathbb{E}}[f(X_i)\,|\,V_i, f]$ and $\widehat{\mathbb{E}}[f(Y_i)\,|\,W_i, f]$, respectively. Similarly, we define $\widehat{\sigma}^{\dagger 2}_{1,X,f}$ and $\widehat{\sigma}^{\dagger 2}_{2,X,f}$ as the counterparts of $\widehat{\sigma}^{2}_{1,X,f}$ and $\widehat{\sigma}^{2}_{2,X,f}$ with the estimated conditional expectations.

Since we have already established that $\overline{T}_{\text{oracle}}$ is asymptotically $N(0,1)$ as $n = n_1 \wedge n_2 \to \infty$, it suffices to prove

$$\overline{T}_{\text{cross}} - \overline{T}_{\text{oracle}} = o_P(1).$$

Once this convergence is verified, $\overline{T}_{\text{cross}}$ will also converge to $N(0,1)$ by Slutsky's theorem. To this end, denote $\overline{T}_{\text{cross}} = \frac{N_{\text{cross}}}{D_{\text{cross}}}$ and $\overline{T}_{\text{oracle}} = \frac{N_{\text{oracle}}}{D_{\text{oracle}}}$. Then

$$\begin{aligned}
\overline{T}_{\text{cross}} - \overline{T}_{\text{oracle}} &= \frac{N_{\text{cross}}}{D_{\text{cross}}} - \frac{N_{\text{cross}}}{D_{\text{oracle}}} + \frac{N_{\text{cross}}}{D_{\text{oracle}}} - \frac{N_{\text{oracle}}}{D_{\text{oracle}}} \\
&= \frac{N_{\text{cross}} - N_{\text{oracle}}}{D_{\text{oracle}}}\left(\frac{D_{\text{oracle}}}{D_{\text{cross}}} - 1\right) + \underbrace{\frac{N_{\text{oracle}}}{D_{\text{oracle}}}}_{=O_P(1)}\left(\frac{D_{\text{oracle}}}{D_{\text{cross}}} - 1\right) + \frac{N_{\text{cross}} - N_{\text{oracle}}}{D_{\text{oracle}}}.
\end{aligned}$$

Hence it suffices to show the following two claims hold:

$$\text{(i)}\ \ \frac{N_{\text{cross}} - N_{\text{oracle}}}{D_{\text{oracle}}} = o_P(1) \quad \text{and} \quad \text{(ii)}\ \ \frac{D_{\text{oracle}}}{D_{\text{cross}}} - 1 = o_P(1),$$

which are proved below.

Proof of claim (i). We define $D^{\star}_{\text{oracle}}$ as the population standard deviation. Note that $(D_{\text{oracle}})^2 = \widehat{\sigma}^2_{X,f} + \widehat{\sigma}^2_{Y,f}$ and let $(D^{\star}_{\text{oracle}})^2 = \sigma^2_{X,f} + \sigma^2_{Y,f}$ where $\sigma^2_{X,f} = \frac{1}{n_1}\sigma^2_{1,X} + \frac{1}{n_1+m_1}\sigma^2_{2,X}$ and $\sigma^2_{Y,f} = \frac{1}{n_2}\sigma^2_{1,Y} + \frac{1}{n_2+m_2}\sigma^2_{2,Y}$. From our previous result in Theorem 2.3, we have

$$\left(\frac{D_{\text{oracle}}}{D^{\star}_{\text{oracle}}}\right)^2 - 1 = \frac{\widehat{\sigma}^2_{X,f} + \widehat{\sigma}^2_{Y,f}}{\sigma^2_{X,f} + \sigma^2_{Y,f}} - 1 = o_P(1).$$

By the continuous mapping theorem, we have

$$\frac{D^{\star}_{\text{oracle}}}{D_{\text{oracle}}} - 1 = o_P(1)$$

and consequantly,

$$\frac{N_{\text{cross}} - N_{\text{oracle}}}{D_{\text{oracle}}} = \frac{N_{\text{cross}} - N_{\text{oracle}}}{D^{\star}_{\text{oracle}}}\{1 + o_P(1)\}.$$

On the other hand,

$$\left|\frac{N_{\text{cross}} - N_{\text{oracle}}}{D^{\star}_{\text{oracle}}}\right| \le \frac{|R_X|}{\sigma_{X,f}} + \frac{|R_Y|}{\sigma_{Y,f}},$$

where

$$
\begin{aligned}
R_X = \frac{1}{n_1}\sum_{i=1}^{n_1}&\{\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]-\mathbb{E}[f(X_i)\,|\,V_i,f]\}\\
&+\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\{\mathbb{E}[f(X_i)\,|\,V_i,f]-\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]\},\quad\text{and}\\
R_Y = \frac{1}{n_2}\sum_{i=1}^{n_2}&\{\widehat{\mathbb{E}}[f(Y_i)\,|\,W_i,f]-\mathbb{E}[f(Y_i)\,|\,W_i,f]\}\\
&+\frac{1}{n_2+m_2}\sum_{i=1}^{n_2+m_2}\{\mathbb{E}[f(Y_i)\,|\,W_i,f]-\widehat{\mathbb{E}}[f(Y_i)\,|\,W_i,f]\}.
\end{aligned}
$$

Without loss of generality, we focus on the first term $|R_X|/\sigma_{X,f}$ and the other term $|R_Y|/\sigma_{Y,f}$ can be handled similarly.

Observe that R_X takes the form $\frac{1}{n}\sum_{i=1}^{n} g(V_i)-\frac{1}{n+m}\sum_{j=1}^{n+m} g(V_j)$ for some function g. This form is invariant under location shifts in the function g. Specifically, for any constant $c\in\mathbb{R}$, we have $\frac{1}{n}\sum_{i=1}^{n}\{g(V_i)+c\}-\frac{1}{n+m}\sum_{j=1}^{n+m}\{g(V_j)+c\}=\frac{1}{n}\sum_{i=1}^{n} g(V_i)-\frac{1}{n+m}\sum_{j=1}^{n+m} g(V_j)$. Then, we choose $c=-\mathbb{E}[g(V)]$ where $g(V_j)+c$ has expectation of 0. Therefore, without loss of generality, we may assume that R_X has zero mean, i.e., $\mathbb{E}\big[\mathbb{E}[f(X_i)\,|\,V_i,f]-\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]\big]=0$ for all i. Then by Chebyshev's inequality, it can be seen that

$$
R_X^2 = O_P\Big(n_1^{-1}\mathbb{E}\big[\{\widehat{\mathbb{E}}[f(X)\,|\,V,f]-\mathbb{E}[f(X)\,|\,V,f]\}^2\,|\,f\big]\Big).
$$

Combining this with the lower bound (15) for $\sigma^2_{X,f}$,

$$
\frac{R_X^2}{\sigma^2_{X,f}} \le \left(\frac{1}{n_1}+\frac{1}{n_1+m_1}\right)^{-1}\frac{R_X^2}{\sigma^2_{1,X,f}\wedge\sigma^2_{2,X,f}} = O_P\left(\frac{\mathbb{E}\big[\{\widehat{\mathbb{E}}[f(X)\,|\,V,f]-\mathbb{E}[f(X)\,|\,V,f]\}^2\,|\,f\big]}{\sigma^2_{1,X,f}\wedge\sigma^2_{2,X,f}}\right).
$$

Hence, under the condition (2), we have

$$
\frac{R_X^2}{\sigma^2_{X,f}} = o_P(1).
$$

A similar argument applies to $R_Y^2/\sigma^2_{Y,f}$, which proves that claim (i) holds.

Proof of claim (ii). By the continuous mapping theorem, it is sufficient to prove that

$$
\left(\frac{D_{\text{cross}}}{D_{\text{oracle}}}\right)^2 - 1 = o_P(1),
$$

which, in turn, is implied by

$$
\begin{aligned}
&\frac{1}{n_1}\frac{|\widehat{\sigma}^{\dagger 2}_{1,X,f}-\widehat{\sigma}^2_{1,X,f}|}{\widehat{\sigma}^2_{X,f}} = o_P(1),\ \ \frac{1}{n_1+m_1}\frac{|\widehat{\sigma}^{\dagger 2}_{2,X,f}-\widehat{\sigma}^2_{2,X,f}|}{\widehat{\sigma}^2_{X,f}} = o_P(1),\quad\text{and}\\
&\frac{1}{n_2}\frac{|\widehat{\sigma}^{\dagger 2}_{1,Y,f}-\widehat{\sigma}^2_{1,Y,f}|}{\widehat{\sigma}^2_{Y,f}} = o_P(1),\ \ \frac{1}{n_2+m_2}\frac{|\widehat{\sigma}^{\dagger 2}_{2,Y,f}-\widehat{\sigma}^2_{2,Y,f}|}{\widehat{\sigma}^2_{Y,f}} = o_P(1).
\end{aligned}
$$

Without loss of generality, we focus on $\widehat{\sigma}^{\dagger 2}_{1,X,f}$ and $\widehat{\sigma}^{\dagger 2}_{2,X,f}$. Using the Cauchy–Schwarz inequality, we observe that

$$
\begin{aligned}
\frac{1}{n_1}\frac{|\widehat{\sigma}^{\dagger 2}_{1,X,f}-\widehat{\sigma}^2_{1,X,f}|}{\widehat{\sigma}^2_{X,f}} \ \le\ & \frac{1}{n_1}\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\big\{\mathbb{E}[f(X_i)\,|\,V_i,f]-\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]\big\}^2}{\widehat{\sigma}^2_{X,f}}\\
&+\frac{2}{n_1}\sqrt{\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\big\{\mathbb{E}[f(X_i)\,|\,V_i,f]-\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]\big\}^2}{\widehat{\sigma}^2_{X,f}}}.
\end{aligned}
$$

This becomes $o_P(1)$ when

$$\frac{1}{n_1}\frac{\mathbb{E}\big[\{\mathbb{E}[f(X)\,|\,V,f]-\widehat{\mathbb{E}}[f(X)\,|\,V,f]\}^2\,|\,f\big]}{\sigma^2_{X,f}} \le \left(\frac{1}{n_1}+\frac{1}{n_1+m_1}\right)^{-1}\frac{1}{n_1}\frac{\mathbb{E}\big[\{\mathbb{E}[f(X)\,|\,V,f]-\widehat{\mathbb{E}}[f(X)\,|\,V,f]\}^2\,|\,f\big]}{\sigma^2_{1,X,f}\wedge\sigma^2_{2,X,f}}$$
$$= o_P(1),$$

which is satisfied under the condition (2). On the other hand, we again use the Cauchy–Schwarz inequality to observe that

$$\begin{aligned}\frac{1}{n_1+m_1}\frac{|\widehat{\sigma}^{\dagger 2}_{2,X}-\widehat{\sigma}^2_{2,X}|}{\widehat{\sigma}^2_{X,f}} \le\ & \frac{1}{n_1+m_1}\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}A_i^2}{\widehat{\sigma}^2_{X,f}} \\ &+2\sqrt{\frac{1}{n_1+m_1}\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}A_i^2}{\widehat{\sigma}^2_{X,f}}}\underbrace{\sqrt{\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}B_i^2}{\widehat{\sigma}^2_{2,X}}}}_{=1}\end{aligned}$$

where

$$A_i=\{\mathbb{E}[f(X_i)\,|\,V_i,f]-\widehat{\mathbb{E}}[f(X_i)\,|\,V_i,f]\}-\frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\{\mathbb{E}[f(X_j)\,|\,V_j,f]-\widehat{\mathbb{E}}[f(X_j)\,|\,V_j,f]\},$$
$$B_i=\mathbb{E}[f(X_i)\,|\,V_i,f]-\frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\mathbb{E}[f(X_j)\,|\,V_j,f].$$

The above upper bound becomes $o_P(1)$ when

$$\begin{aligned}&\frac{1}{n_1+m_1}\frac{\mathbb{E}\big[\{\mathbb{E}[f(X)\,|\,V,f]-\widehat{\mathbb{E}}[f(X)\,|\,V,f]\}^2\,|\,f\big]}{\sigma^2_{X,f}} \\ &\le\left(\frac{1}{n_1}+\frac{1}{n_1+m_1}\right)^{-1}\frac{1}{n_1+m_1}\frac{\mathbb{E}\big[\{\mathbb{E}[f(X)\,|\,V,f]-\widehat{\mathbb{E}}[f(X)\,|\,V,f]\}^2\,|\,f\big]}{\sigma^2_{1,X,f}\wedge\sigma^2_{2,X,f}}=o_P(1),\end{aligned}$$

which is also satisfied under the condition (2). A similar argument applies to $\widehat{\sigma}^{\dagger 2}_{1,Y,f}$ and $\widehat{\sigma}^{\dagger 2}_{2,Y,f}$, which proves that claim (ii) holds. Therefore, under the condition (2), we conclude that $\overline{T}_{\text{cross}}$ is asymptotically $N(0,1)$ as $n=n_1\wedge n_2\to\infty$.

F.3 Proof of Theorem 3.4

Before presenting the formal proof of Theorem 3.4, we first provide the explicit mathematical formulation of the cross-fitted test statistic $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ as defined in (4).

Recall that to compute the cross-fitted estimators, we partition the labeled dataset $\mathcal{L}_{XV}$ into two disjoint subsets: $\mathcal{L}_{XV,a}=\{(X_i,V_i):i\in\mathcal{I}_a\}$ and $\mathcal{L}_{XV,b}=\{(X_i,V_i):i\in\mathcal{I}_b\}$, where $\mathcal{I}_a$ and $\mathcal{I}_b$ represent the odd and even indices of $\{1,\ldots,n_1\}$, respectively. Similarly, we partition the unlabeled dataset $\mathcal{U}_V$ into $\mathcal{U}_{V,a}=\{V_i:i\in\mathcal{J}_a\}$ and $\mathcal{U}_{V,b}=\{V_i:i\in\mathcal{J}_b\}$, where $\mathcal{J}_a$ and $\mathcal{J}_b$ represent the odd and even indices of $\{n_1+1,\ldots,n_1+m_1\}$, respectively. Let $\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]$ denote the conditional expectation estimator trained on $\mathcal{L}_{XV,a}$ if $i\in\mathcal{I}_b\cup\mathcal{J}_b$ (even indices), and on $\mathcal{L}_{XV,b}$ if $i\in\mathcal{I}_a\cup\mathcal{J}_a$ (odd indices). We apply an analogous partition to $\mathcal{L}_{YW}$ and $\mathcal{U}_W$ and define $\widehat{\mathbb{E}}[\widehat{f}(Y_i)\,|\,W_i]$.

Based on the obtained estimator, the cross-fitted semi-supervised mean estimator for X, denoted as $\hat{\mu}^{\dagger}_{X,\widehat{f}}$, is defined as

$$\hat{\mu}^{\dagger}_{X,\widehat{f}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{\widehat{f}(X_i)-\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]\Big\}+\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i].$$

The estimator $\hat{\mu}^{\dagger}_{Y,\widehat{f}}$ is defined analogously using $\mathcal{L}_{YW}$ and $\mathcal{U}_W$.

Similarly, the cross-fitted variance estimator $\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}}$ is defined as $\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} = n_1^{-1}\widehat{\sigma}^{\dagger 2}_{1,X,\widehat{f}} + (n_1+m_1)^{-1}\widehat{\sigma}^{\dagger 2}_{2,X,\widehat{f}}$, where each term represents the empirical sample variance of the cross-fitted components:

$$\begin{aligned} \widehat{\sigma}^{\dagger 2}_{1,X,\widehat{f}} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{\widehat{f}(X_i) - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]\Big\}^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{\widehat{f}(X_i) - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]\Big\}\right)^2, \\ \widehat{\sigma}^{\dagger 2}_{2,X,\widehat{f}} &= \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\Big\{\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]\Big\}^2 - \left(\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i]\right)^2. \end{aligned} \tag{16}$$

The variance $\widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}$ is computed symmetrically using $\mathcal{L}_{YW}$ and $\mathcal{U}_W$.

The final cross-fitted semi-supervised MMD test statistic is then given by:

$$\mathrm{xss}\widehat{\mathrm{MMD}}^2 = \frac{\hat{\mu}^{\dagger}_{X,\widehat{f}} - \hat{\mu}^{\dagger}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}.$$

With the precise formulation of the test statistic established, we now proceed to the main proof of Theorem 3.4. We first note that the tests are defined using the $(1-\alpha)$-quantile of the standard normal distribution. Hence, it suffices to demonstrate that under Assumption 3.1 and Assumption 3.2, the asymptotic normality of $\mathrm{x}\widehat{\mathrm{MMD}}^2$ and $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ holds under the null, and it further holds under the alternative when Assumption 3.3 is satisfied.

The asymptotic normality of $\mathrm{x}\widehat{\mathrm{MMD}}^2$ under Assumption 3.1 has already been established by Shekhar et al. (Theorem 5, 2022). Thus, it remains to show that this result also holds under the alternative and that $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ asymptotically follows $N(0,1)$ as $n\to\infty$ under the both null and alternative given the considered conditions.

To this end, we first prove that $\mathrm{x}\widehat{\mathrm{MMD}}^2$ and the oracle statistic, $\mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ} := \overline{T}_{\mathrm{oracle}}$ are asymptotically $N(0,1)$. Then, we prove that $\mathrm{xss}\widehat{\mathrm{MMD}}^2 - \mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ} = o_P(1)$ to conclude the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$. Finally, we compare their asymptotic power under the alternative to complete the proof of Theorem 3.4.

Step 1: Asymptotic Normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ}$. We proceed in similar steps as we have done in the proof of Theorem 2.3. We first show the asymptotic normality when using the true variance, then show the same result is valid with the sample variance.

For the oracle version, we prove the asymptotic normality for general distributions P_X and P_Y, which ensures that the result holds under both the null and alternative hypotheses.

Recall from the proof of Theorem 2.3 that

$$\mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ} = \frac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}} - \mathbb{E}[\widehat{f}(X)|\widehat{f}] + \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{s_{n+m}} = \frac{\sum_{i=1}^{n_1+m_1} G_i - \sum_{i=1}^{n_2+m_2} H_i}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}},$$

where

$$G_i = \begin{cases} \frac{1}{n_1}\big\{\widehat{f}(X_i) - \mathbb{E}[\widehat{f}(X_i)\,|\,\widehat{f}]\big\} - \frac{m_1}{n_1(n_1+m_1)}\big\{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \mathbb{E}[\widehat{f}(X_i)\,|\,\widehat{f}]\big\} & \text{if } 1\le i\le n_1, \\ \frac{1}{n_1+m_1}\big\{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \mathbb{E}[\widehat{f}(X_i)\,|\,\widehat{f}]\big\} & \text{if } n_1+1\le i\le n_1+m_1, \end{cases}$$

and

$$H_i = \begin{cases} \frac{1}{n_2}\big\{\widehat{f}(Y_i) - \mathbb{E}[\widehat{f}(Y_i)\,|\,\widehat{f}]\big\} - \frac{m_2}{n_2(n_2+m_2)}\big\{\mathbb{E}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}] - \mathbb{E}[\widehat{f}(Y_i)\,|\,\widehat{f}]\big\} & \text{if } 1\le i\le n_2, \\ \frac{1}{n_2+m_2}\big\{\mathbb{E}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}] - \mathbb{E}[\widehat{f}(Y_i)\,|\,\widehat{f}]\big\} & \text{if } n_2+1\le i\le n_2+m_2. \end{cases}$$

The denominator s_{n+m} is recalled as

$$
\begin{aligned}
s_{n+m}^2 &= \mathrm{Var}(\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}} \,|\, \widehat{f}) = \sum_{i=1}^{n_1+m_1} \mathrm{Var}(G_i \,|\, \widehat{f}) + \sum_{i=1}^{n_2+m_2} \mathrm{Var}(H_i \,|\, \widehat{f}) \\
&= \underbrace{\frac{1}{n_1}\sigma_{1,X,\widehat{f}}^2 + \frac{1}{n_1+m_1}\sigma_{2,X,\widehat{f}}^2}_{\sigma_{X,\widehat{f}}^2} + \underbrace{\frac{1}{n_2}\sigma_{1,Y,\widehat{f}}^2 + \frac{1}{n_2+m_2}\sigma_{2,Y,\widehat{f}}^2}_{\sigma_{Y,\widehat{f}}^2}.
\end{aligned}
$$

To ensure that $\mathrm{xss}\widehat{\mathrm{MMD}}_{\circ}^2$ is asymptotically $N(0,1)$ distributed, it suffices to show that Lyapunov's condition (12) is satisfied. For simplicity, we take $\delta = 2$ and show the following convergence holds under the given conditions:

$$
\sum_{i=1}^{n_1+m_1} \mathbb{E}\biggl[\biggl|\frac{G_i}{s_{n+m}}\biggr|^4 \,\bigg|\, \widehat{f}\biggr] + \sum_{i=1}^{n_2+m_2} \mathbb{E}\biggl[\biggl|\frac{H_i}{s_{n+m}}\biggr|^4 \,\bigg|\, \widehat{f}\biggr] = o_P(1). \tag{17}
$$

By symmetry, we focus on the first term involving G_i values. Letting $r_1 = \frac{m_1}{n_1+m_1}$, we obtain an upper bound for the first term as

$$
\begin{aligned}
\sum_{i=1}^{n_1+m_1} \mathbb{E}\biggl[\biggl|\frac{G_i}{s_{n+m}}\biggr|^4 \,\bigg|\, \widehat{f}\biggr] \;&\lesssim\; \frac{1}{n_1^3}\mathbb{E}\biggl[\frac{|\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]|^4}{s_{n+m}^4} \,|\, \widehat{f}\biggr] \\
&+ \frac{r_1^4}{n_1^3}\mathbb{E}\biggl[\frac{|\mathbb{E}[\widehat{f}(X) \,|\, V, \widehat{f}] - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]|^4}{s_{n+m}^4} \,|\, \widehat{f}\biggr] \\
&+ \frac{r_1}{(n_1+m_1)^3}\mathbb{E}\biggl[\frac{|\mathbb{E}[\widehat{f}(X) \,|\, V, \widehat{f}] - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]|^4}{s_{n+m}^4} \,|\, \widehat{f}\biggr] \\
&\lesssim\; \biggl(\frac{1+r_1^4}{n_1^2} + \frac{r_1 n_1}{(n_1+m_1)^3}\biggr)\mathbb{E}\biggl[\frac{|\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]|^4}{n_1 s_{n+m}^4} \,|\, \widehat{f}\biggr],
\end{aligned} \tag{18}
$$

where we used conditional Jensen's inequality for the last inequality.

On the other hand, by the law of total variance, s_{n+m}^2 term is lower bounded as

$$
\begin{aligned}
s_{n+m}^2 \geq \sigma_{X,\widehat{f}}^2 &= \frac{1}{n_1}\sigma_{1,X,\widehat{f}}^2 + \frac{1}{n_1+m_1}\sigma_{2,X,\widehat{f}}^2 \\
&\geq \frac{1}{n_1+m_1}\sigma_{1,X,\widehat{f}}^2 + \frac{1}{n_1+m_1}\sigma_{2,X,\widehat{f}}^2 = \frac{1}{n_1+m_1}\mathrm{Var}(\widehat{f}(X) \,|\, \widehat{f}) = \frac{1}{n_1+m_1}\sigma_{X,\widehat{f}}^2.
\end{aligned}
$$

Hence in order to show that the first term in (17) is $o_P(1)$ under $n_1 \asymp m_1$, it suffices to show that the two claims hold:

$$
\text{(i)} \;\; \frac{\mathbb{E}[\{\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]\}^4]}{n_1\{\mathbb{E}[\sigma_{X,\widehat{f}}^2]\}^2} = o_P(1) \quad \text{and} \quad \text{(ii)} \;\; \frac{\mathbb{E}[\sigma_{X,\widehat{f}}^2]}{\sigma_{X,\widehat{f}}^2} = O_P(1).
$$

We shall prove these two claims in order.

Proof of claim (i) Denote $X \sim P_X$ and $Y \sim P_Y$. Given a kernel k and its feature map ψ so that $k(x,y) \coloneqq \langle \psi(x), \psi(y) \rangle_{\mathcal{H}_k}$ (we will drop the dependence on $\mathcal{H}_k$ for brevity), we define its centered version $\overline{k}_X$ with respect to P_X as

$$
\begin{aligned}
\overline{k}_X(x_1,x_2) &= k(x_1,x_2) - \mathbb{E}[k(x_1,X)] - \mathbb{E}[k(X,x_2)] + \mathbb{E}[k(X_1,X_2)] \\
&= \langle \psi(x_1) - \mathbb{E}_{P_X}[\psi(X)], \psi(x_2) - \mathbb{E}_{P_X}[\psi(X)] \rangle \\
&= \sum_{i=1}^{\infty} \lambda_i \phi_i(x_1)\phi_i(x_2),
\end{aligned}
$$

where we use spectral decomposition to denote the centered kernel $\overline{k}_X(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$ with orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ and corresponding eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$. Similarly, we define the centered kernel with respect to P_Y as $\overline{k}_Y(x,y) \coloneqq \langle \psi(x) - \mathbb{E}[\psi(Y)], \psi(y) - \mathbb{E}[\psi(Y)] \rangle = \sum_{i=1}^{\infty} \check{\lambda}_i \check{\phi}_i(x)\check{\phi}_i(y)$.

We express the witness function in terms of the inner product of feature maps as follows:

$$\widehat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} k(X_i', x) - \frac{1}{n_2}\sum_{i=1}^{n_2} k(Y_i', x) = \langle \bar{\psi}_X - \bar{\psi}_Y, \psi(x)\rangle,$$

where we denote the sample mean of the feature map $\bar{\psi}_X \coloneqq \frac{1}{n_1}\sum_{i=1}^{n_1} k(X_i, \cdot)$ and $\bar{\psi}_Y \coloneqq \frac{1}{n_1}\sum_{i=1}^{n_2} k(Y_i, \cdot)$ as $\bar{\psi}_X$ and $\bar{\psi}_Y$, respectively. We also let

$$\overline{k}_{Y,X}(y,x) = \langle \psi(y) - \mathbb{E}_{P_Y}[\psi(Y)], \psi(x) - \mathbb{E}_{P_X}[\psi(X)]\rangle$$

from which we observe that

$$\mathrm{MMD}^2 = \mathbb{E}[\overline{k}_X(Y,Y')] = \mathbb{E}[\overline{k}_Y(X,X')] = \sum_{i=1}^{\infty} \lambda_i \mathbb{E}[\phi_i(Y)]^2 = \sum_{i=1}^{\infty} \check{\lambda}_i \mathbb{E}[\check{\phi}_i(X)]^2.$$

Given the notation and denoting $\overline{\phi}_{i,X}$ and $\overline{\phi}_{i,X}$ as the sample mean of $\{\phi_i(X_j')\}_{j=1}^{n_1}$ and $\{\phi_i(Y_j')\}_{j=1}^{n_2}$, respectively, we obtain the upper bound for the numerator as

$$\begin{aligned}
&\mathbb{E}[\{\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]\}^4] = \mathbb{E}\left[\left\{\sum_{i=1}^{\infty} \lambda_i(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})\phi_i(X)\right\}^4\right] \\
&= \mathbb{E}\left[\left\{\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{k}_X(X_i, X) - \frac{1}{n_2}\sum_{i=1}^{n_2} \overline{k}_X(Y_i, X)\right\}^4\right] \\
&= \mathbb{E}\left[\left\{\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{k}_X(X_i, X) - \frac{1}{n_2}\sum_{i=1}^{n_2} \overline{k}_{Y,X}(Y_i, X) + \langle \mathbb{E}_{P_X}[\psi(X)] - \mathbb{E}_{P_Y}[\psi(Y)], \psi(X) - \mathbb{E}_{P_X}[\psi(X)]\rangle\right\}^4\right] \\
&\lesssim \mathbb{E}\left[\left\{\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{k}_X(X_i, X)\right\}^4\right] + \mathbb{E}\left[\left\{\frac{1}{n_2}\sum_{i=1}^{n_2} \overline{k}_{Y,X}(Y_i, X)\right\}^4\right] \\
&\qquad\qquad + \mathbb{E}\big[\langle \mathbb{E}_{P_X}[\psi(X)] - \mathbb{E}_{P_Y}[\psi(Y)], \psi(X) - \mathbb{E}_{P_X}[\psi(X)]\rangle^4\big] \\
&\lesssim \frac{1}{n_1^3}\mathbb{E}[\overline{k}_X(X_1,X_2)^4] + \frac{1}{n_1^2}\mathbb{E}[\overline{k}_X(X_1,X_2)^2\overline{k}_X(X_1,X_3)^2] + \frac{1}{n_2^3}\mathbb{E}[\overline{k}_{Y,X}(Y,X)^4] \\
&\qquad\qquad + \frac{1}{n_2^2}\mathbb{E}[\overline{k}_{Y,X}(Y_1,X)^2\overline{k}_{Y,X}(Y_2,X)^2] + \mathrm{MMD}^4 \times \mathbb{E}[\overline{k}_X(X,X)^2].
\end{aligned}$$

Next, we compute the conditional variance of $\widehat{f}(X)$ as

$$\begin{aligned}
\mathrm{Var}[\widehat{f}(X) \,|\, \widehat{f}] &= \mathbb{E}[\langle \bar{\psi}_X - \bar{\psi}_Y, \psi(X) - \mathbb{E}_P[\psi(X)]\rangle^2 \,|\, \bar{\psi}_X, \bar{\psi}_Y] \\
&= \mathbb{E}[\langle \bar{\psi}_X - \mathbb{E}[\psi(X)] + \mathbb{E}[\psi(X)] - \bar{\psi}_Y, \psi(X) - \mathbb{E}_P[\psi(X)]\rangle^2 \,|\, \bar{\psi}_X, \bar{\psi}_Y] \\
&= \mathbb{E}\left[\left\{\sum_{i=1}^{\infty} \lambda_i\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\phi_i(X_j')\right)\phi_i(X) - \sum_{i=1}^{\infty} \lambda_i\left(\frac{1}{n_2}\sum_{j=1}^{n_2}\phi_i(Y_j')\right)\phi_i(X)\right\}^2 \,\middle|\, (X_j'), (Y_j')\right] \\
&= \mathbb{E}\left[\left\{\sum_{i=1}^{\infty} \lambda_i(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})\phi_i(X)\right\}^2 \,\middle|\, (X_j'), (Y_j')\right] \\
&= \sum_{i=1}^{\infty} \lambda_i^2(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2.
\end{aligned}$$

Moreover, we denote

$$\overline{g}_X(x,y) = \mathbb{E}[\overline{k}_X(x,X)\overline{k}_X(y,X)] = \sum_{i=1}^{\infty} \lambda_i^2 \phi_i(x)\phi_i(y) \quad \text{and}$$

$$\overline{g}_Y(x,y) = \mathbb{E}[\overline{k}_Y(x,Y)\overline{k}_Y(y,Y)] = \sum_{i=1}^{\infty} \check{\lambda}_i^2 \check{\phi}_i(x)\check{\phi}_i(y),$$

and compute the lower bound for the denominator as

$$
\begin{aligned}
\mathbb{E}[\sigma^2_{X,\widehat{f}}] = \mathbb{E}\{\mathrm{Var}[\widehat{f}(X) \mid \widehat{f}]\} &= \sum_{i=1}^{\infty} \lambda_i^2 \mathbb{E}\big[(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2\big] = \sum_{i=1}^{\infty} \lambda_i^2 \bigg(\frac{1}{n_1} + \mathbb{E}\big[\overline{\phi}^2_{i,Y}\big]\bigg) \\
&= \sum_{i=1}^{\infty} \lambda_i^2 \bigg(\frac{1}{n_1} + \frac{\mathrm{Var}[\phi_i(Y)]}{n_2} + \mathbb{E}[\phi_i(Y)]^2\bigg) \\
&= \frac{\mathbb{E}[\overline{k}_X(X_1,X_2)^2]}{n_1} + \frac{\mathbb{E}[\overline{g}_X(Y,Y)] - \mathbb{E}[\overline{g}_X(Y_1,Y_2)]}{n_2} + \mathbb{E}[\overline{g}_X(Y_1,Y_2)] \\
&\gtrsim \frac{\mathbb{E}[\overline{g}_X(X,X)]}{n_1} + \frac{\mathbb{E}[\overline{g}_X(Y,Y)]}{n_2} + \mathbb{E}[\overline{g}_X(Y_1,Y_2)].
\end{aligned}
$$

Combining these and letting $n_1 \leq n_2$, it suffices to show that the following convergence results hold:

$$
\begin{aligned}
&\text{(a)} \;\; \frac{\mathrm{MMD}^4 \mathbb{E}[\overline{k}_X(X,X)^2]}{n_1\{(n_1^{-1} + n_2^{-1})\mathbb{E}[\overline{g}_X(X,X)] + \mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} = o_P(1), \\
&\text{(b)} \;\; \frac{\mathbb{E}[\overline{k}_{Y,X}(Y,X)^4]}{n_1 n_2^3\{n_1^{-1}\mathbb{E}[\overline{g}_X(X,X)] + n_2^{-1}\mathbb{E}[\overline{g}_X(Y,Y)] + \mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} = o_P(1) \\
&\text{(c)} \;\; \frac{\mathbb{E}[\overline{k}_{Y,X}(Y_1,X)^2 \overline{k}_{Y,X}(Y_2,X)^2]}{n_1 n_2^2\{n_1^{-1}\mathbb{E}[\overline{g}_X(X,X)] + n_2^{-1}\mathbb{E}[\overline{g}_X(Y,Y)] + \mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} = o_P(1), \\
&\text{(d)} \;\; \frac{\mathbb{E}[\overline{k}_X(X_1,X_2)^4]}{n_1^2\{\mathbb{E}[\overline{g}_X(X,X)]\}^2} = o_P(1), \quad \text{and} \quad \text{(e)} \;\; \frac{\mathbb{E}[\overline{k}_X(X_1,X_2)^2 \overline{k}_X(X_1,X_3)^2]}{n_1\{\mathbb{E}[\overline{g}_X(X,X)]\}^2} = o_P(1).
\end{aligned}
\tag{19}
$$

Let us verify that these convergence results hold. With $n_1 \leq n_2$, we obtain from Assumption 3.3

$$
\frac{\mathrm{MMD}^4 \mathbb{E}[\overline{k}_X(X,X)^2]}{n_1\{(n_1^{-1} + n_2^{-1})\mathbb{E}[\overline{g}_X(X,X)] + \mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} \leq \frac{\mathrm{MMD}^4 \mathbb{E}[\overline{k}_X(X,X)^2]}{\{n_1\mathbb{E}[\overline{g}_X(X,X)] + n_1^2\mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} = o_P(1),
$$

which implies that (a) holds.

Since we assume that P_X and P_Y have density functions p and q and $\|p/q\|_{L_\infty} \vee \|q/p\|_{L_\infty} \leq C$,

$$
\mathbb{E}[\overline{g}_X(Y,Y)] \asymp \mathbb{E}[\overline{g}_X(X,X)] \quad \text{and} \quad \mathbb{E}[\overline{k}_{Y,X}(Y,X)^4] \lesssim \mathrm{MMD}^4 \mathbb{E}[\overline{k}_X(X,X)^2].
$$

Combining these results with Assumption 3.3, we show that (b) and (c) hold. Lastly, from Assumption 3.1, (d) and (e) are satisfied. Hence we prove that the claim (i) holds.

Proof of claim (ii) We show that the ratio converges to one in probability

$$
\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\sigma^2_{X,\widehat{f}}} = 1 + o_P(1).
$$

which directly shows that the claim (ii) holds.

Recall

$$
\begin{aligned}
&\sigma^2_{X,\widehat{f}} = \sum_{i=1}^{\infty} \lambda_i^2 (\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2 \quad \text{and} \\
&\mathbb{E}[\sigma^2_{X,\widehat{f}}] \gtrsim \frac{\mathbb{E}[\overline{g}_X(X,X)]}{n_1} + \frac{\mathbb{E}[\overline{g}_X(Y,Y)]}{n_2} + \mathbb{E}[\overline{g}_X(Y_1,Y_2)].
\end{aligned}
$$

Letting $Z \sim N(0,1)$, we have for any $\epsilon > 0$ that

$$
\begin{aligned}
&\mathbb{P}\Bigg(\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\sigma^2_{X,\widehat{f}}} \geq \epsilon\Bigg) = \mathbb{P}\Bigg(\frac{\sigma^2_{X,\widehat{f}}}{\mathbb{E}[\sigma^2_{X,\widehat{f}}]} \leq \epsilon^{-1}\Bigg) \\
&\leq \mathbb{P}\Bigg(\frac{\lambda_1^2(\overline{\phi}_{1,X} - \overline{\phi}_{1,Y})^2}{\mathbb{E}[\sigma^2_{X,\widehat{f}}]} \leq \epsilon^{-1}\Bigg) = \mathbb{P}\Bigg((\overline{\phi}_{1,X} - \overline{\phi}_{1,Y})^2 \leq \frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2}\Bigg) \\
&= \mathbb{P}\Bigg(-\sqrt{\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}} \leq \frac{\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}}{\sqrt{\mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}} \leq \sqrt{\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}}\Bigg) \\
&\overset{\text{(a)}}{\leq} \mathbb{P}\Bigg(-\sqrt{\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}} \leq Z - \frac{\mu_1}{\sqrt{\mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}} \leq \sqrt{\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}}\Bigg) + C\sqrt{\frac{\mathbb{E}[\phi_1(X)^4]}{n_1}} \\
&\overset{\text{(b)}}{\lesssim} \sqrt{\frac{1}{\epsilon} \times \frac{\sum_{i=1}^\infty \lambda_i^2(1 + n_1\mu_i^2)}{\lambda_1^2}} + C\sqrt{\frac{\mathbb{E}[\phi_1(X)^4]}{n_1}} \\
&\overset{\text{(c)}}{\lesssim} \sqrt{\frac{1}{\epsilon} \times \frac{\sum_{i=1}^\infty \lambda_i^2(1 + n_1\mu_i^2)}{\lambda_1^2}} + \frac{\sum_{i=1}^\infty \lambda_i^2}{\lambda_1^2} \times o(1),
\end{aligned}
$$

where step (a) uses the Berry–Esseen bound and step (b) and (c) hold by the following reasoning: First of all, we use the observation that for $b \geq 0$

$$
\sup_{a \in \mathbb{R}} \int_{a-b}^{a+b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \int_{-b}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \leq \frac{2b}{\sqrt{2\pi}},
$$

which can be verified by calculus. Hence, the first term in step (a) can be bounded above by

$$
\sqrt{\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]}} \quad \text{up to a constant.}
$$

Hence step (b) follows since

$$
\begin{aligned}
\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]} &\lesssim \frac{n_1^{-1}\mathbb{E}[\overline{g}_X(X,X)] + \mathbb{E}[\overline{g}_X(Y_1,Y_2)]}{\epsilon\lambda_1^2(n_1^{-1}\mathrm{Var}[\phi_1(X)] + n_2^{-1}\mathrm{Var}[\phi_1(Y)])} \\
&\lesssim \frac{\mathbb{E}[\overline{g}_X(X,X)] + n_1\mathbb{E}[\overline{g}_X(Y_1,Y_2)]}{\epsilon\lambda_1^2} = \frac{1}{\epsilon} \times \frac{\sum_{i=1}^\infty \lambda_i^2(1 + n_1\mu_i^2)}{\lambda_1^2}
\end{aligned}
$$

where $\mu_i = \mathbb{E}[\phi_i(Y)]$. Step (c) uses the observation that

$$
\Bigg(\frac{\lambda_1^2}{\sum_{i=1}^\infty \lambda_i^2}\Bigg)^2 \frac{\mathbb{E}[\phi_1(X)^4]}{n_1} \leq \frac{\mathbb{E}[\overline{k}_X(X_1,X_2)^2\overline{k}_X(X_1,X_3)^2]}{n_1\{\mathbb{E}[\overline{g}_X(X,X)]\}^2} = o_P(1),
$$

and thus

$$
\sqrt{\frac{\mathbb{E}[\phi_1(X)^4]}{n_1}} = \frac{\sum_{i=1}^\infty \lambda_i^2}{\lambda_1^2} o(1),
$$

where the first convergence is derived from Assumption 3.1. Combining these results, we prove that the claim (ii) is valid.

A similar argument applies to the second term of each equation in (17). Therefore, assuming that Assumption 3.1 and Assumption 3.3 hold, we prove that $\mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ}$ is asymptotically $N(0,1)$ under the alternative.

Based on the previous results, we now focus on the null hypothesis and verify the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}_{\circ}^{2}$ under the null when Assumption 3.1 holds. Through the same reasoning, it suffices to prove that claims (i) and (ii) are valid.

To show that claim (i) holds, we prove that condition (19) is satisfied. Under the null, $\mathrm{MMD} = 0$ which satisfies (a), (b), and (c). From Assumption 3.1, (d) and (e) are satisfied. Therefore, we prove that the claim (i) holds.

Next, we show that the claim (ii) holds. Under the null, we obtain the upper bound

$$\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\epsilon\lambda_1^2 \mathrm{Var}[\overline{\phi}_{1,X} - \overline{\phi}_{1,Y}]} \lesssim \frac{\mathbb{E}[\overline{g}_X(X,X)]}{\epsilon\lambda_1^2} = \frac{1}{\epsilon} \times \frac{\sum_{i=1}^{\infty}\lambda_i^2}{\lambda_1^2}$$

which leads to

$$\mathbb{P}\left(\frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\sigma^2_{X,\widehat{f}}} \geq \epsilon\right) \lesssim \sqrt{\frac{1}{\epsilon} \times \frac{\sum_{i=1}^{\infty}\lambda_i^2}{\lambda_1^2}} + \frac{\sum_{i=1}^{\infty}\lambda_i^2}{\lambda_1^2} o(1) \quad \text{for sufficiently large } n_1.$$

Hence we show that the claim (ii) is valid. Therefore, we conclude that $\mathrm{xss}\widehat{\mathrm{MMD}}^{2}$ is asymptotically normal under the null hypothesis as well.

A similar argument applies to the second term of each equation in (17). Therefore, assuming that Assumption 3.1 hold, we prove that $\mathrm{xss}\widehat{\mathrm{MMD}}_{\circ}^{2}$ is asymptotically $N(0,1)$ under the null.

In addition, we discuss the asymptotic normality of $\mathrm{x}\widehat{\mathrm{MMD}}^{2}$ under the alternative. Recall that in Appendix D.1, $\mathrm{x}\widehat{\mathrm{MMD}}^{2}$ with the same witness function $\widehat{f}$ is defined as

$$\mathrm{x}\widehat{\mathrm{MMD}}^{2} = \frac{\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}}{\sqrt{\widetilde{\sigma}^2_{X,\widehat{f}} + \widetilde{\sigma}^2_{Y,\widehat{f}}}}.$$

Similar to the previous proof, we show that the Lyapunov condition (12) is satisfied. For simplicity, we take $\delta = 2$ and show

$$\frac{1}{\sigma^4_{n_1,n_2}}\left[\frac{1}{n_1^4}\sum_{i=1}^{n_1}\mathbb{E}[\{\widehat{f}(X_i) - \mathbb{E}[\widehat{f}(X) \mid \widehat{f}]\}^4 \mid \widehat{f}] + \frac{1}{n_2^4}\sum_{i=1}^{n_2}\mathbb{E}[\{\widehat{f}(Y_i) - \mathbb{E}[\widehat{f}(Y) \mid \widehat{f}]\}^4 \mid \widehat{f}]\right] = o_P(1)$$

where we denote

$$\sigma^2_{n_1,n_2} = \widetilde{\sigma}^2_{X,\widehat{f}} + \widetilde{\sigma}^2_{Y,\widehat{f}}.$$

In the previous proof, we have already shown that the following convergence holds under Assumption 3.1 and Assumption 3.3:

$$\frac{\mathbb{E}[\{\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \mid \widehat{f}]\}^4]}{n_1\{\mathbb{E}[\sigma^2_{X,\widehat{f}}]\}^2} + \frac{\mathbb{E}[\{\widehat{f}(Y) - \mathbb{E}[\widehat{f}(Y) \mid \widehat{f}]\}^4]}{n_2\{\mathbb{E}[\sigma^2_{Y,\widehat{f}}]\}^2} = o_P(1) \quad \text{and} \quad \frac{\mathbb{E}[\sigma^2_{X,\widehat{f}}]}{\sigma^2_{X,\widehat{f}}} + \frac{\mathbb{E}[\sigma^2_{Y,\widehat{f}}]}{\sigma^2_{Y,\widehat{f}}} = O_P(1).$$

Hence the asymptotic normality of $\mathrm{x}\widehat{\mathrm{MMD}}^{2}$ also follows under the alternative.

Step 2: Asymptotic Normality with Sample Variance. In this step, we show that the ratio of the sample variance to the population variance converges to 1 in probability. Following the same approach as in Step 2 of the proof of Theorem 2.3, it suffices to show that (14) holds.

Without loss of generality, we focus on the first convergence result. From the definition of $\widehat{\sigma}^2_{X,\widehat{f}}$ and $\widehat{\sigma}^2_{Y,\widehat{f}}$, we obtain

$$\frac{\left|\widehat{\sigma}^2_{X,\widehat{f}} - \sigma^2_{X,\widehat{f}}\right|}{\sigma^2_{X,\widehat{f}}} \leq \underbrace{\frac{1}{n_1}\frac{\left|\widehat{\sigma}^2_{1,X,\widehat{f}} - \sigma^2_{1,X,\widehat{f}}\right|}{\sigma^2_{X,\widehat{f}}}}_{\text{(I)}} + \underbrace{\frac{1}{n_1+m_1}\frac{\left|\widehat{\sigma}^2_{2,X,\widehat{f}} - \sigma^2_{2,X,\widehat{f}}\right|}{\sigma^2_{X,\widehat{f}}}}_{\text{(II)}}.$$

Since $\mathbb{E}[\widehat{\sigma}^2_{1,X,\widehat{f}} \,|\, \widehat{f}] = \sigma^2_{1,X,\widehat{f}}$, the first term above satisfies (I) $= o_P(1)$ if

$$\frac{1}{n_1^2}\frac{\mathbb{E}[\{\widehat{\sigma}^2_{1,X,\widehat{f}} - \sigma^2_{1,X,\widehat{f}}\}^2 \,|\, \widehat{f}]}{\sigma^4_{X,\widehat{f}}} = o_P(1).$$

This can be seen using the pieces established before in step 1. When showing the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ$, we have proved the the both claims (i) and (ii) hold. From this, we obtain that

$$\frac{1}{n_1^2}\frac{\mathbb{E}[\{\widehat{\sigma}^2_{1,X,\widehat{f}} - \sigma^2_{1,X,\widehat{f}}\}^2 \,|\, \widehat{f}]}{\sigma^4_{X,\widehat{f}}} \lesssim \underbrace{\frac{1}{n_1^3}\frac{\mathbb{E}[\{\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]\}^4]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}}_{o_P(1)} \times \underbrace{\frac{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}{\sigma^4_{X,\widehat{f}}}}_{O_P(1)} = o_P(1).$$

For the second term (II), we may similarly proceed using conditional Jensen's inequality as

$$\begin{aligned}
\text{(II)} &\lesssim \frac{1}{n_1^3}\frac{\mathbb{E}[\{\mathbb{E}[\widehat{f}(X) \,|\, V, \widehat{f}] - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]\}^4 \,|\, \widehat{f}]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2} \times \frac{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}{\sigma^4_{X,\widehat{f}}} \\
&\lesssim \frac{1}{n_1^3}\frac{\mathbb{E}[\{\widehat{f}(X) - \mathbb{E}[\widehat{f}(X) \,|\, \widehat{f}]\}^4]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2} \times \frac{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}{\sigma^4_{X,\widehat{f}}} = o_P(1).
\end{aligned}$$

Combining the results, we use Slutsky's theorem to conclude that

$$\mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ = \frac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}} = \frac{\widehat{\mu}_{X,\widehat{f}} - \widehat{\mu}_{Y,\widehat{f}}}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}} \times \frac{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}} \xrightarrow{D} N(0,1).$$

Step 3: Asymptotic Normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$.

The aim of this subsection is to identify condition on these estimators under which

$$\mathrm{xss}\widehat{\mathrm{MMD}}^2 - \mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ = o_P(1).$$

Once this condition is fulfilled, $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ converges to $N(0,1)$ by Slutsky's theorem. Denote $\mathrm{xss}\widehat{\mathrm{MMD}}^2 = \frac{\widehat{N}}{\widehat{D}}$ and $\mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ = \frac{N_\circ}{D_\circ}$. Then

$$\begin{aligned}
\mathrm{xss}\widehat{\mathrm{MMD}}^2 - \mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ &= \frac{\widehat{N}}{\widehat{D}} - \frac{\widehat{N}}{D_\circ} + \frac{\widehat{N}}{D_\circ} - \frac{N_\circ}{D_\circ} \\
&= \frac{\widehat{N} - N_\circ}{D_\circ}\left(\frac{D_\circ}{\widehat{D}} - 1\right) + \underbrace{\frac{N_\circ}{D_\circ}}_{=O_P(1)}\left(\frac{D_\circ}{\widehat{D}} - 1\right) + \frac{\widehat{N} - N_\circ}{D_\circ}
\end{aligned}$$

where we have proved that $\mathrm{xss}\widehat{\mathrm{MMD}}^2_\circ = N_\circ/D_\circ = O_P(1)$ in the previous step.

Hence it suffices to show that two claims hold:

$$\text{(i}'\text{)}\ \frac{\widehat{N} - N_\circ}{D_\circ} = o_P(1) \quad \text{and} \quad \text{(ii}'\text{)}\ \frac{D_\circ}{\widehat{D}} - 1 = o_P(1).$$

Proof of claim (i′). Note that $D_\circ^2 = \widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}$ using the sample variance and let $D_\star^2 = \sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}$ using the population variance where $\sigma^2_{X,\widehat{f}} = \frac{1}{n_1}\sigma^2_{1,X} + \frac{1}{n_1+m_1}\sigma^2_{2,X}$ and $\sigma^2_{Y,\widehat{f}} = \frac{1}{n_2}\sigma^2_{1,Y} + \frac{1}{n_2+m_2}\sigma^2_{2,Y}$. From our previous result obtained in Step 2,

$$\frac{D_\circ^2}{D_\star^2} = \frac{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}} \xrightarrow{P} 1.$$

By the continuous mapping theorem,

$$\frac{D_\star}{D_\circ} \xrightarrow{P} 1,$$

and thus

$$\frac{\widehat{N} - N_\circ}{D_\circ} = \frac{\widehat{N} - N_\circ}{D_\star}\{1 + o_P(1)\}.$$

On the other hand,

$$\left|\frac{\widehat{N} - N_\circ}{D_\star}\right| \le \frac{|R_X|}{\sigma_{X,\widehat{f}}} + \frac{|R_Y|}{\sigma_{Y,\widehat{f}}}$$

where

$$\begin{aligned}
R_X &= \frac{1}{n_1}\sum_{i=1}^{n_1}\{\widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\} \\
&\qquad + \frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\}, \\
R_Y &= \frac{1}{n_2}\sum_{i=1}^{n_2}\{\widehat{\mathbb{E}}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}] - \mathbb{E}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}]\} \\
&\qquad + \frac{1}{n_2+m_2}\sum_{i=1}^{n_2+m_2}\{\mathbb{E}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(Y_i)\,|\,W_i,\widehat{f}]\}.
\end{aligned}$$

Using the fact that R_X and R_Y are location-shift invariant , it can be seen that

$$\begin{aligned}
R_X^2 &= O_P\Big(n_1^{-1}\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(X)\,|\,V,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]\}^2\,|\,\widehat{f}\big]\Big), \\
R_Y^2 &= O_P\Big(n_2^{-1}\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(Y)\,|\,W,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(Y)\,|\,W,\widehat{f}]\}^2\,|\,\widehat{f}\big]\Big).
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{R_X^2}{\sigma^2_{X,\widehat{f}}} &= O_P\left(\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(X)\,|\,V,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]\}^2\,|\,\widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(X)\,|\,\widehat{f}\}}\right), \\
\frac{R_Y^2}{\sigma^2_{Y,\widehat{f}}} &= O_P\left(\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(Y)\,|\,W,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(Y)\,|\,W,\widehat{f}]\}^2\,|\,\widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(Y)\,|\,\widehat{f}\}}\right).
\end{aligned}$$

Since Assumption 3.2 holds, both terms converges to zero with in probability, which proves that the claim (i′) holds.

Proof of claim (ii′). Note that $\widehat{D}^2 = \widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}$. By the continuous mapping theorem, it is sufficient to prove that

$$\frac{\widehat{D}^2}{D_\circ^2} - 1 = o_P(1),$$

which, in turn, is implied by

$$\begin{aligned}
\frac{1}{n_1}\frac{|\widehat{\sigma}^{\dagger 2}_{1,X} - \widehat{\sigma}^2_{1,X}|}{\widehat{\sigma}^2_{X,\widehat{f}}} &= o_P(1), \quad \frac{1}{n_1+m_1}\frac{|\widehat{\sigma}^{\dagger 2}_{2,X} - \widehat{\sigma}^2_{2,X}|}{\widehat{\sigma}^2_{X,\widehat{f}}} = o_P(1), \\
\frac{1}{n_2}\frac{|\widehat{\sigma}^{\dagger 2}_{1,Y} - \widehat{\sigma}^2_{1,Y}|}{\widehat{\sigma}^2_{Y,\widehat{f}}} &= o_P(1), \quad \frac{1}{n_2+m_2}\frac{|\widehat{\sigma}^{\dagger 2}_{2,Y} - \widehat{\sigma}^2_{2,Y}|}{\widehat{\sigma}^2_{Y,\widehat{f}}} = o_P(1).
\end{aligned}$$

Without loss of generality, we focus on the terms on the first row with $\widehat{\sigma}^{\dagger 2}_{1,X}$ and $\widehat{\sigma}^{\dagger 2}_{2,X}$.

Observe that by the Cauchy–Schwarz inequality,

$$\frac{1}{n_1}\frac{|\widehat{\sigma}^{\dagger 2}_{1,X}-\widehat{\sigma}^{2}_{1,X}|}{\widehat{\sigma}^2_{X,\widehat{f}}} \;\leq\; \frac{1}{n_1}\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}A_i^2}{\widehat{\sigma}^2_{X,\widehat{f}}} + 2\sqrt{\frac{1}{n_1}\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}A_i^2}{\widehat{\sigma}^2_{X,\widehat{f}}}}\underbrace{\sqrt{\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}B_i^2}{\widehat{\sigma}^2_{1,X}}}}_{=1}$$

where A_i and B_i are defined as

$$A_i = \{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\} - \frac{1}{n_1}\sum_{j=1}^{n_1}\{\mathbb{E}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}]\},$$

$$B_i = \{\widehat{f}(X_i) - \mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\} - \frac{1}{n_1}\sum_{j=1}^{n_1}\{\widehat{f}(X_j) - \mathbb{E}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}]\}.$$

Note that $\frac{1}{n_1}\sum_{i=1}^{n_1}A_i^2 \leq \frac{1}{n_1}\sum_{i=1}^{n_1}\{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\}^2$. Thus, this term becomes $o_P(1)$ when

$$\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(X)\,|\,V,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]\}^2\,|\,\widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(X)\,|\,\widehat{f}\}} = o_P(1).$$

Similarly, for the second term, we have

$$\frac{1}{n_1+m_1}\frac{|\widehat{\sigma}^{\dagger 2}_{2,X}-\widehat{\sigma}^{2}_{2,X}|}{\widehat{\sigma}^2_{X,\widehat{f}}} \;\leq\; \frac{1}{n_1+m_1}\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widetilde{A}_i^2}{\widehat{\sigma}^2_{X,\widehat{f}}} + 2\sqrt{\frac{1}{n_1+m_1}\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widetilde{A}_i^2}{\widehat{\sigma}^2_{X,\widehat{f}}}}\underbrace{\sqrt{\frac{\frac{1}{n_1+m_1}\sum_{i=1}^{n_1+m_1}\widetilde{B}_i^2}{\widehat{\sigma}^2_{2,X}}}}_{=1}$$

where $\widetilde{A}_i$ and $\widetilde{B}_i$ are analogously defined as

$$\widetilde{A}_i = \{\mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}]\} - \frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\{\mathbb{E}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}]\},$$

$$\widetilde{B}_i = \mathbb{E}[\widehat{f}(X_i)\,|\,V_i,\widehat{f}] - \frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\mathbb{E}[\widehat{f}(X_j)\,|\,V_j,\widehat{f}].$$

By the same logic, this also becomes $o_P(1)$ when

$$\frac{\mathbb{E}\big[\{\mathbb{E}[\widehat{f}(X)\,|\,V,\widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}]\}^2\,|\,\widehat{f}\big]}{\mathrm{Var}\{\widehat{f}(X)\,|\,\widehat{f}\}} = o_P(1).$$

A similar argument applies to $\widehat{\sigma}^{\dagger 2}_{1,Y}$ and $\widehat{\sigma}^{\dagger 2}_{2,Y}$, which proves the claim (ii) holds. Thus, assuming Assumption 3.2 holds, we conclude that $\mathrm{xss}\widehat{\mathrm{MMD}}^2 - \mathrm{xss}\widehat{\mathrm{MMD}}^2_{\circ} = o_p(1)$, implying the asymptotic normality of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$.

Step 4: Power Comparison.

Assuming that Assumption 3.1, Assumption 3.2, and Assumption 3.3 hold, we showed that both $\mathrm{x}\widehat{\mathrm{MMD}}^2$ and $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ converges to a normal distribution under the null and alternative. This derives the explicit expression of the asymptotic power of each test statistics. Recall the definition of each test statistic. The power function of the xMMD test approximates that

$$\Phi\left(z_\alpha + \frac{\mathbb{E}[\widehat{f}(X)] - \mathbb{E}[\widehat{f}(Y)]}{\sqrt{n_1^{-1}\mathrm{Var}(\widehat{f}(X)) + n_2^{-1}\mathrm{Var}(\widehat{f}(Y))}}\right) \quad \text{as } n \to \infty.$$

On the other hand, the power function of the xssMMD test approximates

$$\Phi\left(z_\alpha + \frac{\mathbb{E}[\widehat{f}(X)] - \mathbb{E}[\widehat{f}(Y)]}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}}\right) \quad \text{as } n \to \infty.$$

Observe that the only difference results in the denominator which consists of the variance. Hence it suffices to show that

$$\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}} \leq \mathrm{Var}(\widehat{f}(X)) + \mathrm{Var}(\widehat{f}(Y)).$$

Using the total law of variance, we show

$$\begin{aligned}
\sigma^2_{X,\widehat{f}} &= \frac{1}{n_1}\mathbb{E}[\mathrm{Var}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] + \frac{1}{n_1+m_1}\mathrm{Var}[\mathbb{E}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] \\
&\leq \frac{1}{n_1}\mathbb{E}[\mathrm{Var}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] + \frac{1}{n_1}\mathrm{Var}[\mathbb{E}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] \\
&= \frac{1}{n_1}\mathrm{Var}(\widehat{f}(X)).
\end{aligned}$$

A similar computation applies to $\sigma^2_{X,\widehat{f}}$ and $\frac{1}{n_2}\mathrm{Var}(\widehat{f}(X))$. Therefore, we conclude that the power of the xssMMD test is asymptotically greater than or equal to that of the xMMD test.

F.4 Proof of Lemma 3.5

The proof of Lemma 3.5 follows the same lines of argument as in the proof of Shekhar et al. (2022, Theorem 8) with the additional observation that the numerator of $\mathrm{xss}\widehat{\mathrm{MMD}}^2$ has an expectation equal to $\mathrm{MMD}(P_X, P_Y)^2$. Hence we omit the proof here.

F.5 Proof of Theorem D.1

We will prove that under the condition (9), the following two inequalities hold:

$$\mathbb{E}_{P_{X,n},P_{Y,n}}\big[\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}\big] \lesssim \mathbb{E}_{P_{X,n},P_{Y,n}}\big[\widetilde{\sigma}^{2}_{X,\widehat{f}} + \widetilde{\sigma}^{2}_{Y,\widehat{f}}\big] \quad \text{and} \tag{20}$$

$$\mathrm{Var}_{P_{X,n},P_{Y,n}}\big[\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \widehat{\mu}^{\dagger}_{Y,\widehat{f}}\big] \lesssim \mathrm{Var}_{P_{X,n},P_{Y,n}}\big[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\big], \tag{21}$$

which directly implies the claim of Theorem D.1. For simplicity, we will omit the dependence on $P_{X,n}$ and $P_{Y,n}$ in subsequent expressions. As mentioned in the main text, we focus on the case where n_1, n_2, m_1, m_2 are even for simplicity and the general case can be handled similarly with minor modifcations.

Verification of (20)**.** Starting with the first inequality (20), it suffices to prove that $\mathbb{E}[\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}}] \lesssim \mathbb{E}[\widetilde{\sigma}^{2}_{X,\widehat{f}}]$ or equivalently,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}\big\{\widehat{f}(X_i) - \widehat{u}_X(V_i)\big\}^2 + \frac{1}{(n_1+m_1)^2}\sum_{i=1}^{n_1+m_1}\left\{\widehat{u}_X(V_i) - \frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\widehat{u}_X(V_j)\right\}^2\right] \\
\lesssim\ &\mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}\left\{\widehat{f}(X_i) - \frac{1}{n_1}\sum_{j=1}^{n_1}\widehat{f}(X_j)\right\}^2\right] = \frac{n_1-1}{n_1^2}\mathbb{E}\big[\mathrm{Var}\{\widehat{f}(X)\,|\,\widehat{f}\}\big] \lesssim \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\{\widehat{f}(X)\,|\,\widehat{f}\}\big].
\end{aligned}$$

Under the condition (9), we have

$$\mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}\big\{\widehat{f}(X_i) - \widehat{u}_X(V_i)\big\}^2\right] \lesssim \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big]$$

and

$$\mathbb{E}\Big[\frac{1}{(n_1+m_1)^2}\sum_{i=1}^{n_1+m_1}\Big\{\widehat{u}_X(V_i)-\frac{1}{n_1+m_1}\sum_{j=1}^{n_1+m_1}\widehat{u}_X(V_j)\Big\}^2\Big] \lesssim \frac{1}{n_1+m_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)\,|\,\widehat{u}_X\big\}\big]$$
$$\lesssim \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)\,|\,\widehat{u}_X\big\}\big] = \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)-\widehat{f}(X)+\widehat{f}(X)\,|\,\widehat{f},\widehat{u}_X\big\}\big]$$
$$\lesssim \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)-\widehat{f}(X)\,|\,\widehat{f},\widehat{u}_X\big\}\big] + \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big] \lesssim \frac{1}{n_1}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big].$$

This completes the proof of the first inequality (20).

Verification of (21)**.** For the second inequality (21), the law of total variance gives

$$\mathrm{Var}\big[\widehat{\mu}^{\dagger}_{X,\widehat{f}}-\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\big] = \mathbb{E}\big[\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}-\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\,|\,\widehat{f}\big\}\big] + \mathrm{Var}\big[\underbrace{\mathbb{E}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}-\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\,|\,\widehat{f}\big\}}_{=\mathbb{E}\{\widetilde{\mu}_{X,\widehat{f}}-\widetilde{\mu}_{Y,\widehat{f}}\,|\,\widehat{f}\}}\big].$$

Therefore it suffices to show that

$$\mathbb{E}\big[\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}-\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\,|\,\widehat{f}\big\}\big] \lesssim \mathbb{E}\big[\mathrm{Var}\big\{\widetilde{\mu}_{X,\widehat{f}}-\widetilde{\mu}_{Y,\widehat{f}}\,|\,\widehat{f}\big\}\big].$$

Using independence between $\widehat{\mu}^{\dagger}_{X,\widehat{f}}$ and $\widehat{\mu}^{\dagger}_{Y,\widehat{f}}$ (also $\widetilde{\mu}_{X,\widehat{f}}$ and $\widetilde{\mu}_{Y,\widehat{f}}$) conditional on $\widehat{f}$, we have

$$\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}-\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\,|\,\widehat{f}\big\} = \mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f}\big\} + \mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{Y,\widehat{f}}\,|\,\widehat{f}\big\} \quad \text{and}$$
$$\mathrm{Var}\big\{\widetilde{\mu}_{X,\widehat{f}}-\widetilde{\mu}_{Y,\widehat{f}}\,|\,\widehat{f}\big\} = \mathrm{Var}\big\{\widetilde{\mu}_{X,\widehat{f}}\,|\,\widehat{f}\big\} + \mathrm{Var}\big\{\widetilde{\mu}_{Y,\widehat{f}}\,|\,\widehat{f}\big\}.$$

Without loss of generality, we focus on $\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f}\big\}$ and $\mathrm{Var}\big\{\widetilde{\mu}_{X,\widehat{f}}\,|\,\widehat{f}\big\}$, and show that $\mathbb{E}\big[\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f}\big\}\big] \lesssim \mathbb{E}\big[\mathrm{Var}\big\{\widetilde{\mu}_{X,\widehat{f}}\,|\,\widehat{f}\big\}\big]$. The other terms can be handled similarly. Since the cross-fit estimator has smaller variance than the single-split estimator and a constant factor is not of interest, we may assume that $\widehat{\mathbb{E}}[\widehat{f}(X)\,|\,V,\widehat{f}] \coloneqq \widehat{u}_X(V)$ is trained on an auxiliary dataset. Under this simplification, another application of the law of total variance gives

$$\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f}\big\} = \mathbb{E}\big[\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f},\widehat{u}_X\big\}\,|\,\widehat{f}\big] + \mathrm{Var}\big[\underbrace{\mathbb{E}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f},\widehat{u}_X\big\}}_{=\mathbb{E}\{\widetilde{\mu}_{X,\widehat{f}}\,|\,\widehat{f}\}}\,|\,\widehat{f}\big]$$

and thus we focus on the expectation of the conditional variance above. A direct calculation yields

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{\mu}^{\dagger}_{X,\widehat{f}}\,|\,\widehat{f},\widehat{u}_X\big\}\,|\,\widehat{f}\big] = \frac{1}{n_1}\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\} &+ \frac{m_1}{n_1(n_1+m_1)}\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\,|\,\widehat{f}\big]\\ &- \frac{2m_1}{n_1(n_1+m_1)}\mathbb{E}\big[\mathrm{Cov}\big\{\widehat{f}(X),\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\,|\,\widehat{f}\big].\end{aligned}$$

Now, in order to prove the second inequality (21), we need to ensure that

$$\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\big] - 2\mathbb{E}\big[\mathrm{Cov}\big\{\widehat{f}(X),\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\big] \lesssim \mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big].$$

This follows from the condition (9) as we have

$$\begin{aligned}&\mathbb{E}\big[\mathrm{Var}\big\{\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\} - 2\mathrm{Cov}\big\{\widehat{f}(X),\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\big]\\ &= \mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)-\widehat{u}_X(V)\,|\,\widehat{f},\widehat{u}_X\big\}\big] - \mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big]\\ &\leq \mathbb{E}\big[\big\{\widehat{f}(X)-\widehat{u}_X(V)\big\}^2\big] - \mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big]\\ &\lesssim \mathbb{E}\big[\mathrm{Var}\big\{\widehat{f}(X)\,|\,\widehat{f}\big\}\big].\end{aligned}$$

Thus, the proof of Theorem D.1 is complete.

F.6 Proof of Corollary D.2

To prove this result, we will show that in the case of fixed alternative with $P_{X_n} = P_X$ and $P_{Y_n} = P_Y$ with $\mathrm{MMD}(P_X, P_Y) = \gamma > 0$ for all $n \geq 1$, we have

$$\max\left\{\mathbb{E}_{P_X}[k(X_1, X_1)],\ \mathbb{E}_{P_Y}[k(Y_1, Y_1)]\right\} < \infty \quad \Longrightarrow \quad \text{Equation (8).} \tag{22}$$

This fact, along with Theorem D.1 implies Corollary D.2. The proof of the above implication is essentially the same as that of Shekhar et al. (2022, Theorem 7). However, we include the details for completeness and also because the fourth moment condition required by Shekhar et al. (2022, Theorem 7) is unnecessary.

Since $\gamma > 0$ for all n, the conditions in (8) are equivalent to the following (we drop the subscripts $P_{X,n}$ and $P_{Y,n}$ from $\mathbb{E}$ and Var to simplify the notation)

$$\lim_{n_1\to\infty} \mathbb{E}\left[\widetilde{\sigma}^2_{X,\widehat{f}}\right] = 0, \quad \lim_{n_2\to\infty} \mathbb{E}\left[\widetilde{\sigma}^2_{Y,\widehat{f}}\right] = 0, \quad \text{and} \quad \lim_{n_1\wedge n_2\to\infty} \mathrm{Var}\left[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\right] = 0.$$

We will first show that $\lim_{n_1\to\infty}\mathbb{E}[\widetilde{\sigma}^2_{X,\widehat{f}}] = 0$. To do this, we introduce the notation $\widehat{g}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} k(X_i, \cdot)$, and observe the following:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\sigma}^2_{X,\widehat{f}}\right] &= \mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}(\widehat{f}(X_i) - \widetilde{\mu}_{X,f})^2\right] = \mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}(\langle k(X_i,\cdot), \widehat{f}\rangle_{\mathcal{H}_k} - \langle \widehat{g}_X,\ \widehat{f}\rangle_{\mathcal{H}_k})^2\right] \\
&= \frac{1}{n_1}\mathbb{E}\left[\left(\langle k(X_1,\cdot) - \widehat{g}_X,\ \widehat{f}\rangle_{\mathcal{H}_k}\right)^2\right] \stackrel{(i)}{\leq} \frac{1}{n_1}\mathbb{E}\left[\|k(X_1,\cdot) - \widehat{g}_X\|^2_{\mathcal{H}_k}\|\widehat{f}\|^2_{\mathcal{H}_k}\right] \\
&\stackrel{(ii)}{\lesssim} \frac{1}{n_1}\mathbb{E}\left[\|k(X_1,\cdot)\|^2_{\mathcal{H}_k} + \|\widehat{g}_X\|^2_{\mathcal{H}_k}\right]\mathbb{E}\left[\|\widehat{f}\|^2_{\mathcal{H}_k}\right].
\end{aligned} \tag{23}$$

Here (i) uses the Cauchy–Schwarz inequality, and (ii) follows from the independence of $k(X_1,\cdot) - \widehat{g}_X$ and $\widehat{f}$. Now, $\|k(X_1,\cdot)\|^2_{\mathcal{H}_k} = k(X_1, X_1)$ by the reproducing property, and $\|\widehat{g}_X\|^2_{\mathcal{H}_k} \leq \frac{1}{n_1}\sum_{i=1}^{n_1} k(X_i, X_i)$ by Jensen's inequality and the convexity of $\|\cdot\|^2_{\mathcal{H}_k}$. Thus, $\mathbb{E}[\|k(X_1,\cdot)\|^2_{\mathcal{H}_k}] + \mathbb{E}[\|\widehat{g}_X\|^2_{\mathcal{H}_k}] \lesssim \mathbb{E}[k(X_1, X_1)]$. Now, consider the term $\|\widehat{f}\|^2_{\mathcal{H}_k}$, and observe that

$$\begin{aligned}
\|\widehat{f}\|^2_{\mathcal{H}_k} &= \left\|\frac{1}{n_1}\sum_{i=1}^{n_1} k(\widetilde{X}_i,\cdot) - \frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i,\cdot)\right\|^2_{\mathcal{H}_k} \lesssim \left\|\frac{1}{n_1}\sum_{i=1}^{n_1} k(\widetilde{X}_i,\cdot)\right\|^2_{\mathcal{H}_k} + \left\|\frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i,\cdot)\right\|^2_{\mathcal{H}_k} \\
&\leq \frac{1}{n_1}\sum_{i=1}^{n} k(\widetilde{X}_i, \widetilde{X}_i) + \frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i, \widetilde{Y}_i),
\end{aligned}$$

where the last inequality again uses Jensen's inequality along with the convexity of the mapping $x \mapsto \|x\|^2_{\mathcal{H}_k}$. This implies $\mathbb{E}[\|\widehat{f}\|^2_{\mathcal{H}_k}] \lesssim \mathbb{E}[k(X_1, X_1)] + \mathbb{E}[k(Y_1, Y_1)]$. Plugging this back into (23), we get

$$\mathbb{E}\left[\widetilde{\sigma}^2_{X,\widehat{f}}\right] \lesssim \frac{1}{n_1}\left(\mathbb{E}[k(X_1, X_1)]\right)\left(\mathbb{E}[k(X_1, X_1)] + \mathbb{E}[k(Y_1, Y_1)]\right). \tag{24}$$

Thus, under the assumption that $\mathbb{E}[k(X_1, X_1)] < \infty$ and $\mathbb{E}[k(Y_1, Y_1)] < \infty$, the above inequality implies that $\lim_{n_1\to\infty}\mathbb{E}[\widetilde{\sigma}^2_{X,\widehat{f}}] = 0$. An exactly analogous argument implies that $\lim_{n_2\to\infty}\mathbb{E}[\widetilde{\sigma}^2_{Y,\widehat{f}}] = 0$.

We now show that $\lim_{n_1\wedge n_2\to\infty}\mathrm{Var}\big[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\big] = 0$. To show this, we introduce some new notation: let $\mu_X = \mathbb{E}[k(X,\cdot)]$ and $\mu_Y = \mathbb{E}[k(Y,\cdot)]$ denote the kernel mean embeddings corresponding to P_X and P_Y, and let $\widehat{g}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} k(Y_i,\cdot)$. Then, we have the following:

$$\begin{aligned}
\mathrm{Var}\left[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\right] &= \mathbb{E}\left[\left(\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}} - \langle \mu_X - \mu_Y, \mu_X - \mu_Y - \widehat{f} + \widehat{f}\rangle_{\mathcal{H}_k}\right)^2\right] \\
&= \mathbb{E}\left[\left(\langle \widehat{g}_X - \mu_X, \widehat{f}\rangle_{\mathcal{H}_k} - \langle \widehat{g}_Y - \mu_Y, \widehat{f}\rangle_{\mathcal{H}_k} + \langle \widehat{f} - (\mu_X - \mu_Y), \mu_X - \mu_Y\rangle_{\mathcal{H}_k}\right)^2\right] \\
&\lesssim \mathbb{E}\left[\langle \widehat{g}_X - \mu_X, \widehat{f}\rangle^2_{\mathcal{H}_k}\right] + \mathbb{E}\left[\langle \widehat{g}_Y - \mu_Y, \widehat{f}\rangle^2_{\mathcal{H}_k}\right] + \mathbb{E}\left[\langle \widehat{f} - (\mu_X - \mu_Y), \mu_X - \mu_Y\rangle^2_{\mathcal{H}_k}\right].
\end{aligned}$$

Applying the Cauchy–Schwarz inequality on all the three terms, we get

$$\operatorname{Var}\left[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\right] \lesssim \mathbb{E}[\|\widehat{f}\|_{\mathcal{H}_k}^2]\left(\mathbb{E}[\|\widehat{g}_X - \mu_X\|_{\mathcal{H}_k}^2] + \mathbb{E}[\|\widehat{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2]\right) \\ + \|\mu_X - \mu_Y\|_{\mathcal{H}_k}^2 \mathbb{E}\left[\|\widehat{f} - (\mu_X - \mu_Y)\|_{\mathcal{H}_k}^2\right].$$

Now, we can break up $\widehat{f}$ into $\widetilde{g}_X - \widetilde{g}_Y$, with $\widetilde{g}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} k(\widetilde{X}_i, \cdot)$ and $\widetilde{g}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} k(\widetilde{Y}_i, \cdot)$, and get the following bound $\mathbb{E}\left[\|\widehat{f} - (\mu_X - \mu_Y)\|_{\mathcal{H}_k}^2\right] \lesssim \mathbb{E}\left[\|\widetilde{g}_X - \mu_X\|_{\mathcal{H}_k}^2\right] + \mathbb{E}\left[\|\widetilde{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2\right]$. To summarize, we have proved that

$$\operatorname{Var}\left[\widetilde{\mu}_{X,\widehat{f}} - \widetilde{\mu}_{Y,\widehat{f}}\right] \lesssim \mathbb{E}[\|\widehat{f}\|_{\mathcal{H}_k}^2]\left(\mathbb{E}[\|\widehat{g}_X - \mu_X\|_{\mathcal{H}_k}^2] + \mathbb{E}[\|\widehat{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2]\right) \\ + \gamma^2\left(\mathbb{E}[\|\widetilde{g}_X - \mu_X\|_{\mathcal{H}_k}^2] + \mathbb{E}[\|\widetilde{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2]\right).$$

We have already proved that $\mathbb{E}[\|\widehat{f}\|_{\mathcal{H}_k}^2] \lesssim \mathbb{E}[k(X_1, X_1)] + \mathbb{E}[k(Y_1, Y_1)] < \infty$ under the assumptions of this corollary. Thus, to complete the proof, we need to show that

$$\lim_{n_1\to\infty} \mathbb{E}[\|\widehat{g}_X - \mu_X\|_{\mathcal{H}_k}^2] = 0, \; \lim_{n_2\to\infty} \mathbb{E}[\|\widehat{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2], \; \lim_{n_1\to\infty} \mathbb{E}[\|\widetilde{g}_X - \mu_X\|_{\mathcal{H}_k}^2] = 0, \text{ and } \lim_{n_2\to\infty} \mathbb{E}[\|\widetilde{g}_Y - \mu_Y\|_{\mathcal{H}_k}^2]. \tag{25}$$

We present the details of the first of these four conditions, since the steps for proving the other three are exactly the same.

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_X - \mu_X\|_{\mathcal{H}_k}^2\right] &= \mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\langle k(X_i,\cdot) - \mu_X,\, k(X_j,\cdot) - \mu_X\rangle_{\mathcal{H}_k}\right] \\ &= \frac{1}{n_1^2}\left(\sum_{i=1}^{n_1}\mathbb{E}[\langle k(X_i,\cdot) - \mu_X, k(X_i,\cdot) - \mu_X\rangle_{\mathcal{H}_k}] + \sum_{i\neq j}\mathbb{E}\langle k(X_i,\cdot) - \mu_X, k(X_j,\cdot) - \mu_X\rangle_{\mathcal{H}_k}]\right).\end{aligned} \tag{26}$$

Observe that by the Cauchy–Schwarz inequality, we have $\mathbb{E}[\langle k(X_i,\cdot) - \mu_X, k(X_i,\cdot) - \mu_X\rangle_{\mathcal{H}_k}] \lesssim \mathbb{E}[k(X_1, X_1)] + \gamma^2$. Furthermore, for any $i \neq j$, we have $\mathbb{E}\langle k(X_i,\cdot) - \mu_X, k(X_j,\cdot) - \mu_X\rangle_{\mathcal{H}_k}] = 0$. Plugging these back into (26), we get

$$\lim_{n_1\to\infty} \mathbb{E}\left[\|\widehat{g}_X - \mu_X\|_{\mathcal{H}_k}^2\right] \leq \lim_{n_1\to\infty} \frac{1}{n_1}\left(\mathbb{E}[k(X_1, X_1)] + \gamma^2\right) = 0,$$

under the assumption that $\mathbb{E}[k(X_1, X_1)] < \infty$. The remaining three terms in (25) can also be shown to go to zero similarly. This completes the proof of Corollary D.2.

F.7 Proof of Corollary D.4

We show that using a linear operator, the condition (10) results in Assumption 3.2. Without loss of generality, we focus on the first convergence result of (5). Leveraging the linearity of the estimators of conditional expectations, combined with the spectral decomposition of the centered kernel $\overline{k}$, yields:

$$\mathbb{E}[\widehat{f}(X) \,|\, V, \widehat{f}] - \widehat{\mathbb{E}}[\widehat{f}(X) \,|\, V, \widehat{f}] = \sum_{i=1}^{\infty} \lambda_i(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})\{\mathbb{E}[\phi_i(X) \,|\, V] - \widehat{\mathbb{E}}[\phi_i(X) \,|\, V]\},$$

$$\operatorname{Var}[\widehat{f}(X) \,|\, \widehat{f}] = \sum_{i=1}^{\infty} \lambda_i^2(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2.$$

Write

$$a_i = \lambda_i(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y}), \; b_i = \overline{\phi}_{i,X} - \overline{\phi}_{i,Y} \quad \text{and} \quad \Delta_{X,i} = \mathbb{E}[\phi_i(X) \,|\, V] - \widehat{\mathbb{E}}[\phi_i(X) \,|\, V].$$

Then the first convergence condition in (5) is equivalent to

$$\frac{\mathbb{E}[\{\sum_{i=1}^{\infty} a_i \Delta_{X,i}\}^2 \mid (a_i)_{i=1}^{\infty}]}{\sum_{i=1}^{\infty} a_i^2} = o_P(1).$$

We decompose the above ratio into two terms as

$$\frac{\mathbb{E}[\{\sum_{i=1}^{\infty} a_i \Delta_{X,i}\}^2 \mid (a_i)_{i=1}^{\infty}]}{\sum_{i=1}^{\infty} a_i^2} = \frac{\sum_{i=1}^{\infty} a_i^2 \mathbb{E}[\Delta_{X,i}^2]}{\sum_{i=1}^{\infty} a_i^2} + \frac{\sum_{i \neq j} a_i a_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} a_i^2},$$

and show that each term converges to zero in probability. For the first term, we have

$$\frac{\sum_{i=1}^{\infty} a_i^2 \mathbb{E}[\Delta_{X,i}^2]}{\sum_{i=1}^{\infty} a_i^2} \leq \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2] \times \frac{\sum_{i=1}^{\infty} a_i^2}{\sum_{i=1}^{\infty} a_i^2} = \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2] = o_P(1),$$

where the last equality follows by $\sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2] = o(1)$.

Next we decompose the second term into

$$\frac{\sum_{i \neq j} a_i a_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} a_i^2} = \frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} a_i a_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2} \times \frac{n_1 + n_2}{n_1 n_2} \frac{\sum_{i=1}^{\infty} \lambda_i^2}{\sum_{i=1}^{\infty} a_i^2}. \tag{27}$$

We first show that the first term of (27) converges to zero in probability

$$\frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} a_i a_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2} = \frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} \lambda_i \lambda_j b_i b_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2} = o_P(1).$$

By including the constant function 1 as an eigenfunction corresponding to the eigenvalue zero, it can be shown that the expectation of the above expression is zero:

$$\mathbb{E}\left[\frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} \lambda_{X,i} \lambda_{X,j} b_i b_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2}\right] = \frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} \lambda_{X,i} \lambda_{X,j} \mathbb{E}[b_i b_j] \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2} = 0.$$

On the other hand, the variance satisfies

$$\begin{aligned}
\mathbb{E}\left[\left\{\frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} \lambda_i \lambda_j b_i b_j \mathbb{E}[\Delta_{X,i} \Delta_{X,j}]}{\sum_{i=1}^{\infty} \lambda_i^2}\right\}^2\right] &= \frac{n_1^2 n_2^2}{(n_1 + n_2)^2} \frac{\sum_{i \neq j} \lambda_i^2 \lambda_j^2 \mathbb{E}[b_i^2 b_j^2] \{\mathbb{E}[\Delta_{X,i} \Delta_{X,j}]\}^2}{(\sum_{i=1}^{\infty} \lambda_i^2)^2} \\
&\leq \frac{n_1^2 n_2^2}{(n_1 + n_2)^2} \frac{\sum_{i \neq j} \lambda_i^2 \lambda_j^2 \mathbb{E}[b_i^2 b_j^2]}{(\sum_{i=1}^{\infty} \lambda_i^2)^2} \times \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2].
\end{aligned}$$

The expectation of the product of b_i^2 and b_j^2 can be bounded as

$$\begin{aligned}
\mathbb{E}[b_i^2 b_j^2] &= \mathbb{E}[(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2 (\overline{\phi}_{j,X} - \overline{\phi}_{j,Y})^2] \\
&\lesssim \mathbb{E}[\overline{\phi}_{i,X}^2 \overline{\phi}_{j,X}^2] + \mathbb{E}[\overline{\phi}_{i,Y}^2 \overline{\phi}_{j,Y}^2] + \mathbb{E}[\overline{\phi}_{i,X}^2] \mathbb{E}[\overline{\phi}_{j,Y}^2] + \mathbb{E}[\overline{\phi}_{j,X}^2] \mathbb{E}[\overline{\phi}_{i,Y}^2] \\
&\lesssim \left(\frac{1}{n_1^3} + \frac{1}{n_2^3}\right) \mathbb{E}[\phi_i^2(X) \phi_j^2(X)] + \frac{1}{n_1^2} + \frac{1}{n_2^2} + \frac{1}{n_1 n_2}.
\end{aligned}$$

Thus the variance is bounded above by

$$\begin{aligned}
&\mathbb{E}\left[\left\{\frac{n_1 n_2}{n_1 + n_2} \frac{\sum_{i \neq j} \lambda_i \lambda_j b_i b_j \mathbb{E}[\Delta_i \Delta_j]}{\sum_{i=1}^{\infty} \lambda_i^2}\right\}^2\right] \\
\lesssim\ & \frac{n_1^2 n_2^2}{(n_1 + n_2)^2} \left(\frac{1}{n_1^3} + \frac{1}{n_2^3}\right) \frac{\sum_{i \neq j} \lambda_i^2 \lambda_j^2 \mathbb{E}[\phi_i^2(X) \phi_j^2(X)]}{(\sum_{i=1}^{\infty} \lambda_i^2)^2} \times \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2] \\
+\ & \frac{n_1^2 n_2^2}{(n_1 + n_2)^2} \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} + \frac{1}{n_1 n_2}\right) \frac{\sum_{i \neq j} \lambda_i^2 \lambda_j^2}{(\sum_{i=1}^{\infty} \lambda_i^2)^2} \times \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2] \\
\lesssim\ & \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{\sum_{i \neq j} \lambda_i^2 \lambda_j^2 \mathbb{E}[\phi_i^2(X) \phi_j^2(X)]}{(\sum_{i=1}^{\infty} \lambda_i^2)^2} + 1\right] \times \sup_{i \geq 1} \mathbb{E}[\Delta_{X,i}^2].
\end{aligned}$$

Moreover, under Assumption 3.1, we have

$$\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{\sum_{i\neq j}\lambda_i^2\lambda_j^2\mathbb{E}[\phi_i^2(X)\phi_j^2(X)]}{(\sum_{i=1}^\infty \lambda_i^2)^2} \le \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{\mathbb{E}[\overline{k}^2(X_1,X_3)\overline{k}^2(X_2,X_3)]}{\{\mathbb{E}[\overline{k}(X_1,X_2)]\}^2} \to 0,$$

which implies that the variance term converges to zero. Thus, the second term is also $o_P(1)$ by Chebyshev's inequality.

Next we show the second term of (27) is bounded in probability

$$\frac{n_1+n_2}{n_1n_2}\frac{\sum_{i=1}^\infty \lambda_i^2}{\sum_{i=1}^\infty a_i^2} = O_P(1).$$

As in Kim and Ramdas (2024, page 55), let $\widetilde{\lambda}_i = \lambda_i^2/\sum_{i'=1}^\infty \lambda_{i'}^2 \ge 0$, the reciprocal of the above is

$$\sum_{i=1}^\infty \widetilde{\lambda}_i \frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2,$$

whose expectation is

$$\sum_{i=1}^\infty \widetilde{\lambda}_i \frac{n_1n_2}{n_1+n_2}\mathbb{E}[(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2] = \sum_{i=1}^\infty \widetilde{\lambda}_i = 1.$$

On the other hand, its variance satisfies

$$\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i \frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2\right) \lesssim \frac{n_1^2n_2^2}{(n_1+n_2)^2}\left[\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i\overline{\phi}_{i,X}^2\right)+\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i\overline{\phi}_{i,Y}^2\right)\right]$$

Letting $g(x,y)=\mathbb{E}[\overline{k}(x,X)\overline{k}(y,X)]$, we may see

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i\overline{\phi}_{i,X}^2\right)+\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i\overline{\phi}_{i,Y}^2\right) &\lesssim \frac{\mathbb{E}[\overline{k}^2(X_1,X_3)\overline{k}^2(X_2,X_3)]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}\left(\frac{1}{n_1^3}+\frac{1}{n_2^3}\right)\\ &\quad + \frac{\mathbb{E}[g^2(X_1,X_2)]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right).\end{aligned}$$

Hence, we obtain the upper bound of the variance as

$$\begin{aligned}&\mathrm{Var}\left(\sum_{i=1}^\infty \widetilde{\lambda}_i \frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2\right)\\ \lesssim\ & \underbrace{\frac{\mathbb{E}[\overline{k}^2(X_1,X_3)\overline{k}^2(X_2,X_3)]}{\{\mathbb{E}[\overline{k}^2(X_1,X_2)]\}^2}\left(\frac{n_1+n_2}{n_1n_2}\right)}_{\to 0} + \underbrace{\frac{\sum_{i=1}^\infty \lambda_i^4}{\{\sum_{i=1}^\infty \lambda_i^2\}^2}}_{\le \tilde{\lambda}_1}.\end{aligned}$$

On the other hand,

$$\frac{\sum_{i=1}^\infty \lambda_i^2}{\sum_{i=1}^\infty \lambda_i^2\frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{i,X}-\overline{\phi}_{i,Y})^2} \le \frac{1}{\widetilde{\lambda}_1\frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{1,X}-\overline{\phi}_{1,Y})^2}.$$

Using this inequality, for any $t>0$,

$$\mathbb{P}(T\ge t) \le \mathbb{P}\left(\frac{n_1n_2}{n_1+n_2}(\overline{\phi}_{1,X}-\overline{\phi}_{1,Y})^2 \le \widetilde{\lambda}_1^{-1}t^{-1}\right).$$

Letting

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}(\overline{\phi}_{1,X}-\overline{\phi}_{1,Y}) = \sum_{i=1}^{n_1+n_2} Z_i,$$

where the summands are mutually independent, given as

$$Z_i = \begin{cases} \sqrt{\frac{n_2}{n_1(n_1+n_2)}}\phi_1(X_i) & \text{if } 1 \le i \le n_1, \\ \sqrt{\frac{n_1}{n_2(n_1+n_2)}}\phi_1(Y_{i-n_1}) & \text{if } n_1+1 \le i \le n_1+n_2, \end{cases}$$

we may expect that it converges to a normal distribution. In particular, using the Berry–Esseen bound for independent but not identically distributed summands (Berry, 1941), we have

$$\begin{aligned} &\mathbb{P}\biggl(\frac{n_1 n_2}{n_1+n_2}(\overline{\phi}_{1,X} - \overline{\phi}_{1,Y})^2 \le \widetilde{\lambda}_1^{-1} t^{-1}\biggr) \\ \lesssim\ &\mathbb{P}(\xi^2 \le \widetilde{\lambda}^{-1} t^{-1}) + \frac{\mathbb{E}[|Z_1|^3]}{\sqrt{n_1}\{\mathrm{Var}[Z_1]\}^{3/2}} + \frac{\mathbb{E}[|Z_{n_1+1}|^3]}{\sqrt{n_2}\{\mathrm{Var}[Z_{n_1+1}]\}} \\ \lesssim\ &\mathbb{P}(\xi^2 \le \widetilde{\lambda}^{-1} t^{-1}) + \mathbb{E}[|\phi_1(X)|^3]\sqrt{\frac{1}{n_1^2} + \frac{1}{n_2^2}} \end{aligned} \tag{28}$$

where $\xi \sim N(0,1)$. Using the fact that

$$\begin{aligned} &\widetilde{\lambda}_1^2 \mathbb{E}[\phi_1^4(X)]\biggl(\frac{1}{n_1} + \frac{1}{n_2}\biggr) \to 0 \quad \text{and} \\ &\mathbb{E}[|\phi_1(X)|^3] = \mathbb{E}[|\phi_1(X)||\phi_1(X)|^2] \le \{\mathbb{E}[\phi_1^2(X)]\}^{1/2}\{\mathbb{E}[\phi_1^4(X)]\}^{1/2} = \{\mathbb{E}[\phi_1^4(X)]\}^{1/2}, \end{aligned}$$

we observe the convergence rate of the upper bound of the second term of (28)

$$\mathbb{E}[|\phi_1(X)|^3]\sqrt{\frac{1}{n_1^2} + \frac{1}{n_2^2}} \le \sqrt{\mathbb{E}[|\phi_1(X)|^4]\biggl(\frac{1}{n_1^2} + \frac{1}{n_2^2}\biggr)} \le \sqrt{\mathbb{E}[|\phi_1(X)|^4]\biggl(\frac{1}{n_1} + \frac{1}{n_2}\biggr)} = o(\widetilde{\lambda}_1^{-1}).$$

We then follow the proof of Kim and Ramdas (2024, Theorem 4.2) and show that

$$\frac{n_1+n_2}{n_1 n_2}\frac{\sum_{i=1}^{\infty}\lambda_i^2}{\sum_{i=1}^{\infty}\lambda_i^2(\overline{\phi}_{i,X} - \overline{\phi}_{i,Y})^2} = O_P(1).$$

Consequently, the first convergence condition in (5) is satisfied. A similar argument applies to Y and W, demonstrating that the second convergence condition is also satisfied. Thus, we establish that if condition (10) and Assumption 3.1 hold, then Assumption 3.2 is valid as claimed.

F.8 Proof of Theorem D.6

Using the test statistic, power function could be written as

$$\begin{aligned} \mathbb{P}(\mathrm{xss}\widehat{\mathrm{MMD}}^2 > z_{1-\alpha}) &= \mathbb{P}\Biggl(\frac{\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \widehat{\mu}^{\dagger}_{Y,\widehat{f}}}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}} > z_{1-\alpha}\Biggr) \\ &= \mathbb{P}\Biggl(\frac{\bigl(\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \mathbb{E}[\widehat{f}(X)|\widehat{f}]\bigr) - \bigl(\widehat{\mu}^{\dagger}_{Y,\widehat{f}} - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]\bigr)}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}} > z_{1-\alpha} - \frac{\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}\Biggr) \end{aligned}$$

Assume that we have bilinear kernel $k(x,y) = x^{\top}y$. We first show

$$T = \frac{\bigl(\widehat{\mu}^{\dagger}_{X,\widehat{f}} - \mathbb{E}[\widehat{f}(X)|\widehat{f}]\bigr) - \bigl(\widehat{\mu}^{\dagger}_{Y,\widehat{f}} - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]\bigr)}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}$$

follows asymptotically normal distribution. It suffices to show that the above satisfies Assumption 3.1 and Assumption 3.3.

From the Gaussianity of the data and the independence of X_1, X_2, X_3, we compute the key moments of the centralized kernel:

$$\begin{aligned}
&\mathbb{E}[\overline{g}_X(X,X)] = \mathbb{E}[\overline{k}(X,X_1)^2] = \mathrm{tr}(\Sigma_{11}^2) \asymp d,\\
&\mathbb{E}[\overline{k}(X_1,X_2)^4] = 2\mathrm{tr}(\Sigma_{11}^4) + (\mathrm{tr}(\Sigma_{11}^2))^2 \asymp d^2,\\
&\mathbb{E}[\overline{k}(X_1,X_2)^2\overline{k}(X_1,X_3)^2] = 2\mathrm{tr}(\Sigma_{11}^4) + 2(\mathrm{tr}(\Sigma_{11}^2))^2 \asymp d^2,
\end{aligned}$$

since the eigenvalues of Σ_{11} are bounded from Assumption D.5.

We substitute these results into the following term

$$\frac{\mathbb{E}[\overline{k}(X_1,X_2)^4] + n_1\mathbb{E}[\overline{k}(X_1,X_2)^2\overline{k}(X_1,X_3)^2]}{n_1^2\{\mathbb{E}[\overline{g}_X(X,X)]\}^2} \asymp \frac{1}{n^2} + \frac{1}{n} = o_P(1),$$

which implies that Assumption 3.1 is satisfied.

Similarly, we compute other moments:

$$\begin{aligned}
&\mathrm{MMD}(P_X,P_Y)^2 = 2\mathrm{tr}(\Sigma_{11}) \asymp d,\\
&\mathbb{E}[\overline{g}_X(Y_1,Y_2)] = \mathbb{E}[\overline{k}(X,Y_1)\overline{k}(X,Y_2)] = \mathrm{tr}(\Sigma_{11}^2) \asymp d.
\end{aligned}$$

Combining these results with the previous computation, we obtain the following convergence

$$\frac{\mathrm{MMD}(P_X,P_Y)^4 \times \mathbb{E}[\overline{k}_X(X,X)^2]}{\{n_1\mathbb{E}[\overline{g}_X(X,X)] + n_1^2\mathbb{E}[\overline{g}_X(Y_1,Y_2)]\}^2} \asymp \frac{d^4}{(nd^2+n^2d)^2} = \frac{\tau^2}{\tau^2+2\tau+1}\frac{1}{n^2} = o_P(1),$$

which shows that Assumption 3.3 holds.

Therefore, together with Assumption 3.2, we conclude that the test statistic T follows asymptotically $N(0,1)$.

Next, we analyze asymptotic behavior of

$$\frac{\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}.$$

Since Assumption 3.1, 3.2, and 3.3 holds, we know that

$$\frac{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}} \xrightarrow{P} 1, \quad \text{and} \quad \frac{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}} \xrightarrow{P} 1,$$

from the proof of Step 2 and Step 3, claim (ii′) in Theorem 3.4, respectively. From these facts, it follows that

$$\begin{aligned}
\frac{\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}} &= \frac{\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}} \times \frac{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}}{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}} \times \frac{\sqrt{\widehat{\sigma}^2_{X,\widehat{f}} + \widehat{\sigma}^2_{Y,\widehat{f}}}}{\sqrt{\widehat{\sigma}^{\dagger 2}_{X,\widehat{f}} + \widehat{\sigma}^{\dagger 2}_{Y,\widehat{f}}}}\\
&= \frac{\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}]}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}} \times \{1 + o_P(1)\}.
\end{aligned}$$

Note that with normal distribution assumption, $\overline{X} - \overline{Y} \sim N(\mu_X - \mu_Y, 2n^{-1}\Sigma_{11})$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. We compute

$$\begin{aligned}
\mathbb{E}[\widehat{f}(X)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y)|\widehat{f}] &= (\overline{X} - \overline{Y})^\top(\mu_X - \mu_Y)\\
&= (\mu_X - \mu_Y)^\top(\mu_X - \mu_Y) + O_P\big(\sqrt{(\mu_X - \mu_Y)^\top\Sigma(\mu_X - \mu_Y)}\big).
\end{aligned}$$

On the other hand, we compute $\sigma^2_{X,\widehat{f}}$:

$$
\begin{aligned}
\sigma^2_{X,\widehat{f}} =& \frac{1}{n}\mathbb{E}[\mathrm{Var}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] + \frac{1}{2n}\mathrm{Var}[\mathbb{E}\{\widehat{f}(X)\,|\,V,\widehat{f}\}\,|\,\widehat{f}] \\
=& \ \frac{1}{n}\mathbb{E}[(\overline{X}-\overline{Y})^\top \mathrm{Var}(X\,|\,V)(\overline{X}-\overline{Y})] + \frac{1}{2n}\mathrm{Var}((\overline{X}-\overline{Y})^\top \mathbb{E}[X\,|\,V]) \\
=& \ (\overline{X}-\overline{Y})^\top \Lambda(\overline{X}-\overline{Y}) \\
=& \ 2n^{-1}\mathrm{tr}(\Lambda\Sigma_{11}) + (\mu_X-\mu_Y)^\top\Lambda(\mu_X-\mu_Y) \\
& \qquad + O_P\Big(\sqrt{8n^{-2}\mathrm{tr}(\Lambda\Sigma_{11}\Lambda\Sigma_{11}) + 8n^{-1}(\mu_X-\mu_Y)^\top\Lambda\Sigma_{11}\Lambda(\mu_X-\mu_Y)}\Big)
\end{aligned}
$$

where $\Lambda = n^{-1}\Sigma_{11} - (2n)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Assuming that $(X,V)^\top$ and $(V,W)^\top$ have equal sample sizes and equal covariancce matrices, a similar computation derives the same result for $\sigma^2_{Y,\widehat{f}}$:

$$
\begin{aligned}
\sigma^2_{Y,\widehat{f}} =& \ 2n^{-1}\mathrm{tr}(\Lambda\Sigma_{11}) + (\mu_X-\mu_Y)^\top\Lambda(\mu_X-\mu_Y) \\
& \qquad + O_P\Big(\sqrt{8n^{-2}\mathrm{tr}(\Lambda\Sigma_{11}\Lambda\Sigma_{11}) + 8n^{-1}(\mu_X-\mu_Y)^\top\Lambda\Sigma_{11}\Lambda(\mu_X-\mu_Y)}\Big).
\end{aligned}
$$

Hence we obtain

$$
\begin{aligned}
\frac{\mathbb{E}[\widehat{f}(X_i)|\widehat{f}] - \mathbb{E}[\widehat{f}(Y_i)|\widehat{f}]}{\sqrt{\sigma^2_{X,\widehat{f}} + \sigma^2_{Y,\widehat{f}}}} &= \frac{(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y) + O_P(n^{-3/4})}{\sqrt{4n^{-1}\mathrm{tr}(\Lambda\Sigma_{11}) + O_P(n^{-3/2})}}\{1+o_P(1)\} \\
&= \frac{(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y)}{\sqrt{4n^{-1}\mathrm{tr}(\Lambda\Sigma_{11})}} + o_P(1),
\end{aligned}
$$

using the fact that $(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y) = O_P(n^{-1/2})$ and $\mathrm{tr}(\Lambda\Sigma_{11}) = O_P(1)$ from Assumption D.5.

Combining with the normal approximation, we conclude that

$$
\begin{aligned}
\mathbb{P}(T^\star > z_{1-\alpha}) &= \Phi\left(z_\alpha + \frac{(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y)}{\sqrt{4n^{-1}\mathrm{tr}(\Lambda\Sigma_{11})}}\right) + o(1) \\
&= \Phi\left(z_\alpha + \frac{n(\mu_X-\mu_Y)^\top(\mu_X-\mu_Y)}{\sqrt{4\mathrm{tr}(\Sigma_{11}^2) - 2\mathrm{tr}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11})}}\right) + o(1).
\end{aligned}
$$