UAI 2025 Xunye Rebuttal

Reviewer faca (Rating 4, Confidence 2)

Thanks for your valuable time to provide insightful feedbacks, and we will respond to the following concerns that worries you accordingly.

Weakness 1:

> *Can the learned representation reflect the original distribution distance? The paper uses nonlinear feature maps (like autoencoders) that may change the data distribution, and while it argues this does not affect Type-I error under permutation tests, it doesn't formally characterize the induced distributions in feature space.*

Response 1:

Since we are conducting the permutation test on the learned representations, it can **guarantee the validity of the Type-I error rate under exchangeability conditions**. Our autoencoder-based feature map is learned in an unsupervised manner, independent of the sample labels. Hence, the learned nonlinear transformation applies symmetrically to both samples. This symmetry preserves the exchangeability required for the permutation test, ensuring valid control of Type-I error regardless of nonlinear transformations.

Weakness 2:

> *Novelty. The proposed RL-TST framework builds upon MMD-D and classifier-based two-sample testing methods. The main novelty appears to be the use of features learned from an autoencoder (or its variants) as an intermediate representation for testing. Could the authors elaborate on the theoretical contributions beyond prior work, and clarify whether the empirical improvements stem primarily from this architectural change or from other practical innovations?*

Response 2:

Motivation: Our main novelty lies in the proposed **general framework** (RL-TST), which flexibly integrates inherent representations (IRs, learned unsupervised from unlabeled test data) and discriminative representations (DRs), surpassing limitations of existing purely unsupervised or purely supervised representation learning approaches. Additionally, we highlight that the **framework itself is flexible**: beyond autoencoders, other unsupervised or semi-supervised representation learning approaches could be utilized (in Table 2), as long as the data strictly follow the assumptions of methods (e.g., smoothness or cluster assumptions in semi-supervised learning).

Theoretical contribution: Theoretically, our approach **reduces the effective hypothesis space** for feature extraction in learning DR, enhancing the likelihood of finding optimal discriminative representations. Although the it is a black-box framework and the specific Type-I error control formally depends on downstream DR learning methods, our framework **does not negatively affect the exchangeability**.

Weakness 3:

> *The method relies on the multi-stage training, which makes the approach challenging to scale up to larger dataset.*

Response 3:

We acknowledge the reviewer's concern that our proposed RL-TST framework requires multiple stages (learning inherent representations, followed by discriminative feature learning). However, the extra IR pre-training step is **computationally manageable**, since it is performed in an unsupervised manner and can be efficiently executed offline. Moreover, By first pre-training inherent representations, the subsequent discriminative representation learning step usually converges **faster and more robustly**. Thus, the extra computational effort in the initial IR step is partially offset by faster convergence and reduced training epochs needed in later stages. In practical, fine-tuning a pre-trained encoder in IR learning step would have higher performance and efficiency in subsequent DR learning.

Weakness 4:

> *The paper focuses on controlling Type-I error (false positives) using permutation testing. But it does not analyze Type-II error (false negatives), nor does it explore conditions under which RL-TST fails to detect distribution differences, especially under misspecified manifolds or noisy data.*

Response 4:

We sincerely thank the reviewer for raising this point. We want to clarify that we indeed analyze and report **Type-II error** in our experiments. As in the context of two-sample testing, the test power is defined as (1 - Type-II error). We will explicitly clarify this equivalence in the manuscript.

Moreover, we emphasize the standard theoretical result in non-parametric kernel two-sample testing (**consistency**): Under mild assumptions (such as characteristic kernels for MMD or appropriately flexible classifiers), the test power converges to 1 as the sample size increases, formally ensuring:

$$\lim_{n\to\infty} \text{Power}(n) = 1, \text{given } \mathbb{P} \neq \mathbb{Q}.$$

Lastly, we agree that explicitly discussing potential conditions under which RL-TST may face challenges (e.g., misspecified manifolds, extreme noise levels) is valuable. We will clarify that our data assumptions follow the setting of only the mainstream non-parametric two-sample testing literatures [1,2,3,4,5,6].

[1]. Arthur Gretton, Karsten M Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. Journal of Machine Learning Research, 13:723–773, 2012a.
[2]. David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In ICLR, 2017.
[3]. Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In ICML, 2020
[4]. Antonin Schrab, Ilmun Kim, Melisande Albert, Beatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD Aggregated Two-Sample Test. Journal of Machine Learning Research, 2023.
[5]. Felix Biggs, Antonin Schrab, and Arthur Gretton. MMDFUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting. In NeurIPS, 2023.

Reviewer Qgpm (Rating 5, Confidence 3)

Weakness 1:

> *I did not see any discussion on Type I error control. How can you ensure that the proposed method maintains valid Type I error rates? From the simulation results, the Type I error appears to be somewhat inflated compared to existing methods, especially for RL-C2ST-L when N = 4000. It would be helpful for the authors to address this issue and provide more analysis or theoretical justification regarding the Type I error behavior.*

Response 1:

Reviewer qT5G (Rating 6, Confidence 4)

Weakness 1:

> *There is an effort in providing a theoretical analysis based on existing results (not from the authors), although it is in the supplementary. The authors provide a uniform probabilistic control of the classification error. The arguments are based on generalization error for classification-based algorithms having finite VC-dimension, which is very classical. However, it is not clear to me whether this key assumption is fulfilled in the present framework. I would suggest to add a remark on that matter.*

Response 1:

Weakness 2:

> *Although I understand that the paper tends to a methodological contribution, I would be interested to see whether theoretical guarantees would possibly be proved, and in particular for probabilistic uniform control of the type-I error, and if that can be distribution-free (permutation test is computationally a drawback for this large datasets).*

Response 2:

Weakness 3:

> *Regarding the experiments, although the pipeline is diverse, I would suggest to add some sensitivity analysis (at least std) in the tables and graphs to fully compare the methods that have, in fact, very similar power (cf table 3). In addition, Fig 4b, is not vey useful as it is only at 0.05: it does not guarantee that the tests have level for all alphas. In addition, the empirical results are tested on low-dimensional data (d=2,10): this is misleading regarding the arguments of section 3. And small datasets, which would imply high empirical variance (that is not reported) and does not guarantee consistency of the procedure, in the particular as in relies on three steps.*

Response 3:

Weakness 4:

> *Lastly, the manifold assumption assumes a "sparse" structure of the data (low signal-to-noise ratio), is this the case in the experimental setting? I would be interested to see it clearly written in the disccusion. In particular as some methods are known to perform well only in one of the two settings.*

Response 4:

Comment to author:

> *I would suggest to add a discussion on how to optimally choose the subsample sizes. (I know it is mentioned in the supplementary, but in light of my remark on the theoretical assumptions, it would require some further discussion).*
>
> *I would change section 3.2. as a remark, and I haven't seen any typo.*
>
> *Lastly, causal discovery mainly uses conditional independence testing (in the introduction paragraph 2).*
>
> *Some references are missing in the literature for two sample testing, e.g., Biau Gyorfi 2005, Deb and Sen 2019, Clemencon Limnios and Vacates (2023), Bach Moulines and Harchaoui (2008).*

Response to review:

Reviewer qofF (Rating 7, Confidence 4)

Weakness 1:

> *Discussion of alternatives in Section 3 and Section 3.2 is a bit redundant to me.*

Response 1:

Weakness 2:

> *It will be great if the authors can explicitly point out that the SSL assume the labels are noiseless, while the alternative hypothesis in the two-sample testing typically indicate the labels are noisy.*

Response 2: