We would like to thank the meta reviewer and four reviewers for the care with which you handle the submission and for your professional and constructive comments. We have made every effort to address the concerns.

Response to Reviewer 1's comments

1 **Weaknesses**: (1) To the best of our knowledge, our work is the first approach that can achieve a "full SDR mean subspace" instead of a partial SDR mean subspace obtained by alternative approaches. Theorem 1 provides a theoretical justification for the success of our method's empirical performance, as all the alternatives can only recover part of the SDR mean subspace, which may be far away from the true one. (2) The classification problem in Experiment 3 is a favorable case for dimension reduction; thus, it warrants the asymptotic convergence of all dimension reduction methods. We use such a problem to demonstrate that even for such a simple setup, our method can work as well as any other methods. Furthermore, as demonstrated in the simulation that does not favor conventional dimension reduction approaches, our method significantly outperforms all the other approaches. (3) We will add a comparison of CPU time in the revision. (4) We extended the proposed method for binary response cases (illustrated in line 152-157), of which the consistency of our SDR estimates is ensured in Theorem 1, to multi-class response cases, detailed in Algorithm 1. (5) The $\Lambda \in \mathbb{R}^{nk \times d}$ in Algorithm 1 is used to generate the displacement matrix of OTP for multi-class response cases using one-vs-rest strategy. We found using the one-vs-one strategy yields similar results. Theorem 1 ensures that the right singular vector of Λ is consistent for the SDR subspace. We will provide more discussion of theoretical properties of Λ. (6) Both [41] and our method aims to optimize some objective functions with respect to the displacement matrix. However, [41] focuses on finding a surrogate of displacement matrix for OT such that the estimate of Wasserstein distance is robust, while our method aims to find the SDR subspace using OT. Thus, the theoretical results of our work and [41] rooted in completely different lands.

2 **Additional feedback**: (1) The Wasserstein dictionary learning is an unsupervised approach; however, our method is a supervised one. They are not directly related. (2) The q is the true dimensionality of SDR space, while r is the estimated dimensionality. If $Y \perp\!\!\!\perp X|Y^*$, we can show $q = r$, otherwise $q \geq r$. Thus, they are not exchangeable. (3) Slicing in our method is not used to estimate OT but the SDR. Thus, these two approaches are not related. (4) The time complexities will be added. (5) We agree that that the effect of the regulation parameter may play an essential role in OT estimate. However, in our paper, we used the EMD method that does not include a regulation parameter. We may include this type of study in our future publications. (6) We will cite the Brenier theorem that demonstrates the equivalence of Monge and the Kantorovich formulations under certain conditions, which are not required for our work. (7) Max-pooling is used only for numerical convenience. The accuracy will not change significantly without down-sampling. (8) Yes. These two estimates are the same.

Response to Reviewer 2's comments

3 **Strengths**: We highly appreciate the insightful comments and future-research suggestions.

4 **Weaknesses**: (1) We apologize for the confusion. Our theorem only require $\hat{\phi}^*$ in line 183 to be a consistent estimator for OT. The existence of the Monge map is not required. We have revised the manuscript accordingly. (2) We will include more detailed discussion and corresponding references of the OT convergence rate in the revision. (3-4) Suggestions will be taken.

5 **Correctness**: We used the EMD method in python `POT` package instead of Sinkhorn to calculate the OTP. Thus, the results are reproducible. More details will be provided.

6 **Reproducibility**: We will upload codes with detailed comments to NeurIPS and GitHub.

7 **Additional feedback**: Typos will be corrected. Suggestions will be taken.

Response to Reviewer 3's comments

8 **Weaknesses**: (1) Please see 1:(1) and 1:(4)–(5) for your weakness concerns.

9 **Clarity and Additional feedback**: (1) All results are not susceptible to the order of the transport or label switching. (2) Fig.2 illustrates how our proposed method outperformed the conventional moment-based SDR approaches, which estimate the space that is sufficient for conditional moments of $X|Y$. For example, SIR estimates SDR using the first conditional moments while SAVE estimates SDR using the second conditional moment. More discussion will be included in the revised manuscript. (3) Typos have been corrected.

Response to Reviewer 4's comments

10 **Weaknesses**: (1) Suggestion will be taken. (2) (H.1)–(H.2) in Theorem 1 are necessary conditions for the consistency of the estimated SDR subspace. Violations of the conditions will forfeit the consistency of any SDR methods. For example, if the SDR subspace is not unique as required by (H.1)–(H.2), the true SDR subspace is not well defined, not to mention the consistency of the estimated SDR subspace. However, Conditions (H.3)–(H.5) are only sufficient conditions which ensure the convergence of the empirical OT.

11 **Correctness**: (1) When data are extremely imbalanced, (H.5) is violated because the number of observations respecting to different classes is not at the same order, in which case Theorem 2 will fail. However, this problem can be fixed by adding weight to each of the classes when calculating the SDR directions. Such a generalization is not trivial and is beyond the scope of the current manuscript. (2) Suggestions will be taken.

12 **Clarity**: (1) The typo will be fixed. (2) Please see 1:(4)–(5) for your concern. (3) We apologize for the low quality of the figures. High-resolution figures will be used in the revision.