We thank the reviewers for the detailed and positive feedback on our work. We address the specification clarifications requested by the reviewers below, with an explanation of how each of these will be addressed in the final paper. We believe that these clarifications will resolve all of the reviewer concerns, but would be happy to take any additional suggestions into account.

General response regarding experiments (R3 & R4): Our experiments are designed specifically to study the problem with mutual information estimation that we discuss in Sec 2.1 (see also discussion in Sec 5.3). Datasets with high mutual information make it difficult to learn complete representations with KL-divergence-based methods like CPC. Not all tasks have this issue, and we expect that tasks where this is not an issue would have similar performance for CPC and WPC. We will clarify this, explaining that we are not proposing a general improvement across all tasks, but a specific improvement that aims to address the issue that we motivate in Sec 2.1 and confirm experimentally in Fig 2 (bottom row).

R4: *Q: Difference in comparison of experimental setup vs. CPC? A:* As we state above, our experiments are chosen to study settings with high mutual information, as motivated in Sec 2.1. We illustrate that this problem happens in practice (see Fig 2, bottom row), and then show that WPC mitigates the problem. We had to design new tasks to allow us to vary the mutual information in the data. Because of this, our experimental setup is different from the CPC paper – we are not trying to show that WDM improves over CPC in all cases, but that it mitigates the particular problem motivated in the paper. We will clarify this in the final version.

R4: *Q: code? A:* We were unable to prepare the code for release in time for submission, but will release it soon.

R4: *Q: Evaluation is specialized on images. A:* We agree that the method is general, and not specific to images. However, it is very common for representation learning work to be evaluated on images. Images provide a number of challenges and considerable breadth, and there are many standard datasets, making the visual domain well-suited for comparing representation learning methods.

R4: *Q: Abstract's "real-world tasks"? A:* We will revise the abstract to replace "real-world tasks" with "a number of tasks with synthetic and realistic images." We would be happy to make other revisions that the reviewers might suggest.

R3: *Q: How is the problem of exponential sample size resolved? A:* We will clarify the discussion regarding exponential sample size. We do not have a formal result showing that the WDM actually resolves the exponential sample size issue fully. The exponential sample size discussion is meant to motivate seeking an alternative to the KL-divergence for representation learning, but the utility of our proposed WDM solution is verified empirically, with experiments conducted on a set of tasks that are intentionally selected to have high mutual information between the context and predicted variable. The experimental results show that WDM produces better performance under these conditions. We will clarify that the primary evidence for the efficacy of our method is empirical, rather than theoretical.

R3: *Q: top5 and top1 results on ImageNet? A:* Unfortunately, we did not have time to conduct this experiment during the rebuttal, though we will try to add it in the final. However, as we discuss above, the goal of our experiments is to evaluate settings with high mutual information in a controlled way, rather than show improvement over CPC in all cases.

R3: *Q: Directly estimating the WDM without a bound and without contrastive/negative samples? A:* We are not aware of any way to estimate Wasserstein distance exactly without employing a bound or a variational approximation for high dimensional inputs, though this would certainly be an improvement to the method. We will discuss this in the paper.

R3: *Q: Base metric, discussion of CPC? A:* We will revise the wording on line 116 to address your suggestion. We will also add a more complete summary of CPC to ensure that the discussion in the paper stands on its own.

R3: *Q: Summation over j in Equation 3. A:* This notation can definitely be improved. The outer j (in the expectation) and the inner one (in the log-sum-exp) are different, but running over the same set. The inner one is the one observed in the data, the outer one goes over all possible values. We will change the notation to clarify.

R3: *Q: Equations 2 and 3. A:* Both equations have negative samples (from the product of marginals), albeit their scores are accounted in a different way, either as a simple average (Eqn 2) or as a log-sum-exp. Will clarify in the text.

R2: *Q: Limitations, and when does WPC break down as compared to CPC? A:* In general, we did not observe settings where WPC "breaks down" as compared to CPC. This is likely because the difference in the actual methods is not very large: while the theoretical basis is quite different, implementation-wise WPC amounts to CPC augmented with an additional regularizer. The regularizer is meant to ensure that the function class is 1-Lipschitz. A variety of regularizers can be used to get this effect, though we use gradient penalty (GP). In principle, GP might also lead to underfitting if the regularizer has a large coefficient and the function class is too small, as with any other regularizer, but we did not observe this problem experimentally. We will endeavor to conduct additional experiments on larger datasets for the final version to identify other potential corner cases or limitations, and discuss the tradeoffs of introducing regularization in the discussion section.