We thank all the reviewers for their time, valuable and encouraging feedback and recommendations for improvement. In the following, we address their concerns and questions.

Motivations, intuitions, and formulation lead to better semantic coherence and downstream tasks (**R1, R2**): In short, we can justify the better semantic coherence of learned topics using underlying mechanism of word embeddings and optimizer in Eq. (5) which is $\beta_k = \arg\min_\beta \sum_i \frac{\pi_{ik}}{b_k} \sum_{u,v} \gamma^i_{uv} c_{uv}$. Our intuitive explanation is as follows. The word embeddings are resulted from learning co-occurring of words in documents, therefore when two words w_u and w_v appear more frequently in the corpus, their embeddings are more similar, i.e. c_{uv} is small. Our proposed model aims to optimize Eq. (5) which will put a higher value on γ^i_{uv}. As a consequence, the pair of w_u and w_v usually gets higher weight in topic β_k. When computing topic coherence, we usually choose top words with high weights, it is more like this pair of words to present in the top words list which may produce bigger the numerator of coherence formula. Another property that our proposed model process is clustering characteristic which means closer documents in terms of optimal transport (aka word mover distance - WMD) will have similar topic proportion vectors. We also knew that WMD provides good distance for documents in text classification in [1]. Our results in downstream classification task are orthogonal with their results.

Tuning the regularizer parameter lambda (**R1,R2**): In fact, we did **not** heavily tune the regularizer parameter. We ran the model with four settings of the regularizer parameter is to check the sensitivity of our model. On the contrary, for the baseline approaches, we chose the best-reported values in their papers which we think had already gone through a tuning procedure.

The advantage of considering word counts in our proposed model (**R1,R3**): When dealing with a varied document length corpus, thank to word counts consideration, our model can up-weight longer documents while down-weighting shorter ones. We have demonstrated that our model can handle short document datasets such as 20NGshort or Tweets better. Word count weighting also provides the connection between our proposed model and LDA.

Some minor suggestions, notations, and typos (**R1,R2, R3**): We appreciate your pointing and constructive suggestions. We will improve the manuscript with the suggestions.

R1: *Exploration of entropic regularization parameters*: There are different gamma entropic regularization parameters. As we mentioned we did not tune these parameters when comparing to baseline approaches. We leave the investigation of the effects the entropic regularization parameters as future work.

Not to compare against other Wasserstein-based approaches: There are two Wasserstein-based approaches which can solve the problem namely WDL and DWL. Unfortunately, the code of DWL is not publicly available, we are not able to compare with. One of our baselines is DWL which we called WNMF in the paper.

Times/complexity for this model: In comparison with WNMF, our model is much faster since using Sinkhorn-based algorithms to learn while WNMF runs Sinkhorn-based algorithms (forward) then compute the gradient to update the model (backward). In comparison with neural topic models like ETM, our model is slower since neural topic models use amortized variational inference to learn. We did not include running time since the code are implemented in different platforms or programming languages. For instance, WNMF is implemented with Matlab/C++, ETM is coded with PyTorch while our model is implemented with plain python using POT library.

L184: It is L_1 norm. We will clarify it.

R2: *Only 20-100 topics were used*: Choosing the number of topics in topic models is a challenging task which is not the main focus in our experiments. Our strategy for selecting the number of topics for each dataset follows existing work in the literature. Moreover, in practice, learned topics are usually inspected by a human for the use of visualization or understanding, it is impractical to deal with a very large number of topics.

I found the notation very confusing: We will clarify symbols and notations in the revised version. In particular, thanks for your suggestion to include a table to summarize the notations, we will implement that.

R3: *Datasets are not large*: We agree that the datasets we used are not considered as modern datasets. In this paper, we would like to demonstrate a novel tool to solve the topic modeling problem. Scaling up the current model to massive datasets is one of our future work.

Embeddings: We did mention the use of word2vec embeddings in our experiments in lines 257–258.

Qualitative results and UCI metric: Given space restriction, we had to make a choice to balance the theory and experimental results, we shall aim to improve the post-analysis results as well as more topics in the supplementary materials

[1] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.