We thank the reviewers for their in-depth reviews, and will use them to make the final version as clear as possible. Our code, which was included in our supplementary material submission, will be made publicly available for reproducibility. We will correct the missing notations and typos, we apologize for these oversights. Thank you for suggesting other related work: we will expand our discussion with these.

Re: discussion/comparison: We have tried to provide both experimental and analytical discussion about word-level topic modeling in conjunction with auto-regressive models without marginalizing out the topics. Our theorem shows that under specific conditions our model reduces to a recurrent LM and our experimental results show that our topic coherency outperforms variational LDA and Mallet LDA, covering both topic modeling and auto-regressive aspects.

We have reported perplexity per word consistent with prior work; we wil clarify the equation. We used a fixed max length of 90 words per doc, with one layer and 200 hidden units in the recurrent structure. For all the experiments the Dirichlet prior distribution parameter is 0.5. We will report these in the experiments section. As "RNN" can be used to describe a number of different specific recurrent cells, we will add experimental results where we specifically compare to a traditional RNN cell, an LSTM cell, and a GRU cell.

[R1]: We will reorganize the equations to make the analytic comparisons between our approach and previous work more clear. Thank you for this suggestion. In our model definition, we are not sampling the discrete latent variable from the posterior variational distribution and the discrete expectations are calculated in closed form. Without resorting to the reparametrization trick, in our theorem (page 6 on top), we have shown that our model reduces to a simple RNN just by assuming all the tokens are non-thematic words. We should note that our proof for this relies on the closed form calculation and not sampling. We aim to design a language model that can preserve this word-level topic information. The remarkable difference between our model and the TopicRNN paper is preservering this topic information. We have imposed the simple uniform probability assumption just for the non-thematic words. This mechanism not only helps the topic model distinguish between the thematic and non-thematic words, but also it leads to stable training, since in the topic model part the gradients for the non-thematic words are zero and just the RNN part would be updated.

[R2]: We have included both variational LDA and Mallet LDA results for the switchP part to show that VRTM can also be considered as a topic model in terms of topic coherency, we will add the results for LDA+LSTM in the final version. However, we note that in LDA+LSTM an LDA model is trained and then used in a recurrent LM; this is in contrast to our approach which allows both the topic & language model to be learned and updated jointly. We agree that examining the generative capabilities of these types of recurrent models is important, but we believe that doing rigorously and comprehensively needs its own study and is beyond the scope of this particular work. We provide output sentences as examples so readers may make their own qualitative assessments on the strengths and limitations of our methods.

[R3]: We reconsidered core decisions made by TopicRNN, such as not marginalizing out topics and the doc prior. The Dirichlet can *easily* be parameterized to generate sparse samples just by tuning the prior distribution parameters.

We will update our discussion to include textTOvec. We note that its topic modeling component h^{DN} does not marginalize over the topics in a traditional sense, as it passes topic parameters through an activation function, effectively compressing the topic signal prior to any word generation. While not precisely the same as previous efforts, this is in contrast to what we advocate, which is specifically conditioning a word's generation on a particular topic (and as topics aren't observed, marginalizing after any activation function).

Re: implementation/results: Although we have employed 400 dimensional input embeddings, our word embeddings are learned from scratch, which do not have the massive pretraining of other methods. Our early experiments showed that our model and results were consistent even with other sizes like 300; we can include these numbers in the camera ready. Our masked embedding is obtained by both masking the non-thematic words and multiplying the thematic word embedding to their frequency. Since the topic models are mostly trained based on the frequency of (thematic) words our aim is to define the embeddings in a way that (i) the topic model part neglects the non-thematic words, so the coefficient for these words is zero. (ii) the thematic words with higher frequencies are emphasized. By adding this coefficient the gradients flow will increase for the thematic words. We will clarify the notation. We have reported SwitchP rather than other topic modeling metrics is that it makes the very intuitive yet simple assumption that "good" topics will exhibit a type of inertia. We will update the paper with this motivation and add additional information to example output.

[R4]: We will add a diagram to clarify the encoding/decoding process. We view using the simpler mean field approximation as a benefit and are excited to explore more expressive approximations in future work. Although we have used the plain Bernoulli random variable in the generative process but in the joint probability definition the parameters of this Bernoulli are learned in an auto-regressive manner where the information from the previous tokens play the role to draw a thematic or non-thematic word. "Neural Variational Inference" is from [25]. We will clarify this point & apologize for any confusion. Re: $q(z_t)$ and $p(z_t)$: For both p and q we have assumed the non-thematic words have the uniform distribution. This construction motivates the topic model part to distinguish between thematic and non-thematic words, and is used in the proof of the theorem.