We thank the reviewers for their valuable comments and suggestions. We first respond to all: 1) We set the same network structure for all models in supervised experiments: 200, 200-100, and 200-100-50 for one-, two-, and three-layer models, respectively. 2) We fix the hyperparameters of our models as $\tau_0 = 1, \epsilon_0 = 0.1, \gamma_0 = 0.1, \eta = 0.05$ for all experiments; the performance is not sensitive to these hyperparameters, an usual advantage of hierarchical Bayesian models.

Methods	ELEC	
	Accuracy	NMI
PGBN	71.4	60.8
CPGBN	77.8	65.1
CPGDS	78.6	66.2

Table 1: Clustering performance comparison.

To R1: Thank you for your positive feedback, which really encourages us to continue our efforts along this promising direction! **To R2 & R4:** 1) To verify the efficiency of our generative model, additional tasks on document clustering and sentence and document likelihood evaluations have been included. Following Cai et al. (TPAMI 2011), we use accuracy and NMI to evaluate document clustering performance, as shown in Table 1 (we only include dataset ELEC given space constraint; we will add more methods on more datasets), which further verifies the advantages of CPGDS. We estimate the likelihood of sentence with shuffled word order. Fig. 1 (left) shows the likelihood decreases as the shuffling rate increases, indicating CPGDS provides a higher confidence on real sentences than orderless ones. We further estimate the likelihood of document with shuffled sentence order and observe similar behaviors in Fig. 1 (right).

To R2: Example sentences in Fig. 4 are provided to illustrate the point that introducing the relationships between different sentences can help improve the accuracy of the sentiment-level judgments for the whole document, which confirms our motivations.

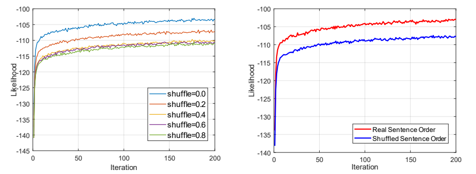

Figure 1: left: likelihood of shuffled sentence; right: likelihood of shuffled document

To R3: 1) Combining CFPA and PGDS into a coherent statistical model requires addressing several technical challenges, such as how to handle variable sentence lengths, avoid cutting off backward message passing, and speed up Gibbs sampling. 2) First, we have compared our model with a wide variety of topic models and unsupervised generative models in unsupervised experiments. To the best of our knowledge, except for DocNADE that is already included for comparison, there are few deep NN based probabilistic models for unsupervised document modeling. Second, in supervised experiments, we have compared to a wide variety of deep NN based models (CNNs, RNNs, hierarchical NNs, and Transformers based models).

To R4: 1) Regardless of which MCMC method is used, the need to perform a sampling based iterative procedure (e.g., hundreds of MCMC iterations) for each test document limits the efficiency for out-of-sample prediction. In addition, if restricting to Gibbs sampling, it is difficult to incorporate label information into the model. Thus, we develop an encoder network to map the observations directly to their latent representations. We also introduce a hybrid SG-MCMC/VI for inference. While [15] has validated hybrid SG-MCMC/VI empirically, we acknowledge there is still theoretical gap to fill to validate the practice of sampling from a variational posterior in lieu of the exact conditional posterior, and rolling these approximate samples into a Markov chain. This presents an interesting theoretical question (including analysis of convergence and mixing), which, however, is beyond the scope of this paper. The reason why we develop a parallelized Gibbs sampler as well as a hybrid SG-MCMC/VI is that both of them have their own advantages. The use of encoder makes our model fast in testing time, but leads to a tradeoff in accuracy, as shown in Table 2. In addition, the encoder network enables our model to directly incorporate side information.

Methods	Accuracy				Testing time in seconds			
	Reuters	ELEC	IMDB-2	IMDB-10	Reuters	ELEC	IMDB-2	IMDB-10
bi-conv-PGDS (TLASGR-MCMC)	78.0 ± 0.7	84.5 ± 0.8	84.0 ± 0.8	37.9 ± 0.4	17.35	15.07	25.61	27.52
bi-conv-PGDS (hybrid SG-MCMC and VI)	76.8 ± 1.0	83.7 ± 1.2	82.6 ± 1.1	36.9 ± 0.7	0.15	0.14	0.17	0.20

Table 2: Comparison of the testing times (seconds) with batch-size 25.

2) Note in each topic, the words assigned with negligible weights are not important, as shown in Fig. 2; the weights of these noted meaningless phrases are: "lap ($2e^{-4}$) guess ($9e^{-3}$) none ($3e^{-4}$)", "rarely ($2e^{-4}$) though ($3e^{-4}$) packaged ($2e^{-4}$)", "pleased (0.35) roll ($2e^{-4}$) recorded ($2e^{-4}$)". 3) The validation set is not used by our models to select parameters for unsupervised learning (see discussion at the very beginning); it is used to select the step size in supervised learning. We use Adam to update the encoder of our model and use ELBO as the convergence criteria. All code can be found in corresponding papers. 4) We will add the standard deviations in Fig. 1. 5) CPGBN is not a multilayer convolutional model, but a coupling of CPFA and GBN via a probabilistic document-level pooling layer. It extends CPFA to capture the hierarchical relationships of different phrases. Comparing with CPGBN, the proposed CPGDS focuses on the structural improvement at the sentence level by capturing the relationships of different sentences. They are two complementary ideas. In addition, comparing with CPGBN, our model has greater advantages in multi-category data, like IMDB-10 in Table 1 and yelp14 in Table 2, which are multi-level sentiment classification problems that need to consider the relationships between sentences. 6) We list more attention visualization of different datasets in Figs. 6 and 7, and they are not "cherry picked" examples; we note similar visualizations can be found in [28]. 7) We will provide more clear and simplified notation in our revision. 8) Gamma$(a, 1/b)$ in our paper have mean a/b. 9) To exploit a rich set of tools developed for count data analysis, we first link sequential binary vectors to sequential count vectors via the Bernoulli-Poisson link. This can be seen as an auxiliary variable trick to arrive at a Poisson-gamma structure that is amenable to posterior inference. 10) To utilize the reparameterization trick motivates the choice of Weibull distribution, which exhibits a similar probability density function as the gamma distribution that is not reparameterizable (see [15] for more details). 11) We will correct Line 169 and use λ to replace ξ in line 198.

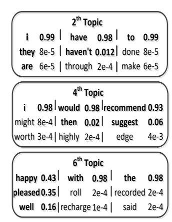

Figure 2: Example topics.