A Datasets & Model Details

In the following, we provide details of data preprocessing and the experimental setups used in the experiments. For both Yelp Reviews and arXiv Abstracts datasets, we truncate the original paragraph to the first five sentences (split by punctuation marks including *comma*, *period* and *point* symbols), where each sentence contains at most 25 words. Therefore, each paragraph has at most 125 words. We remove those sentences that contain less than 30 words. The statistics of both datasets are detailed in Table 10. The average length of paragraphs considered here are much larger than previous generative models for text (Bowman et al., 2016; Yu et al., 2017; Hu et al., 2017; Zhang et al., 2017), since these works considered text sequences that contain only one sentence with at most twenty words.

Dataset	Train	Test	Vocabulary	Aver. Length
Yelp Reviews	244748	18401	12461	48
arXiv Abstracts	504268	28016	32487	59

Table 10: Summary statistics for the datasets used in the generic text generation experiments.

In all the VAE and extentions, the dimension of the latent variable z is set to 300. The dimensions of both the sentence-level and word-level LSTM decoders are set to 512. For the generative networks, to infer the bottom-level latent variable (*i.e.*, modeling $p(z_1|z_2)$), we first feed the sampled latent codes from z_2 to two MLP layers, which is followed by two linear transformation to infer the mean and variance of z_1, respectively.

The model is trained using Adam (Kingma and Ba, 2014) with a learning rate of 3×10^{-4} for all parameters, with a decay rate of 0.99 for every 3000 iterations. Dropout (Srivastava et al., 2014) is employed on both word embedding and latent variable layers, with rates selected from $\{0.3, 0.5, 0.8\}$ on the validation set. We set the mini-batch size to 128. Following (Bowman et al., 2016) we adopt the KL cost annealing strategy to stabilize training: the KL cost term is increased linearly to 1 until 10,000 iterations. All experiments are implemented in Tensorflow (Abadi et al., 2016), using one NVIDIA GeForce GTX TITAN X GPU with 12GB memory.

B Additional Generated Samples from *ml*-VAE-D vs *flat*-VAE

We provide additional examples for the comparison between *ml*-VAE-D vs *flat*-VAE in Table 11, as a continuation of Table 1.

C Retrieved closest training instances of generated samples (Yelp Reviews Dataset)

We provide samples of retrieved instances from the Yelp Review training dataset which are closest to the generated samples. Table 12 shows the closest training samples of each generated Yelp review. The first column indicates the intermediate generated sentences produced from linear transition from a point A to another point B in the prior latent space. The second column on the right are the real sentences retrieved from the training set that are closest to the ones generated on the left (determined by BLEU-2 score). We can see that the retrieved training data is quite different from the generated samples, indicating that our model is indeed generating samples that it has never seen during training.

D Human evaluation setup and details

Some properties of the generated paragraphs, such as (topic) coherence or non-redundancy, can not be easily measured by automated metrics. Therefore, we conduct human evaluation based on 100 samples randomly generated by each model (the models are trained on the Yelp Reviews dataset for this evaluation). We consider *flat*-VAE, adversarial autoencoders (AAE) and real samples from the test set to compare with our proposed *ml*-VAE-D model. The same hyperparameters are employed for the different model variants to ensure fair comparison. We evaluate the quality of these generated samples with a blind heads-to-head comparison using Amazon Mechanical Turk. Given a pair of generated reviews, the judges are asked to select their preferences ("no difference between the two reviews" is also an option) according to the following 4 evaluation criteria: (1) *fluency & grammar*, the one that is more grammatically correct and fluent; (2) *consistency*, the one that depicts a sequence of topics and events that is more consistent; (3) *non-redundancy*, the one that is better at non-redundancy (if a review repeats itself, this can be taken into account); and (4) *overall*,

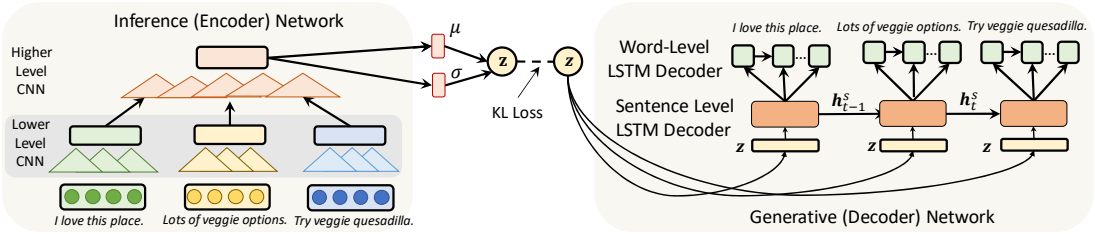

Figure 3: Schematic diagram of the proposed *multi-level* VAE with **single** latent variable.

the one that more effectively communicates reasonable content. These different criteria help to quantify the impact of the hierarchical structures employed in our model, while the non-redundancy and consistency metrics could be especially correlated with the model's plan-ahead abilities. The generated paragraphs are presented to the judges in a random order and they are not told the source of the samples. Each sample is rated by three judges and the results are averaged across all samples and judges.

E More Samples on Attribute Vector Arithmetic

We provide more samples for sentiment manipulation, where we intend to alter sentiment of negative Yelp reviews with "attribute vector arithmetic", as a continuation of Table 9.

F Comparison with the "utterance drop" strategy

To resolve the "posterior collapse" issue of training textual VAEs, (Park et al., 2018) also introduced a strategy called *utterance drop* (u.d). Specifically, they proposed to weaken the autoregressive power of hierarchical RNNs by dropping the utterance encoder vector with a certain probability. To investigate the effectiveness of their method relative to our strategy of employing a hierarchy of latent variables, we conduct a comparative study. Particularly, we utilize *ml*-VAE-S as the baseline model and apply the two strategies to it respectively. The corresponding results on language modeling (Yelp dataset) are shown in Table 14. Their u.d strategy indeed allows better usage of the latent variable (indicated by a larger KL divergence value). However, the NLL of the language model becomes even worse, possibly due to the weakening of the decoder during training (similar observations have also been reported in Table 2 of (Park et al., 2018)). Our hierarchical prior strategy yields larger KL terms as well as lower NNL value, indicating the advantage of our strategy to mitigate the "posterior collapse" issue.

Model	NLL	KL	PPL
ml-VAE-S	160.8	3.6	46.6
ml-VAE-S (with u.d)	161.3	5.6	47.1
ml-VAE-D	160.2	6.8	45.8

Table 14: Comparison with the *utterance drop* strategy.

***ml*-VAE**	***flat*-VAE**
i would give this place zero stars if i could , the guy who was working the front desk was rude and unprofessional , i have to say that i was in the wrong place , and i m not sure what i was thinking , this is not a good place to go to .	this is a great little restaurant in vegas , i had the shrimp scampi and my wife **had the shrimp scampi**, and my husband **had the shrimp scampi** , it was delicious , i **had the shrimp scampi** which was delicious and seasoned perfectly .
my wife and i went to this place for dinner , we were seated immediately , the food was good , i ordered the shrimp and grits , which was the best part of the meal .	**very good chinese food, very good chinese food**, the service was very slow, i guess that s what they were doing, very slow to get a quick meal.
we got a gift certificate from a store, we walked in and were greeted by a young lady who was very helpful and friendly, so we decided to get a cut, I was told that they would be ready in 15 minutes.	we go there for eakfast, i ve been here 3 times and it s always good, the hot dogs are delicious, and **the hot dogs are delicious**, i ve been there for eakfast and it is so good.
the place was packed, chicken was dry, tasted like a frozen hot chocolate, others were just so so, i wouldn t recommend this place.	do not go here, their food is terrible, they were very slow, in my opinion.
went today with my wife, and received a coupon for a free appetizer, we were not impressed, we both ordered the same thing, and **we were not impressed**.	the wynn is a great place to eat, the food was great and i had the linguine, and it was so good, i **had the linguine** and clams, (i was so excited to try it).
recently visited this place for the first time, i live in the area and have been looking for a good local place to eat, we stopped in for a quick bite and a few beers, always a nice place to sit and relax, wonderful and friendly staffs.	i came here for a quick bite before heading to a friend s recommendation, the place was packed, but the food was delicious, i am a fan of the place, and the place is **packed** with a lot of people.
best haircut i ve had in years, friendly staff and great service, he made sure that i was happy with my hair cut, just a little pricey but worth it, she is so nice and friendly.	had a great experience here today, the delivery was friendly and efficient and the food was good, i would recommend this place to anyone who will work in the future, will be back again.
great place to go for a date night, first time i went here, service is good, the staff is friendly, 5 stars for the food.	best place to get in vegas, ps the massage here is awesome, if you want to spend your money, then go there, ps **the massage is great**.

Table 11: Samples randomly generated from *ml*-VAE-D and *flat*-VAE, which are both trained on the Yelp review dataset. The repetitive patterns within the generated reviews are highlighted.

	Generated samples	Closest instance (in the training dataset)
A	the service was great, the receptionist was very friendly and the place was clean, we waited for a while, and then our room was ready .	i ve only been here once myself , and i wasn t impressed , the service was great , staff was very friendly and helpful , we waited for nothing
•	same with all the other reviews, this place is a good place to eat, i came here with a group of friends for a birthday dinner, we were hungry and decided to try it, we were seated promptly.	i really love this place , red robin alone is a good place to eat , but the service here is great too not always easy to find , we were seated promptly , ought drinks promptly and our orders were on point .
•	this place is a little bit of a drive from the strip, my husband and i were looking for a place to eat, all the food was good, the only thing i didn t like was the sweet potato fries.	after a night of drinking , we were looking for a place to eat , the only place still open was the grad lux , its just like a cheesecake factory , the food was actually pretty good .
•	this is not a good place to go, the guy at the front desk was rude and unprofessional, it s a very small room, and the place was not clean.	the food is very good , the margaritas hit the spot , and the service is great , the atmosphere is a little cheesy but overall it s a great place to go .
•	service was poor, the food is terrible, when i asked for a refill on my drink, no one even acknowledged me, they are so rude and unprofessional.	disliked this place , the hostess was so rude , when i asked for a booth , i got attitude , a major .
B	how is this place still in business, the staff is rude, no one knows what they are doing, they lost my business .	i can t express how awful this store is , don t go to this location , drive to any other location , the staff is useless , no one knows what they are doing .

Table 12: Using the *ml*-VAE-D model trained on the Yelp Review dataset, intermediate sentences are produced from linear transition between two points (**A** and **B**) in the prior latent space. Each sentence in the left panel is generated from a latent point on a linear path, and each sentence on the right is the closet sample to the left one within the entire training set (determined by BLEU-2 score).

Original: papa j s is expensive and inconsistent , the ambiance is nice but it doesn t justify the prices , there are better restaurants in carnegie .	**Transferred**: love the food , the prices are reasonable and the food is great , it s a great place to go for a quick bite .
Original: i had a lunch there once , the food is ok but it s on the pricy side , i don t think i will be back .	**Transferred**: i had a great time here , the food is great and the prices are reasonable , i ll be back .
Original: i have to say that i write this review with much regret , because i have always loved papa j s , but my recent experience there has changed my mind a bit , from the minute we were seated , we were greeted by a server that was clearly inexperienced and didn t know the menu .	**Transferred**: i have to say , the restaurant is a great place to go for a date , my girlfriend and i have been there a few times , on my last visit , we were greeted by a very friendly hostess .
Original: a friend recommended this to me , and i can t figure out why , the food was underwhelming and pricey , the service was fine , and the place looked nice .	**Transferred**: a friend of mine recommended this place , and i was so glad that i did try it , the service was great , and the food was delicious .
Original: this is a small , franchise owned location that caters to the low income in the area , selection is quite limited throughout the store with limited quantities on the shelf of the items they do carry , because of the area in which it is located , the store is not 24 hours as most giant eagle s seem to be .	**Transferred**: this is a great little shop, easy to navigate , and they are always open , their produce is always fresh , the store is clean and the staff is friendly .

Table 13: Sentiment transfer results with attribute vector arithmetic.

Original: you have no idea how badly i want to like this place, they are incredibly vegetarian vegan friendly , i just haven t been impressed by anything i ve ordered there , even the chips and salsa aren t terribly good , i do like the bar they have great sangria but that s about it .
Transferred: this is definitely one of my favorite places to eat in vegas , they are very friendly and the food is always fresh, i highly recommend the pork belly , everything else is also very delicious, i do like the fact that they have a great selection of salads .
Original: my boyfriend and i are in our 20s , and have visited this place multiple times , after our visit yesterday , i don t think we ll be back , when we arrived we were greeted by a long line of people waiting to buy game cards .
Transferred: my boyfriend and i have been here twice , and have been to the one in gilbert several times too , since my first visit , i don t think i ve ever had a bad meal here , the servers were very friendly and helpful .

Table 15: An example sentiment transfer result with attribute vector arithmetic.