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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Test</th>
<th>Vocabulary</th>
<th>Aver. Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yelp Reviews</td>
<td>244748</td>
<td>18401</td>
<td>12461</td>
<td>48</td>
</tr>
<tr>
<td>arXiv Abstracts</td>
<td>504268</td>
<td>28016</td>
<td>32487</td>
<td>59</td>
</tr>
</tbody>
</table>

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

In all the VAE and extensions, 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 $3 \times 10^{-4}$ 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,
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>NLL</th>
<th>KL</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml-VAE-S</td>
<td>160.8</td>
<td>3.6</td>
<td>46.6</td>
</tr>
<tr>
<td>ml-VAE-S (with u.d)</td>
<td>161.3</td>
<td>5.6</td>
<td>47.1</td>
</tr>
<tr>
<td>ml-VAE-D</td>
<td>160.2</td>
<td>6.8</td>
<td>45.8</td>
</tr>
</tbody>
</table>

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

• 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.

• 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.

• 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.

• 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.

B how is this place still in business, the staff is rude, no one knows what they are doing, they lost my business.

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.

• 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.

• 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.

• 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.

• disliked this place, the hostess was so rude, when I asked for a booth, I got attitude, a major.

• 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.

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 pricey 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 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 closest sample to the left one within the entire training set (determined by BLEU-2 score).

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.