A Surprisingly Effective Fix for Deep Latent Variable Modeling of Text

Bohan Li\textsuperscript{1}, Junxian He\textsuperscript{1}, Graham Neubig\textsuperscript{1}, Taylor Berg-Kirkpatrick\textsuperscript{2}, Yiming Yang\textsuperscript{1}
\textsuperscript{1}Language Technologies Institute, Carnegie Mellon University
\textsuperscript{2}Department of Computer Science and Engineering, University of California San Diego
{bohanli, junxianh, gneubig, yiming}@cs.cmu.edu, tberg@eng.ucsd.edu

Abstract

When trained effectively, the Variational Autoencoder (VAE) is both a powerful language model and an effective representation learning framework. In practice, however, VAEs are trained with the evidence lower bound (ELBO) as a surrogate objective to the intractable marginal data likelihood. This approach to training yields unstable results, frequently leading to a disastrous local optimum known as posterior collapse. In this paper, we investigate a simple fix for posterior collapse which yields surprisingly effective results. The combination of two known heuristics, previously considered only in isolation, substantially improves held-out likelihood, reconstruction, and latent representation learning when compared with previous state-of-the-art methods. More interestingly, while our experiments demonstrate superiority on these principle evaluations, our method obtains a worse ELBO. We use these results to argue that the typical surrogate objective for VAEs may not be sufficient or necessarily appropriate for balancing the goals of representation learning and data distribution modeling.\textsuperscript{1}

1 Introduction

Latent variable models attempt to model observed data \( x \) given latent variables \( z \), both for purposes of modeling data distributions \( p(x) \) (e.g. language modeling) and learning representations \( z \) for a particular \( x \) (e.g. sentence embedding). Variational Autoencoders (VAEs) (Kingma and Welling, 2014) are a powerful framework for learning latent variable models using neural networks. The generative model of VAEs first samples a latent vector \( z \) from a prior \( p(z) \), then applies a neural decoder \( p(x|z) \) to produce \( x \) conditioned on the latent code \( z \). VAEs are trained to maximize the evidence lower bound (ELBO) of the intractable log marginal likelihood:

\[
\mathbb{E}_{z \sim q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - D_{KL}(q_{\phi}(z|x)\|p(z)),
\]

where \( q_{\phi}(z|x) \) represents an approximate posterior distribution (i.e. the encoder or inference network) and \( p_{\theta}(x|z) \) is the generative distribution (i.e. the decoder).

However, modeling text with VAEs has proven to be challenging, and is an open research problem (Yang et al., 2017; Xu and Durrett, 2018; Kim et al., 2018; Dieng et al., 2018; He et al., 2019; Pelsmaeker and Aziz, 2019). When a strong decoder (e.g. the LSTM (Hochreiter and Schmidhuber, 1997)) is employed, training often falls into a trivial local optimum where the decoder learns to ignore the latent variable and the encoder fails to encode any information. This phenomenon is referred to as “posterior collapse” (Bowman et al., 2016). Existing efforts tackling this problem include re-weighting the KL loss (Bowman et al., 2016; Kingma et al., 2016; Liu et al., 2019), changing the model (Yang et al., 2017; Semeniuta et al., 2017; Xu and Durrett, 2018), and modifying the training procedure (He et al., 2019).

After conducting an empirical examination of the state-of-the-art methods (Section 2), we find that they have difficulty striking a good balance between language modeling and representation learning. In this paper, we present a practically effective combination of two simple heuristic techniques for improving VAE learning: (1) pretraining the inference network using an autoencoder objective and (2) thresholding the KL term in the ELBO objective (also known as “free bits” (Kingma et al., 2016)). The former technique initializes VAE training with an inference network that encodes useful information about \( x \), biasing learning away from local optima where \( x \) is ignored (i.e. posterior collapse). The latter tech-

\footnotetext{\textsuperscript{1}Equal contribution.}\
\footnotetext{\textsuperscript{2}Code is available at https://github.com/bohanli/vae-pretraining-encoder.
nique modifies the ELBO objective to prevent the KL term from dominating the encoder’s role in reconstruction (Alemi et al., 2018), again biasing learning to avoid posterior collapse (Chen et al., 2017; Zhao et al., 2019).

In experiments we find that these two techniques do not perform well in isolation. However, when combined, they substantially outperform all baselines across various metrics that evaluate both language modeling and representation learning capabilities. Finally, we find that our method tends to achieve a superior language modeling results in terms of perplexity but an inferior ELBO value, and we use these results to argue that ELBO is suboptimal for language modeling even though it provides a formal lower bound on log marginal likelihood. Thus, we suggest that future research in this direction should be careful to monitor the gap between ELBO and log marginal likelihood, and reconsider using ELBO as the surrogate, especially for evaluation.

2 Analysis of Existing Methods

In this section, we first analyze and compare state-of-the-art solutions to posterior collapse to obtain a holistic view of current progress and remaining challenges. Based on these observations, we then propose and evaluate a method that demonstrates substantially improved performance.

2.1 Evaluation

In this paper we focus on the standard VAE setting where both the prior $p(z)$ and the posterior $p_{\theta}(z|x)$ are factorized Gaussians. We conduct preliminary experiments on the English Penn Treebank (PTB) (Marcus and Marcinkiewicz), a standard dataset for benchmarking VAE that has been used extensively in previous work (Bowman et al., 2016; Xu and Durrett, 2018; Liu et al., 2019). We use 32-dimension latent codes (full setup details can be found in Appendix A). We evaluate using the following metrics:

**Perplexity (PPL).** To compute perplexity, we first estimate the log marginal likelihood with 1000 importance weighted samples (Burda et al., 2016). Note that it is inappropriate to estimate PPL with ELBO directly since the gap between ELBO and log marginal likelihood might be large, particularly when the posterior does not collapse.

**Reconstruction loss (Recon).** Reconstruction loss is equivalent to the negative reconstruction term in ELBO: $-\mathbb{E}_{z \sim q_{\phi}(z|x)} \log p_{\theta}(x|z)$. It characterizes how well the latent code can be used to recover the input.

**Number of active units (AU, Burda et al. (2016)).** Active units correspond to the dimensions of $z$ that covary with observations after the model is trained. More active units usually indicate richer latent representations (Burda et al., 2016). Specifically, a dimension is “active” when it is sensitive to the change in observations $x$. Here we follow (Burda et al., 2016) and classify a latent dimension $z$ as active if $\text{Cov}(x, E_{z \sim q(z|x)}[z]) > 0.01$.

In addition to the metrics above, we include KL between prior and posterior approximation, as well as the negative ELBO, for reference – though we find that these quantities are only partially descriptive of model quality. In Section 3.2 we also evaluate the latent space of learned models with specialized metrics such as reconstruction BLEU and classification accuracy.

2.2 Baselines

We experiment with several state-of-the-art techniques to mitigate posterior collapse, including several KL reweighting methods and the recently proposed aggressive training (He et al., 2019).

**KL annealing (Bowman et al., 2016).** KL annealing might be the most common method for reweighting. During annealing, the weight of the KL term is increased from a small value to 1.0 in the beginning of training.

**Cyclic annealing (Liu et al., 2019).** Cyclic annealing is another reweighting scheme proposed recently. It changes the weight of the KL term in a cyclic fashion, rather than monotonically increasing the weight.\footnote{We use the default cyclic annealing schedule from \url{https://github.com/haofuml/cyclic_annealing} in our codebase for a fair comparison.}

**KL Thresholding / Free Bits (FB) (Kingma et al., 2016).** FB replaces the KL term in ELBO with a hinge loss term that maxes each component
of the original KL with a constant:

\[ \sum_i \max[\lambda, D_{KL}(q(\mathbf{z}_i | \mathbf{x}) \parallel p(\mathbf{z}_i))] \]  

(1)

Here, \( \lambda \) denotes the target rate, and \( z_i \) denotes the \( i^{th} \) dimension in the latent variable \( \mathbf{z} \). Using the FB objective causes learning to give up trying to drive down KL for dimensions of \( \mathbf{z} \) that are already beneath the target rate. Pelsmaeker and Aziz (2019) conduct a comprehensive experimental evaluation of related methods and conclude that the FB objective is able to achieve comparable or superior performance (in terms of both language modeling and reconstruction) in comparison with other top-performing methods, many of which are substantially more complex. We notice that Pelsmaeker and Aziz (2019) experiment with a slightly different version of FB where the threshold is applied to the entire KL term directly, rather than on each dimension’s KL separately. We examine both versions here and refer to the single threshold version as “FBP” and the multiple threshold version (Eq. 1) as “FB”. For both FB and FBP, we vary the target rate \( \lambda \) and report the setting with the best validation PPL and the setting with the best balance between PPL and reconstruction loss.\(^5\)

**Aggressive training (He et al., 2019).** He et al. (2019) observe that when posterior collapse occurs, the inference network often lags behind the generator during training. In contrast with the KL reweighting methods described above, He et al. (2019) propose an aggressive training schedule which iterates between multiple encoder update steps and one decoder update step to mitigate posterior collapse.\(^6\)

**Autoencoder (AE).** We also include an autoencoder\(^7\) as a reference for reconstruction loss.

We show the results of these baselines trained on PTB in Table 1, where we find that it is difficult to balance language modeling (PPL) and representation learning (Recon and AU) – the systems with relatively good reconstruction (FBP, \( \lambda = 7 \)) or higher AU (FB baselines) have suboptimal PPL, and the best PPL is achieved with sacrifice of Recon and AU. Note that good PPL indicates good LM, but good recon and AU indicates good representation learning. Without both, we do not really

---

\(^5\)It is subjective to judge “balance”, thus we also report complete results for different target rates in Appendix C.

\(^6\)We use the public code at https://github.com/jxhe/vae-lagging-encoder.

\(^7\)Here, AE denotes the VAE trained without the KL term.

---

### Table 1: Results on PTB test set for various baselines.

<table>
<thead>
<tr>
<th>Method</th>
<th>PPL ↓</th>
<th>Recon ↓</th>
<th>AU ↑</th>
<th>KL</th>
<th>-ELBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>70.36</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VAE</td>
<td>101.39</td>
<td>101.27</td>
<td>0.00</td>
<td>101.27</td>
<td></td>
</tr>
<tr>
<td>+ anneal</td>
<td>101.40</td>
<td>101.28</td>
<td>0.00</td>
<td>101.28</td>
<td></td>
</tr>
<tr>
<td>+ cyclic</td>
<td>108.97</td>
<td>101.85</td>
<td>1.37</td>
<td>103.22</td>
<td></td>
</tr>
<tr>
<td>+ aggressive</td>
<td>99.83</td>
<td>100.26</td>
<td>4.93</td>
<td>101.19</td>
<td></td>
</tr>
<tr>
<td>+ FBP (( \lambda = 7 ))</td>
<td>102.82</td>
<td>95.63</td>
<td>7.05</td>
<td>102.67</td>
<td></td>
</tr>
<tr>
<td>+ FBP (( \lambda = 3 ))</td>
<td>99.62</td>
<td>98.52</td>
<td>2.95</td>
<td>101.48</td>
<td></td>
</tr>
<tr>
<td>+ FB (( \lambda = 7 ))</td>
<td>104.06</td>
<td>98.97</td>
<td>6.74</td>
<td>105.72</td>
<td></td>
</tr>
<tr>
<td>+ FB (( \lambda = 3 ))</td>
<td>100.50</td>
<td>99.94</td>
<td>2.96</td>
<td>102.90</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Results on PTB test with encoder pretraining.

<table>
<thead>
<tr>
<th>Method</th>
<th>PPL ↓</th>
<th>Recon ↓</th>
<th>AU ↑</th>
<th>KL</th>
<th>-ELBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>70.36</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VAE</td>
<td>101.39</td>
<td>101.27</td>
<td>0.00</td>
<td>101.27</td>
<td></td>
</tr>
<tr>
<td>+ pretrain</td>
<td>102.26</td>
<td>101.46</td>
<td>0.00</td>
<td>101.46</td>
<td></td>
</tr>
<tr>
<td>+ pretrain + anneal</td>
<td>97.74</td>
<td>99.67</td>
<td>1.01</td>
<td>100.68</td>
<td></td>
</tr>
</tbody>
</table>

have a strong probabilistic model of language that captures latent factors.

We make two observations from the results in Table 1. First, reconstruction loss for an AE is substantially better than all VAE methods, which is intuitive since reconstruction is the only goal of training an AE. Second, models with high ELBO do not necessarily have good PPL (e.g. VAE+anneal); ELBO is not an ideal surrogate for evaluating language modeling performance.

### 2.3 Autoencoder-based Initialization

Based on the observations above we hypothesize that VAEs might benefit from initialization with an non-collapsed encoder, trained via an AE objective. Intuitively, if the encoder is providing useful information from the beginning of training, the decoder is more likely to make use of the latent code. In Table 2 we show the results of exploring this hypothesis on PTB. Even with encoder pretraining, we see that posterior collapse occurs immediately after beginning to update both encoder and decoder using the full ELBO objective. This indicates that the gradients of ELBO point towards a collapsed local optimum, even with biased initialization. When pretraining is combined with annealing, PPL improves substantially. However, the pretraining and anneal combination only has 2 active units and has small KL value – the latent representation is likely unsatisfactory. We speculate that this is because the annealing schedule eventually returns to the full ELBO objective which guides learning towards a (nearly) collapsed latent space. In the next section, we present an alternate approach using the KL thresholding / free bits method.
2.4 Our Method

In our proposed method, we initialize the inference network with an encoder that is pretrained using an autoencoder objective, as described above. Then, we train the VAE using the FB objective, also described above. We use the original FB which thresholds along each dimension. Thus, we combine two approaches so far considered independently: pretraining and KL thresholding. In this way, however, the VAE would start with a large KL and is thus trained with the full ELBO objective in the initial stage, which is bad as observed in Section 2.3. To remedy this, we apply an annealing weight to Eq. 1. We use the simplest linear annealing schedule that increases the weight from 0 to 1 in the first 10 epochs for all of our experiments. This approach can be viewed in connection with KL annealing: we train with zero KL weight until convergence, then reset the decoder and start increasing the KL weight with the free bits objective. Next we conduct comprehensive experiments across different datasets to validate our method.

3 Experiments

In this section we work with three text datasets: PTB (Marcus and Marcinkiewicz), Yahoo (Yang et al., 2017), and a downsamplled version of SNLI (Bowman et al., 2015). We demonstrate the effectiveness of our method through language modeling, text reconstruction, and quality of the learned latent space. Complete experimental setup details can be found in Appendix A.

3.1 Language Modeling

For language modeling, we only report the results for PTB and Yahoo due to the space limit, but include the SNLI results in Appendix C. Since we have already shown that FBP outperforms FB on PTB in Section 2 (without pretraining), here we only include FBP as a baseline. For both FBP and our method we vary the target rate and report the settings that achieve competitive validation PPL.

As shown in Table 3, our method with different target rates is able to consistently outperform all the baselines in terms of PPL. Meanwhile, for representation learning, our method beats all the baselines by a large margin on reconstruction loss with all latent units active. We also note that our method is not very sensitive to the target rate λ.

It is worth noticing that in some cases our method (e.g. λ = 8 in PTB and λ = 8, 9 in Yahoo) is able to outperform all the baselines but produce a bad ELBO (actually the worst on PTB). This suggests that ELBO might be a suboptimal surrogate for the log marginal likelihood sometimes, especially when KL is large, where the gap between ELBO and the marginal tends to be large as well.

3.2 Probing the Latent Space

We assess quality of learned latent space with SNLI through several metrics. For our method and FBP, we use the target rate where the best reconstruction loss is achieved while maintaining comparable PPL with aggressive training.

Text Reconstruction We use greedy decoding and compute the BLEU score of the reconstructed sentence with the original one as the reference. The result is shown in Table 4. Unsurprisingly,
the autoencoder achieves the highest BLEU score, meanwhile our method beats other VAE baselines.

Table 4: Reconstruction

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>60.80</td>
</tr>
<tr>
<td>VAE</td>
<td>1.82</td>
</tr>
<tr>
<td>+ anneal</td>
<td>2.51</td>
</tr>
<tr>
<td>+ cyclic</td>
<td>4.39</td>
</tr>
<tr>
<td>+ aggressive</td>
<td>2.95</td>
</tr>
<tr>
<td>+ FBP (λ = 7)</td>
<td>8.07</td>
</tr>
<tr>
<td>Ours (λ = 4)</td>
<td><strong>8.62</strong></td>
</tr>
</tbody>
</table>

Table 5: Smoothness

<table>
<thead>
<tr>
<th>Method</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>0.620</td>
</tr>
<tr>
<td>VAE</td>
<td>0.039</td>
</tr>
<tr>
<td>+ anneal</td>
<td>0.009</td>
</tr>
<tr>
<td>+ cyclic</td>
<td>0.482</td>
</tr>
<tr>
<td>+ aggressive</td>
<td>0.209</td>
</tr>
<tr>
<td>+ FBP (λ = 7)</td>
<td>0.242</td>
</tr>
<tr>
<td>Ours (λ = 4)</td>
<td><strong>0.683</strong></td>
</tr>
</tbody>
</table>

**Smoothness of Latent Space** A major difference between VAE and AE is that VAE can learn a smooth latent space through the regularization from the Gaussian prior. In a smooth latent space, latent codes of similar sentences should be close to each other and vice versa. Therefore, we randomly sample 100k sentence pairs and evaluate the Pearson Correlation Coefficient (PCC) between the \( \ell_2 \) distances of latent codes and edit word distances. As shown in Table 5, our method achieves a much higher PCC compared to the baselines.

Zhao et al. (2018) argue that a smooth latent space is beneficial for reconstructing noisy inputs. We follow their experiments and introduce noise to the input by randomly swapping words \( k \) times. As shown in Table 6, while AE achieves the best reconstruction when the noise is small (\( k = 1 \)), its reconstruction deteriorates dramatically when \( k > 1 \), which suggests AE fails to learn a smooth latent space. In contrast, our method outperforms all the baselines by a large margin when \( k > 1 \).

**Interpolation** As illustrated in (Bowman et al., 2016), linear interpolation between latent variables is an intuitive way to qualitatively evaluate the smoothness of the latent space. We sample two latent codes \( z_0 \) and \( z_1 \) from the prior \( p(z) \) (Table 7) and do linear interpolation between the two with even divided intervals.\(^{11}\) For each interpolated point, we decode it greedily. Our method is able to generate grammatically plausible and semantically consistent interpolation in both cases.

3.3 Classification

To further evaluate the quality of the latent codes, we train a Gaussian mixture model (for unsupervised clustering) or a one-layer linear classifier (for supervised classification) on the pretrained latent codes. We work with a downsampled version of Yelp sentiment dataset collected by Shen et al. (2017). We vary the number of labeled data\(^{12}\) and the results are shown in Table 8. For FBP and our method, the target rate is selected in terms of the average validation accuracy. Our method consistently yields the best results on all settings – remarkably, its performance with only 100 labeled samples already surpasses others with 10k labels.

4 Conclusion

In this paper, we propose a simple training fix to tackle posterior collapse in VAEs. Extensive experiments demonstrate the effectiveness of our method on both representation learning and language modeling.

\(^{11}\)Interpolation examples from posterior samples \( q_\phi(z|x) \) are provided in Appendix 13.

\(^{12}\)Specifically, we train a VAE model on the Yelp dataset to obtain the latent codes, then we use the latent codes of labeled data to train the classifier.

3598
References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In Proceedings of ICML.


Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In Proceedings of NeurIPS.


A Details of Experimental Setup

For SNLI, we randomly downsample a subset of it, which contains 100K/10K/10K sentences as training/validation/test. For the Yelp sentiment dataset, we also randomly downsample 100K/10K/10K sentences for training/validation/test, respectively. We use an one-layer LSTM for both the encoder and decoder and a latent vector size of 32.

We follow Kim et al. (2018); He et al. (2019) and use a single-layer LSTM the encoder and decoder in all of our experiments. The sizes of word embeddings and hidden states for different datasets are given in Table 9. We initialize the LSTM parameters with a uniform distribution $U(-0.01, 0.01)$, and embeddings with another uniform distribution $U(-0.1, 0.1)$.

Just as in Kim et al. (2018); He et al. (2019), for the decoder, we use dropout of 0.5 on both the input word embedding and the last dense before logits. During training, we use the SGD optimizer without momentum. Initialized with 0.5, the learning rate is decayed by $\times 0.5$ with a patience of 2 if the validation loss has not improved in the past 2 epochs. The maximum number of epochs is 100, but the training will stop early after 5 learning rate decays. For our VAE + anneal baseline, we use the simplest linear annealing schedule that increases the weight from 0 to 1 in the first 10 epochs, just the same as in our method (stated in Section 2.4).

Table 9: The sizes of word embeddings and hidden states for PTB, SNLI and Yahoo.

<table>
<thead>
<tr>
<th></th>
<th>PTB</th>
<th>SNLI</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Embedding Size</td>
<td>256</td>
<td>128</td>
<td>512</td>
</tr>
<tr>
<td>Hidden Size of Encoder</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td>Hidden Size of Decoder</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
</tbody>
</table>

B Copying Behaviour

We check to make sure that our model is not simply learning to copy sentences from the training set. We test for copying behaviour on the PTB dataset. Specifically, we sampled 300 sentences from the prior and retrieved their nearest neighbors in the training set. The average edit distance between the samples and their nearest neighbors from our method is 14.11 (the average training sentence length is 22.10), versus 13.68 from the collapsed VAE. This means that there is no obvious copying behaviour when KL grows in our method.

C Additional Results of Language Modeling

In Table 10, we provide a more detailed version of the language modeling results on PTB, SNLI and Yahoo. The full results of all the target rate trial for both FBP and our method are included. In addition to the metrics in Table 3, we also report NLL, Mutual Information $I_q$ between $z$ and $x$ under $q_\phi(z|x)$ (MI)$^{13}$, and the perplexity computed by ELBO (ELBO PPL).

---

$^{13}$For the estimation of $I_q$, we follow the same method as in He et al. (2019)
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>NLL↓</th>
<th>PPL↓</th>
<th>Recon↓</th>
<th>MI↑</th>
<th>AU↑</th>
<th>-ELBO</th>
<th>ELBO</th>
<th>PPL</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB</td>
<td>LSTM-LM</td>
<td>101.04</td>
<td>100.47</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AE</td>
<td>-</td>
<td>70.36</td>
<td>8.22</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VAE</td>
<td>101.23</td>
<td>101.39</td>
<td>101.27</td>
<td>0.01</td>
<td>0</td>
<td>101.27</td>
<td>101.58</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ anneal</td>
<td>101.24</td>
<td>101.40</td>
<td>101.28</td>
<td>0.00</td>
<td>0</td>
<td>101.28</td>
<td>101.62</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ cyclic (reported)</td>
<td>-</td>
<td>100.51</td>
<td>-</td>
<td>-</td>
<td>102.46</td>
<td>107.25</td>
<td>1.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ cyclic</td>
<td>102.81</td>
<td>108.97</td>
<td>101.85</td>
<td>1.27</td>
<td>5</td>
<td>103.22</td>
<td>111.03</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ cyclic (SGD)</td>
<td>102.06</td>
<td>105.28</td>
<td>102.14</td>
<td>0.00</td>
<td>0</td>
<td>102.14</td>
<td>105.67</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ aggressive</td>
<td>100.89</td>
<td>99.83</td>
<td>100.26</td>
<td>0.83</td>
<td>4</td>
<td>101.19</td>
<td>101.17</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ pretrained enc</td>
<td>101.42</td>
<td>102.26</td>
<td>101.46</td>
<td>0.00</td>
<td>0</td>
<td>101.46</td>
<td>102.45</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ anneal enc</td>
<td>100.43</td>
<td>97.74</td>
<td>99.67</td>
<td>0.97</td>
<td>2</td>
<td>100.68</td>
<td>98.88</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 9)</td>
<td>101.95</td>
<td>104.73</td>
<td>94.66</td>
<td>7.22</td>
<td>6</td>
<td>103.59</td>
<td>112.88</td>
<td>8.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 8)</td>
<td>101.64</td>
<td>103.30</td>
<td>95.39</td>
<td>7.11</td>
<td>6</td>
<td>103.43</td>
<td>111.03</td>
<td>8.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 7)</td>
<td>101.54</td>
<td>102.82</td>
<td>95.63</td>
<td>6.52</td>
<td>4</td>
<td>102.67</td>
<td>108.26</td>
<td>7.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 6)</td>
<td>101.68</td>
<td>103.47</td>
<td>96.66</td>
<td>7.16</td>
<td>4</td>
<td>102.14</td>
<td>105.67</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ IW AE (k = 10)</td>
<td>100.86</td>
<td>99.69</td>
<td>100.89</td>
<td>0.05</td>
<td>0</td>
<td>100.89</td>
<td>99.81</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ pretrained enc</td>
<td>100.43</td>
<td>97.74</td>
<td>99.67</td>
<td>0.97</td>
<td>2</td>
<td>100.68</td>
<td>98.88</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 9)</td>
<td>101.09</td>
<td>100.71</td>
<td>92.00</td>
<td>7.78</td>
<td>32</td>
<td>104.43</td>
<td>117.32</td>
<td>12.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 8)</td>
<td>100.51</td>
<td>98.07</td>
<td>92.60</td>
<td>7.49</td>
<td>32</td>
<td>103.59</td>
<td>112.72</td>
<td>10.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 7)</td>
<td>101.06</td>
<td>100.60</td>
<td>93.25</td>
<td>7.46</td>
<td>32</td>
<td>103.74</td>
<td>113.68</td>
<td>10.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 6)</td>
<td>100.83</td>
<td>99.62</td>
<td>98.52</td>
<td>2.86</td>
<td>3</td>
<td>101.48</td>
<td>102.52</td>
<td>2.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 5)</td>
<td>100.12</td>
<td>96.35</td>
<td>94.52</td>
<td>6.30</td>
<td>32</td>
<td>102.67</td>
<td>108.25</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 4)</td>
<td>100.23</td>
<td>96.86</td>
<td>95.87</td>
<td>5.31</td>
<td>32</td>
<td>102.41</td>
<td>106.97</td>
<td>6.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 3)</td>
<td>100.85</td>
<td>99.62</td>
<td>98.52</td>
<td>2.86</td>
<td>3</td>
<td>101.48</td>
<td>102.52</td>
<td>2.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 2)</td>
<td>100.14</td>
<td>100.96</td>
<td>99.37</td>
<td>1.91</td>
<td>2</td>
<td>101.36</td>
<td>101.98</td>
<td>1.99</td>
<td></td>
</tr>
<tr>
<td>SNLI</td>
<td>LSTM-LM</td>
<td>32.97</td>
<td>21.44</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AE</td>
<td>-</td>
<td>-</td>
<td>8.68</td>
<td>9.18</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VAE</td>
<td>33.09</td>
<td>21.67</td>
<td>33.08</td>
<td>0.03</td>
<td>1</td>
<td>33.12</td>
<td>21.73</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ anneal</td>
<td>33.01</td>
<td>21.50</td>
<td>31.66</td>
<td>1.45</td>
<td>2</td>
<td>33.07</td>
<td>21.63</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ cyclic</td>
<td>34.04</td>
<td>23.67</td>
<td>30.69</td>
<td>3.60</td>
<td>5</td>
<td>34.32</td>
<td>24.29</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ cyclic (SGD)</td>
<td>33.07</td>
<td>21.62</td>
<td>30.89</td>
<td>2.33</td>
<td>4</td>
<td>33.25</td>
<td>21.99</td>
<td>2.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ aggressive</td>
<td>32.83</td>
<td>21.16</td>
<td>31.53</td>
<td>1.38</td>
<td>5</td>
<td>32.95</td>
<td>21.39</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 9)</td>
<td>33.28</td>
<td>22.05</td>
<td>25.26</td>
<td>8.06</td>
<td>6</td>
<td>34.25</td>
<td>24.13</td>
<td>8.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 8)</td>
<td>33.26</td>
<td>22.02</td>
<td>26.07</td>
<td>7.35</td>
<td>7</td>
<td>34.08</td>
<td>23.75</td>
<td>8.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 7)</td>
<td>33.07</td>
<td>21.62</td>
<td>26.65</td>
<td>6.76</td>
<td>6</td>
<td>33.78</td>
<td>23.11</td>
<td>7.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 6)</td>
<td>33.09</td>
<td>21.68</td>
<td>27.54</td>
<td>5.95</td>
<td>7</td>
<td>33.59</td>
<td>22.71</td>
<td>6.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 5)</td>
<td>33.04</td>
<td>21.58</td>
<td>28.38</td>
<td>4.95</td>
<td>6</td>
<td>33.49</td>
<td>22.48</td>
<td>5.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 4)</td>
<td>33.04</td>
<td>21.57</td>
<td>29.25</td>
<td>4.06</td>
<td>4</td>
<td>33.36</td>
<td>22.22</td>
<td>4.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 3)</td>
<td>33.04</td>
<td>21.56</td>
<td>30.19</td>
<td>3.00</td>
<td>4</td>
<td>33.31</td>
<td>22.11</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 2)</td>
<td>32.99</td>
<td>21.46</td>
<td>31.04</td>
<td>2.11</td>
<td>3</td>
<td>33.16</td>
<td>21.80</td>
<td>2.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 9)</td>
<td>33.42</td>
<td>22.33</td>
<td>22.30</td>
<td>8.80</td>
<td>32</td>
<td>35.70</td>
<td>27.61</td>
<td>13.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 8)</td>
<td>33.47</td>
<td>22.45</td>
<td>22.65</td>
<td>8.76</td>
<td>32</td>
<td>35.62</td>
<td>27.40</td>
<td>12.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 7)</td>
<td>33.25</td>
<td>22.00</td>
<td>23.36</td>
<td>8.48</td>
<td>32</td>
<td>35.11</td>
<td>26.14</td>
<td>11.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 6)</td>
<td>33.17</td>
<td>21.84</td>
<td>24.06</td>
<td>8.24</td>
<td>32</td>
<td>34.83</td>
<td>25.47</td>
<td>10.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 5)</td>
<td>33.07</td>
<td>21.64</td>
<td>24.94</td>
<td>7.71</td>
<td>32</td>
<td>34.47</td>
<td>24.63</td>
<td>9.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 4)</td>
<td>32.88</td>
<td>21.24</td>
<td>26.52</td>
<td>6.65</td>
<td>32</td>
<td>34.11</td>
<td>23.83</td>
<td>7.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 3)</td>
<td>32.87</td>
<td>21.23</td>
<td>28.02</td>
<td>5.25</td>
<td>32</td>
<td>33.87</td>
<td>23.31</td>
<td>5.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 2)</td>
<td>32.79</td>
<td>21.07</td>
<td>29.75</td>
<td>3.32</td>
<td>32</td>
<td>33.42</td>
<td>22.35</td>
<td>3.67</td>
<td></td>
</tr>
</tbody>
</table>
### Table 11: Additional results of language modeling on Yahoo.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>NLL</th>
<th>PPL</th>
<th>Recon</th>
<th>MI</th>
<th>AU</th>
<th>ELBO</th>
<th>ELBO PPL</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>LSTM-LM</td>
<td>328.00</td>
<td>60.75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AE</td>
<td>-</td>
<td>-</td>
<td>278.76</td>
<td>9.26</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VAE</td>
<td>329.00</td>
<td>61.52</td>
<td>329.10</td>
<td>0.00</td>
<td>0</td>
<td>329.10</td>
<td>61.59</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>+ anneal</td>
<td>328.60</td>
<td>61.21</td>
<td>328.80</td>
<td>0.00</td>
<td>0</td>
<td>328.80</td>
<td>61.36</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>+ cyclic</td>
<td>335.74</td>
<td>66.93</td>
<td>333.80</td>
<td>2.77</td>
<td>4</td>
<td>336.63</td>
<td>67.69</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>+ cyclic (SGD)</td>
<td>332.48</td>
<td>64.26</td>
<td>332.65</td>
<td>0.00</td>
<td>1</td>
<td>332.68</td>
<td>64.42</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>+ aggressive</td>
<td>326.70</td>
<td>59.77</td>
<td>322.70</td>
<td>2.9</td>
<td>15</td>
<td>328.40</td>
<td>61.06</td>
<td>5.70</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 9)</td>
<td>330.38</td>
<td>62.59</td>
<td>322.91</td>
<td>8.21</td>
<td>6</td>
<td>331.99</td>
<td>63.86</td>
<td>9.08</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 8)</td>
<td>330.80</td>
<td>62.92</td>
<td>324.03</td>
<td>7.54</td>
<td>6</td>
<td>332.09</td>
<td>63.94</td>
<td>8.05</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 7)</td>
<td>330.60</td>
<td>62.76</td>
<td>324.66</td>
<td>6.76</td>
<td>5</td>
<td>331.69</td>
<td>63.62</td>
<td>7.03</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 6)</td>
<td>331.05</td>
<td>63.12</td>
<td>325.87</td>
<td>5.94</td>
<td>5</td>
<td>332.01</td>
<td>63.88</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 5)</td>
<td>330.62</td>
<td>62.78</td>
<td>326.26</td>
<td>5.00</td>
<td>3</td>
<td>331.32</td>
<td>63.33</td>
<td>5.07</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 4)</td>
<td>331.06</td>
<td>63.13</td>
<td>327.55</td>
<td>4.00</td>
<td>3</td>
<td>331.66</td>
<td>63.60</td>
<td>4.11</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 3)</td>
<td>330.75</td>
<td>62.88</td>
<td>328.13</td>
<td>2.99</td>
<td>2</td>
<td>331.19</td>
<td>63.23</td>
<td>3.06</td>
</tr>
<tr>
<td></td>
<td>+ FBP (λ = 2)</td>
<td>331.30</td>
<td>63.32</td>
<td>329.60</td>
<td>1.98</td>
<td>1</td>
<td>331.63</td>
<td>63.58</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 2)</td>
<td>326.34</td>
<td>59.50</td>
<td>322.55</td>
<td>5.35</td>
<td>32</td>
<td>328.51</td>
<td>61.14</td>
<td>5.96</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 3)</td>
<td>326.12</td>
<td>59.34</td>
<td>321.29</td>
<td>6.41</td>
<td>32</td>
<td>328.73</td>
<td>61.31</td>
<td>7.44</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 4)</td>
<td>326.01</td>
<td>59.26</td>
<td>319.49</td>
<td>7.58</td>
<td>32</td>
<td>329.03</td>
<td>61.54</td>
<td>9.54</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 5)</td>
<td>326.04</td>
<td>59.28</td>
<td>318.55</td>
<td>8.08</td>
<td>32</td>
<td>329.31</td>
<td>61.76</td>
<td>10.76</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 6)</td>
<td>325.97</td>
<td>59.23</td>
<td>317.39</td>
<td>8.51</td>
<td>32</td>
<td>329.48</td>
<td>61.89</td>
<td>12.09</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 7)</td>
<td>326.08</td>
<td>59.31</td>
<td>316.42</td>
<td>8.78</td>
<td>32</td>
<td>329.76</td>
<td>62.10</td>
<td>13.34</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 8)</td>
<td>326.35</td>
<td>59.51</td>
<td>315.31</td>
<td>8.99</td>
<td>32</td>
<td>330.33</td>
<td>62.55</td>
<td>15.02</td>
</tr>
<tr>
<td></td>
<td>Ours (λ = 9)</td>
<td>326.47</td>
<td>59.60</td>
<td>313.99</td>
<td>9.03</td>
<td>32</td>
<td>330.58</td>
<td>62.75</td>
<td>15.49</td>
</tr>
</tbody>
</table>
D Additional Qualitative Examples

D.1 Interpolation between Prior Samples
We randomly sample 10 pair of source and target latent code from the standard Gaussian prior and do linear interpolation. For the sampled and interpolated latent codes, we do greedy decoding. The results are shown in Table 12.

D.2 Interpolation between Posterior Samples
We randomly sample 10 pairs of source and target input sentences from the test set of SNLI. For each input sentence, we randomly sample a latent code from the approximated posterior $q(z|x)$. Then we linearly interpolate between each pair of sampled source/target latent code. For the sampled and interpolated latent codes, we do greedy decoding. The results are shown in Table 13.
<table>
<thead>
<tr>
<th>EXAMPLE 1</th>
<th>EXAMPLE 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>the kids are playing hide and seek in the classroom</td>
<td>a man is jumping off a rock into the air</td>
</tr>
<tr>
<td>the girl is about to play the drums</td>
<td>a man is sitting on a bench with a red umbrella</td>
</tr>
<tr>
<td>the girl is about to play in the sandbox</td>
<td>a man is sitting on a bench with a red umbrella</td>
</tr>
<tr>
<td>the girls are watching tv in the classroom</td>
<td>a man is sitting on a bench in front of a building</td>
</tr>
<tr>
<td>the girl is eating cake in the kitchen</td>
<td>a young man is sitting on a bench outside</td>
</tr>
<tr>
<td>the women are watching tv in the bar</td>
<td>a large group of people are taking pictures of a building</td>
</tr>
<tr>
<td>the women are eating lunch</td>
<td>a large group of people are taking pictures in the street</td>
</tr>
<tr>
<td>the women are eating dinner</td>
<td>a large group of people are taking pictures in the street</td>
</tr>
<tr>
<td>a woman is eating in a restaurant</td>
<td>a woman is sitting on a bench in front of a building</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXAMPLE 2</th>
<th>EXAMPLE 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>a man with a cane is walking down the street</td>
<td>two men are swimming in the ocean</td>
</tr>
<tr>
<td>a man with a cane is walking down the street</td>
<td>two men are swimming in the ocean</td>
</tr>
<tr>
<td>a man in a hat and jeans is walking down a sidewalk</td>
<td>two men are on the beach</td>
</tr>
<tr>
<td>man in a blue shirt is eating food</td>
<td>two people are at the beach</td>
</tr>
<tr>
<td>people are eating food</td>
<td>the man is at the park</td>
</tr>
<tr>
<td>people walk through a city street</td>
<td>a man is at the beach</td>
</tr>
<tr>
<td>people walk in a city</td>
<td>a man is at the beach</td>
</tr>
<tr>
<td>people walk in a city</td>
<td>a man is at the beach</td>
</tr>
<tr>
<td>people are outside in a city</td>
<td>a man is taking pictures of the ocean</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXAMPLE 3</th>
<th>EXAMPLE 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>the man is going to the bathroom</td>
<td>two people in bathing suits are in a park</td>
</tr>
<tr>
<td>the man is going to the bathroom</td>
<td>two people in bathing suits are in a park</td>
</tr>
<tr>
<td>the man is going to the bathroom</td>
<td>two people in blue shirts are in a field</td>
</tr>
<tr>
<td>the man is playing music in the living room</td>
<td>two people in blue shirts are in a field</td>
</tr>
<tr>
<td>the man is playing with his dog</td>
<td>two people are at a beach</td>
</tr>
<tr>
<td>the man is playing with a cat</td>
<td>two people are at a beach</td>
</tr>
<tr>
<td>the man is playing with a ball</td>
<td>a man in a blue shirt is eating food</td>
</tr>
<tr>
<td>the man is playing with a ball</td>
<td>a man in a blue shirt is playing basketball</td>
</tr>
<tr>
<td>the man is playing with a ball</td>
<td>a man in a blue shirt is playing basketball</td>
</tr>
<tr>
<td>the man is playing with a ball</td>
<td>a man sits at a carnival</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXAMPLE 4</th>
<th>EXAMPLE 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>a person is about to get a picture taken</td>
<td>two women are outside</td>
</tr>
<tr>
<td>a person is about to get a picture taken</td>
<td>two women are outside</td>
</tr>
<tr>
<td>a person is about to get a picture taken</td>
<td>two women are outside</td>
</tr>
<tr>
<td>a person is about to get a picture taken</td>
<td>two women are sitting outside</td>
</tr>
<tr>
<td>a person is waiting for a friend to come</td>
<td>a couple is sitting outside</td>
</tr>
<tr>
<td>a person is trying to find a speech</td>
<td>a couple is sitting outside</td>
</tr>
<tr>
<td>the people are playing monopoly</td>
<td>a man is reading a newspaper in a park</td>
</tr>
<tr>
<td>there are people playing monopoly</td>
<td>a man is reading a book in a park</td>
</tr>
<tr>
<td>there are people performing surgery</td>
<td>a man is carrying a bag of food</td>
</tr>
<tr>
<td>there are no people in this picture</td>
<td>a man is carrying a bag of food</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXAMPLE 5</th>
<th>EXAMPLE 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>a girl sits on a bench in front of a large crowd</td>
<td>the man is climbing the mountain</td>
</tr>
<tr>
<td>a girl sits on a bench in front of a large crowd</td>
<td>the man is making a noise</td>
</tr>
<tr>
<td>a man sits on a bench in front of a large crowd</td>
<td>the man is making a noise</td>
</tr>
<tr>
<td>a woman wearing a blue shirt is walking on the beach</td>
<td>the people are enjoying the sunshine</td>
</tr>
<tr>
<td>a woman wearing a blue shirt is walking on the beach</td>
<td>the people are watching a movie</td>
</tr>
<tr>
<td>the woman is wearing a blue shirt</td>
<td>the women are eating a meal</td>
</tr>
<tr>
<td>the woman is wearing a blue shirt</td>
<td>the men are watching a movie</td>
</tr>
<tr>
<td>the woman is wearing a blue shirt</td>
<td>two women sit in a circle together</td>
</tr>
<tr>
<td>the woman is moving her legs</td>
<td>two women sit in a circle together</td>
</tr>
</tbody>
</table>
Table 13: Interpolation between posterior samples on SNLI.

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Example 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Input</strong></td>
<td>a child is eating with utensils.</td>
</tr>
<tr>
<td><strong>Target Input</strong></td>
<td>a youth wearing a blue and red jersey and yellow helmet is crouching in a football position.</td>
</tr>
<tr>
<td><strong>Interpolation</strong></td>
<td>a little girl is eating at a restaurant. a little girl is playing with a toy. a little girl is playing with a ball. a little girl is wearing a pink shirt. a little girl is wearing a pink shirt and holding a popsicle. a man is wearing a blue shirt and a hat. a man is wearing a black shirt and black pants.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Source</strong></td>
<td>a little girl is eating at a restaurant. the two men are wearing jeans and a blue shirt. a woman are holding a rope.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Target</strong></td>
<td>a man wearing a black shirt and black pants is standing on a sidewalk.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 2</th>
<th>Example 7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Input</strong></td>
<td>the men are feeling competitive. both men are wearing similar colors.</td>
</tr>
<tr>
<td><strong>Target Input</strong></td>
<td>a young woman is sitting in a field.</td>
</tr>
<tr>
<td><strong>Interpolation</strong></td>
<td>a man is eating a pizza. a man is wearing a blue shirt. a man is wearing a blue shirt. a man is wearing a blue shirt.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Source</strong></td>
<td>a little girl is eating at a restaurant. the two men are wearing jeans and a blue shirt. a woman are holding a rope.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Target</strong></td>
<td>a woman is standing in front of a large crowd.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
<th>Example 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Input</strong></td>
<td>the animals are near the water. a crowded city street in asia.</td>
</tr>
<tr>
<td><strong>Target Input</strong></td>
<td>a young woman is sitting in a field. a huge animal surrounding a small child in the background.</td>
</tr>
<tr>
<td><strong>Interpolation</strong></td>
<td>two boys are at a beach. two boys are at a beach. two boys are at a beach. two boys are at a beach.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Source</strong></td>
<td>two boys are at a beach. a bunch of people are in a park. a woman is looking at a man in a wheelchair. the children are at the beach. the children are looking at the sky.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Target</strong></td>
<td>two boys are at a beach. a bunch of people are in a park. a woman is looking at a man in a wheelchair. the children are at the beach. the children are looking at the sky.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 4</th>
<th>Example 9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Input</strong></td>
<td>one young child in a swimsuit jumping off a blue inflatable slide. a man with 5 small children.</td>
</tr>
<tr>
<td><strong>Target Input</strong></td>
<td>a girl swings from a rope swing in front. a man works on a circuit as he monitors the progress on a tablet device.</td>
</tr>
<tr>
<td><strong>Interpolation</strong></td>
<td>a young man in a blue shirt and black pants is standing by a large rock formation. a woman in a pink shirt and black shorts is jumping into a pool. a woman in a pink shirt and black shorts is jumping into a pool. a woman in a pink shirt and black shorts is jumping into a pool.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Source</strong></td>
<td>a young man in a blue shirt and black pants is standing by a large rock formation. a woman in a pink shirt and black shorts is jumping into a pool. a woman in a pink shirt and black shorts is jumping into a pool. a woman in a pink shirt and black shorts is jumping into a pool.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Target</strong></td>
<td>a woman watches a man play a game of soccer.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 5</th>
<th>Example 10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Input</strong></td>
<td>a girl with glasses next red white and blue flags. a girl watching a game.</td>
</tr>
<tr>
<td><strong>Target Input</strong></td>
<td>a girl sitting at a table. the girl is drinking water with a bucket. the girl is using a camera.</td>
</tr>
<tr>
<td><strong>Interpolation</strong></td>
<td>the girl is drinking milk with the camera. the girl is drinking milk with her hands. the girl is drinking water with a bucket. the girl is using a camera.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Source</strong></td>
<td>the girl is drinking milk with the camera. the girl is drinking milk with her hands. the girl is drinking water with a bucket. the girl is using a camera.</td>
</tr>
<tr>
<td><strong>Posterior-Sampled Target</strong></td>
<td>two girls are outside with a blue umbrella. two girls are outside with a blue umbrella. two girls are outside with a dog.</td>
</tr>
</tbody>
</table>

3605