Dear Reviewers,

Thanks a lot for your helpful and insightful comments. We read them carefully and tried our best to address the issues raised in the comments. Below are our responses to the key issues in the reviews:

To Reviewer #1:

1. (For the problem that "How to ensure the classifier has the optimal set of parameters as z is changed") The classifier's parameters are kept fixed as z is changed. We train the autoencoder and the classifier on the training set, which is diverse and contains texts of varying degrees of attributes, reflected by the different confidence values given by the classifier. Thus we believe when the text in the test set is encoded into the same latent space as the training set, the classifier can continue to provide the optimal direction of modification, even when z changes.

2. (For the problem that "It's unclear what the notion 'degree of transfer' means") The 'degree of transfer' here means the apparent extent of the desired attributes embodied in the transfer results. Taking sentiment transfer as example, it is reflected in the sentimental intensity felt from the text (e.g., weak positive, moderate positive and strong positive). Different from most previous work that only provides binary control over attributes, one advantage of our model is the ability to give control over the degree of attribute transfer desired. Following the evaluation settings of previous work, we use various automatic evaluation indicators (ie, Acc, BLEU, and PPL in Section 4.1) to automatically evaluate transfer results from different aspects. Particularly, 'Acc' is used to evaluate the attribute's accuracy. In Figure 2, by showing the values of three automatic indicators under different fixed modification weights, we demonstrate the influence of the weight on controlling the transfer results. Sorry for the misunderstanding, we will add more words to explain this.

3. Thanks for your suggestion, we will add details about the interaction between the autoencoder and the classifier into Algorithm 1.

4. (For the problem that "What does 'this may undermine the integrity' mean ?") The word "integrity" here means naturality. For example, some phrase-based methods directly delete/replace/insert some sentimental words/phrases, which may result in the generated sentences that are unnatural and not as likely to be written by the human (can be partly reflected by the automatic indicator 'PPL'). Sorry for the misunderstanding, we will add more explanations.

5. Thank you for your careful review, we will further verify our model on more involved Multi-Aspect Sentiment Transfer dataset and multi-field conditioned generation dataset by Ficler and Goldberg 2017. We will improve our paper by adding more explanations about "degree of transfer" and "Automatic Evaluation" results, adding more comprehensive references, and carefully proofreading the paper.

To Reviewer #2:

1. (For concerns #1 and #2) We will compare our model with some simpler autoencoder architectures (e.g., LSTM, GRU), and show the BLEU scores between input and reconstructions. We will add more ablation&comparison experiments, to show the gains of different parts of our model.

2. (For concerns #3) Thanks for your insightful suggestions, we will add references about activation maximization including PPGN, DeVise, etc., and provide a more comprehensive related work section.

To Reviewer #3:

1. (For the problem that "how human evaluation was performed") For each test sample, we hired 3 workers to annotate and average the scores. Moreover, we divided the test samples of each model on each task into the same number of small sets, and the same person annotated the same task for all the models. Due to a large number of samples to be labeled, we did not perform an A/B test, and we will add it if necessary. There was some test data collected for acquire human references for attribute transfer, and the references written by human for BLEU calculation are provided by previous work (Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer.). Thanks for your suggestions, we will provide a clearer explanation to explain this.

2. (For the problem that "The argument about the correlation between human evaluation and automatic evaluation") The results of human evaluation and automatic evaluation are not always consistent with each other. In this study, we follow the evaluation settings of previous works to provide comprehensive results with various indicators. Thank you for your careful review, we will remove this argument about the correlation but provide more detailed evidence and analysis.