We would like to thank all the reviewers for your insightful and constructive reviews and your commendation for the theoretical soundness, novelty, and effectiveness of our method. Here are the responses to some of your concerns:

Reviewer #1: (1) **Concerns on the quality of generated explanations.** Good point! The cohen's kappa score of 0.51 is moderate. Kappa scores for human evaluation of natural language generation tasks are generally moderate (between 0.4-0.6) as it is relatively difficult for human annotators to accurately rate machine-generated text. We also conducted another quantitative human evaluation based on pairwise comparison and observe similar results. We have presented some qualitative results (generated explanations) in the Appendix. (2) **Merge the results of supervised and semi-supervised setting.** Thanks for the suggestion. We will merge this in the revised version. (3) **Data Splits, SoTA, and Baselines.** We follow the dataset split used in [7] throughout all experiments. We chose the baselines that use the same backbone model and did not include the ones with additional resources such as knowledge bases. Our model compares similarly or slightly worse compared with the SoTA results that leverage additional resources or larger backbone models. We will add the SoTA results for reference in the revised version. (4) **Clarification for Fig2&3 .** The results in Fig2&3 are conducted on the Restaurant datasets of the ASC task and are compared against the most competitive baselines. (5) **More clarification on the experiments**. orig + rand exp refers to the model with the combination of original explanations and the corrupted explanations. The corrupted explanations (w. 80% Rand Word) is worse than the baseline model, showing the importance of good explanations. The BERT+self-training baseline denotes the baseline BERT-base model trained with the self-training method.

Reviewer #2: (1) **Technical Contributions.** We develop a novel EM algorithm to jointly train a natural language explanation generation module and explanation-augmented prediction module, which mutually enhance each other. Specifically, the explanation-based classifier can help improve the explanation generator as it is able to identify some high quality of explanations for training the generator (see Eq.(4)). (2) **Small samples of explanations.** The number of annotated explanations is indeed small because we want to focus on the more realistic scenario where only a few annotated explanations are available. However, the annotated explanations cover all the relation types/sentiments in RE/ASC tasks so that annotated explanations are still representative. (3) **Details on the explanation generation.** For generating NL explanation after prediction, we use the variational distribution $q(e|x, y)$ for generating the explanation given the input sentence x and predicted label y. The same language model backbone and beam search are used for the generation.

Reviewer #3: (1) **Experiments on Data Sets e-SNLI [3] or Cos-E [4]**. Thanks very much for pointing out these very relevant data sets. We will add the results on these data sets in the revised version. (2) **Performance w.r.t. Number of Retrieved Explanations.** This is a good point. We change the number of retrieved explanations to 5 and 15, and test them on the Restaurant dataset. The F1 are 77.3 and 75.1 respectively, which are worse than 77.8 with 10 explanations reported in this paper. We will add the sensitivity analysis w.r.t. the number of retrieved explanations in the revised version.

Reviewer #4: (1) **Explanations as Latent Variables**. Note that our goal lies in both generating explanations and providing extra supervision through natural language explanations. Treating explanations as latent variables allow the learning and inference process to be connected in a principled way via explanation generation and explanation-augmented prediction. The label-based explanation used during training is actually used for approximating the posterior distribution of explanations for training the explanation generator. The predictor is based on retrieved explanations that are not based on the label of the input. The final explanation of a prediction is ad-hoc (generated based on both the input and the prediction), this enables generating more informative explanations. The explanation retrieval process is important for the predictor in M-step since it conveys additional supervision. Therefore, explanations not only serve the role of explaining but also serve as additional supervision. For the human evaluation of the quality of generated explanations, a kappa coefficient of 0.51 is ok for evaluating natural language generation results as it is hard for human annotators to perfectly rate machine-generated texts. Thanks for the caveat, we admit the necessity of investigating other aspects of explainability to ensure our model's alignment with societal desiderata and leave debiasing, risks mitigation and adversarial defense as our future work. (2) **Overclaiming on explanations**: In our generic explainable framework, we use existing explanations in [7] for simplicity. Following works like [7,13], we use "logic rules" and "labeling function" interchangeably since they convey information about labels. Our model could potentially learn to explain during testing in the example you provided if faithful natural language explanations in more complicated logical forms are provided, e.g. sent: "anything here is wonderful but the food isn't", exp: The word "wonderful" and "but" occur before the term "food" and "isn't" occurs after the term "food". We'll clarify this issue in the final version to avoid overclaiming. Thanks for the comments and we have revised the label in Fig 1 accordingly in our next draft.