Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations

Jun Gao[1]*, Yuhan Liu[1]*, Haolin Deng[1], Wei Wang[3],
Yu Cao[4], Jiachen Du[1], Ruifeng Xu[1,2]†
[1]Harbin Institute of Technology (Shenzhen), China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Shenzhen International Graduate School, Tsinghua University
[4]School of Computer Science, The University of Sydney
{jgao95,yhanliu}@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Current approaches to empathetic response generation focus on learning a model to predict an emotion label and generate a response based on this label, and have achieved promising results. However, the emotion cause, an essential factor for empathetic responding, is ignored. The emotion cause is a stimulus for human emotions. Recognizing the emotion cause is helpful to better understand human emotions to generate more empathetic responses. To this end, we propose a novel framework that improves empathetic response generation by recognizing emotion cause in conversations. Specifically, an emotion reasoner is designed to predict a context emotion label and a sequence of emotion cause-oriented labels, which indicate whether the word is related to the emotion cause. Then we devise both hard and soft gated attention mechanisms to incorporate the emotion cause into response generation. Experiments show that incorporating emotion cause information improves the performance of the model on both emotion recognition and response generation.

1 Introduction

In recent years, open-domain dialogue systems are becoming increasingly ubiquitous and have been extensively leveraged for mental healthcare and entertainment (Oh et al., 2017; Zhou et al., 2020; Sharma et al., 2020). In part, this progress is driven by advances in neural response generation models (Vinyals and Le, 2015; Li et al., 2016a,c; Gao et al., 2019a,b) which have shown success in generating fluent and relevant responses, given a wide variety of user inputs. However, people can still feel a clear gap between humans and machines when conversing with them. One of the primary reasons is that existing dialogue systems lack emotion understanding and empathy (Rashkin et al., 2019). Empathetic responding is a desirable communicative skill that can make more natural communication in daily conversations (Callender, 2015). Table 1 shows an example of empathetic responding from *empathetic-dialogues* dataset (Rashkin et al., 2019). A speaker is talking about a situation that happened to him/her related to a lonely feeling and a listener needs to respond with an appropriate emotion. Therefore, empathy is important in conversations. However, endowing dialogue systems with the capability of emotion understanding and empathetic responding is challenging.

Emotion: Lonely
Context:
Speaker: I feel so lonely sometimes because all my friends live in a different country
Listener: Oh, I'm sure you are lonely. Maybe you can join some kind of club that lets you meet new friends?
Speaker: I was thinking about it! I wanted to join a group for local moms
Target: That's a good idea! This way you can meet friends for yourself, also maybe for your children!

Table 1: An example of empathetic responding from *empathetic-dialogues* dataset. An empathetic dialogue model is required to generate an appropriate response given the dialogue context. The utterance highlighted in blue contains the emotion cause.

Most of the existing approaches improve empathetic response generation from two directions. The first usually promotes the model's emotion understanding (Lubis et al., 2018; Rashkin et al., 2019; Lin et al., 2019; Li et al., 2020b). In this line of work, models are often trained to predict an emotion state of the speaker and generate a response based on the emotion state. The second focuses on improving response generation strategy (Welivita and Pu, 2020; Shin et al., 2020; Majumder et al., 2020). For example, Shin et al. (2020) proposes to use the look-ahead of user emotion to model empathetic response generation and improve the empathetic responding model via Reinforcement Learning. Majumder et al. (2020) presents an ap-

*Equal Contribution
†Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 807–819
November 7–11, 2021.

proach to mimic the emotion of the speaker while accounting for their affective polarity.

However, both kinds of existing methods only consider using the surface information of emotions such as emotion labels to improve the quality of generated responses. The emotion cause, an essential factor for empathetic responding, is ignored. We argue that such surface information of emotions is not sufficient for empathetic responding. The model can better understand human emotions and respond empathetically if it has the ability to perform reasoning about emotions in conversations, which means it needs to identify the cause of a certain emotion. For example in Table 1, given the dialogue context, we need to recognize not only the emotion "lonely" of the speaker, but also the emotion cause behind the emotion. We can see that the speaker is lonely due to the event "... all friends live ... different country". Here, we could infer that the speaker's emotion is caused by the first utterance containing the aforementioned event. With such deep emotional information, we can generate more relevant and empathetic responses.

To this end, we propose a novel framework to improve empathetic response generation by endowing the empathetic dialogue model with the ability to reason about human emotions in conversations. Specifically, our model is able to identify the cause behind the emotions in addition to the types of emotions. Our framework involves two components, an emotion reasoner and a response generator. The emotion reasoner first performs a context-level emotion prediction and a word-level emotion cause detection, providing emotional information for response generation. The response generator then makes use of such deep emotional information to generate empathetic responses. To incorporate emotion cause information into the response generator, we devise a gated attention mechanism and explore both hard and soft gating strategies to allow the model to focus more on words related to the emotion cause. For model training, we use multi-task learning to build the connection between the emotion reasoner and the response generator.

Our contributions can be summarized as follows:

- An emotion reasoner is designed to recognize the context emotion of the speaker and the emotion cause behind the emotion, providing deep emotional information for response generation. To the best of our knowledge, this is the first work that investigates emotion cause in empathetic response generation.
- To incorporate emotion cause into response generation, we devise a gated attention mechanism and explore both hard and soft gating strategies, which allow the model to focus on emotion cause related words.
- Experimental results show that our proposed models benefit from the emotion cause and significantly outperform other compared methods, resulting in more empathetic responses.

2 Related Work

In recent years, neural approaches to open-domain dialogue systems have achieved great progress (Serban et al., 2016; Wolf et al., 2019; Zhang et al., 2020b; Zhou et al., 2020; Xu et al., 2020; Wang et al., 2021). Especially, incorporating personality and emotional features can make dialogue systems more human-like. For emotion-aware response generation, it aims at generating responses corresponding to specific emotions. Several methods are proposed to tackle this task (Zhou et al., 2018; Huang et al., 2018; Colombo et al., 2019; Song et al., 2019; Shen and Feng, 2020; Xu et al., 2021; Majumder et al., 2021).

Empathetic response generation is a sub-task of emotion-aware response generation, Rashkin et al. (2019) first proposes a standard benchmark that contains large-scale empathetic conversations. Lin et al. (2020) adapts GPT2 (Radford et al., 2019) to generate empathetic responses via transfer learning and continues to improve its response quality via active learning and negative training. Welivita and Pu (2020) develops a taxonomy of empathetic listener intents by human judges to generate more controlled and interpretable responses. Shin et al. (2020) utilizes reinforcement learning to improve the empathetic responding model, in which the model is rewarded with an estimated user sentiment look-ahead. Lin et al. (2019) models empathy in conversations through Mixture of Experts (Shazeer et al., 2017) and gets final output based on emotion distribution. Majumder et al. (2020) argues that empathetic response generation can mimic the emotion of the speaker, and introduces the emotion stochastic sampling strategy during training. Li et al. (2020b) leverages multi-type knowledge to enrich the dialogue history so that the model can

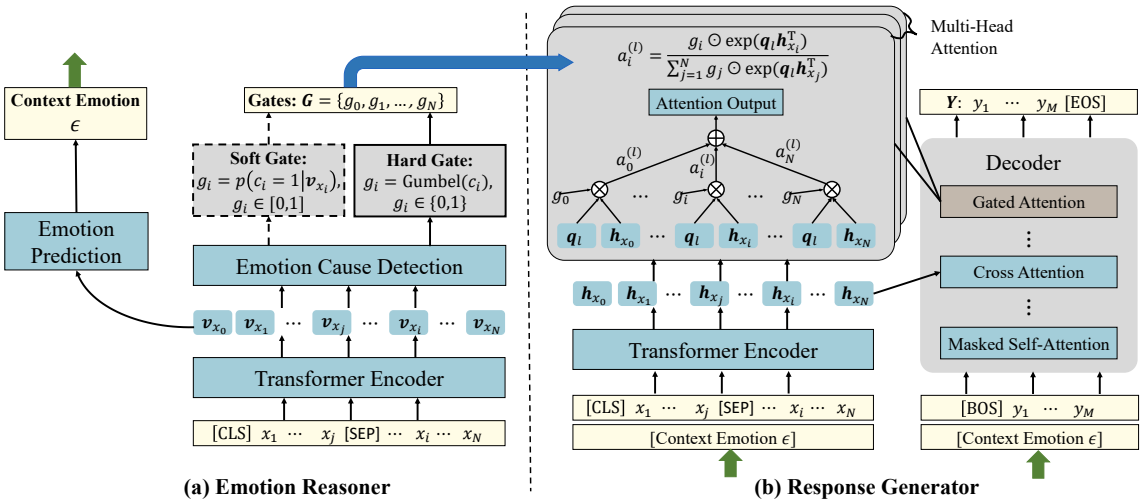

Figure 1: Architecture of the proposed framework. Our framework contains two components: an emotion reasoner (a) and a response generator (b). The emotion reasoner is used to predict a context emotion label and locate words related to the emotion cause, based on the dialogue context. The response generator makes use of the emotional information obtained from the emotion reasoner to generate the response. Specifically, a gated attention mechanism is designed to incorporate emotion cause information into the response generator.

accurately perceive and respond to implicit emotions. Li et al. (2020a) exploits user feedback and multi-granularity emotion, and introduces an adversarial learning framework to capture the nuances of user emotion.

Emotion cause extraction (ECE), aims at exploring the reason for emotion change and what causes a certain emotion. Lee et al. (2010); Chen et al. (2010) first define it as a word-level and clause-level task respectively. Gui et al. (2016) proposes the first open dataset for ECE, and it serves as a standard benchmark up till now. Xia and Ding (2019) reforms ECE into emotion-cause pair extraction task. Similar to ECE, Poria et al. (2020) first introduces the task of recognizing emotion cause in conversations.

3 Task Formulation

We formulate the task of empathetic response generation as follows. Given a dialogue context $\boldsymbol{\mathcal{M}} = \{\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_L\}$ of L utterances and each utterance $\boldsymbol{U}_i = \{w_1^i, w_2^i \cdots, w_K^i\}$ consists of K tokens. Following the previous work (Lin et al., 2019; Shin et al., 2020), we concatenate the L utterances together as input. Specifically, we separate utterances by [SEP] tokens and insert a special token [CLS] at the start of the sequence to form an input sequence $\boldsymbol{X} = \{x_0, x_1, \cdots, x_N\}$ (See Figure 1 for example). Therefore, given an input sequence $\boldsymbol{X}$, our goal is to generate an empathetic response $\boldsymbol{Y} = \{y_0, y_1, \cdots, y_M\}$ that is emotionally appropriate and relevant to the dialogue context.

4 Approach

Our framework that explicitly considers the emotion cause for empathetic response generation is shown in Figure 1. Our framework contains two components: an emotion reasoner and a response generator. The emotion reasoner is used to predict a context emotion label and locate words related to the emotion cause, based on the dialogue context. The response generator is responsible for incorporating the information obtained from the emotion reasoner then generating the response. Below we first introduce how we construct training samples for emotion cause detection, then we describe the two components in detail.

4.1 Emotion Cause Annotation

Since we do not have readily available data with emotion cause information on the empathetic dialogue dataset, we leverage an existing emotion cause detection model (Poria et al., 2020) for identifying emotion causes at utterance level in conversations. The model is trained on an open-domain emotional dialogue dataset, namely RECCON (Poria et al., 2020). Given a dialogue context consisting of L utterances and a context emotion label, the goal of emotion cause detection model is to identify which utterance in the dialogue context contains the emotion cause. Note that an emotion may have multiple cause-correlated utterances.

To verify the transfer performance of the detection model on the empathetic dialogue dataset used in our work, we randomly selected 100 dialogue samples from the test set and asked 3 human annotators to assign a label $\in \{0, 1\}$ to each utterance in the dialogue context, indicating whether it is a cause-correlated utterance. The final verdict on each sample is determined by majority voting. On these annotated samples, The emotion cause annotation model finally achieved an accuracy of 89%, indicating that the annotation model has a reliable performance.

In our work, we use an emotion reasoner to perform a word-level emotion cause detection. To achieve this, we automatically assign each word in the dialogue context with a binary label. If the word is in a causal utterance, we annotate it with 1, otherwise 0.

4.2 Emotion Reasoner

The emotion reasoner aims to recognize a context emotion given a dialogue context, as well as the cause behind the emotion. It can be decomposed into two tasks: context emotion prediction and emotion cause detection.

Context Emotion Prediction: The context emotion prediction is a classification problem, aiming at predicting a context emotion label ε based on the dialogue context. Specifically, given an input sequence $\boldsymbol{X}$, we first construct a representation for each word by summing the corresponding word and position embeddings. The word representations are then fed into a transformer encoder to obtain a sequence of contextualized word representation $\boldsymbol{V} = \{\boldsymbol{v}_{x_0}, \boldsymbol{v}_{x_1}, \cdots, \boldsymbol{v}_{x_M}\}$. The context emotion distribution is finally computed based on the representation $\boldsymbol{v}_{x_0}$ of the first special token ([CLS]) as follows:

$$\mathcal{P}(\varepsilon|\boldsymbol{X}) = \text{softmax}(\boldsymbol{W}_e \boldsymbol{v}_{x_0} + b_e), \quad (1)$$

where $\boldsymbol{W}_e$ and b_e are trainable parameters.

Emotion Cause Detection: In our work, we perform a word-level emotion cause detection, which can provide word-level emotional features for response generation. We formulate the emotion cause detection as a sequence labeling problem, where each word in the sequence is labeled with an emotion cause-oriented label $\in \{0, 1\}$, indicating whether the word is related to the emotion cause. Note that the [CLS] token is always labeled with 1. The sequence of emotion cause-oriented labels will later be used as gating controllers to select the emotion cause-related words in the input sequence to attend to for the response generator.

Formally, given an input sequence $\boldsymbol{X} = \{x_0, x_1, \cdots, x_N\}$, the output of this task is a sequence of emotion cause-oriented labels $\boldsymbol{C} = \{c_0, c_1, \cdots, c_N\}$. We compute the probability c_i of the i-th word related to the emotion cause with a linear layer coupled with a softmax function:

$$\mathcal{P}(c_i|\boldsymbol{v}_{x_i}) = \text{softmax}(\boldsymbol{W}_c \boldsymbol{v}_{x_i} + \boldsymbol{b}_c), i \in N \quad (2)$$

where $\boldsymbol{W}_c$ and $\boldsymbol{b}_c$ are trainable parameters. To jointly model context emotion prediction and emotion cause detection, the objective is formulated as:

$$\mathcal{P}(\varepsilon, \boldsymbol{C}|\boldsymbol{X}) = \mathcal{P}(\varepsilon|\boldsymbol{v}_{x_0}) \prod_{i=1}^{N} \mathcal{P}(c_i|\boldsymbol{v}_{x_i}) \quad (3)$$

The parameters of the emotion reasoner can be learned by optimizing a negative log likelihood (NLL) loss defined as:

$$\mathcal{L}_r = -\log \mathcal{P}(\varepsilon|\boldsymbol{v}_{x_0}) - \sum_{i=1}^{N} \log \mathcal{P}(c_i|\boldsymbol{v}_{x_i}) \quad (4)$$

4.3 Response Generator

With the predicted context emotion ε and the emotion cause-oriented labels $\boldsymbol{C}$ obtained from the emotion reasoner, the response generator aims to generate an empathetic response $\boldsymbol{Y} = \{y_1, \cdots, y_M\}$ that is emotionally appropriate and relevant to the dialogue context through maximizing the probability $\mathcal{P}(\boldsymbol{Y}|\boldsymbol{X}, \varepsilon, \boldsymbol{C})$. The basis for our response generator is a Transformer network, which consists of an encoder and a decoder. Next, we describe how we incorporate the emotional information including the context emotion ε and the emotion cause-oriented labels $\boldsymbol{C}$ into the response generator.

Input Representation: To fuse the context emotion label ε into the response generator, we leverage trainable emotion embeddings $\boldsymbol{E}_\varepsilon \in \mathbb{R}^{n_{emo} \times d_{\text{model}}}$ to represent each context emotion label, where $n_{emo} = 32$. Then each input word of the encoder and the decoder is represented as a sum of three embeddings: word embedding $\boldsymbol{E}_w$, positional embedding $\boldsymbol{E}_p$ and emotion embedding $\boldsymbol{E}_\varepsilon$. We feed the representations of the input sequence $\boldsymbol{X}$ into the encoder to obtain contextualized word representations of the input sequence $\boldsymbol{H} = \{\boldsymbol{h}_{x_0}, \boldsymbol{h}_{x_1}, \cdots, \boldsymbol{h}_{x_N}\}$, which provide context information for the decoder.

Applications of Attention In Transformer: As proposed by Vaswani et al. (2017), a multi-head attention function maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The multi-head attention is typically used in two different ways: (1) both the encoder and the decoder contain "Self Attention" layers, where the quires, keys and values come from the output of the previous layer in the encoder/decoder; (2) in a "Cross Attention" layer, the queries come from the previous self attention layer, and the output $\boldsymbol{H}$ of the encoder are used as the keys and values. The "Cross Attention" layer is only used by the decoder.

Gated Attention Mechanism: In our work, to leverage emotion cause information, we devise a "Gated Attention" layer on top of the cross attention layer in the decoder, where the queries come from the cross attention layer and the keys and values come from the output $\boldsymbol{H}$ of the encoder. The gated attention mechanism utilizes a sequence of gates $\boldsymbol{G} = \{g_0, g_1, \cdots, g_N\}$ to dynamically select elements related to the cause from input, then the decoder is forced to pay more attention to these selected elements, which give important information on the context emotion. We will later describe how we obtain the sequence of gates $\boldsymbol{G}$. For a single-head attention layer of the l-th block in the decoder, the gated attention weight $a_i^{(l)}$ for i-th position can be computed by:

$$a_i^{(l)} = \frac{g_i \odot \exp(\boldsymbol{q}_l \boldsymbol{h}_{x_i}^\top)}{\sum_{j=1}^{N} g_j \odot \exp(\boldsymbol{q}_l \boldsymbol{h}_{x_j}^\top)}, \tag{5}$$

where g_i is the gate for i-th position, $\boldsymbol{q}_l$ is the output of the l-th cross attention layer and $\boldsymbol{h}_{x_i}$ is the contextualized word representation at i-th position.

The sequence of gates $\boldsymbol{G}$ is used to force the decoder to pay more attention to important words from the input. A straightforward way is to use a binary gate $g_i \in \{0, 1\}$ to decide whether the decoder should pay attention to the i-th word. For the position with $g_i = 1$, the attention weights $a_i^{(l)}$ are non-zeros. On the other hand, for the positions with $g_i = 0$, we have $a_i^{(l)} = 0$. We refer to this gating strategy as "hard gating strategy". However, the hard gating strategy is rather rigid. If the model chooses the wrong words, then important information will be ignored. An alternative method is to use "soft gating strategy" where each gate g_i is a continuous value ranging from 0 to 1, indicating how much information of the contextualized word representations at i-th position should be used. The soft gating strategy is more flexible compared with the hard gating strategy. In our work, we explore both soft and hard gating strategies.

Next, we introduce how we compute the sequence of gates $\boldsymbol{G}$. In the soft gating strategy, the i-th gate g_i in the $\boldsymbol{G}$ is defined by $g_i = \mathcal{P}(c_i = 1|\boldsymbol{v}_{x_i})$, which is the probability that the i-th word being related to the emotion cause. The value for soft gating is continuous, ranging from 0 to 1.

In the hard gating strategy, the i-th gate $g_i \in \{0, 1\}$ is a binary label obtained by $g_i = c_i$. To overcome the problem of the inability for back-propagating, we resort to the Gumbel-Softmax trick (Jang et al., 2017). It is a procedure for sampling a categorical one-hot value from the Gumbel distribution, instead of direct sampling from a categorical distribution.

The final loss for the response generator is:

$$\mathcal{L}_g = -\sum_{i=0}^{M} \mathcal{P}(y_i|y_{<i}, \boldsymbol{X}, \varepsilon, \boldsymbol{C}) \tag{6}$$

4.4 Model Training

Our proposed approach consists of two components: the emotion reasoner and the response generator. To better explore their interaction, we solve both tasks together by multi-task learning. The full-fledged loss of the two tasks is computed as:

$$\mathcal{L}_{ml} = \mathcal{L}_r + \mathcal{L}_g \tag{7}$$

We pretrain the emotion reasoner using the objective as defined in Eq. 4 before joint training the two components.

5 Experimental Setup

5.1 Dataset

We use *empathetic-dialogues* (Rashkin et al., 2019) for experiments. The dataset comprises 24,850 open-domain multi-turn conversations between two participators. Specifically, each conversation is grounded by a situation description and a fine-grained emotion. There are 32 emotion categories in total. We use the 8:1:1 train/valid/test subset split following the original dataset definitions.

5.2 Comparison Methods

The following models are selected as baselines: 1) **MoEL** (Lin et al., 2019): a transformer-based seq2seq model which uses several decoders to generate different outputs and softly combines them

Method	BLEU	P_{BERT}	R_{BERT}	F_{BERT}	Dist-1	Dist-2	Accuracy
EmpDG	1.506±0.155	0.116±0.005	0.112±0.012	0.115±0.006	0.010±0.002	0.082±0.020	29.2±0.4
MoEL	1.610±0.041	0.144±0.005	0.123±0.004	0.134±0.001	0.008±0.000	0.074±0.004	36.2±1.1
MIME	1.578±0.068	**0.150**±0.006	0.120±0.007	0.135±0.005	0.008±0.000	0.064±0.005	38.3±1.6
MK-EDG	1.376±0.062	0.144±0.004	0.114±0.002	0.129±0.002	0.008±0.000	0.058±0.002	36.4±1.4
Ours(Hard)	1.734±0.083	0.143±0.003	0.125†±0.002	0.134†±0.001	**0.018**±0.001	**0.090**±0.009	42.3†±0.3
Ours(Soft)	**1.774**±0.063	0.145†±0.003	**0.127**†±0.004	**0.136**†±0.003	0.017±0.003	0.084±0.018	**42.4**†±0.3

Table 2: Results on Automatic Evaluation. For each method, we repeated 5 runs with different seeds and average the results. Standard deviations are given in the small text. The numbers marked with †means the results are statistically significant at $p < 0.01$. All results of different methods can be found in Appendix B.

Method	Fluency	Relevance	Empathy
EmpDG	4.378	2.414	2.444
MoEL	4.422	2.310	2.354
MIME	4.426	2.352	2.394
MK-EDG	4.432	2.422	2.494
Ours(Hard)	4.560	2.904	3.006
Ours(Soft)	**4.584**	**3.096**	**3.244**

Table 3: Results on human ratings. Fleiss kappa of the results is 0.35, which constitutes a fair level of agreement.

according to emotion distributions. 2) **MIME** (Majumder et al., 2020): Another extension of transformer-based model which considers emotion clustering and emotional mimicry. Besides, it also introduces sampling stochasticity during training. 3) **EmpDG** (Li et al., 2020a): an adversarial model which applies two discriminators for interacting with the user feedback. It exploits both coarse-grained dialogue-level and fine-grained token-level emotions for generation. 4) **MK-EDG** (Li et al., 2020b): A contextual-enhanced empathetic dialogue generator that leverages multi-type external knowledge and emotional signal distilling for response generation.

We explore our model using the hard gating strategy and the soft gating strategy, as introduced in Sec 4.3, denoted as **Ours(Hard)** and **Ours(Soft)**. Detailed information about the implementations is covered in Appendix A.

5.3 Evaluation metrics

Automatic Evaluation: Four kinds of automatic metrics are applied for evaluation: 1) BLEU (Papineni et al., 2002) calculates the co-occurrence frequency of n-grams between candidates and references. Following MIME and MoEL, we use BLEU-4. 2) BERTscore (Zhang et al., 2020a) uses embeddings from pre-trained language models to compute a weighted cosine similarity of reference and the generated sentence. We use matching precision, recall and F1 score (R_{BERT}, P_{BERT} and F_{BERT}) in our experiments. 3) Dist-{1,2} (Li et al., 2016b) are diversity metrics aiming at measuring text diversity by calculating the proportion of different grams in the text. 4) To evaluate the model capabilities for emotion understanding, we adopt emotion classification accuracy (Accuracy) to further evaluate model performance.

Human Ratings: Evaluating open-domain dialogue systems is challenging since the lack of reliable automatic evaluation metrics (Gao et al., 2021), thus human judgements are necessary. Following previous works, we randomly sample 100 dialogues and the corresponding generated responses for different models and then ask 5 professional annotators to give each response a rating score from Fluency aspect, Relevance aspect, and Empathy aspect. Each aspect is on a scale from 1 to 5, where 1, 3, and 5 indicate unacceptable, moderate, and excellent performance respectively. In order to keep the anonymization of compared methods, the response order in each sample is totally shuffled.

Human A/B Test: Human A/B test is also conducted. We re-sample another 100 samples and form them into A-vs-B types, where A is our model and B is another baseline. Another 3 annotators are asked to choose the better response for each instance. They can also choose a Tie if both are good or bad. To make sure fairness, each group of A/B test uses a distinct dialogue context.

6 Experimental Results

6.1 Main results

Automatic Evaluation: Table 2 reports the evaluation results on automatic metrics. For each method, we repeated 5 runs with different seeds and average the results. Standard deviations are given

Method	BLEU	P_{BERT}	R_{BERT}	F_{BERT}	Dist-1	Dist-2	Accuracy
Ours(Hard)	1.734±0.083	**0.143**±0.003	0.125±0.002	**0.134**±0.001	**0.018**±0.001	**0.090**±0.009	**42.3**±0.3
w/o ml	1.682±0.053	0.136±0.006	0.121±0.004	0.129±0.002	0.017±0.001	0.087±0.006	42.1±0.2
w/o el	1.716±0.148	0.137±0.003	0.124±0.005	0.131±0.001	0.015±0.002	0.084±0.013	42.1±0.3
w/o ec	1.676±0.104	0.139±0.005	0.119±0.006	0.130±0.004	0.017±0.002	0.083±0.010	38.5±0.5
vs. eLex	**1.770**±0.099	0.132±0.005	**0.126**±0.005	0.130±0.003	0.015±0.002	0.087±0.010	40.8±0.4
Ours(Soft)	**1.774**±0.063	**0.145**±0.003	**0.127**±0.004	**0.136**±0.003	0.017±0.003	0.084±0.018	**42.4**±0.3
w/o ml	1.724±0.076	0.138±0.002	0.123±0.003	0.131±0.002	0.017±0.020	0.086±0.013	42.2±0.2
w/o el	1.656±0.155	0.128±0.017	0.119±0.007	0.124±0.012	0.017±0.020	**0.094**±0.017	42.0±0.2
w/o ec	1.676±0.104	0.139±0.005	0.119±0.006	0.130±0.004	0.017±0.002	0.083±0.010	38.5±0.5
vs. eLex	1.676±0.143	0.130±0.009	0.124±0.008	0.127±0.002	0.017±0.002	0.090±0.016	40.7±0.3

Table 4: Results on ablation study. Here ml, el, ec and eLex are short for multi-task learning, emotion label, emotion cause and emotion lexicon respectively. Note that "Ours(Hard) w/o ec" and "Ours(Soft) w/o ec" are the same model. For each method, we repeated 5 runs with different seeds and average the results. Standard deviations are given in the small text.

Method	Win %	Loss %	Tie %
Ours(Hard) vs EmpDG	38.00	20.33	41.67
Ours(Hard) vs MoEL	40.67	24.33	35.00
Ours(Hard) vs MIME	46.67	21.67	31.67
Ours(Hard) vs MK-EDG	41.00	23.67	31.67
Ours(Soft) vs EmpDG	53.33	17.33	29.33
Ours(Soft) vs MoEL	54.33	19.00	26.67
Ours(Soft) vs MIME	59.00	19.67	21.33
Ours(Soft) vs MK-EDG	52.67	22.33	25.00
Ours(Soft) vs Ours(Hard)	33.00	29.33	37.67

Table 5: Results on A/B test. Fleiss kappa of the results is 0.63, which falls within a generally accepted range of rater agreement.

in the small text. As can be seen from the table, our proposed models Ours(Hard) and Ours(Soft) have a clear advantage over the baseline models in terms of all metrics except the P_{BERT}. This demonstrates that our model generates more appropriate and informative responses by recognizing emotion cause in conversations. We also observe that the difference in performance between Ours(Soft) and Ours(Hard) is not significant, yet each has its own focus. Ours(Soft) outperforms Ours(Hard) on BLEU and BERTScores, while Ours(Hard) has better performance on Dist-1 and Dist-2 ratios. It seems that Ours(Soft) sacrifices diversity for relevance gains.

Human Evaluation: Table 3 presents all the results in terms of human ratings of Fluency, Relevance, and Empathy. We observed in Table 3 that Ours(Soft) and Ours(Hard) significantly outperform most of the baselines in terms of all the three criteria, achieving best and second-best results respectively. This indicates that trying to recognize emotion cause in conversations is beneficial for improving emotional understanding and generating more empathetic responses. Besides, we can see that using soft gating mechanism achieves better performance than using the hard gating mechanism. This can be explained by the fact that the hard gating mechanism is rigid in controlling information, and there is a chance that important information will be ignored. The soft gating mechanism, on the other hand, has more flexible control over the information.

The results on A/B Test which is shown in Table 3 also confirms that the responses generated by Ours(Hard) and Ours(Soft) are preferred by humans compared to the results of other baseline models, and the results of Ours(Soft) are slightly better than those of Ours(Hard).

6.2 Ablation Analysis

In order to assess contributions toward the final performances, we perform a series of ablation studies for Ours(Hard) and Ours(Soft).

Effect of Emotion Label: To assess the contribution of the predicted emotion label which is incorporated into the response generator, the emotion embedding is removed from the input of the encoder and decoder in the response generator. As shown in Table 4, we can see that for both models, removing the emotion embedding causes performance degradation and Ours(Soft) suffers from more pronounced performance degradation in terms of BLEU and BERTScores. This indicates that the information provided by the emotion label has importance in improving response quality.

Effect of Multi-task Learning: Multi-task learning is used to build the connection between the emotion reasoner and the response generator. As shown

Emotion: Excited	Emotion: Impressed
Context: Speaker: i am so happy it 's a boy Listener: you just had a baby ? ! i did n't even know you were expecting ! Speaker: he 'll be here in about 3 months	**Context:** Speaker: my friend purchased a new car . the ride is sweet ! Listener: nice what kind of car is it they purchased Speaker: they ended up getting a honda prelude .
EmpDG:that is a big deal with .	**EmpDG:**wow that is so exciting!
MoEL:that is really nice of you !	**MoEL:**oh wow! that is amazing!
MIME: i am sure you were able to get the kids .	**MIME:** i hope you get your friends .
MK-EDG: i bet he was so excited	**MK-EDG:** wow that is a lot of money
Ours(Hard):i am happy for you. he must be a good boy!	**Ours(Hard):** that's great. i hope you will have one soon!
Ours(Soft): that's so nice . i bet you are going to be a great parent .	**Ours(Soft):** wow, that's cool. i hope they get a great vehicle.

Table 6: Example responses generated by different models, as well as a visualization of the emotion cause detection in our models. Words highlighted in blue are the results of emotion cause detection for Ours(Hard), while words highlighted in purple are the results of emotion cause detection for Ours(Soft). Darker color indicates the higher probability that the word being related to the emotion cause.

in Table 4, the two models trained with multi-task learning achieve better performance in response generation, compared with the two models without using multi-task learning. At the same time, we can find that multi-task training is not very helpful for emotion recognition, and the models only get a small improvement.

Effect of Emotion Cause: To investigate the impact of emotion cause on emotion recognition and empathetic response generation, We remove the emotion-cause related part in the emotion reasoner and the response generator at the same time. The emotion reasoner only performs the emotion recognition task and we remove the gated attention mechanism which is used to incorporate emotion cause information from the response generator. Looking at Table 4, we can clearly see that removing the emotion cause part causes a significant decrease in the performance of both models in terms of response generation and emotion recognition. In particular, the accuracy of emotion recognition drops from 42.4% to 38.5%. This indicates that emotion cause plays an important role in promoting the understanding of emotions, confirming our insights about the emotion cause. The gated attention mechanism can be seen as a denoising technique that allows the model to acquire important information relatively easily.

Emotion Cause vs. Emotion Lexicon: Emotion lexicon also plays an important role in sentiment analysis and empathetic response generation (Li et al., 2020b). To further demonstrate the superiority of the emotion cause, we compare the importance of the emotion cause versus the emotion lexicon. Similarly, we assign a label to each word in the input sequence using NRC-VAD, indicating whether the word is an emotion lexicon. The emotion reasoner performs both emotion recognition and emotion lexicon detection, and the information is then used for response generation. The results shown in the Table 4 indicate that the information provided by emotion cause is more useful for helping the model understand emotions and dialogue context than the surface information of emotions such as emotion classes.

6.3 Case Study

We also present some example responses generated by our models and baseline models in Table 6. As shown in the first example, Ours(Hard) does a good job of identifying words that are relevant to emotion causes. In addition, both Ours(Hard) and Ours(Soft) appear to generate responses that are more empathetic and contextually relevant to the conversation than other baseline models. In the second example, Ours(Soft) again is successful in locating the words associated with the emotion cause. The responses generated by Ours(Hard) and Ours(Soft) are more informative and have a richer expression of affections, while the responses generated by other models are monotonous and lack empathy.

7 Conclusion

In this paper, we presented a novel framework that can incorporate emotion cause information into empathetic response generation. Our approach consists of an emotion reasoner and a response generator. The emotion reasoner first predicts a context emotion label and locating the words in the dia-

logue context which are associated with the emotion cause. The response generator then generates a response with the predicted context emotion label and the emotion cause information. To incorporate the emotion cause information into response generation, we devise a gated attention mechanism and explore both hard and soft gating strategies. Automatic and manual evaluations show that our proposed models can generate more meaningful and empathetic responses.

8 Ethical Considerations

The *empathetic-dialogues* dataset (Rashkin et al., 2019) used in our paper is annotated through Amazon Mechanical Turk, which means it totally protects the privacy of real users. Besides, we make sure anonymization in the emotion cause annotation of this dataset and human evaluation process. We believe our research work meets the ethics of EMNLP.

9 Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (61632011, 61876053, 62006062, 62176076), the Guangdong Province Covid-19 Pandemic Control Research Funding (2020KZDZX1224), the Shenzhen Foundational Research Funding (JCYJ20180507183527919 and JCYJ20200109113441941), China Postdoctoral Science Foundation (2020M670912), Joint Lab of HITSZ and China Merchants Securities.

References

J. Callender. 2015. Being amoral: Psychopathy and moral incapacity. *British Journal of Psychiatry*, 207:274 – 275.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China. Coling 2010 Organizing Committee.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019a. Generating multiple diverse responses for short-text conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6383–6390.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019b. A discrete CVAE for response generation on short-text conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1898–1908, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Wei Bi, Ruifeng Xu, and Shuming Shi. 2021. REAM♯: An enhancement approach to reference-based evaluation metrics for open-domain dialog generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2487–2500, Online. Association for Computational Linguistics.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020a. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qintong Li, Piji Li, Zhumin Chen, and Z. Ren. 2020b. Towards empathetic dialogue generation over multi-type knowledge. *arXiv: Computation and Language*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *AAAI*.

Nurul Lubis, S. Sakti, Koichiro Yoshino, and S. Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *AAAI*.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2021. Exemplars-guided empathetic response generation controlled by the elements of human communication. *arXiv preprint arXiv:2106.11791*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Kyo-Joong Oh, D. Lee, ByungSoo Ko, and H. Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 371–375.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2020. Recognizing emotion cause in conversations. *arXiv preprint arXiv:2012.11820*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*.

Lei Shen and Yang Feng. 2020. CDL: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566, Online. Association for Computational Linguistics.

Jamin Shin, P. Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ArXiv*, abs/1506.05869.

Wei Wang, Piji Li, and Hai-Tao Zheng. 2021. Generating diversified comments via reader-aware topic modeling and saliency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Chen Xu, Jianyu Zhao, Rang Li, Changjian Hu, and Chuangbai Xiao. 2021. Change or not: A simple approach for plug and play language models on sentiment control. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. In *ECAI 2020*, pages 2244–2251. IOS Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

L. Zhou, J. Gao, Di Li, and H. Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46:53–93.

A Implementation Details

Our models are implemented using Pytorch and Texar-PyTorch, which is a modularized, versatile, and extensible toolkit for machine learning and text generation tasks. We used 300 dimensional word embedding and 300 hidden size everywhere in our experiments. the word embedding is initialize using pre-trained Glove vectors. We initialize transformer encoder with one layer and one attention head for the emotion reasoner and remove the position embedding in our emotion reasoner. A Transformer network with 6 layers and 8 attention heads is used for the response generator. We train our models using Adam optimization with a learning rate of 0.0005 and the maximum number of tokens per batch is set to 8192. Early stopping is applied during training. The training time of our models is 2 hours for around 80 epochs on a single Tesla V100 GPU. All results of different methods are generated with top-K sampling, and the K is set to 3 in our experiments.

B Results

In our experiments, we repeated 5 runs with different seeds (1024, 2048, 3170, 4096 and 5120) and average the results. The full results of different methods are presented in Table 7.

Method	BLEU	P_{BERT}	R_{BERT}	F_{BERT}	Dist-1	Dist-2	Accuracy
EmpDG	1.470	0.113	0.120	0.117	0.011	0.093	28.9
	1.500	0.121	0.117	0.119	0.012	0.100	29.0
	1.660	0.109	0.115	0.112	0.007	0.056	29.7
	1.630	0.117	0.118	0.118	0.012	0.096	28.9
	1.270	0.121	0.090	0.106	0.008	0.065	29.6
MoEL	1.670	0.146	0.119	0.133	0.008	0.074	36.0
	1.610	0.135	0.128	0.132	0.009	0.080	38.0
	1.560	0.149	0.121	0.135	0.008	0.073	36.0
	1.590	0.146	0.121	0.134	0.008	0.072	35.0
	1.620	0.143	0.124	0.134	0.008	0.071	36.0
MIME	1.630	0.155	0.124	0.140	0.008	0.066	33.1
	1.660	0.147	0.125	0.137	0.008	0.060	34.0
	1.560	0.142	0.116	0.129	0.008	0.069	30.0
	1.550	0.153	0.124	0.139	0.008	0.057	31.6
	1.490	0.154	0.109	0.132	0.009	0.067	33.1
MK-EDG	1.370	0.146	0.113	0.130	0.008	0.056	35.2
	1.400	0.146	0.111	0.129	0.008	0.055	35.7
	1.450	0.136	0.115	0.126	0.008	0.058	36.6
	1.380	0.146	0.116	0.132	0.008	0.060	38.8
	1.280	0.145	0.112	0.129	0.008	0.059	35.7
Ours(Hard)	1.790	0.144	0.122	0.133	0.018	0.095	41.8
	1.770	0.138	0.128	0.133	0.018	0.093	42.3
	1.640	0.146	0.125	0.136	0.016	0.077	42.4
	1.820	0.143	0.126	0.135	0.017	0.085	42.5
	1.650	0.143	0.125	0.134	0.020	0.101	42.4
Ours(Soft)	1.790	0.141	0.130	0.136	0.019	0.098	41.9
	1.800	0.148	0.128	0.138	0.013	0.065	42.7
	1.670	0.145	0.125	0.135	0.015	0.074	42.2
	1.770	0.148	0.130	0.139	0.015	0.076	42.7
	1.840	0.142	0.121	0.132	0.021	0.109	42.3

Table 7: All results from different methods.