ECC: Synergizing Emotion, Cause and Commonsense for Empathetic Dialogue Generation

Xu Wang[1,2], **Bo Wang**[1,*], **Yihong Tang**[3],
Dongming Zhao[4], **Jing Liu**[4], **Ruifang He**[1], **Yuexian Hou**[1]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Beijing Wenge Technology Co., Ltd., Beijing, China
[3]School of New Media and Communication, Tianjin University, Tianjin, China
[4]AI Lab, China Mobile Communication Group Tianjin Co., Ltd.
{3191981070@qq.com, bo_wang@tju.edu.cn}

Abstract

Empathy improves human-machine dialogue systems by enhancing the user's experience. While traditional models have aimed to detect and express users' emotions from dialogue history, they neglect the crucial and complex interactions among emotion, emotion causes, and commonsense. To address this, we introduce the ECC (Emotion, Cause, and Commonsense) framework, which leverages specialized encoders to capture the key features of emotion, cause, and commonsense and collaboratively models these through a Conditional Variational Auto-Encoder. ECC further employs novel loss functions to refine the interplay of three factors and generates empathetic responses using an energy-based model supported by ODE sampling. Empirical results on the EmpatheticDialogues dataset demonstrate that ECC outperforms existing baselines, offering a robust solution for empathetic dialogue generation.

1 Introduction

Empathy is the ability to understand and share the emotional states of others during social interactions. In human-machine dialogue systems, empathy enables the system to perceive and respond to the user's emotions (Hoffman, 2000), thereby optimizing the user's experience.

To enhance empathetic expression, traditional approaches focus on detecting and expressing users' emotions based on dialogue history (Majumder et al., 2020a; Li et al., 2019; Lin et al., 2019; Rashkin et al., 2019). Despite their success, achieving human-like empathy remains a challenge, because emotions are not isolated in dialogue but interact with other factors (Davis, 1983). The close connection between affection and cognition (Mischel and Shoda, 1995) was first noted, with affection representing an individual's affective state, such as happiness or sadness, and cognition reflecting an individual's experiences and realities. As research on cognition deepens, commonsense knowledge and emotion cause are widely introduced as parts of cognition to enhance empathetic expression. Sabour et al. (2022) exploited the generic commonsense knowledge to help understand the user's situation and Zhou et al. (2022) aligned coarse-to-fine cognitive and affective factors to refine the empathetic responses. Meanwhile, the introduction of emotion causes has led to substantial improvements (Kim et al., 2021; Gao et al., 2021; Zhao et al., 2022; Zhang et al., 2022), as the causes are the sources of emotions, closely related to the interlocutor's experiences, enhancing the understanding of the interlocutor's situation and enabling of the speaker's emotional state.

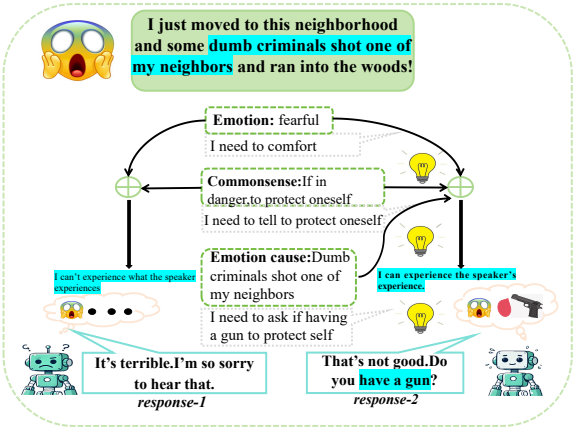

Figure 1: An example from the EMPATHETICDIALOGUES dataset. The combination of emotion, commonsense, and emotion cause enables a conscious experience of the speaker's experiences.

However, previous work all treats commonsense and emotion cause in isolation. Actually, the two factors often interact with each other in human empathetic dialogue. Davis (1983) indicates that human empathetic responses often stem from consciously experiencing others' circumstances. The circumstances are often a joint effect of the two

**Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5475–5485
January 19–24, 2025.

factors. On the one hand, emotion causes stimulates human emotions, reflecting people's logical judgments about events. Then, commonsense knowledge provides insight into these causes, further aiding the speaker's understanding of the interlocutor's situation. On the other hand, understanding commonsense helps humans reason about and expand on emotions, representing specific experiential content (Schutz, 1962). To precisely reason commonsense, the speaker also needs to consciously experience other's experiences with the recognition of emotion cause. Therefore, an empathetic dialogue system requires a combination of both, which focuses on identifying the emotion causes, accurately predicts emotional state, and generates empathetic responses by integrating commonsense to provide a satisfactory emotional response.

Previous works have overlooked the trilateral connection among the **emotion, emotion cause, and commonsense**, resulting in limited empathetic expression capabilities. For one example in Figure 1, the left model only combines the fearful emotion and generic commonsense knowledge. However, due to the absence of emotion cause, the left model can't be further based on commonsense to deduce the speaker's "shot" experiences and to expand on the emotion, leading to a weak empathy response like **response-1**: *I'm so sorry to hear that*. The right model not only incorporates the commonsense, but also leverages the emotion cause, which is heavily related to the speaker's experiences and helps the right model to experience the speaker's experiences consciously. Then generate an informative and empathetic response like **response-2**: *That's not good. Do you have a gun?*.

To address this challenge, we synergize the three factors-emotion, cause, and commonsense(**ECC**)-by adaptively fusing three features from dialogue history. "Emotion" extracts the emotional state manifested in the dialogue history. "Emotion cause" is a conscious recognition, aiding in experiencing the user's experience. Commonsense helps expand on the emotion and reflect the specific experiential content. For encoding the above three factors, we train three special encoders according to their different characteristics. Then, we leverage the Conditional Variational Auto-Encoder (CVAE) (Li et al., 2018) to collaboratively model the three factors and then we model their interaction by designing novel loss functions. Further, to efficiently sample and generate informative and empathetic dialogues, we leverage an energy-based model (EBM) (Khalifa et al., 2020a), which enables the flexible combination of these factors. By assigning lower energy to responses that align with the specified aspects, ECC facilitates empathetic dialogue generation. This is further supported by an ODE sampling method, allowing for a flexible synergy and fine-grained control over the interaction of the three factors.

Our contributions are summarized as follows:

(1) We conceptualize empathy as a conscious experience driven by the interaction of emotion, emotional causes, and commonsense, instead of treating them isolated. This approach significantly enhances the ability to comprehend and respond to user context with greater accuracy and depth.

(2) We develop a novel framework ECC to simulate this conscious experience by controlling the interaction of the three factors within a compact continuous latent space. Three specialized encoders, each trained with distinct preferences, are employed to capture the three factors within the dialogue independently.

(3) Experiments demonstrate the superiority of the ECC in both automatic and human evaluations, underscoring its effectiveness in generating empathetic and contextually appropriate responses.

2 Related work

Empathetic dialogue Expressing empathy toward others is a key trait that builds up seamless relationship with others when communicating. Thus, endowing chatbot with humanized empathy is crucial for building a trustful communicative environment for human-AI dialogue. In psychology, empathy is a deep definition that includes both aspects: affection and cognition(Davis, 1983). With the aid of the two aspects, empathetic dialogue system aims to fully experience interlocutor's experiences, understand the interlocutor's situation and feelings, finally response empathically. Although most exiting work has made progress with much attention attracted to the empathetic dialogues, there are still some issues to be addressed. First, some traditional work only pays attention to detecting affective factors(e.g., emotional state) to enhance the empathetic expression(Fu et al., 2023; Majumder et al., 2020a). Second, some work has realized the importance of cognition with thorough research of cognition, incorporating commonsense as part of cognition(Sabour et al., 2022).Meanwhile, Kim

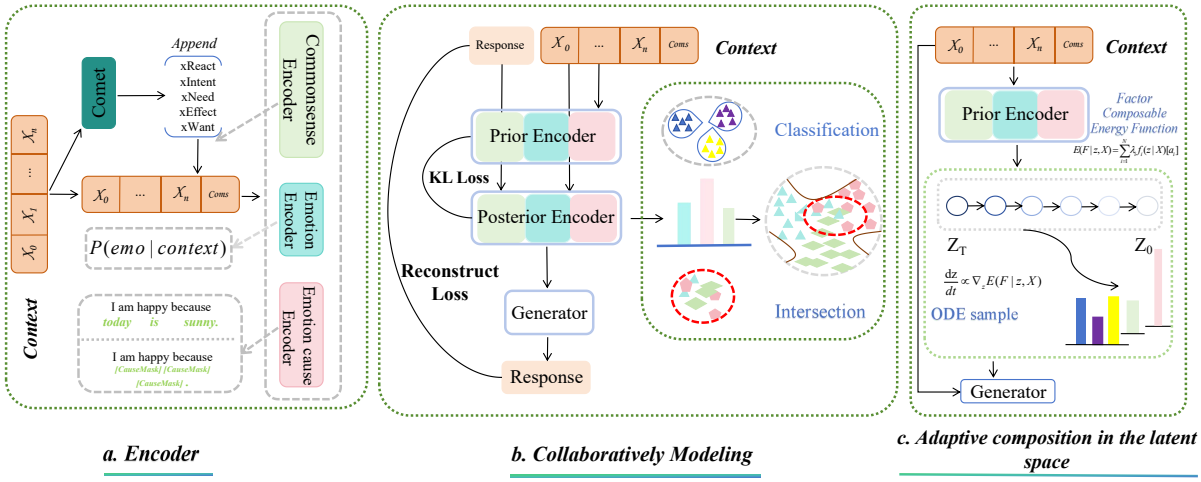

Figure 2: The framework of the ECC. **a:** we train three encoders to capture the factors' features. **b:** we introduce CVAE to collaboratively model the three factors, designing two novel loss functions to model the three factors' interaction in the latent space. **c:** we introduce the energy-based model to composite the three factors adaptively, and with ODE sampling, achieve flexible factors control and generate informative and empathetic responses.

et al. (2021) introduces emotion cause to enhance the empathetic dialogues. However, they all ignore the interaction and combination of the three factors-(**emotion, emotion cause, commonsense**), which limits the system's ability of experiencing the user's experience, thus leading to weak empathy.

Energy-based models The energy-based model (EBM) has been introduced as a flexible generative framework(Khalifa et al., 2020b). Due to its strong ability that allows for the incorporation of arbitrary functions into energy function, many researchers have introduced the EBM into controllable generation work. For example, some work utilizes EBM to perform multiple control factors from the image generation work(Nie et al., 2021) to controllable text generation work(Liu et al., 2022; Qin et al., 2022), resulting in a flexible composition of factors and diverse text. Inspired by them, we employ EBM in our empathetic dialogue generation work to implement the composition of emotion, emotion cause and commonsense, which draw more information about the user's experiences, further enhancing the empathetic responses.

3 Method

3.1 Task formulation

Our task is to require a dialogue model to play the role of the listener and express empathetic responses to the speaker's experiences. Formally, the dialogue context $X = [X_0, ..., X_n]$, where n denotes the $n+1$-th utterance. Our goal is to generate the next utterance following the $n+1$-th utterance, which needs to be empathetic, informative and consistent with the user's experiences.

3.2 Framework

As shown in Figure 3, our framework consists of four steps: encoder, collaboratively modeling, adaptive composition in the latent space, and generator. The first step trains three special encoders to encode the three factors respectively: emotion, emotion cause and commonsense. Then, we employ a Conditional Variational Auto-Encoder (CVAE)(Li et al., 2018) to collaboratively model the three factors within a compact continuous latent space, and next, we also introduce novel loss functions to model the three factors' interactions. Subsequently, following Liu et al. (2023), we introduce EBM with ODE sampling during inference, which allows ECC to adaptively acquire the factors' distribution and permits arbitrary factors' composition and efficient sampling. Finally, the sampled vector is fed into generator to generate empathetic responses.

3.3 Encoder

Firstly, ECC integrates an Encoder module that distills the key features of the three factors from the dialogue context. Specifically, we introduce three special encoders based on the three factors' characteristics. The three encoders fine-tune starts with a pre-trained model in different ways.

Emotion Encoder The role of the emotion en-

coder is to perceive more emotionally relevant content in the context of the dialogue. To this end, we adopt the fine-tuning task of emotion classification to improve the model's ability to model the context emotion state. Finally we get Enc_{emo}:

$$\begin{aligned} h_{emo} &= Enc_{emo}(X), \\ h'_{emo} &= Enc_{emo}(X, r). \end{aligned} \tag{1}$$

Emotion cause Encoder For the emotion cause, to enhance the ability to identify emotion cause words in dialogues, we integrate the Masked Language Model (Devlin et al., 2018) into ECC. Adopting the BERT (Devlin et al., 2018) pre-training objective, we mask the words about emotion cause in the utterance $X_{cau} = [x_0, x_1, ..., [MASK], ..., x_n]$, where these words are replaced with a [MASK] token waiting for prediction in fine-tuning. After fine-tuning, we obtain the representation Enc_{cause}, capable of capturing the relevant emotion cause. Subsequently, the encoder is employed to encode the utterance X:

$$\begin{aligned} h_{cause} &= Enc_{cause}(X), \\ h'_{cause} &= Enc_{cause}(X, r). \end{aligned} \tag{2}$$

Commonsense Encoder Given the dialogue context X, following Sabour et al. (2022), we focus on five key generic commonsense aspects: react, want, need, intent, and effect. Each aspect is respectively represented by tokens(xReact, xWant, xNeed, xIntent, xEffect) that are concatenated to the last utterance. Subsequently, we invoke COMET(Bosselut et al., 2019)to predict corresponding commonsense. Finally, we obtain a sequence of commonsense:$COMS = coms_1 \bigoplus coms_2 \bigoplus coms_3 \bigoplus coms_4 \bigoplus coms_5$, which are encoded as follows:

$$\begin{aligned} h_{coms} &= Enc_{coms}(COMS), \\ h'_{coms} &= Enc_{coms}(COMS, r). \end{aligned} \tag{3}$$

Through this process, ECC acquires a deeper and more abstracted understanding of the three factors that underpin the user's situation, enabling to capture the key features of the three factors.

3.4 Collaboratively modeling

Based on the three encoders obtained in Chapter 3.3, then we introduce CVAE to collaboratively model the three factors. Further, due to the assumption that CVAE framework's prior and recognition distribution follow isotropic multivariate Gaussian distribution, we calculate the mean u, $u'$and the variance σ^2, σ'^2 as follows:

$$\begin{aligned} h &= h_{emo} + h_{cause} + h_{coms}, \\ h' &= h'_{emo} + h'_{cause} + h'_{coms}. \\ \begin{bmatrix} \mu \\ \log \sigma^2 \end{bmatrix} &= \text{Layer}(h), \\ \begin{bmatrix} \mu' \\ \log \sigma'^2 \end{bmatrix} &= \text{Layer}'(h'). \end{aligned} \tag{4}$$

Then, in order to sample z_{coms}, z_{cause}, z_{emo} from the prior and recognition network during training, we utilize reparameterization technique(Kingma and Welling, 2013) and we have:

$$\begin{aligned} z &= u + \sigma\xi, \quad \xi \sim N(0, 1), \\ z' &= u' + \sigma'\xi', \quad \xi' \sim N(0, 1). \end{aligned} \tag{5}$$

Further, we refer to $p(z|X)$ as prior network. To approximate $p(z|X)$, we refer to $q(z|X, r)$ as recognition network, which allows for capturing factors' characteristics given the context X. Subsequently, we leverage $P(r|z, X)$ as a generator, and we train this CVAE through Stochastic Gradient Variational Bayes(Kingma and Welling, 2013), and the optimization goal is to maximize the variational lower bound of conditional log-likelihood. Formally, the ELBO function can be written as:

$$\begin{aligned} \text{ELBO} &= E_{p_\theta(z|X,r)}[\log p_\phi(r|X, z)] \\ &\quad - \text{KL}(p_{\theta'}(r|z)||p_\theta(z|X, r)), \\ \mathcal{L}_{\text{VAE}} &= -\text{ELBO}, \end{aligned} \tag{6}$$

which allows for the consistency between the prior and recognition distribution, and ensures the generative quality of the recognition $p_\theta(z|X, r)$.

Meanwhile, to model the three factors'(**emotion, emotion cause, commonsense**) interplay, force the encoder to distinguish the different aspects'(e.g., happy, sad) features from the same factor(e.g., emotion) and perceive the different factors' internal intersections(Gu et al., 2022), we introduce the aspect classification loss: $\mathcal{L}_c$ and the factor distance loss: $\mathcal{L}_d$. We define the $\mathcal{L}_c$ as follow:

$$\mathcal{L}_c = \sum_{i=1}^{F} y^{(i)} \log(\hat{y}^{(i)}), \tag{7}$$

where $y^{(i)}$ is the ground truth label in the factor i, and F is the number of factors. By minimizing this loss, we encourage the encoder to differentiate the different aspects from the same factor, leading to a

more fine-grained sampling. The $\mathcal{L}_d$ is:

$$\mathcal{L}_d = \sum_{m,n}^{F} \left\| \frac{1}{B}\sum_{i}^{B} h_m^{(i)} - \frac{1}{B}\sum_{j}^{B} h_n^{(j)} \right\|, \quad (8)$$

where $h_m^{(i)}$ and $h_n^{(j)}$ are from the current batch B aiming to bridge the gap among the factors and enable ECC to understand the three factors'(**emotion, emotion cause, commonsense**) internal intersections. Thus, we have ECC's training objective:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_c + \mathcal{L}_d. \quad (9)$$

3.5 Adaptive composition in the latent space

However, in order to produce appropriate, empathetic responses and adaptively compose the different aspects' distribution within the latent space, we leverage EBM to estimate the factor's richness expressed in the response. Specifically, we leverages special classifiers [1] f_i to assess how well the aspect $K_i^{a_i}$ is carried in the latent vector z.

$$p(\mathbf{F}|z, X) = \frac{exp(-E(\mathbf{F}|z, X))}{Z}, \quad (10)$$

where Z is a normalizing factor, and its energy function is designed to compose the three different factors into a comprehensive representation of overall user's situation.

$$\begin{aligned} E(\mathbf{F}|z, X) &= \sum_{i=1}^{N} E_i(F_i|z, C) \\ &= \sum_{i=1}^{N} \lambda_i f_i(z|X)[a_i], \end{aligned} \quad (11)$$

where λ_i is the weight of desired factor $F_i^{a_i}$. $[a_i]$ is the index of desired factor $F_i^{a_i}$. In equation11, the energy-based function $E(\mathbf{F}|z, X)$ can be introduced as the combination of the factors' density. Therefore, sampling from this EBM with low energy can generate informative and empathetic response. It is necessary for us to state that the energy-based formulation is only used during inference, enabling adaptive factors' composition without the fine-tuning for composition. Further, due to the intractable normalization factor Z, we derive the ODE sampling method in the latent space, sampling from EBM rather than directly calculating it. Moreover, according to Lu et al. (2023); Song et al. (2021), the ODE method in the latent space accords with the followings:

$$\frac{dz}{dt} = \frac{1}{2}\beta(t)\left[\nabla_z \sum_{i=1}^{N} \lambda_i f_i(z|X)[a_i]\right]. \quad (12)$$

The ODE is resolved in reverse time progression, commencing from T towards 0. To create a customized sample r that resonates with required factors F, the methodology involves initially sampling $z(T)$ from a normal distribution $N(z|X)$. It employs a generic ODE solver (Chen et al., 2018)to solve for $z(0)$ in the stipulated equation. After solving for z(0), this result is decoded and translated back into the textual space, thereby producing empathetic response. Through this process, we can draw a z within more factor-abundant places by letting $\frac{dz}{dt} \propto \nabla_z f_i(z|X)[a_i]$, leading a more empathetic and informative responses.

3.6 Generator

We use a pre-trained model as the generator, which uses the final latent vector z obtained above to generate an empathetic response:

$$\hat{R} = Decoder_{gnerator}(X, z). \quad (13)$$

4 Experiment

4.1 Experimental Setup

Dataset We conduct the widely used Empathetic-Dialogues(Rashkin et al., 2018), where the user tells personal experience and the listener infers the user's emotional state and expresses appropriate empathy. Following Rashkin et al. (2018), we split the train/valid/test set by $8:1:1$.

Implementation Details We implemented all the models using the PyTorch and Transformers framework. The response generator is based on the medium version of DialoGPT. The AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ is used for training. The training sets the mini-batch size to 16. The maximum learning rate is 1e-4. We use kernel sampling (Holtzman et al., 2020) as our decoding strategy with $top-p = 0.9$ and $temperature\tau = 0.7$. Please refer to the published project for additional details, which is publicly available[2].

4.2 Baselines

We choose Transformer-based baselines, which all don't take -(**emotion, emotion cause, commonsense**)-together into consideration. Meanwhile, since ECC is based on DialoGPT, we also select PLM-based models.

[1] We train these classifiers by Equation 7

[2] https://github.com/uuaaaaaaa/WX-ECC

Models	PPL	Dist-1	Dist-2	Acc	BLEU-1	BLEU-2
Transformer-based Models						
Transformer	37.65	0.47	2.05	–	18.07	8.34
MIME	37.33	0.41	1.62	0.296	18.60	8.39
KEMP	37.32	0.55	2.31	0.341	18.19	8.15
CEM	36.86	0.64	2.84	0.373	16.12	7.29
CASE	35.37	0.74	4.01	0.402	17.90	8.69
PLM-based Models						
LEMPEx	26.37	1.41	14.66	0.432	19.18	8.46
DialoGPT	18.74	2.71	12.01	–	18.69	8.58
BrenderBot	16.71	2.58	16.20	0.470	19.79	9.33
EmpGPT-3	–	3.15	**18.63**	–	16.38	7.67
EmpCRL	16.91	4.33	16.32	0.411	20.77	9.85
PECER	16.79	3.69	16.83	–	21.23	10.14
ECC(ours)	**16.23**	**4.86**	17.09	**0.484**	**28.63**	**14.76**

Table 1: Results of automatic evaluation.

4.2.1 Transformer-based Models

(1).Transformer (Vaswani et al., 2017): A response generator based on transformer. **(2).MIME** (Majumder et al., 2020b): An empathy dialogue model that mimics the detected user's emotion in response, only paying attention to emotion. **(3).KEMP** (Li et al., 2022): An empathetic dialogue model combining external commonsense knowledge with emotion vocabulary, ignoring emotion cause. **(4).CEM**(Sabour et al., 2022): An empathetic dialogue model incorporating commonsense as part of cognition, ignoring emotion cause. **(5).CASE**(Zhou et al., 2022): A chatting model that aligns affection with commonsense knowledge, ignoring emotional cause.

4.2.2 PLM-based Models

(1)LEMPEx(Majumder et al., 2022):A commonsense-aware model exploiting human communication's elements. **(2)DialoGPT**(Zhang et al., 2019):A dialogue generation model and we select the medium version. **(3)BlenderBot**(Roller et al., 2020):A dialogue model for pre-trained communication skills and we select the 400M version. **(4)EmpGPT-3**(Lee et al., 2022):An empathetic dialogue model based prompt-based in-context learning. **(5)EmpCRL**(Cai et al., 2024a): An empathetic dialogue model via In-Context Commonsense Reasoning and Reinforcement Learning. **(6)PECER**(Cai et al., 2024b): An empathetic dialogue model via dynamic personality extraction and contextual emotional reasoning.

4.3 Automatic Evaluation

We explored the widely used Perplexity(**PPL**), **BLEU-1/2**(Papineni et al., 2002) and Distinct-1/2(**Dist**-1/2) (Li et al., 2015) for evaluation. Perplexity represents the generative quality. The BLEU-1/2 measure the similarity between generated responses and factual responses, which we believe can indirectly measure the coherence of dialogues. Dist-n measures the proportion of unique n-grams in the responses as the generation diversity. We also report the accuracy(**Acc**) for the emotion classification.

Table 1 shows the automatic evaluation results. **ECC** outperforms most baselines on the all metrics. **First**, ECC accomplishes the lowest perplexity compared with other baselines, which suggests that **ECC** could generate responses of higher quality. **Second**, the results for Dist-1 and Dist-2 indicates that **ECC** ensures the diverse response. **Third**, **ECC**'s accuracy of emotion prediction is the highest among baselines, and this shows that with the synergize of the three key factors, **ECC** has a better understanding of the user's experience and predicts emotional state more accurately. Higher BLEU-1/2 score also denotes the outstanding performance of **ECC** in coherence. EmpGPT-3 adopts GPT-3's capabilities to produce responses through prompting mechanisms, and it calculates the PPL in a manner apart from other baselines. Therefore, EmpGpt-3's PPL is not shown. Transformer, DialoGPT, EmpGPT-3 and PECER don't have the ability of emotional classification, so their Acc scores are

Models	PPL	Dist-1	Dist-2	Acc	BLEU-1	BLEU-2
ECC	**16.23**	**4.86**	**17.09**	**0.484**	**28.63**	**14.76**
w/o Emc.	16.58	4.71	16.12	0.310	25.25	13.28
w/o Coms.	16.38	4.75	16.33	0.325	26.51	12.91
w/o Post.	16.41	4.69	16.71	0.320	23.33	11.95
w/o EBM	16.31	4.60	15.01	0.463	25.37	10.94
w/o L_d	16.57	4.79	16.25	0.472	27.11	14.39
w/o L_c	16.32	4.83	16.66	0.330	26.51	13.54

Table 2: Results of ablation study. To easily observe, we respectively abbreviate Emotion cause, Commonsense, and Posterior as Emc, Coms, and Post.

also not shown.

4.4 Ablation Studies

We conducted ablation studies to verify the effectiveness of each component in **ECC**, in Table 2. We designed six variants of **ECC**:

(1) w/o Emotion cause: The emotion cause encoder is removed. The emotion prediction accuracy decreases significantly. This indicates that the emotion cause encoder plays a crucial role in capturing the underlying reasons and contextual information behind emotions. Without this encoder, the model lacks the ability to effectively experience user's experiences, leading to low prediction accuracy.

(2) w/o Commonsense: The commonsense encoder is removed. The emotion prediction accuracy and BLEU-1/2 also decrease. This suggests that the commonsense encoder, which helps reason about and expand on emotion, is helpful in identifying the user's emotions and situations. Without this encoder, the model misses essential commonsense knowledge, which is necessary for inferring and recognizing complex emotional states and contexts.

(3) w/o Posterior: The CVAE posterior encoder is removed. There is a significant decrease across all metrics. This suggests that if without the guidance of $P(z|X, r)$, the prior encoder cannot learn the latent internal relationships between dialogue history X and response r.

(4)w/o EBM: The part of EBM is removed, and the **Dist** score decreases, which suggests that the allowance for flexible composition in EBM helps ECC to efficiently synergize the key factors, leading to a diverse and empathetic response.

(5) w/o L_d: When L_d was removed, and the BLEU score decreases. This indicates that L_d plays a role in mitigating conflicts among factors, thereby enhancing the overall performance of the model.

(6) w/o L_c: The L_c is removed, and the emotion prediction accuracy decreases, which suggests that the L_c develops ECC's to classify different aspects within the same factor, making contributions to correctly detecting user's emotional state.

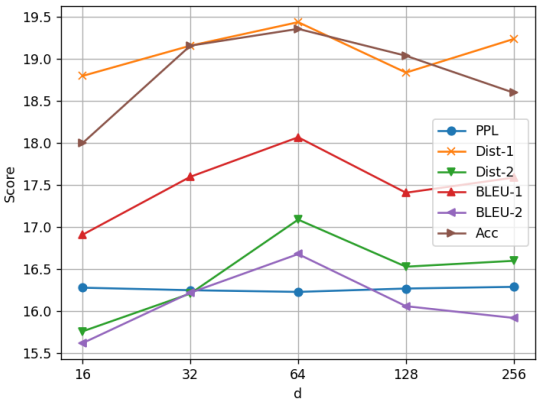

Figure 3: Experiments with the different d for the dimensions of the latent variable. For ease of viewing, Dist-1, Acc, and BLEU-2 are multiplied by a factor of 4, 40, and 2, respectively.

Comparisons	Aspects	Win	Lose	κ
ECC vs. CASE	Coh.	**53.9**‡	33.6	0.56
	Emp.	**55.8**‡	32.1	0.57
	Inf.	**56.2**‡	31.9	0.55
ECC vs. EmpCRL	Coh.	**51.1**†	37.5	0.50
	Emp.	**53.2**‡	36.7	0.51
	Inf.	**51.9**‡	36.1	0.53
ECC vs. PECER	Coh.	**50.1**‡	46.3	0.51
	Emp.	**49.6**‡	41.5	0.49
	Inf.	**51.7**‡	43.2	0.47

Table 3: Percentage results of human evaluations for ECC and baseline methods. The kappa statistic κ, falling within the range [0.41, 0.6], is employed to indicate moderate agreement. Symbols † and ‡ denote statistically significant improvements with p-values less than 0.1 and 0.05, respectively.

4.5 Effect of Latent Dimension Coefficients

In ECC, the latent vectors' dimension d reflects how precise it is to capture the factors' information. Intuitively, the factor d will affect the model's performance, and we report this effect in Figure**??**. The model cannot play well when the d is small since it doesn't capture enough information. When d is set too large, the model can also not generate an accurate response due to the increased noise caused by too many dimensions, making the information too scattered. This result is consistent with the observation that texts in the dataset typically include a limited number of factors. Therefore, the optimal dimension, as indicated by our experimental

Context	Yeah about 10 years ago I had **a horrifying experience**. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost **ran me off the road**.	I recently **spoke with my ex-girlfriend** on the phone. The conversation went pretty well and it **reminded me of my past experiences** with her.
Emotion	**Terrified**	**Nostalgic**
CEM	This is a bad experience and I hope you are good now.	I am sorry to hear that.
CASE	I would be too! I would have been terrified by this terrible experience!	I hope you are good now.
EmpCRL	So terrible! I hope you will get better soon!	I am sure you will have a great time.
PECER	Oh, that's really bad to have such an experience. Are you better now?	I am so sorry to hear that and did you recover?
Commonsense	**To have a bad experience**, **Get into a fight**, To be in danger, To look bad	**To have good time** , To have fun, **To be sad** , **To miss**
EmotionCause	**A horrifying experience**, **Ran me off the road**	**Spoke with my ex-girlfriend**, **Reminded me of my past experiences with her**
ECC	That is **too bad**, and do you **have any injury?**	Oh! Did you **end with happiness**? I hope **everything is good**!
Ground-Truth	Did you suffer any injuries?	Did things end amicably? I hope everything is good now!

Table 4: Case study of the generated responses by ECC and the baselines.

findings, is $d = 64$.

4.6 Human Evaluation

These 200 contexts are randomly selected, and following Sabour et al. (2022), three crowdsourcing workers choose the better response(**Win**) by considering the three aspects: **Coherence(Coh.)** response is more coherent in content and more related to the context. **Empathy(Emp.)** response expresses more understanding of the user's situation and shows more empathy. **Informativeness(Inf.)** responses carry more information related to the context.

As shown in Table 3, the results indicate that ECC outperforms the three more competitive baselines on the above three aspects. Especially, **ECC** outperforms baselines significantly in terms of empathy and informativeness, which shows the superiority of the combination of **emotion, emotion cause, and commonsense**.

4.7 Case Study

Two cases from four models are selected in Table 4, among which **ECC** tends to express more informative responses in a highly empathetic tone. This is mainly beneficial from two advantages:

(1) The combination of emotion, emotion cause, and commonsense. For example, in the first case, **ECC** assumes the speaker's terrified emotional state by consciously experiencing the speaker's horrifying experience: "*Ran me off the road*". Further, based on commonsense knowledge, **ECC** expands on the terrified emotion and then generates a more empathetic and informative response.

In the second case, **ECC** recognizes that the speaker is nostalgic through experiencing "*Spoke with my ex-girlfriend*". Then, based on the "To have a good time" knowledge, **ECC** generates a precise response: "*Oh! I hope everything is good!*".

(2) The adaptive composition and efficient sample. In **ECC**, we leverage EBM to compose three factors and process efficient samples with ODE adaptively. For example, in the first case, **ECC** understands that the speaker tells more about the horrifying experience by EBM, so **ECC** responds not only in an empathetic tone but also in a more meaningful response: "*Do you have any injury?* ", which is beneficial from ODE method to sample factor-abundant vectors.

In the second case, the speaker talks about past experiences and a good conversation with an ex-girlfriend, actually in a nostalgic and not too sad atmosphere. Finally, with the factors sampled wisely, **ECC** responds that: "*Oh! I hope everything is good!*".

5 Conclusion and Future Work

In this paper, to respond empathetically, we propose **ECC** framework, which simulates humans' conscious experiences of others by combining emotion, emotional cause, and commonsense. We first train three special encoders about emotion, emotion cause, and commonsense, and introduce CVAE to collaboratively model the three factors, then leverage EBM to adaptively compose them and achieve an efficient sample by the ODE method. Experimental results verify the superiority of **ECC** in terms of overall quality and empathy performance.

Our work will encourage future work to simulate more interaction among factors related to empathy.

Limitations

The main limitation of our work is that the scores of the automatic evaluation metrics don't align with the results of the human evaluations. The automatic evaluation metrics mainly focus on the quality of generated responses and the accuracy of emotion prediction. The lack of a generalized evaluation method for empathy hinders the effective evaluation of the generation of empathetic dialogue.

Ethics Statement

Our experiments are based on the widely-used EmpatheticDialogues dataset, which has undergone thorough filtering to remove sensitive and personal information during its construction. We ensure that no personal or identifiable data is included in the dataset. Additionally, human evaluations are conducted with full anonymity to protect the privacy of evaluators. We have strictly followed ethical guidelines for both dataset usage and human evaluation, ensuring that no harm, bias, or privacy violations occur in any part of our study.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376188, 62272340, 62276187, 62376192, 62166022).

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024a. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *International Conference on Language Resources and Evaluation*.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024b. Pecer: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10631–10635. IEEE.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-core: Emotion correlation enhanced empathetic dialogue generation. *arXiv preprint arXiv:2311.15016*.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. *arXiv preprint arXiv:2210.02889*.

Martin L. Hoffman. 2000. Empathy and moral development : implications for caring and justice. *Contemporary Sociology*, 30:487.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020a. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*.

Muhammad Khalifa, Hady ElSahar, and Marc Dymetman. 2020b. A distributional approach to controlled text generation. *ArXiv*, abs/2012.11635.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3890–3900.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2022. Composable text controls in latent space with odes. *arXiv preprint arXiv:2208.00638*.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023. Composable text controls in latent space with ODEs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16543–16570, Singapore. Association for Computational Linguistics.

Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, Singapore. Association for Computational Linguistics.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020a. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020b. 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.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102 2:246–68.

Weili Nie, Arash Vahdat, and Anima Anandkumar. 2021. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510.

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.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Alfred Schutz. 1962. Common-sense and scientific interpretation of human action. In *Collected papers I: The problem of social reality*, pages 3–47. Springer.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. *Preprint*, arXiv:2011.13456.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Duzhen Zhang, Zhenfei Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. Tsam: A two-stream attention model for causal emotion entailment. In *International Conference on Computational Linguistics*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2022. Knowledge-bridged causal interaction network for causal emotion entailment. *ArXiv*, abs/2212.02995.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.