CoMIF: Modeling of Complex Multiple Interaction Factors for Conversation Generation

Yuxuan Chen[1,2], Wei Wei[1,2] *, Shixuan Fan[1,2], Kaihe Xu[2,3], Dangyang Chen[2,3]
[1] Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory,
School of Computer Science and Technology, Huazhong University of Science and Technology
[2] Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL),
[3] Ping An Property & Casualty Insurance company of China, Ltd.
{chenyuxuan232, weiw, fanshixuan}@hust.edu.cn
xukaihenupt@gmail.com
chendangyang273@pingan.com.cn

Abstract

Highly realistic human-machine interaction is challenging for open-domain dialogue systems. Although existing methods have achieved notable progress by leveraging various interaction factors (e.g., emotion, personality, topic) for delivering human-like (e.g., empathetic, personalized and semantically-consistent) responses, they typically model such factors alone and thus easily suffer from low-quality response generation issue, due to the overlook of implicit correlations among factors. Furthermore, different factors may alternately dominate token-level response generation during decoding, making it harder to generate high-quality responses by applying various factors at the sentence level. To address the issue, we present a unified response generation framework, which is capable of simultaneously modeling Complex Multiple Interaction Factors (named CoMIF) to generate human-like conversations. To model the implicit correlations among factors, CoMIF first employs a *dynamic perception* module to construct a directed *collaborative*-graph to jointly learn the dynamics over time of each factor, as well as the cross-dependencies among them. Additionally, we also design a scalable post-adaptation module to introduce token-level factor signals to generate more human-like responses with appropriately multiple factors. Extensive experiments over multiple datasets demonstrate that the proposed method achieves the superior performance in generating more human-like responses with appropriate multiple-factors, as compared to the state-of-the-art methods.

1 Introduction

Open-domain dialogue systems aim to achieve highly realistic human-machine interactions, playing increasingly vital roles not only in daily life but also in industrial production, such as question answering (Wei et al., 2011). To enhance the quality of generated dialogues, an increasing number of factors are being integrated into dialogue systems. Recent research has delved into various aspects, such as speaker emotions (Wei et al., 2019; Yan et al., 2023; Zhao et al., 2023b), persona consistency (Song et al., 2021; Liu et al., 2022), and topic diversity (Sevegnani et al., 2021; Xu et al., 2021), to better emulate human dialogue. These investigations have significantly advanced the development of human-machine dialogue systems.

Figure 1: An example of introducing speaker's persona into dialogue and considering its interaction with factors such as emotion and topic.

However, most existing works focus on the isolated effects of different interaction factors on response generation, which is incomplete as the interaction factors between humans and chatbots influence each other. As illustrated in Figure 1, without considering the speaker's persona, when Speaker A shares negative news about Speaker B's idol, Speaker B may respond positively and continue the conversation. However, when considering Speaker B's persona as a fan, they are more likely to exhibit negative emotions and attempt to end or change

*Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 7355–7366
January 19–24, 2025.

the topic of the conversation. Clearly, people's responses are influenced by multiple interaction factors, so jointly modeling these factors is key to generating high-quality responses.

Another issue to address is the selective dominance of interaction factors in token-level response generation during decoding. As illustrated in the lower half of Figure 1, only a small number of tokens are influenced by interaction factors, and the tokens influenced by each factor vary. Allowing all factors to dominate response generation simultaneously could significantly reduce the quality of the generated responses.

To this end, we proposes a unified response generation framework named CoMIF to simultaneously model the complex multiple interaction factors in conversation generation. Specifically, interaction-related factors such as emotion, topic, and persona are extracted from historical dialogues. Subsequently, we use a directed collaborative graph to jointly model the temporal dynamics within these factors to generate factor-related signals consistent with the speaker's situation. Finally, inspired by (Li et al., 2024), a scalable post-adaptation module is proposed to selectively inject token-level factor-related signals for generating responses. Experimental results on multiple datasets show that our model has significant advantages in dialogue generation.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to propose a multi-factor framework for open-domain dialogue generation. This framework integrates the correlations and cooperative effects among multiple variables to produce more human-like responses, surpassing those generated by previous single-factor approaches.
- We propose a unified multi-factor dialogue framework named CoMIF, to simultaneously model the complex multiple interaction factors for conversation generation.
- Results from extensive experiments on datasets demonstrate that our model is effective in generating responses that are tailored to the speaker's situation.

2 Related Work

2.1 Personality-based Response

To simulate human conversations, a dialogue system or chatbot should generate responses based on a fixed persona. In this context, Zhang et al. (2018) added persona information to the dialogue and proposed the PERSONA-CHAT dataset, which effectively improved the personality consistency of the dialogue system. Kim et al. (2020) and Cheng et al. (2023) extracted personality from the dialogue history and integrated it into the generation process to produce personality-based responses. Song et al. (2021) decomposed personality-based dialogue generation into two subtasks: dialogue response generation and personality consistency understanding, achieving good performance on limited character personality dialogue datasets. In contrast, Xu et al. (2023) automatically generated character personality information based on a preset template and used the generated personality information to produce responses.

2.2 Topic-based Response Generation

A topic is essential for keeping each participant engaged in a conversation and is therefore crucial for a dialogue system. Previous work has explored two approaches to applying topics to dialogue generation. On the one hand, retrieval-based approaches (Tang et al., 2019; Qian and Dou, 2023) aim to obtain responses related to the current topic from a topic-based response repository. On the other hand, generation-based approaches (Chen and Yang, 2020; Xu et al., 2021) predict the topic of the current dialogue from the context and generate appropriate responses that match the current topic. Additionally, other approaches, such as combining knowledge graphs with topics (Wu et al., 2019; Liu et al., 2020), have been used to generate appropriate responses.

2.3 Emotion-based Dialogue Generation

An emotional dialogue system can detect subtle changes in the user's emotions and generate responses with specific emotional tones. Emotions have proven to be a key factor in creating more engaging dialogue systems. Zhou et al. (2017) and Wei et al. (2019) were the first to consider emotional factors in dialogue generation, introducing specific emotional representations in the generation process to produce emotional responses. Later, Wang et al. (2022) modeled fine-grained changes

in sentiment within a dialogue, making the generated responses more realistic. Other works, such as Zhao et al. (2023b), consider information from multiple rounds of sentimental dialogue to ensure smooth transitions between dialogues.

3 Methodology

3.1 Task Definition

In this paper, our task is to dynamically perceive interaction factors such as emotions and topics in the dialogue, incorporate them into the generation process along with the speaker's personas, and generate semantically reasonable, emotionally appropriate, and topic-consistent responses. Hence, our response generation task is defined as follows: given a set of predefined personas $\mathcal{P} = \{\mathrm{p}_1, \mathrm{p}_2, \cdots, \mathrm{p}_n\}$ and historical dialogue sentences $\mathcal{X} = \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_m\}$ with corresponding emotion sequence $E_{his} = \{e_1, e_2, \cdots, e_m\}$ and topic sequence $T_{his} = \{t_1, t_2, \cdots, t_m\}$, the goal is to generate a corresponding response Y that is coherent with the speaker's situation.

3.2 Overview of Architecture

The overall architecture of CoMIF is shown in Figure 2. We divide the entire model into three modules: the encoding module, the dynamics perception module, and the post-adaptation module. First, the historical dialogue and predefined personas are input into the encoding module to obtain various representations in the embedding space, such as the historical dialogue representation x , the predefined personas representation sequence P_{pre}, the historical emotion representation sequence E_{his}, and the historical topic representation sequence T_{his}. Then, in dynamics perception module, the temporal dynamics within each factor are modeled to obtain the token-level factor-related signals P, E and T. Finally, the post-adaptation module modifies the original output of the decoder according to the signals to generate the final response.

3.3 Encoding Module

Personas Representation Sequence. An encoder is used to obtain the corresponding embedding representation of the predefined personas. For each persona $\mathrm{p}_i \in \mathcal{P}$, we encode them separately and finally we get the personas representation sequence $\mathrm{P}_{pre} = \{p_1, p_2, \cdots, p_n\}$.

Historical Dialogue Representation. First, We use the same encoder to encode each utterance x_i in the dialogue and get a sequence of utterance representations $\mathrm{X}^{enc} = \{x_1^{enc}, x_2^{enc}, \cdots, x_m^{enc}\}$.

After that, X^{enc} is fed into a bidirectional RNN to extract the semantic information and we obtain $\mathrm{X}^s = \{x_1^s, x_2^s, \cdots, x_m^s\}$ containing semantic information:

$$\mathrm{X}^s = \mathrm{RNN}_{semantic}(\mathrm{X}^{enc}). \tag{1}$$

Note that x_i^s has the same dimensions as x_i^{enc}.

In order to obtain the historical dialogue representation x, we need to fuse the semantic representations in X^s. Based on the assumption that the target response Y is often highly correlated with the last sentence in the dialogue history (i.e., x_m) and inspired by Luong et al. (2015), we use the method in Luong Attention to calculate the weight of each semantic representation in X^s and perform a weighted sum to obtain x:

$$\alpha_i = \frac{\exp(\mathrm{Score}_i)}{\sum_{j=0}^{m} \exp(\mathrm{Score}_j)}, \tag{2}$$

$$x = \sum_i \alpha_i x_i^s, \tag{3}$$

where $\mathrm{score}_i \in \mathbb{R}^1$ and $x \in \mathbb{R}^d$. The Score function is as follows:

$$\mathrm{Score}_i = \mathrm{sum}(W_2 \cdot \tanh(W_1 \cdot [x_m^{enc}; x_i^s] + b_1)), \tag{4}$$

where $W_1 \in \mathbb{R}^{2d \times d}$, $b_1 \in \mathbb{R}^d$, $W_2 \in \mathbb{R}^{d \times d}$ are the parameters of linear layer.

Historical Representation Sequence. We model the emotion factors and topic factors in the dialogue to ensure the smoothness and consistency of the generated dialogue in terms of emotion and topic. Specifically, X^{enc} is input into another bidirectional RNN to extract specific historical emotion sequence representations $\mathrm{E}_{his} = \{e_1, e_2, \cdots, e_m\}$:

$$\mathrm{E}_{his} = \mathrm{RNN}_{emotion}(\mathrm{X}^{enc}), \tag{5}$$

where $\mathrm{E}_{his} \in \mathbb{R}^{m \times d}$.

During training, e_i is inputted into a linear layer followed by softmax operation to generate the emotion category distribution $\mathrm{P}_{emo}(e_i)$, and we optimize the model by minimizing the cross entropy loss between the emotion category distribution P_{emo} and the true label $\hat{e}_i$:

$$\mathcal{L}_{E_{his}} = -\sum_{i=1}^{m} \log(\mathrm{P}_{emo}(\hat{e}_i)) \tag{6}$$

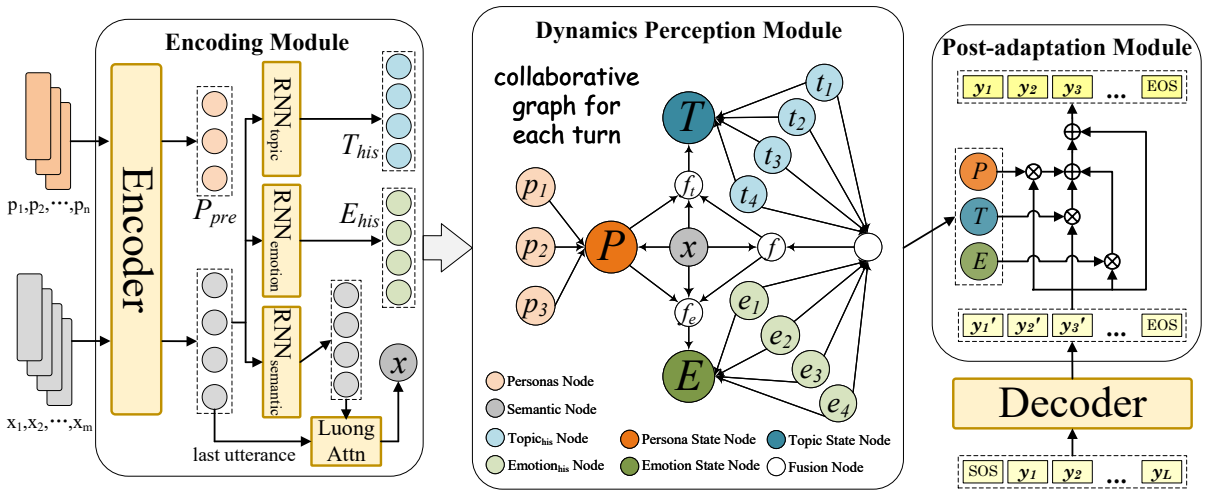

Figure 2: The overall architecture of our proposed CoMIF model, which consists of three modules: Encoding Modulde, Factor Generation Module and Post-adaptation Module

The operation of obtaining T_{his} is similar to that of E_{his}, and will not be described in detail here:

$$\mathrm{T}_{his} = \mathrm{RNN}_{topic}(\mathrm{X}^{enc}), \tag{7}$$

$$\mathcal{L}_{T_{his}} = -\sum_{i=1}^{m} \log(\mathrm{P}_{top}(\hat{t}_i)). \tag{8}$$

3.4 Dynamics perception module

As mentioned above, interaction factors influence each other and change temporally, and it's important to jointly modeling these factors. To achieve this, we makes use of a collaborative graph to jointly model the temporal dynamics within each factor in each turn of conversation.

Graph Construction. As shown in Figure 2, our collaborative graph contains three types of nodes: (1) The first type of nodes includes Personas Nodes p_i, Topic$_{his}$ Nodes t_i, Emotion$_{his}$ Nodes e_i and Semantic Node x, which represent the information extracted from the predefined personality and dialogue history respectively, and are initialized by the output of the encoding module. (2) The second type of nodes includes persona state node P, emotion state node E and topic state node T. These nodes are the final output of the module and are used as the factor-related signals for generating responses. (3) The last type of nodes is called fusion node, which is used to temporarily display the detailed relationship between the other two nodes.

The edges in the graph reveal the dependencies between different nodes, and the details of these dependencies will be discussed later.

Dynamics Perception. We perceive the dynamics in factors and generate the factor-related signals based on the dependencies shown in the collaborative graph.

In order to simplify the dependency details between various factors and improve the scalability of the model, we use a single-layer Transformer Encoder (Vaswani et al., 2017) as a factor fusion module to handle the dependencies between different factors. This module takes two factors as input: the first factor is used as the Query in the attention mechanism, and the second factor is used as both the Key and the Value.

First, we deal with persona factor. The persona displayed by the speaker in the dialogue should be consistent with his or her own personality and highly relevant to the current dialogue. Therefore, the persona state node P in graph should be updated based on the persona nodes P_{pre} and the semantic node x, that is, $\mathrm{P}(P|x, \mathrm{P}_{pre})$:

$$P = \mathrm{TransEnc}(x, \mathrm{P}_{pre}), \tag{9}$$

where $P \in \mathbb{R}^d$.

Next, we take into account two factors: emotion and topic. In contrast to persona, the variations in emotion and topic within a dialogue are more intricate and frequently interrelated. Therefore, we first integrate historical emotions and historical topics in an equal way:

$$f_e^{'} = \mathrm{TransEnc}(\mathrm{T}_{his}, \mathrm{E}_{his}), \tag{10}$$

$$f_t^{'} = \mathrm{TransEnc}(\mathrm{E}_{his}, \mathrm{T}_{his}), \tag{11}$$

$$f = \mathrm{TransEnc}(x, W_3[f_e^{'}; f_t^{'}] + b_3), \tag{12}$$

where $f_e^{'}, f_t^{'} \in \mathbb{R}^{m \times d}$, $f \in \mathbb{R}^d$, $W_3 \in \mathbb{R}^{2d \times d}$ and $b_3 \in \mathbb{R}^d$.

The topic of the target response should be a continuation of the current historical topic sequence, and the speaker's personas, emotions, and the context of the current dialogue will also have a certain degree of influence on the variation of the current topic. Therefore, we update the topic state node as follows:

$$f_t = x + g_t \cdot p + (1 - g_t) \cdot f, \tag{13}$$

$$T = \text{TransEnc}(f_t, \text{T}_{his}), \tag{14}$$

where $f_t, T \in \mathbb{R}^d$, and we get g_t as follows:

$$g_t = \sigma(W_{g_t}[x; P; f] + b_{g_t}). \tag{15}$$

Similar to the topic state node, We get f_e in the same way as f_t and update the emotional state node:

$$E = \text{TransEnc}(f_e, \text{E}_{his}). \tag{16}$$

To ensure that the updated nodes contain the corresponding emotion and topic signals, we input them into linear layers to get the corresponding probability distribution $\text{P}_{emo}(E)$ and $\text{P}_{top}(T)$, and constrain them through the cross entropy loss function:

$$\mathcal{L}_E = -\log(\text{P}_{emo}(\hat{E})), \tag{17}$$

$$\mathcal{L}_T = -\log(\text{P}_{top}(\hat{T})), \tag{18}$$

where $\hat{E}$ and $\hat{T}$ are the corresponding true labels.

3.5 Post-adaptation Module

As mentioned in Fan et al. (2024a), static fusion of different factors results in a fixed trajectory of the output distribution. We design a post-adaptation module to selectively and dynamically inject multiple factor-related signals into the process of response generation.

To generate the target response Y, we first concatenate the historical dialogue into the decoder, and autoregressively generate the hidden state of the original response that do not contain any factor-related information:

$$h_{ori,i} = \text{Decoder}(E_{ori,<i}, E_{\mathcal{X}}) \tag{19}$$

where $E_{ori,<i}$ denotes the embeddings of the generated words before the time step i, $E_{\mathcal{X}}$ is the embeddings of the catenated dialogue history.

Then, to make the generated response more suitable for the speaker's situation, we utilize a unified and scalable method to modify the original hidden state of the response, taking into account various factor-related signals. We fuse the hidden state $h_{ori,i}$ with each signals s separately to obtain the adapted hidden state $h_{s,i}$:

$$w_s = \sigma(W_s[h_{ori,i}; s] + b_s), \tag{20}$$

$$h_{s,i} = w_s \cdot s + (1 - w_s) \cdot h_{ori,i}, \tag{21}$$

where $s \in [P, T, E]$, $W_s \in \mathbb{R}^{2d \times d}$ and $b \in \mathbb{R}^d$ are trainable parameters. The final adapted hidden state $h_{adp,i}$ is the average of $h_{P,i}$, $h_{T,i}$, and $h_{E,i}$.

The distribution over the vocabulary for the t-th token can be obtained by a softmax layer:

$$\text{P}(y_t|y_{<t}, \mathcal{X}) = \text{softmax}(W h_{adp,t}), \tag{22}$$

where $\mathcal{X}$ is the input historical dialogue, $W \in \mathbb{R}^{|V| \times d}$ and V is the vocabulary size.

We use cross entropy loss function to constrain the generation process:

$$\mathcal{L}_{gen} = -\sum_{t=0}^{L} \log(\text{P}(y_t|y_{<t}, \mathcal{X}). \tag{23}$$

Finally, the following joint optimization objective is used to optimize the model parameters:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{gen} \\ + \lambda_2 \mathcal{L}_{E_{his}} + \lambda_3 \mathcal{L}_{T_{his}} \\ + \lambda_4 \mathcal{L}_E + \lambda_5 \mathcal{L}_T \end{aligned} \tag{24}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are hyper-parameters.

4 Experiments

4.1 Dataset

We conduct our experiments on two datasets, Persona-Chat (Zhang et al., 2018) and Synthetic-Persona-Chat (Jandaghi et al., 2023). The Persona-Chat dataset addresses the issue of inconsistent personality in traditional chat models and enhance the model's consistency and appeal by endowing it with a persona. The Synthetic-Persona-Chat dataset is a variant of the Persona-Chat dataset, consisting of two parts. The first part has the same user profile pairs as Persona-Chat but includes new synthetic conversations. The second part contains new synthetic personas and conversations.

Since we need labels such as historical emotions and historical topics in the experiment, the

Dataset	SPC	PC
dialogue	10,680	18,878
utterance	291,553	278,478
P_{avg}	4.49	4.49
T_{avg}	3.12	3.67
E_{avg}	1.83	1.99

Table 1: Statistics of the datasets. P_{avg} is average persona number per user, T_{avg} and E_{avg} are the average number of keywords and emotions per utterance, respectively.

dataset was expanded before the formal experiment. In short, we used two external tools, 'roberta-base-go_emotions' [1] and KeyBERT (Grootendorst, 2020), to generate the emotion labels and topic labels of the dialogue respectively.See Appendix A for more details. Table 1 shows the statistics of the expanded datasets.

4.2 Implementation Details

In this experiment, the encoding module utilizes the RoBERTa-base model (Liu et al., 2019) as the encoder, with its pre-trained parameters serving as the initial settings. During the training phase, the encoder parameters remain fixed and are not subject to updates. The $\text{RNN}_{emotion}$, RNN_{topic}, and $\text{RNN}_{semantic}$ within the encoding module are all implemented as two-layer bidirectional LSTMs (Hochreiter and Schmidhuber, 1997), with random initialization parameters. The decoder used in the post-adaptation module is the smallest variant of GPT-2 (Radford et al., 2019), namely GPT-2-small.

All baselines and our method are implemented in PyTorch and trained on an RTX 4090 24GB GPU. We maintain a maximum of 9 turns of historical dialogue for our method. Throughout the experiments, we utilize the Adam optimizer (Kingma and Ba, 2014). The learning rate is set to 1e-4, and the batch size is 12. The hyperparameters corresponding to each loss function during joint training are: $\lambda_1 = 10$, $\lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 1$. We train the model for up to 10 epochs and employ early stopping when the perplexity (PPL) does not improve on the validation set.

4.3 Baselines

We compare our proposed model with the following competitive baselines.

- **GPT-2** (Radford et al., 2019): A model based on the Transformer-Decoder architecture. We used the smallest variant of GPT-2 in this experiment.

- **PPA** (Huang et al., 2023): An encoder-decoder model with a persona-enhanced attention module. The model uses different Transformer-based encoders to process conversation context and personas, respectively, and generates responses through a Transformer decoder integrated with a persona-enhanced attention mechanism.

- **EmpSOA** (Zhao et al., 2023a): An empathy generation model based on an encoder-decoder architecture comprehensively consider self-other awareness. Through the three stages of self-other differentiation, self-other modulation, and self-other generation, the information of self-other awareness is clearly maintained, regulated, and injected into the process of empathic response generation.

- **Llama 3** (Meta, 2024): The latest generation of llama models in Meta's LLAMA family. In this experiment, the model we used is llama 3 8B Instruct.

4.4 Evaluation Metrics

Automatic Evaluation. We use the following automatic metrics for evaluation: (1) *Perplexity* (PPL) to measure the overall quality of the generated responses; (2) *Distinct-n* (n=1,2) (Li et al., 2016) measures the diversity of generated responses by calculating the proportion of unique n-grams in all n-grams; (3) *ROUGE-n* (n=1,2) (Lin, 2004) is used to measure the relevance between the ground truth and generated response. (4) Since n-gram-based metrics often penalize semantically correct phrases, we use *BertScore* (Zhang et al., 2019) to compensate for this shortcoming.

Human Evaluation. Following previous works (Liu et al., 2022; Fan et al., 2024b), we adopt *Relevance*, *Fluency* and *Informativeness* of the generated utterances with the rating range of [0, 2]. We recruit three experienced annotators to evaluate 100 randomly selected dialogues. The evaluation details are shown in the Appendix B.

[1]This model can get from https://huggingface.co/SamLowe/roberta-base-go_emotions

Dataset	Model	PPL	Distinct-1	Distinct-2	ROUGE-1	ROUGE-2	BertScore	Relevance	Fluency	Informativeness
SPC	GPT-2	**2.81**	7.51	25.24	39.43	25.89	91.62	1.35	1.80	1.60
	PPA	3.18	6.85	19.30	16.97	7.05	89.39	1.46	1.78	1.57
	EmpSOA	4.28	8.20	27.01	40.81	24.34	89.99	1.51	1.82	1.59
	CoMIF	**2.81**	**10.58**	**33.52**	**43.11**	**29.95**	**91.88**	**1.57**	**1.84**	**1.61**
PC	GPT-2	19.42	6.82	21.39	13.77	2.33	**87.43**	0.98	1.54	1.33
	PPA	**13.12**	5.64	15.96	10.18	1.33	87.26	1.18	1.59	1.29
	EmpSOA	36.99	3.19	8.56	14.76	2.79	85.91	1.16	**1.63**	1.31
	CoMIF	19.90	**7.01**	**21.52**	**15.04**	**3.01**	87.23	**1.24**	**1.63**	**1.35**

Table 2: Comparison of CoMIF against baselines on the Synthetic-Persona-Chat (SPC) dataset and Persona-Chat (PC) dataset. Boldface indicates the best result in terms of the corresponding metrics and underline indicates the suboptimal result. Distinct, ROUGE and BertScore are scaled by 10^{-2}.

5 Results

5.1 Automatic Evaluation

We evaluate all methods on the Persona-Chat dataset and Synthetic-Persona-Chat dataset.The results are shown in Table 2.

On the Synthetic-Persona-Chat and the Persona-Chat dataset, CoMIF achieved the best results across most metrics, demonstrating the model's effectiveness. Benefiting from the comprehensive consideration of multiple interaction factors during generation, our model is able to generate high-quality responses with lower perplexity. Additionally, the improvements in *Distinct* highlight our model's superiority in generating more informative responses.

It is worth noting that compared to the backbone model GPT-2, our model incorporates additional factors when generating responses, resulting in more diverse and contextual outputs with nearly the same perplexity. This demonstrates that the performance improvement of our model is due to the consideration of interaction factors rather than solely optimizing the model structure.

5.2 Human Evaluation

In addition to the automatic evaluation, we conducted a human evaluation of the responses generated by the baselines and CoMIF, as shown in Table 2. The consistency metric of the evaluation results, Fleiss' Kappa, is 0.573, indicating moderate agreement.

CoMIF outperforms other baselines in all three aspects. Additionally, while the scores for *Fluency* and *Informativeness* are comparable, our model's score for *Relevance* is significantly higher than those of other models. This suggests that our model can generate responses that are more pertinent to the current speaker's situation by incorporating multiple interaction factors.

5.3 Compaered with LLM

To demonstrate the importance of multiple interaction factors modeling for the Large Language Models (LLMs), we choose Llama3 8B Instruct(Meta, 2024) as the backbone for experiments. We generated responses from prompts using the original model (Llama3), and the finetuned model (Llama3-ft). We also design an enhanced variant of our method (CoMIF*) by using Llama3 8B Instruct as the decoder to generate responses. More details of this experimental settings can be found in the Appendix C.

The experimental results are shown in Table 3. We find that although LLM can generate more fluent and diverse responses, they have difficulty in comprehensively considering the associations between various interaction factors, which leads to lower BertScore and Relevance. CoMIF* achieve optimal performance by modeling complex multiple interaction factors for LLM, indicating that we can cleverly utilize a small post-adapter to improve the initiative and information of LLM generated responses.

5.4 Ablation Study

We conducted ablation studies to assess the impact of various factors on the performance of CoMIF. Specifically, we removed persona, emotion, and topic signals from the post-adaptation module individually and evaluated the model's performance using various metrics. The results are presented in Tabel 4.

After removing the persona signal, the diversity metric Distinct improved across both datasets, while the quality of the model's generated re-

Dataset	Model	PPL	Distinct-1	Distinct-2	ROUGE-1	ROUGE-2	BertScore	Relevance	Fluency	Informativeness
SPC	Llama3(8B)	-	**17.60**	**54.46**	18.02	6.55	88.17	1.33	**2.04**	1.68
	Llama3-ft(8B)	-	13.88	43.49	39.21	25.98	91.58	1.62	1.97	**1.71**
	CoMIF(0.2B)	2.81	10.58	33.52	43.11	29.95	91.88	1.57	1.84	1.61
	CoMIF*(8.1B)	**2.31**	10.27	33.70	**43.29**	**30.71**	**92.26**	**1.68**	1.93	1.70
PC	Llama3(8B)	-	**17.82**	**54.36**	12.43	2.35	86.73	1.16	**1.82**	**1.43**
	Llama3-ft(8B)	-	17.22	52.96	12.53	2.54	87.00	1.21	1.77	1.38
	CoMIF(0.2B)	19.90	7.01	21.52	**15.04**	3.01	87.23	1.24	1.63	1.35
	CoMIF*(8.1B)	**14.46**	9.23	30.36	14.85	**3.87**	**87.69**	**1.31**	1.72	1.39

Table 3: The performance of CoMIF and LLMs. We report the number of parameters of each model in symbol ().

Dataset	Model	D-1	D-2	R-1	R-2	BertScore
SPC	-w/o P	**10.93**	**34.37**	42.06	29.13	91.12
	-w/o E	10.67	33.37	42.43	29.30	91.19
	-w/o T	10.41	32.86	**43.39**	29.83	91.31
	CoMIF	10.58	33.52	43.11	**29.95**	**91.88**
PC	-w/o P	7.18	21.99	14.89	2.99	86.22
	-w/o E	**7.38**	**23.07**	14.86	2.96	86.22
	-w/o T	6.49	19.32	**15.06**	**3.06**	86.38
	CoMIF	7.01	21.52	15.04	3.01	**87.23**

Table 4: Results of ablation study on two datasets.

Personas	$\mathcal{P}_1$: I have a wife and three kids. $\mathcal{P}_2$: I like playing poker. $\mathcal{P}_3$: I do not have many friends. $\mathcal{P}_4$: I am a carpenter. $\mathcal{P}_5$: I used to drink but i stopped five years ago.
Context	A: What do you like to do for fun? B: I like to ride horses. A: That's cool!I've always wanted to learn how to ride a horse. B: It's a lot of fun, but it can be dangerous. A: What do you like about it? B: I like the feeling of freedom and the challenge of it.
Response	**Gold Response:** I can imagine. It must be really exhilarating.
	GPT2: That sounds amazing. I'm sure it's a lot of hard work, but it sounds like it would be a lot of fun.
	PPA: Cool. I've always wanted to learn how to build things.
	EmpSOA: I can see how that would be fun.
	CoMIF: I can imagine. I'm not sure if I could handle the stress of riding a horse, but it sounds like a lot of fun.

Table 5: A case extracted from Synthetic-Persona-Chat.

sponses declined. This suggests that incorporating persona factors effectively constrains the content generated by the model, reducing responses that do not align with the speaker's persona and producing responses that are more consistent with the speaker's situation.

The effect of removing the emotion signal on the model is similar to that of the persona signal. Although the *Distinct-2* (D-2) metric decreases after removing the emotion signal on the Synthetic-Persona-Chat dataset, the overall performance on both datasets suggests that the introduction of emotion factors moderately limits the diversity of the model's generated responses.

After removing the topic factor, the diversity of responses generated by CoMIF decreases, while the quality improves. This aligns with the previous example: considering the topic factor in the generation process makes the model more likely to explore new topics under certain conditions, rather than continuing with existing ones. However, more topics introduce greater uncertainty, which can reduce the quality of the generated responses to some extent.

5.5 Case Study

We provide an example from the Synthetic-Persona-Chat dataset in Table 5. We note that PPA mainly considers the personas of the speaker and thus changes the topic of the current conversation inappropriately. Although the response given by EmpSOA fits the character's emotions, it lacks consideration of the topic of the current conversation, thus giving a relatively generic response. Since GPT does not consider any interaction factors, the response it gives is too emotional which deviates from the speaker's situation, in addition, its response revolves around suboptimal topic (challenge) rather than optimal topic (horse riding). On the contrary, thanks to the comprehensive consideration of multiple interactive factors, CoMIF's response can better reflect the speaker's situation in terms of emotions, topics, etc., and are superior to other baselines.

6 Conclusion and Future Work

In this paper, we propose CoMIF that model complex multiple interaction factors for conversation generation. To achieve this goal, we jointly model the temporal dynamics within each factor based on a directed collaborative graph, and use a post-

adaptation module to dynamically inject various factors-related signals into the generation process. Experimental results on multiple datasets demonstrate the superiority of our model in generating responses that are appropriate to the speaker's situation.

In future research, we will delve deeper into the specific interactions of different interaction factors in order to gain a clearer understanding of how these factors affect dialogue generation.

Limitations

First, our method only uses RNN to predict emotion and topic representations in the dialogue history, without trying more clever sampling methods. Second, although we demonstrate the effectiveness of our method for LLM in Section 5.3, we did not test the effect of this method on other larger LLMs due to experimental costs.

Ethics Statement

In a broad sense, introducing personality information into conversations may indeed lead to user profile privacy leaks and false identity forgery. However, in this work, personality information and responses are limited to the scope of the experiment and are not enough to threaten real conversations. In addition, all models in this paper are trained and evaluated on datasets collected in the public corpus, and the dataset corpus is only used for experimental purposes. The dataset we use does not contain unethical language.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Chenghao Fan, Zhenyi Lu, Wei Wei, Jie Tian, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024a. On giant's shoulders: Effortless weak to strong by dynamic logits fusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shixuan Fan, Wei Wei, Xiaofei Wen, Xian-Ling Mao, Jixiong Chen, and Dangyang Chen. 2024b. Personalized topic selection model for topic-grounded dialogue. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7188–7202, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *Preprint*, arXiv:2312.10007.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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.

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. 2024. Reinforcement learning with token-level feedback for controllable text generation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1704–1719.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yifan Liu, Wei Wei, Jiayi Liu, Xian ling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Hongjin Qian and Zhicheng Dou. 2023. Topic-enhanced personalized retrieval-based chatbot. In *European Conference on Information Retrieval*, pages 79–93. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2492–2504, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Wei, Gao Cong, Xiaoli Li, See-Kiong Ng, and Guohui Li. 2011. Integrating community question and answer archives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1255–1260.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398, Toronto, Canada. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 35, pages 14176–14184.

Ming Yan, Xingrui Lou, Chien Aun Chan, Yan Wang, and Wei Jiang. 2023. A semantic and emotion-based dual latent variable generation model for a dialogue system. *CAAI Transactions on Intelligence Technology*, 8(2):319–330.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023a. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023b. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing-Qian Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI Conference on Artificial Intelligence*.

A Generate emotion labels and topic labels for the dataset

Our model needs to extract the corresponding historical sentiment sequences and historical topic sequences from the conversation history, and then model the interactions between these factors to generate the emotion factors and topic factors contained in the target response.

Since there are no emotion labels and topic labels in the dataset, we need to generate these labels through external tools.

For the emotion labels, we use the 'roberta-base-go_emotions' which based on RoBERTa-base (Liu et al., 2019) model and trained on the go_emotions (Demszky et al., 2020) dataset to classify the emotion of each utterance in the dialogue and obtain the multi-classification labels. According to public information, the model's Precision, Recall, and F1 metrics on the 28 classification labels of the go_emotions dataset are 0.542, 0.577, and 0.541, respectively, which meets our requirements. In this experiment, the threshold optimized by the author of the 'roberta-base-go_emotions' is used as the probability threshold of each emotion label.

For the topic labels, the KeyBERT model is used to extract keywords from the dialogue utterances. Specifically, a pretrained model is used to extract the sentence-level representation of the utterance; then, we use the same model to extract word embeddings of n-grams/phrases; finally, cosine similarity is utilized to identify the words or phrases that most closely match the sentence, and these highly matching terms are deemed the keywords of the sentence.

Relevance
2: Fits the speaker's situation well
1: Fits the speaker's situation in at least one aspect (emotion, topic or persona)
0: Irrelevant with the speaker's situation
Fluency
2: Fluent and easy to read
1: Grammatically formed
0: Not a complete sentence or hard to read
Informativeness
2: Have clear and specific meaning
1: Contain a few informative words
0: Meaningless sentence

Table 6: Criteria of human evaluation

Prompt
You are a chatbot with the following personas: <persona 1>,<persona 2>...<persona n> You need follow the above personas to chat with the user. The topic history is following: <topic 1>, <topic 2>, ..., <topic n> The emotion of each utterance is following: <emotion 1>,<emotion 2>...<emotion n> The conversation is following: Assistant: <utterance 1>, User: <utterance 2>, Assistant: <utterance 3>, ..., User: <utterance n>, Assistant:

Table 7: Prompts used in Llama3 and Llama3-ft

B Human Evaluation

Relevance is used to evaluate whether the generated response is consistent with the speaker's current

situation. *Fluency* is used to measure the fluency of generated utterances. *Informativeness* is used to evaluate whether the generated utterance revolves around the topics and user personas. The detailed scoring criteria are shown in Table 6.

C Experimental settings of LLM

All Llama3-ft models are obtained by fine-tuning the Llama3 8B Instruct model using the Lora (Hu et al., 2021) method for 3 epochs on the corresponding datasets. In the study of CoMIF*, in order to explore the importance of multi-interaction factor modeling on large language models (LLM), we froze all parameters of Llama3 and retrained only the post-adaptation module. The prompts used in Llama3 and Llama3-ft are shown in the Table 7.