CEM MoE: A Lightweight Sparse Mixture of Experts Framework for Empathetic Dialogue Generation

Zhinan Gou*, He Liu, and Yufan Wang
School of Big Data and Intelligent Engineering, Hebei University of Economics and Business, Shijiazhuang, China

***Abstract*—Empathetic dialogue generation requires models to understand user emotions and produce fluent, relevant, and emotionally appropriate responses. Existing lightweight models offer favorable deployment efficiency, but shallow emotion prediction layers limit their ability to represent fine-grained and dynamically changing emotional states. To address this limitation, CEM-MoE, a lightweight sparse mixture of experts (MoE) framework based on the Commonsense-aware Empathetic Chatting Machine (CEM), is proposed. Unlike the original CEM, CEM-MoE employs two independently routed sparse MoE streams to separately model emotional information from dialogue history and the current utterance. A bounded density-aware fusion mechanism integrates the two emotion distributions, where normalized utterance length serves as a lightweight density proxy and a sigmoid gate constrains the fusion coefficient to a valid range. Auxiliary load balancing regularization and label smoothing further improve expert utilization and generation stability. Experiments on the EmpatheticDialogues and ESConv datasets show that CEM-MoE achieves competitive generation performance with only marginal additional overhead. Compared with CEM, CEM-MoE reduces PPL from 37.76 to 36.59 and improves emotion accuracy from 32.10% to 35.47%, while increasing the parameter count by only 0.12M. Expert activation analysis reveals no obvious expert collapse, and human evaluation indicates higher scores in empathy, fluency, and relevance. The results verify that the dual-stream sparse MoE with density-aware fusion is effective for lightweight fine-grained empathetic dialogue modeling. The source code is publicly available at https://github.com/ZhinanGou/CEM-MoE.**

Keywords*—Empathetic dialogue generation; sparse mixture of experts; lightweight dialogue model; emotion modeling; expert routing

I. Introduction

Empathetic dialogue generation aims to enable dialogue systems to understand a user's emotional state and produce responses that are emotionally appropriate, semantically relevant, and linguistically fluent [1]. Unlike open-domain dialogue, empathetic dialogue requires not only grammatically correct responses but also accurate recognition of emotional cues expressed throughout a conversation. Responses are expected to remain consistent with the dialogue context while reflecting appropriate emotional understanding. Such capabilities are valuable for psychological support, intelligent customer service, emotional assistance, and human-computer interaction [2].

Large language models demonstrate strong performance in dialogue generation and emotion understanding [3], [4]. Large-scale pretraining and extensive model capacity enable natural and coherent responses while providing a certain level of understanding of complex emotional expressions. However, inference with large language models requires considerable computational resources, memory, and response latency [5], [6]. Recent LLM-based empathetic dialogue studies also show the potential of commonsense-enhanced generation [7]. The resulting computational and memory demands can make direct deployment infeasible in resource-constrained real-time interaction, edge computing, and privacy-sensitive scenarios. Improving empathetic dialogue generation under limited computational resources therefore remains an important research problem.

Compared with large language models, lightweight empathetic dialogue models are more suitable for practical deployment. Existing methods generally improve emotional appropriateness through emotion classification, emotion embeddings, commonsense knowledge, or multiple decoder architectures. For example, MoEL employs multiple emotion listeners to model different emotional states [8]. MIME enhances empathetic responses through emotion simulation [9]. The Commonsense-aware Empathetic Chatting Machine (CEM) introduces commonsense knowledge to improve the understanding of implicit emotions and user intentions [10]. Although these methods achieve encouraging performance, fine-grained emotion modeling remains insufficient. Many lightweight models still rely on shallow emotion prediction layers or shared representation spaces and therefore struggle to model complex, mixed, and dynamically changing emotional states. When emotions are implicitly expressed or emotional shifts arise across the dialogue history, generated responses tend to become generic, template-based, or emotionally inadequate.

Sparse mixture of experts (MoE) provides a promising solution to this problem. A gating network activates only a small number of experts during inference, which increases representation capacity without introducing substantial computational overhead. Different experts are able to learn complementary emotion patterns, alleviating the limited expressive capability of a single emotion prediction layer. Nevertheless, directly applying sparse mixture of experts introduces new challenges. Without appropriate regularization, the gating network may repeatedly select only a few experts, resulting in severe imbalance among experts. In addition, dialogue history and the current utterance contribute differently to emotion recognition. The current utterance usually provides explicit emotional cues, whereas dialogue history offers emotional evolution and contextual information. Inadequate

modeling may either ignore useful historical information or introduce unnecessary noise.

To address these issues, a lightweight sparse mixture of experts framework, named CEM-MoE, is proposed for empathetic dialogue generation. The framework is built upon the commonsense-enhanced empathetic dialogue model CEM. A dual-stream sparse emotion expert architecture is introduced during emotion modeling. One stream captures historical contextual emotion, while the other models emotional information contained in the current utterance. Each stream activates only the Top K experts through a gating network, and expert outputs are aggregated by weighted combination to improve emotion representation. Auxiliary load balancing regularization is further incorporated to encourage balanced expert utilization, while label smoothing improves generation stability during training.

Emotion fusion also requires careful consideration. Directly constructing fusion weights from utterance length is susceptible to scale variation and activation saturation, and unconstrained fusion weights cannot guarantee a valid probability distribution. A bounded density-aware emotion fusion mechanism is therefore introduced. Utterance length is first normalized to reduce scale variation. Raw density and scaled density features are then extracted and combined through a bounded fusion gate to generate the fusion coefficient. The final emotion distribution forms a valid combination of historical emotion and current emotion distributions.

The current utterance stream models direct emotional cues contained in the current user input. During both training and inference, this branch only receives the current utterance and its encoded representation. Ground truth responses, emotion labels, and any target side information are excluded from the input, ensuring consistent model inputs and avoiding information leakage.

Experiments are conducted on the EmpatheticDialogues dataset through both standard evaluation and controlled ablation studies. Results demonstrate that CEM-MoE achieves competitive generation performance while maintaining a relatively small parameter scale. Ablation studies further show that sparse mixture of experts, auxiliary load balancing regularization, and label smoothing mainly improve generation modeling by reducing perplexity and enhancing training stability. Experimental results also indicate that current utterance information contributes strongly to automatic evaluation on EmpatheticDialogues. Consequently, the role of the dual stream architecture is further examined from the perspectives of historical emotion modeling and expert utilization behavior rather than relying solely on automatic metrics.

Several complementary analyses are conducted. Additional experiments on ESConv evaluate the applicability of CEM-MoE to emotional support dialogue scenarios. Expert activation statistics together with emotion expert activation heatmaps examine routing behavior across both emotion streams and confirm that obvious expert collapse does not occur. Model efficiency is evaluated in terms of parameter size, inference latency, throughput, and peak GPU memory. Human evaluation and qualitative case analysis further assess empathy, fluency, and relevance beyond automatic metrics.

The main contributions are summarized as follows:

- A lightweight sparse mixture of experts framework is proposed for empathetic dialogue generation. The proposed framework enhances emotion representation with marginal additional computation by integrating sparse emotion experts into the CEM backbone and activating only a subset of experts through gated routing.
- A dual-stream emotion modeling architecture is designed to separately model historical contextual emotion and current utterance emotion. A bounded density-aware fusion mechanism is further introduced to integrate both emotion distributions while guaranteeing valid fusion weights and probabilistic consistency.
- Extensive experiments are conducted on EmpatheticDialogues and ESConv. Automatic evaluation, expert activation analysis, efficiency assessment, human evaluation, and case analysis provide a comprehensive evaluation of generation performance, expert utilization behavior, and deployment efficiency.

II. Related Work

A. Empathetic Dialogue Generation

Empathetic dialogue generation aims to enable dialogue systems to understand user emotional states and produce contextually appropriate responses with emotional responsiveness. Unlike general open-domain dialogue generation, empathetic dialogue generation focuses not only on grammatical correctness and semantic coherence, but also on the perception of user emotions, situations, and psychological needs. Therefore, the task usually requires modeling both explicit emotional cues and implicit emotional causes in user utterances.

Early studies on empathetic dialogue mainly rely on emotion labels, emotion embeddings, or retrieval-based methods to control response style. With the introduction of EmpatheticDialogues [1], generative empathetic dialogue models became a major research direction. The dataset provides a standard benchmark for empathetic dialogue generation and promotes the development of emotion-aware generation models. Subsequent studies further explore different emotion modeling strategies. MoEL employs multiple emotion listeners to model different emotional states and dynamically selects emotion-related representations according to the input [8]. MIME attempts to generate responses closer to human empathetic reactions through emotion simulation [9]. CEM introduces commonsense knowledge to improve the understanding of user situations, implicit emotions, and potential intentions [10]. Other studies further improve empathetic response generation through interactive emotion modeling, external knowledge enhancement, commonsense retrieval, and co-curricular learning [11]-[14]. Recent methods

also explore coarse-to-fine cognition-affection alignment and search-enhanced response generation [15], [16].

These methods demonstrate the importance of emotion information and commonsense knowledge for empathetic response generation, but several limitations remain. First, many lightweight models still rely on shallow linear mappings or shared representation spaces during emotion prediction, which limits their ability to capture complex, mixed, and fine-grained emotional states. Second, emotions in multi-turn dialogue do not always remain stable. The current utterance may directly express a new emotional state, while dialogue history provides the background of emotional change. Without an explicit mechanism to distinguish and integrate these two sources of information, models tend to generate generic or emotionally inadequate responses. Based on this observation, sparse mixture-of-experts-based emotion modeling is introduced into CEM to enhance emotion representation in lightweight models.

B. Lightweight Emotion Modeling and Sparse Mixture of Experts

Large language models exhibit strong capabilities in dialogue understanding and emotional generation [3], [4], [7], but large parameter scale and high inference cost limit their use in real-time interaction and edge deployment [5], [6]. In contrast, lightweight empathetic dialogue models have lower computational cost and better deployment feasibility. Improving emotion perception under a limited parameter budget therefore becomes an important issue in empathetic dialogue generation.

Sparse mixture of experts provides a feasible solution for lightweight emotion modeling [17], [18]. A mixture of experts architecture usually consists of multiple expert networks and a gating network. The gating network selects a subset of experts according to input features. Compared with a standard fully connected structure, sparse expert activation expands representation capacity without substantially increasing inference cost. In empathetic dialogue generation, different experts can learn local representations for different emotional patterns, which alleviates the limited expressive ability of a single emotion prediction layer.

Existing expert-based models in large-scale language modeling and dialogue tasks often focus on scalable modeling, feature fusion, or task adaptation [18]-[20]. The use of sparse experts for fine-grained emotion representation in lightweight empathetic dialogue generation remains insufficiently explored. To address this issue, CEM-MoE keeps sparse expert modeling and further analyzes expert activation distributions and emotion expert heatmaps. The analysis observes expert usage in different emotion streams and examines whether obvious expert collapse occurs. Expert activation patterns are further compared across emotion categories to characterize routing behavior. Compared with analysis based only on final classification metrics, such statistics provide additional information about how experts are utilized within the sparse expert structure.

C. Multi-Turn Emotion Modeling and Emotion Fusion

Emotional states in multi-turn empathetic dialogue are dynamic. On the one hand, the current utterance usually contains the most direct emotional cues. On the other hand, dialogue history provides the background of emotion formation and the continuation of dialogue topics. Therefore, modeling the relationship between historical emotion and current emotion is essential for generating high-quality empathetic responses.

Existing methods usually model emotional continuity through emotion embeddings, context encoders, commonsense knowledge, or fixed fusion strategies [10]-[14]. These methods help maintain emotional consistency to a certain extent, but two limitations remain. First, fixed fusion strategies are difficult to adapt to the dynamic importance of historical information and current information across different samples. Second, if the fusion coefficient lacks reasonable constraints, the resulting emotion distribution may not have a clear probabilistic interpretation.

Density-based confidence measurement provides a perspective for characterizing information concentration in representation space [21]. In CEM-MoE, utterance length is instead adopted as a lightweight density proxy, and a bounded fusion gate is introduced to integrate historical and current emotion distributions. The current utterance length is first normalized to reduce the influence of length scale on density features. A bounded fusion gate is then used to generate the fusion coefficient and keep it within a valid range. In this way, the final emotion distribution can be regarded as a valid combination of historical emotion and current emotion distributions. The mechanism provides a more explicit structured fusion strategy for multi-turn emotion modeling.

D. Evaluation and Efficiency Analysis for Empathetic Dialogue

Evaluation of empathetic dialogue generation usually includes automatic evaluation and human evaluation [1], [10]. Automatic metrics are commonly used to measure perplexity, emotion prediction accuracy, and response diversity, such as PPL, Accuracy, and Distinct. These metrics are useful for reproduction and comparison, but they cannot fully reflect empathetic response quality. A response may achieve low perplexity while still lacking empathy or contextual relevance. Therefore, human evaluation is often used as a complement to automatic metrics, assessing whether generated responses are natural and appropriate in terms of empathy, fluency, and relevance.

In addition, a key objective of lightweight empathetic dialogue models is to reduce deployment cost. Reporting only the number of parameters is not sufficient to reflect practical deployment efficiency. Real applications also require attention to inference latency, throughput, and peak GPU memory. Therefore, beyond parameter size, model efficiency is evaluated from inference latency, throughput, and peak memory usage.

III. Proposed Method

The overall architecture of CEM-MoE is shown in Fig. 1. The framework consists of four components: input encoding, dual-stream sparse mixture-of-experts emotion decoding, bounded density-aware emotion fusion, and response generation. Historical dialogue context and the current user

utterance are modeled through two independent emotion streams, whose predicted emotion distributions are integrated by a bounded fusion mechanism. The fused emotion representation is then used to guide empathetic response generation.

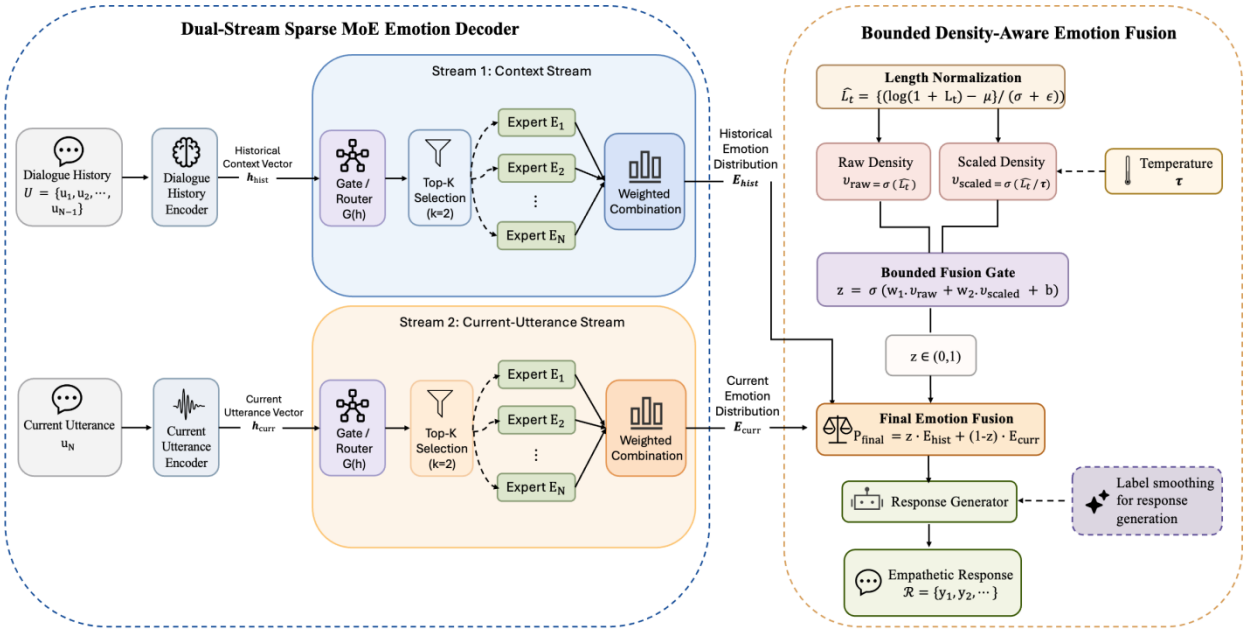

Fig. 1. Overall architecture of CEM-MoE.

A. Problem Definition

Given a multi-turn dialogue history,

$$U = \{u_1, u_2, \ldots, u_{N-1}\} \quad (1)$$

where, u_i denotes the i-th utterance in the dialogue history. The current user utterance is denoted as u_N.

The objective is to generate a semantically relevant and empathetic response according to the dialogue history U and the current utterance u_N:

$$R = \{y_1, y_2, \ldots, y_T\} \quad (2)$$

where, y_t denotes the t-th token in the response. During training, each sample contains the dialogue context, the current utterance, the ground truth response, and the emotion label. The model jointly optimizes two objectives. One is to predict the current emotional state of the user, and the other is to generate a contextually appropriate empathetic response.

The current utterance stream only uses the current user utterance and its encoded representation during both training and inference. Ground truth responses, ground truth emotion labels, and any target side information are not used as inputs. Such a setting avoids information leakage and keeps the input form consistent between training and inference.

B. Input Encoding

The model first encodes the dialogue history and the current utterance separately. Following CEM, both the dialogue history and the current utterance are encoded by Transformer encoders initialized with 300-dimensional pretrained GloVe embeddings. The two encoders use the same architecture but maintain independent parameters to preserve stream-specific emotion information. For the dialogue history U, a dialogue history encoder is used to obtain the historical context representation:

$$h_{hist} = \text{Encoder}_{hist}(U) \quad (3)$$

where, h_{hist} denotes the context vector encoded from the multi-turn dialogue history. The representation captures emotional background, topic continuation, and user state changes appearing in previous turns.

For the current utterance u_N, a current utterance encoder is used to obtain the current utterance representation:

$$h_{curr} = \text{Encoder}_{curr}(u_N) \quad (4)$$

where, h_{curr} denotes direct emotional cues contained in the current user input. Compared with the dialogue history, the current utterance usually expresses the current emotional state more directly. Therefore, a separate current utterance stream is built for subsequent emotion modeling.

Separate encoding explicitly distinguishes the dialogue history from the current utterance, rather than simply concatenating all inputs and feeding them into the same emotion prediction layer. The design has two advantages. First, the dialogue history and the current utterance can learn different emotion representations. Second, the fusion mechanism can dynamically integrate the two kinds of emotional information, instead of using a fixed strategy for multi-turn emotional changes.

C. Dual Stream Sparse Mixture of Experts Emotion Decoder

Traditional lightweight empathetic dialogue models usually convert context representations into emotion distributions

through a single linear mapping. The structure is simple and computationally efficient, but its expressive capacity is limited. Complex, mixed, and fine-grained emotional states are difficult to model with such a shallow structure. To improve emotion representation in lightweight models, a dual-stream sparse mixture-of-experts emotion decoder is introduced, following the general idea of sparse expert activation [17], [18].

The decoder contains two independent emotion streams. The first stream is the context stream, which models emotional background in dialogue history. The second stream is the current utterance stream, which models direct emotional cues in the current user utterance. The two streams share the same structure but use independent parameters.

For an arbitrary input representation h, the sparse mixture of experts module contains N_e expert networks:

$$\{E_1, E_2, \dots, E_{N_e}\}$$

Each expert network is implemented as a lightweight multilayer perceptron and learns local representations for different emotional patterns. The gating network first generates routing scores for all experts according to the input representation:

$$s = W_g h + b_g \tag{5}$$

where, W_g and b_g are learnable parameters, and s denotes the expert routing scores. To reduce computational cost, the model only selects the K experts with the highest scores:

$$S = \mathrm{TopK}(s, K) \tag{6}$$

where, $\mathcal{S}$ denotes the set of activated experts. After Top-K selection, the original routing scores do not directly form normalized weights over the selected experts. Therefore, the weights are recalculated within the selected expert set:

$$\tilde{g}_i = \frac{\exp(s_i)}{\sum_{j \in \mathcal{S}} \exp(s_j)}, \quad i \in \mathcal{S} \tag{7}$$

The final expert output is obtained by weighted summation:

$$o = \sum_{i \in \mathcal{S}} \tilde{g}_i \, E_i(h) \tag{8}$$

Sparse activation involves only a small subset of experts, enriching emotion representations while controlling inference cost. Compared with activating all experts, Top K routing is more suitable for lightweight empathetic dialogue generation.

For the context stream, the input is h_{hist}, and the sparse mixture of experts module produces the historical emotion representation:

$$o_{hist} = \mathrm{MoE}_{hist}(h_{hist}) \tag{9}$$

A classification layer then gives the historical emotion distribution:

$$E_{hist} = \mathrm{Softmax}(W_{hist} o_{hist} + b_{hist}) \tag{10}$$

For the current utterance stream, the input is h_{curr}, and the sparse mixture of experts module produces the current utterance emotion representation:

$$o_{curr} = \mathrm{MoE}_{curr}(h_{curr}) \tag{11}$$

The current utterance emotion distribution is further obtained as

$$E_{curr} = \mathrm{Softmax}(W_{curr} o_{curr} + b_{curr}) \tag{12}$$

where, E_{hist} and E_{curr} denote the emotion distributions predicted from the dialogue history and the current utterance, respectively. The dual stream structure explicitly distinguishes historical contextual emotion from current utterance emotion, and also provides a structural basis for subsequent emotion fusion and expert routing analysis.

D. Auxiliary Load Balancing Regularization

Sparse mixture of experts models tend to suffer from imbalanced expert utilization. The gating network may repeatedly select only a few experts, while other experts are rarely activated. Such a phenomenon leads to insufficient training of some experts and reduces the effectiveness of the sparse expert structure.

To mitigate this issue, auxiliary load balancing regularization is adopted. For each emotion stream, let f_i denote the proportion of samples in a batch that select the i-th expert, and let p_i denote the average routing probability assigned to the same expert. The load balancing loss for one emotion stream is defined as

$$\mathcal{L}_{aux}^{stream} = N_e \sum_{i=1}^{N_e} f_i \, p_i \tag{13}$$

where, N_e denotes the number of experts. The loss encourages more balanced expert usage and prevents routing from concentrating on only a few experts.

Since CEM-MoE contains both the context stream and the current utterance stream, the final auxiliary load balancing loss is formed by the losses of the two streams:

$$\mathcal{L}_{aux} = \frac{1}{2}\left(\mathcal{L}_{aux}^{hist} + \mathcal{L}_{aux}^{curr}\right) \tag{14}$$

where, $\mathcal{L}_{aux}^{hist}$ denotes the load balancing loss of the context stream, and $\mathcal{L}_{aux}^{curr}$ denotes the load balancing loss of the current utterance stream. The regularization mainly improves expert utilization and training stability rather than directly optimizing emotion classification accuracy.

E. Bounded Density Aware Emotion Fusion

Historical emotion and current emotion are both important in multi-turn dialogue. Dialogue history provides emotional background and conversational continuity, while the current utterance usually contains the most direct emotional expression. Therefore, an effective fusion mechanism is required to integrate E_{hist} and E_{curr}.

Directly using current utterance length to construct density features may cause rapid saturation of the Sigmoid function due to scale variation, reducing the discriminative ability of the features. In addition, if the fusion coefficient lacks boundary constraints, the final emotion distribution may fail to remain a valid probability distribution. To address these issues, a bounded density-aware emotion fusion mechanism is adopted.

The length-based feature serves as a lightweight density proxy rather than a direct implementation of semantic density in semantic representation space.

First, the current utterance length L_t is normalized as

$$\hat{L}_t = \frac{\log(1 + L_t) - \mu}{\sigma + \epsilon} \quad (15)$$

where, μ and σ denote the mean and standard deviation of utterance lengths in the training set, ϵ is a smoothing term to avoid division by zero, and $\hat{L}_t$ denotes the normalized length feature. The normalization reduces the influence of raw length scale on the Sigmoid function.

The model then calculates the raw density feature and the scaled density feature:

$$v_{raw} = \sigma(\hat{L}_t) \quad (16)$$

$$v_{scaled} = \sigma(\hat{L}_t/\tau) \quad (17)$$

where, τ is a temperature parameter that adjusts the smoothness of density features. v_{raw} is directly obtained from the normalized length feature, and v_{scaled} is obtained after temperature scaling.

To keep the fusion coefficient within a valid range, a bounded fusion gate generates the coefficient z:

$$z = \sigma(w_1 v_{raw} + w_2 v_{scaled} + b) \quad (18)$$

where, w_1, w_2, and b are learnable parameters. Since z is constrained by the Sigmoid function, it satisfies:

$$z \in (0,1)$$

The final emotion distribution is obtained by fusing the historical emotion distribution and the current utterance emotion distribution:

$$P_{final} = zE_{hist} + (1 - z)E_{curr} \quad (19)$$

Since $z \in (0,1)$, and both E_{hist} and E_{curr} are probability distributions, P_{final} is a convex combination of the two distributions and remains a valid probability distribution. The fusion mechanism integrates historical emotion information and current utterance emotion information in a bounded, differentiable, and probabilistically valid manner.

F. Response Generation

After obtaining the final emotion distribution P_{final}, the model uses it as emotional condition information for the response generator. The generator produces the final response according to the historical context representation, the current utterance representation, and the fused emotion distribution.

At the t-th decoding step, the generator predicts the probability distribution of token y_t:

$$P(y_t|y_{<t}, U, u_N, P_{final}) = \text{Decoder}(y_{<t}, h_{hist}, h_{curr}, P_{final}) \quad (20)$$

where, $y_{<t}$ denotes the previously generated tokens. The generation probability of the whole response is

$$P(R|U, u_N) = \prod_{t=1}^{T} P(y_t|y_{<t}, U, u_N, P_{final}) \quad (21)$$

By introducing the fused emotion distribution, the response generator considers not only semantic context but also emotional state information. The design reduces the tendency to rely entirely on frequent safe responses and helps generate more emotionally responsive text.

G. Label Smoothing for Response Generation

Empathetic dialogue generation often suffers from frequent template responses. A model may overfit common expressions in the training set, such as I am sorry to hear that or That is great. These responses may lead to low perplexity, but they do not necessarily provide sufficient empathy or contextual relevance.

To alleviate overfitting, label smoothing is applied to response generation [22]. Let y denote the one-hot distribution of the ground truth token. The smoothed label distribution is defined as

$$y^{LS} = (1 - \epsilon)y + \frac{\epsilon}{|V|}\mathbf{1} \quad (22)$$

where, ϵ is the smoothing coefficient, $|V|$ denotes the vocabulary size, and $\mathbf{1}$ is an all-one vector over the vocabulary dimension. The generation loss is defined as the cross-entropy between the predicted distribution and the smoothed label distribution:

$$\mathcal{L}_{gen}^{LS} = -\sum_{t=1}^{T}\sum_{w \in V} y_{t,w}^{LS} \log P(y_t = w) \quad (23)$$

Label smoothing reduces overconfidence in a single target token and makes training more stable. Ablation results show that removing label smoothing increases perplexity, while emotion classification accuracy may increase. These results suggest that label smoothing mainly improves generation modeling stability rather than improving all metrics at the same time.

H. Training Objective

The training objective consists of three parts, including generation loss, emotion classification loss, and auxiliary load balancing loss.

The generation loss $\mathcal{L}_{gen}^{LS}$ optimizes response generation quality.

The emotion classification loss supervises the final emotion distribution P_{final} with the ground truth emotion label:

$$\mathcal{L}_{emo} = -\log P_{final}(e) \quad (24)$$

where, e denotes the ground truth emotion label.

The auxiliary load balancing loss $\mathcal{L}_{aux}$ reduces expert routing imbalance.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{gen}^{LS} + \alpha\mathcal{L}_{emo} + \lambda\mathcal{L}_{aux} \quad (25)$$

where, α and λ are loss weighting coefficients. By jointly optimizing these terms, the model learns response generation, emotion recognition, and expert routing allocation at the same time.

Different loss terms correspond to different functional objectives. The generation loss ensures response quality, the emotion classification loss enhances emotion perception, the auxiliary load balancing loss alleviates expert collapse, and label smoothing improves generation stability. The joint objective aligns the training targets with the functional modules in the model architecture.

IV. EXPERIMENTS

CEM-MoE is evaluated from several perspectives, including automatic evaluation, ablation analysis, additional evaluation on ESConv, expert routing, efficiency assessment, human evaluation, and case analysis.

A. Experimental Settings

1) Datasets: Experiments are mainly conducted on EmpatheticDialogues [1]. The dataset is a widely used benchmark for empathetic dialogue generation, containing about 25,000 multi-turn dialogues with 32 emotion categories. Following existing work, the dataset is divided into training, validation, and test sets. The test set is used to evaluate response generation and emotion recognition. Given the dialogue context and the current user utterance, the model generates an empathetic response and predicts the corresponding emotion category.

To further examine model applicability in different emotional dialogue scenarios, an additional evaluation is conducted on ESConv [2]. ESConv is a multi-turn dataset for emotional support conversation and differs from EmpatheticDialogues in dialogue objective, emotion distribution, and response style. The dataset is therefore used to examine whether CEM-MoE remains effective in a different emotional support dialogue setting rather than to evaluate cross-domain transfer. Both Original CEM and CEM-MoE are trained and evaluated on the standard ESConv split using the same optimization and decoding settings unless otherwise stated.

2) Evaluation metrics: Model performance is evaluated with both automatic metrics and human evaluation. Automatic metrics include perplexity, emotion accuracy, and generation diversity.

Perplexity measures the language modeling ability for target responses. A lower value implies tighter alignment to ground truth responses. Emotion accuracy measures the ability to predict user emotion categories. A higher value indicates better emotion recognition. Generation diversity is evaluated by Dist-1 and Dist-2, which measure the proportions of distinct unigrams and bigrams in generated responses. Higher scores correspond to more diverse and less redundant outputs.

Empathetic dialogue generation cannot be fully evaluated by automatic metrics alone. Low perplexity may come from safe, frequent, or template-based responses, which do not necessarily contain sufficient empathy or contextual relevance. Therefore, human evaluation is also conducted. Three aspects are considered, including empathy, fluency, and relevance. Each score ranges from 1 to 5, and a higher score indicates better response quality from the perspective of human evaluators.

3) Implementation details: CEM is used as the base framework, and a dual-stream sparse mixture-of-experts structure is introduced into the emotion modeling stage. The context encoder is initialized with 300-dimensional pretrained GloVe word embeddings [23], and the hidden size is set to 300. The dual-stream sparse mixture of experts emotion decoder contains a context stream and a current utterance stream. Each stream contains 4 expert networks, and Top K is set to 2, which means only the 2 experts with the highest routing scores are activated each time.

The model is implemented with PyTorch [24] and trained on a single NVIDIA RTX 4090 GPU. Adam is used as the optimizer with a learning rate of 1×10^{-4}. The batch size is set to 16, and the model is trained for 20 epochs. The label smoothing coefficient is set to 0.1, and the auxiliary load balancing loss weight λ is set to 0.01. Beam search is used during decoding, with a beam size of 5. The temperature parameter τ is set to 2.0, and the emotion classification loss weight α is set to 1.0.

To evaluate practical deployment overhead, Original CEM and CEM-MoE are compared under the same hardware environment in terms of parameter size, inference latency, throughput, and peak GPU memory. This setting is used to analyze the extra computational cost introduced by sparse experts and to examine whether the model retains lightweight characteristics.

B. Main Results on EmpatheticDialogues

To evaluate the overall performance of CEM-MoE, it is compared with several representative empathetic dialogue generation models, including Transformer [1], CEM [10], KEMP [14], CASE [15], and SeeKeR [16]. The results are shown in Table I.

Table I shows that CEM-MoE obtains the lowest PPL, indicating that sparse expert-based emotion modeling helps improve response generation modeling. Compared with the original CEM, the emotion accuracy increases from 32.10% to 35.47%, which suggests that the emotion expert structure enhances emotion perception to some extent. CEM-MoE also achieves higher Dist-1 and Dist-2, indicating better response diversity.

TABLE I. MAIN RESULTS ON EMPATHETICDIALOGUES

Model	PPL ↓	Acc ↑	Dist-1 ↑	Dist-2 ↑	Params
Transformer	37.80	-	0.53	1.89	12.55M
CEM	37.76	32.10	0.49	3.24	17.49M
KEMP	37.00	34.80	0.55	3.61	17.90M
CASE	36.92	34.55	0.63	3.42	20.55M
SeeKeR	37.09	37.10	0.61	3.23	32.10M
CEM-MoE	36.59	35.47	0.78	3.88	17.61M

CEM-MoE does not achieve the best result across every metric. SeeKeR still obtains higher emotion accuracy, and the PPL improvement of CEM-MoE over CASE is relatively limited. Therefore, CEM-MoE is not described as an overall best model. Instead, the results show that CEM-MoE achieves competitive generation performance with a relatively small parameter scale, demonstrating a favorable balance between performance and efficiency.

C. Ablation Study

To analyze the contribution of each core component, sparse experts, auxiliary load balancing regularization, and label smoothing are removed separately. The results are shown in Table II.

TABLE II. ABLATION RESULTS ON EMPATHETICDIALOGUES

Variant	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑
CEM-MoE	36.59	0.78	3.88	35.47
w/o Sparse MoE	37.24	0.67	3.81	35.78
w/o Aux Loss	37.07	0.59	3.72	36.46
w/o Label Smoothing	36.95	0.46	3.64	37.43

As shown in Table II, removing Sparse MoE increases PPL from 36.59 to 37.24, which indicates that the sparse expert structure helps improve generation modeling. Compared with a single linear emotion prediction layer, multiple experts provide more flexible local representations for different emotional patterns and alleviate the limited emotion representation capacity of lightweight models.

After removing auxiliary load balancing regularization, PPL increases to 37.07, indicating that the regularization contributes to training stability and balanced expert utilization. Although the variant obtains slightly higher Acc, the result does not mean that load balancing is ineffective. The main role of load balancing is not to directly improve emotion classification accuracy, but to avoid routing collapse onto a small subset of experts and to stabilize the generation modeling process.

After removing label smoothing, PPL increases to 36.95, and Dist-2 decreases to 3.64. The result suggests that label smoothing helps reduce overfitting to frequent expressions. The variant obtains higher Acc, which indicates that label smoothing does not consistently boost emotion classification accuracy. Its main contribution lies in improving response generation stability.

Overall, Sparse MoE, Aux Loss, and Label Smoothing have positive effects on PPL and generation diversity. These components mainly improve response generation modeling and training stability. Meanwhile, removing some regularization components increases Acc, indicating that these modules do not directly improve emotion classification accuracy in all cases. Therefore, the ablation results are interpreted conservatively, and no single module is described as producing consistent gains on all metrics.

D. Dual Stream Emotion Modeling and Fusion Behavior Analysis

To further analyze the roles of the context stream and the current utterance stream, three variants are listed separately, including Context only, Current only, and w/o Fusion. The experiment analyzes the influence of different emotion streams and fusion strategies on model behavior. The results are shown in Table III.

TABLE III. ANALYSIS OF EMOTION STREAMS AND FUSION BEHAVIOR

Variant	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑
CEM-MoE	36.59	0.78	3.88	35.47
Context only	36.77	0.73	3.98	33.78
Current only	36.58	0.81	3.92	38.08
w/o Fusion	36.51	0.77	3.79	37.83

Table III shows that Current only and w/o Fusion achieve competitive performance on PPL and Acc. This suggests that the current utterance usually contains direct emotional cues in EmpatheticDialogues and contributes more clearly to automatic metrics. For example, the last user utterance often explicitly expresses emotions such as nervous, sad, happy, or grateful. A model relying only on the current utterance can still predict emotions and generate responses reasonably well.

Context only obtains higher Dist-2 but lower Acc. The result suggests that dialogue history may provide richer bigram combinations, but relying only on historical context is less beneficial for current emotion recognition. In multi-turn dialogue, earlier emotions do not always match the target response. Excessive dependence on historical emotion may affect current response prediction.

The w/o Fusion variant obtains strong PPL and Acc, but both Dist-1 and Dist-2 are lower than those of the full model. The result indicates that removing fusion may lead the model to generate more stable but less diverse responses. Such responses may be favorable for PPL, but they do not necessarily have better empathetic quality. Therefore, PPL and Acc are not used as the only criteria. Dist, human evaluation, expert routing, and case analysis are also considered for comprehensive judgment.

Based on these results, the dual stream structure is better interpreted as a way to organize contextual and current emotional cues and to support expert utilization analysis, rather than as a component that consistently improves all automatic metrics. Expert activation analysis further shows that the context stream and the current utterance stream have different expert usage patterns, indicating that the two streams provide different routing behaviors.

E. Comparison of Different Fusion Strategies

To further evaluate the effectiveness of the proposed bounded density-aware fusion mechanism, CEM-MoE is compared with two alternative fusion strategies: fixed average fusion and learned scalar fusion. Fixed average fusion assigns equal weights to historical and current emotion distributions, while learned scalar fusion introduces a learnable scalar coefficient without the proposed density-aware mechanism. All methods use the same dual-stream emotion representations and training settings. The results are shown in Table IV.

Table IV shows that CEM-MoE achieves the lowest PPL among the three fusion strategies, indicating improved response generation modeling. Learned Scalar Fusion yields higher diversity scores, suggesting that simple adaptive weighting can also boost response variation. Nevertheless, it does not explicitly account for the differing contributions of historical and current emotional information. The proposed bounded density-aware fusion mechanism offers a structured fusion strategy while preserving probabilistic consistency.

TABLE IV. COMPARISON OF DIFFERENT FUSION STRATEGIES

Model	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑
CEM-MoE	36.6896	0.76	3.82	35.47
Fixed Average Fusion	37.4379	0.78	3.84	37.87
Learned Scalar Fusion	37.0666	0.85	4.09	37.98

F. Additional Evaluation on ESConv

To further examine model applicability in different emotional dialogue scenarios, an additional evaluation is conducted on ESConv. Since ESConv differs from EmpatheticDialogues in task formulation, emotion distribution, and dialogue objective, only Original CEM and CEM-MoE are compared. The purpose is to examine whether the proposed architecture remains effective when trained and evaluated in an emotional support dialogue setting. The results are presented in Table V.

TABLE V. RESULTS ON ESCONV

Dataset	Model	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑
ESConv	Original CEM	60.02	0.56	3.71	45.26
ESConv	CEM-MoE	56.98	0.65	3.96	44.53

Table V shows that CEM-MoE reduces PPL from 60.02 to 56.98 on ESConv. Dist-1 increases from 0.56 to 0.65, and Dist-2 rises from 3.71 to 3.96. These results indicate that sparse expert-based emotion modeling still provides generation modeling advantages in emotional support dialogue and improves response diversity.

However, Acc drops from 45.26 to 44.53, which is slightly lower than Original CEM. The decrease suggests that emotion classification performance may vary across datasets because of differences in label distributions and task objectives. ESConv focuses more on emotional support strategies, while EmpatheticDialogues focuses more on responding to emotional experiences. The label spaces and emotional expression patterns are not fully aligned.

Due to these differences, ESConv is used as an additional dataset to evaluate CEM-MoE in an emotional support dialogue setting. The results show that CEM-MoE maintains favorable response generation performance on ESConv, but they are not interpreted as evidence of cross-domain transfer or universally improved emotion classification across datasets.

G. Expert Activation and Routing Analysis

An important issue in sparse mixture of experts models is whether experts are effectively used. If the gating network repeatedly selects only a few experts, expert collapse occurs.

To examine expert routing behavior in CEM-MoE, expert activation counts are collected for both emotion streams, and both utilization entropy and normalized entropy are calculated. The statistics are collected on the EmpatheticDialogues test set, where each count represents the number of times an expert is selected by the Top K router. Let qi denote the proportion of activation counts assigned to the i-th expert. Expert utilization entropy is calculated as $H = -\Sigma i\ qi \log qi$, and normalized entropy is defined as H/log Ne, where Ne = 4. The results are shown in Table VI and Fig. 2.

As shown in Fig. 2, the current stream and the context stream exhibit distinct routing patterns across different emotion categories. The current stream achieves more balanced expert utilization, whereas the context stream shows a stable preference toward certain experts. These observations demonstrate that the dual-stream structure enables differentiated expert allocation behaviors.

The normalized entropy of the current stream is 0.9792, which is close to 1. The result indicates relatively balanced expert utilization in the current utterance stream and no obvious expert collapse. The normalized entropy of the context stream is 0.8880, indicating moderate preference for some experts, but the stream does not collapse into a single expert. This pattern is consistent with the more complex nature of historical context, which may contain longer-term emotional bias. We further visualize an emotion expert activation heatmap to examine expert activation ratios across different emotion categories, where each cell corresponds to the activation ratio for a specific emotion category.

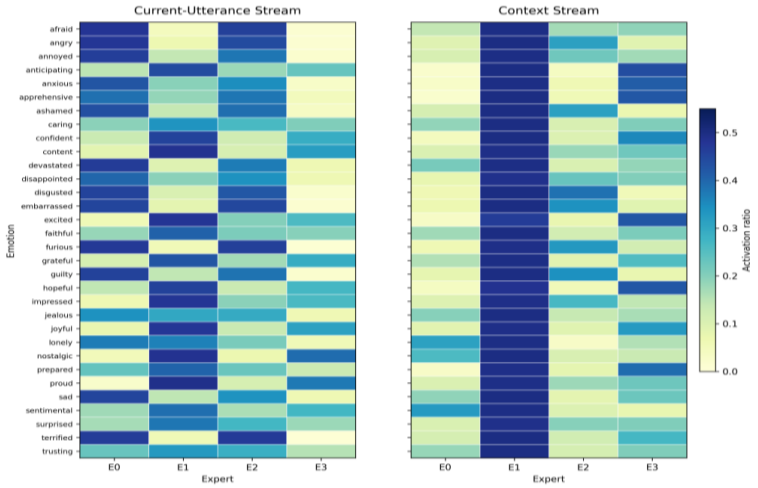

Fig. 2. Emotion expert activation heatmap.

TABLE VI. EXPERT UTILIZATION STATISTICS

Stream	Counts	Entropy	Normalized Entropy
Current stream	{0: 2900, 1: 3106, 2: 2904, 3: 1600}	1.3575	0.9792
Context stream	{0: 1262, 1: 5210, 2: 1690, 3: 2348}	1.2311	0.8880

TABLE VII. EFFICIENCY RESULTS

Model	Params	Latency ↓	Throughput ↑	Peak GPU Mem
Original CEM	17.49M	37.22 ms/sample	275.12 tokens/s	83.85 MB
CEM-MoE	17.61M	37.29 ms/sample	260.27 tokens/s	84.30 MB

H. Efficiency Analysis

One objective of CEM-MoE is to improve emotion modeling while maintaining a lightweight architecture. Therefore, besides generation performance, it is also important to evaluate the additional computational overhead introduced by Sparse MoE. Sparse expert activation aims to improve model capacity while controlling active computation, making it suitable for lightweight dialogue generation models [17], [18]. To provide a comprehensive efficiency evaluation, Original CEM and CEM-MoE are compared in terms of parameter size, inference latency, throughput, and peak GPU memory under the same experimental environment. These metrics reflect both deployment cost and practical inference efficiency. The results are presented in Table VII.

Table VII shows that CEM-MoE increases the parameter size by only 0.12M compared with the original CEM. Inference latency rises from 37.22 ms/sample to 37.29 ms/sample, with an increase of only 0.07 ms/sample. Peak GPU memory grows from 83.85 MB to 84.30 MB, with an increase of only 0.45 MB. Although throughput drops from 275.12 tokens/s to 260.27 tokens/s, the overall inference efficiency remains at a high level.

I. Human Evaluation

Automatic metrics cannot fully reflect the quality of empathetic dialogue generation. Low perplexity or high emotion classification accuracy does not guarantee that a

response is empathetic or contextually appropriate. Therefore, human evaluation is conducted to provide a more comprehensive assessment of response quality. To complement automatic metrics such as PPL, Acc, and Dist, a total of 100 samples are randomly selected from the test set, and three annotators are invited to score the responses generated by Original CEM and CEM-MoE.

The evaluation covers three dimensions: empathy, fluency, and relevance, with scores ranging from 1 to 5. Empathy evaluates whether the response appropriately reflects the user ' s emotional state; fluency measures the naturalness of the generated text; relevance assesses consistency with the dialogue context. During annotation, model identities are blinded, and response order is randomly shuffled to mitigate subjective bias. Krippendorff' s alpha is adopted to compute inter-annotator agreement, yielding values of 0.72, 0.79, and 0.75 for empathy, fluency, and relevance, respectively. The human evaluation results are summarized in Table VIII.

As illustrated in Table VIII, CEM-MoE surpasses the original CEM across all three evaluation dimensions: empathy gains 0.63, fluency rises by 0.66, and relevance increases by 0.55. While CEM-MoE does not attain optimal performance on every automatic metric, its generated responses achieve higher human scores than CEM in empathy, fluency, and relevance.

These findings also demonstrate that automatic evaluation and human evaluation complement each other for empathetic dialogue generation. PPL and Acc quantify model fitting capability and emotion prediction performance, whereas human evaluation more directly reflects whether the generated responses deliver genuine empathetic effects.

TABLE VIII. HUMAN EVALUATION RESULTS

Model	Empathy ↑	Fluency ↑	Relevance ↑
CEM	3.32	3.19	3.32
CEM-MoE	3.95	3.85	3.87

J. Case Analysis

To enable an intuitive comparison of model-generated responses, three representative examples are chosen for case study, covering negative emotion, anxious emotion, and positive emotion. The illustrative cases are listed in Table IX.

TABLE IX. RESPONSE CASES GENERATED BY DIFFERENT MODELS

Emotion / Context	CEM	CEM-MoE	Reference
devastated The hardest time in my life was when my best friend died on Christmas of 2007. She fell off a roof after falling asleep and slipped when she got up.	oh no ! i hope you can find a friend who was in the end of your life .	oh no ! i am so sorry to hear that .	that is horrible . i am really sorry about your friend .
anxious I took a huge test today, and it made me nervous. I studied every day for a week and almost wore myself out studying.	that is great ! i hope you get it back soon .	i hope you do well !	then i am sure you will do well .
impressed My husband let me sleep in, cooked dinner, and helped clean the house. He has been very helpful today.	that is great ! i am sure he will be fine	that is nice of him .	oh wow , that was nice of him .

In the first case, the user describes the death of a close friend. CEM generates a semantically incoherent response, while CEM-MoE directly responds to the sadness and expresses sympathy. In the second case, the user expresses anxiety concerning a critical examination. The response generated by CEM fails to align closely with this exam scenario, while CEM-MoE provides more relevant encouragement. In the third case, the user describes a positive experience in which the husband helps with housework. CEM tends to generate a generic template response, while CEM-MoE responds more directly to the event.

These cases qualitatively demonstrate that CEM-MoE is capable of generating responses with greater emotional appropriateness and contextual relevance under representative scenarios. Nevertheless, case analysis merely offers qualitative evidence and cannot replace automatic metrics or human evaluation. It also should not be interpreted as strict proof that the model is superior across all aspects.

V. DISCUSSION

The experimental results uncover several characteristics and limitations of CEM-MoE. First, the dual stream architecture does not consistently outperform the current-only variant on all automatic metrics. Current-only achieves competitive PPL and higher emotion accuracy on EmpatheticDialogues, indicating that the current utterance often contains strong direct emotional cues. Therefore, the context stream is better regarded as a complementary source of historical emotional information rather than a component that guarantees improvement in every metric. The dual stream design mainly provides a structured way to distinguish current emotional cues from historical context and allows their contributions to be analyzed separately.

Second, the length-based feature used in the bounded fusion mechanism is a lightweight heuristic proxy rather than a direct measure of semantic density. Utterance length cannot fully characterize semantic information concentration, but it provides a simple signal without introducing an additional semantic encoder or confidence estimator. The proposed

mechanism therefore focuses on maintaining low computational cost and a bounded probabilistic fusion process. More informative semantic- or confidence-based fusion signals can be explored in future work.

Third, CEM-MoE retains the 300-dimensional GloVe-based representation setting of the CEM backbone to maintain a lightweight model size and a controlled comparison with the original framework. Static word embeddings, however, have limited ability to represent context-dependent and implicit emotional meanings. Compact contextual encoders may provide richer semantic representations, but they may also introduce additional computational overhead. Investigating the trade-off between contextual representation ability and deployment efficiency remains an important direction.

Finally, expert activation statistics and routing heatmaps provide information about expert utilization patterns, but they do not constitute a full explanation of model decision-making. Different activation patterns indicate that the context stream and the current utterance stream adopt distinct expert allocation strategies. However, the current analysis establishes no explicit mapping between individual experts and specific human-interpretable emotion concepts. The routing analysis should therefore be viewed as an examination of expert utilization behavior rather than a causal interpretation of response generation.

VI. Conclusion and Future Work

CEM-MoE is proposed as a lightweight sparse mixture of experts framework to address the limitations of lightweight empathetic dialogue generation models in fine-grained emotion modeling and dynamic emotion fusion. Built upon CEM, CEM-MoE introduces a dual-stream sparse expert structure during emotion modeling. The context stream models emotional information from dialogue history, while the current utterance stream captures emotional cues in the current user utterance. Through Top-K routing, only a small number of emotion experts are activated for computation, which enhances emotion representation while keeping additional overhead limited. Auxiliary load balancing regularization is adopted to alleviate imbalanced expert utilization. A bounded density-aware fusion mechanism is also used to integrate historical emotion and current emotion distributions, endowing the fusion process with a clearer probabilistic interpretation.

Experimental results on EmpatheticDialogues show that CEM-MoE achieves competitive generation performance. Ablation experiments indicate that the sparse expert structure, auxiliary load balancing regularization, and label smoothing mainly help reduce perplexity and improve generation stability. Additional experiments on ESConv further indicate that CEM-MoE remains effective for emotional support conversation. Expert activation statistics and emotion expert heatmaps indicate that the model does not suffer from obvious expert collapse and that the two emotion streams exhibit different expert utilization patterns. Efficiency experiments show that introducing the sparse expert module only brings negligible increases in parameter scale, latency, and GPU memory. Human evaluation and case analysis further show that responses generated by CEM-MoE are perceived as more empathetic, fluent, and relevant by human evaluators.

Overall, CEM-MoE provides a practical framework for lightweight empathetic dialogue generation while maintaining competitive generation quality, low computational overhead, and observable expert utilization behavior. Future work will focus on exploring richer emotional cues and extending CEM-MoE to multimodal empathetic dialogue generation by incorporating nonverbal information such as speech prosody, facial expressions, and visual behaviors. Fine-grained human feedback can also be incorporated to further improve emotional appropriateness and interaction quality.

Acknowledgment

This work is funded by the Scientific Research and Development Program of Hebei University of Economics and Business under Grant No. 2026ZD11.

References

[1] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, “Towards Empathetic Open-domain Conversation Models,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5370–5381.

[2] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang, “Towards Emotional Support Dialog Systems,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 3469–3483.

[3] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[4] D. Guo, D. Yang, H. Zhang, et al., “DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning,” Nature, vol. 645, pp. 633–638, 2025.

[5] Z. Zhou, X. Ning, K. Hong, T. Fu, J. Xu, S. Li, Y. Lou, L. Wang, Z. Yuan, X. Li, et al., “A Survey on Efficient Inference for Large Language Models,” arXiv preprint arXiv:2404.14294, 2024.

[6] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with PagedAttention,” in Proceedings of the 29th Symposium on Operating Systems Principles, 2023, pp. 611–626.

[7] L. Wang, J. Li, C. Yang, Z. Lin, H. Tang, H. Liu, Y. Cao, J. Wang, and W. Wang, “Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference,” in Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 123–140.

[8] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, “MoEL: Mixture of Empathetic Listeners,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 121–132.

[9] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, “MIME: MIMicking Emotions for Empathetic Response Generation,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 8968–8979.

[10] S. Sabour, C. Zheng, and M. Huang, “CEM: Commonsense-aware empathetic response generation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11229–11237.

[11] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, “EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4454–4466.

[12] Y. Deng, W. Zhang, Z. Li, et al., “Knowledge-Enhanced Empathetic Dialogue Generation with Adaptive Graph Convolutional Networks,” Knowledge-Based Systems, vol. 260, p. 110156, 2023.

[13] Y. Qian, W. Zhang, Z. Wei, et al., “Think Twice: A Graph-based Commonsense Knowledge Retrieval for Empathetic Response Generation,” in ICASSP 2023, 2023, pp. 1–5.

[14] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen, "Knowledge bridging for empathetic dialogue generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 10993–11001.

[15] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang, "CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 8223–8237.

[16] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston, "Language models that seek for knowledge: Modular search and generation for dialogue and prompt completion," in Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 373–393.

[17] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," Journal of Machine Learning Research, vol. 23, no. 120, pp. 1–39, 2022.

[18] B. Zoph, I. Bello, S. Kumar, et al., "ST-MoE: Designing Stable and Transferable Sparse Expert Models," arXiv preprint arXiv:2202.08906, 2022.

[19] A. Q. Jiang, A. Sablayrolles, A. Roux, et al., "Mixtral of Experts," arXiv preprint arXiv:2401.04088, 2024.

[20] Y. Tian, F. Xia, and Y. Song, "Dialogue Summarization with Mixture of Experts based on Large Language Models," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024, pp. 7143–7155.

[21] X. Qiu and R. Miikkulainen, "Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space," in Advances in Neural Information Processing Systems, vol. 37, 2024.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems, vol. 32, 2019.