EmoDial-Reason: Unveiling Affective Reasoning in Speech-Emotion Dialogue

Shubei Tang[1,2], Lu Xiang[1,2], Yaping Zhang[1,2], and Chengqing Zong[1,2] (✉)

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
tangshubei2023@ia.ac.cn, {lu.xiang, yaping.zhang, cqzong}@nlpr.ia.ac.cn

Abstract. Understanding the affective reasoning behind empathetic responses is critical for developing reliable speech-emotion dialogue systems, yet current Large Language Models remain black boxes. To address this, we propose EmoDial-Reason, a novel dataset in which each example is paired with two reasoning paths: free-form reasoning in which the model "thinks aloud" without constraints, and template-guided reasoning that follows a structured cognitive pathway. Upon this, we explore whether explicit reasoning helps and, if so, which reasoning style yields the most benefits. Our findings emphasize that (1) explicit affective reasoning consistently enhances performance and transparency, (2) template-guided reasoning excels on easier scenarios whereas free-form reasoning is superior on complex situations, and (3) a hybrid approach that enables the model to dynamically select between template-guided and free-form reasoning achieves the best overall results. [3]

Keywords: Speech-Emotion Dialogue · Large Speech-Language Model · Reasoning.

1 Introduction

Empathetic dialogue aims to generate responses that both address what a speaker says and resonate with how the speaker feels[20,29]. Speech simultaneously conveys lexical information and rich paralinguistic cues, such as tone, volume and pitch, making it a crucial modality for building supportive voice agents. The rapid progress of large language models (LLMs)[19,10,5,1], along with their multi-modal extensions[16,2,27,22], provides new opportunities for emotion-aware spoken interaction.

Yet the vast majority of spoken dialogue systems remain focused on information delivery or task completion [21,7]. Though some of these works attempt to combine the training task with speech-emotion recognition (SER) to supply the LLM with emotion cues[24,28], they inherit the limitation of lacking explicit

[3] Our dataset is available at https://huggingface.co/datasets/ZoeTang/EmoDial-Reason.

mechanisms to capture and regulate the affective flow between user and model, which hinders emotional continuity in dialogue.

Chain-of-Thought (CoT) prompting[25] offers a potential remedy. By encouraging models to *"think aloud"*, CoT has improved factual reasoning in text dialogue[4] and certain speech tasks[13]. Motivated by this success, we argue that making the affective reasoning process visible could benefit empathy as well: an agent should detect the user's emotion, decide which emotion to convey back, and only then phrase a response. However, *there is little empirical evidence on how to solicit such reasoning from LLMs in speech dialogue.*

In this paper, we propose **EmoDial-Reason**, the first speech-emotion dialogue corpus in which each example is paired with two different explicit reasoning paths. Leveraging this resource, we explore two approaches to elicit reasoning from a single LLM. Free-form reasoning instructs the model to "think aloud and then respond empathetically," enabling it to generate any style or length of internal monologue prior to producing its response. In contrast, template-guided reasoning requires the model to follow a planned reasoning path: (1) inferring the user's current emotion, (2) determining the emotion the agent should express, and (3) generating the final response. By enforcing explicitness in each intermediate decision, the template renders the agent's affective logic both inspectable and directly trainable. Our contributions are summarized as:

- We construct the first publicly available speech-emotion dialogue dataset that pairs each sample with explicit reasoning chains.
- We conduct the first systematic comparison between free-form reasoning and template-guided reasoning for affective response in spoken dialogue, exploring effective reasoning in speech-emotion dialogue.
- We provide empirical evidence that incorporating explicit reasoning mechanisms significantly improves empathetic response generation. Notably, adapting reasoning paths to different scenarios yields better model performance.

2 Related Work

2.1 Emotion-aware Large Speech-Language Models

Many works have attempted to integrate emotional cues from speech into Speech Language Models (SLMs), ranging from cascaded approaches to end-to-end solutions. In order to differentiate model responses to speech inputs that share the same semantics but differ in tone or style, Lin et al.[17] employ prompt engineering to guide the generation of dialogue in specified styles and train the Spoken-LLM, a model that emphasizes the role of paralinguistic features in dialogue generation and improves the model's sensitivity to such cues. However, this cascaded architecture leads to significant cumulative errors and response latency, which hinders real-time user interaction.

In contrast, E-Chat leverages a speech encoder to extract emotional embeddings from audio and integrates them with an LLM, enabling the system to respond appropriately in different emotional contexts[28]. BLSP-EMO adopts a

fully end-to-end architecture[24]. To enable the model to produce empathetic responses, BLSP-EMO proposes a two-stage training approach using existing Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) datasets. It applies the BLSP[23] method to achieve both semantic alignment and emotional alignment. In addition, by explicitly modeling both input and response sentiments in a serialized generation process, the model ParalinGPT further emphasizes the importance of predicting the sentiment of the response, ensuring emotionally coherent responses[18].

While significant progress has been made, these works still depend primarily on the inherent text-generation capabilities of LLMs and lack explicit design for emotion-driven response mechanisms during generation. A more explicit modeling framework is needed to better understand not only the current emotions themselves, but also their causes, dynamic changes and interactions with semantics, enabling more effective empathetic responses.

2.2 Reasoning in Sentiment Analysis and Dialogue

In recent years, prompt-based CoT reasoning and fine-tuning-based internalized reasoning approaches have demonstrated strong capabilities across various domains such as math and question answering[6,15]. These methods have gradually extended from textual tasks to multimodal settings. As a crucial component of emotional speech dialogue, multimodal emotion recognition has seen notable improvements in the comprehension of complex emotional cues when enhanced by reasoning mechanisms[31]. The CoT-based framework DOCTOR enhances response quality in dialogue systems by aggregating key cues dispersed across multiple conversational turns[4]. Similarly, prompt engineering techniques have been employed in speech-based dialogue models to guide the model's output. By directing the generation to follow the sequence of listening, perceiving, and expressing, the approach provides preliminary evidence for the effectiveness of reasoning mechanisms in emotional speech dialogue[26]. However, relying solely on prompts is insufficient to fully activate the model's reasoning potential, and a more systematic integration of reasoning processes is required.

3 Dataset

To investigate the impact of reasoning on speech-emotion dialogue and the role of different reasoning paths, we construct EmoDial-Reason, a speech-to-text daily dialogue dataset that captures reasoning process with different reasoning paths.

Each sample in EmoDial-Reason consists of a dialogue history, a current input, two reasoning processes and corresponding target response. For dialogue history, each utterance is annotated with emotion labels. The current input—typically the last utterance in history—is annotated with paralinguistic information additionally. Two reasoning paths are provided in the dataset: one generated without template guidance and one with it. The construction process of our dataset is shown in Figure 1, which will be elaborated in the following.

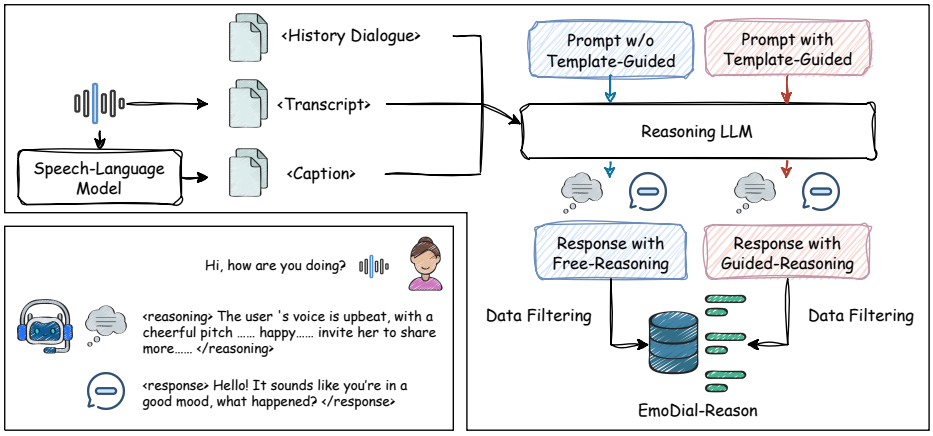


Fig. 1: Dataset collection process of EmoDial-Reason

3.1 Data Collection

Speech dialogue data are firstly collected from Dailytalk[12] and Styletalk[17], which contain multi-turn dialogue and emotion annotations for each turn. Qwen2-Audio[5] is employed to extract captions of the current inputs' speech by prompting it to describe paralinguistic information entailed in the audio, including tone, volume, pitch and so on.

After obtaining the paralinguistic captions, we feed the data into a reasoning model to generate reasoning processes aimed at producing responses that are semantically and emotionally appropriate. DeepSeek-R1[11] is employed to generate reasoning data. By feeding the captions, we extend DeepSeek's reasoning ability from text to a multimodal setting. By designing different prompts, we guide the reasoning model to construct two distinct reasoning paths. The free-form reasoning, where no explicit reasoning trajectory is provided in the prompt, tends to involve more fine-grained and diverse intermediate steps. In contrast, template-guided reasoning, which explicitly follows a fixed sequence—inferring the user's emotion, determining the target response emotion, and then generating the response—exhibits more structured logic and a more concise reasoning process. Table 1 shows the difference between the two reasoning paths.

3.2 Data Filtering and Assessment

To ensure the quality of the captions and reduce errors introduced during the conversion from audio to caption text, keyword-based data filtering is applied. Specifically, for each emotion label, a set of possible paralinguistic descriptors are predefined. We then compare the emotion label of the current input with its corresponding caption, and remove samples in which no corresponding keywords appeared in the caption.

Reasoning Path	Example
	User: <Speech Caption: The speaker seems happy and joyful as indicated by their lively tone, upbeat tempo, cheerful pitch and loud volume.>Hi, how are you doing?
Free-Reason	**Reasoning:** Okay, let's see. The user wants me to generate a response to "Hi! How are you doing?" with appropriate emotion. The friend's voice is happy, upbeat, with a cheerful pitch and volume. So the response should match that positivity. Maybe start with a warm greeting like "Hello!" or "Hey there!" to keep it casual but friendly. Then ask about their day, maybe using "How's your day been?" or "What's been up?" to show interest. Since the friend is happy, the response should sound cheerful too, maybe add a smile in the voice. Avoid any negative words or phrases. Keep it natural, like something you'd say to a friend. Let me check if "How's your day been?" fits. Yeah, that's good. Also, maybe end with a question to keep the conversation going. Like "Anything exciting planned for tonight?" to invite them to share more. Alright, that should work. **Response:** Hello! It sounds like you're in a good mood! How's your day been?
Guided-Reason	**Reasoning:** Okay, let's tackle this step by step. **First, I need to figure out the speaker's emotion in the current utterance.** Now, the current question is "How are you doing?" which is a common way to check on someone's well-being. The voice shows a happy and positive mood. **Next, determining the appropriate emotion for the response.** Since the speaker is asking about their well-being in a friendly manner, the response should match that positivity. Maybe show gratitude or happiness. Keeping it upbeat but sincere makes sense here. **Now, generating the response.** It needs to acknowledge the question positively. Something like, "I'm doing pretty well, thanks for asking! How about you?" This shows appreciation, keeps the conversation going, and invites the other person to share. Making sure it's concise and friendly fits the context. **Response:** I'm doing pretty well, thanks for asking! How about you?

Table 1: An example of different reasoning paths in EmoDial-Reason.

Manual quality evaluation is conducted on the filtered dataset to further ensure reasoning quality. Constrained by available resources, a random sample of 100 instances was selected for human assessment. The evaluation results show that the paralinguistic cues in the captions are consistently leveraged, and both the reasoning trajectories and the responses remain plausible and appropriate.

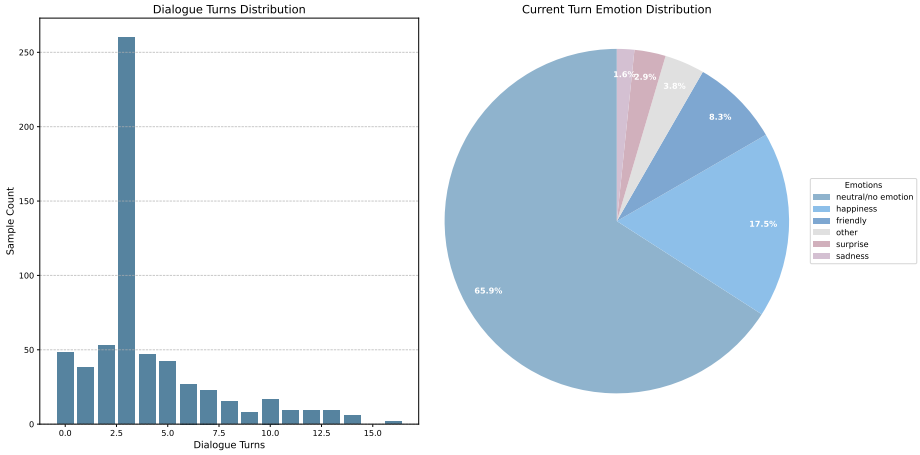


Fig. 2: Distribution of dialogue turns and emotion labels.

3.3 Data Statistics

A total of 2,159 dialogue samples are collected with annotated reasoning processes. After filtering, 613 high-quality samples remain, indicating the importance of filtering noise introduced by captioning. The distribution of dialogue turns and emotion labels can be seen in Figure 2. In the original datasets, the dialogues from StyleTalk are relatively short, therefore, our dataset exhibits an average of 3.99 turns per conversation, with dialogues comprising three turns representing the most frequently, Despite its brevity, such length already provides sufficient information for the model to infer the current context and emotional state. For emotion labels, the emotion distribution is dominated by neutral states, with positive emotions occurring less frequently and negative emotions being rare. Regarding emotion shifts between the input and the response, the most frequent transition is from no emotion to no emotion, followed by transitions such as happiness to happiness and no emotion to cheerful. On average, Free-Reason produces longer reasoning chains (274.1 words) compared to Guided-Reason (242.9 words).

Our dataset suffers from limitations in both scale and label balance. However, prior empirical observations[31] have shown that training with data at this scale is sufficient to trigger the model's reasoning capability. Our subsequent experiments further corroborate this finding. In addition, since the base model has already been pretrained on large-scale corpora, the limited imbalanced data does not significantly impair its classification ability, as it is unlikely to provide the model with substantial new knowledge. Instead, it primarily serves to activate the model's existing reasoning capability developed during pretraining. A larger and more balanced dataset will be considered in future work.

4 Reasoning-Enhanced Speech-Emotion Dialogue

The reasoning-enhanced speech-emotion dialogue task aims to generate emotionally appropriate and contextually coherent responses based on multimodal user input and dialogue history, while explicitly modeling the reasoning process underlying the response. Unlike traditional dialogue systems that produce responses directly, this task emphasizes the incorporation of interpretable reasoning steps, either structured or free-form, to enhance emotional understanding and response quality.

At each dialogue turn t, the model receives a dialogue history H in text and the current utterance A_t in audio. The dialogue history is represented as a sequence of textual utterances $H = \{h_1, h_2, \ldots, h_{t-1}\}$, where each h_i corresponds to a previous speaker turn and adjacent utterances are assumed to come from different speakers, reflecting the natural turn-taking behavior in dialogue.

After the raw audio input A_t is processed by an audio encoder to obtain the corresponding audio embedding, the model takes the dialogue history along with the current utterance embedding as input to the backbone LLM. The LLM then generates a two-part structured output consisting of a reasoning trace R_{reason} and a final response $R_{response}$.

Given the paired input of (H, A_t), the training objective is to first predict the intermediate reasoning process and then generate the final response based on the reasoning. The overall objective is to maximize the following probability:

$$\begin{aligned} P_\theta(T \mid H, A_t) = & P_\theta(R_{reason} \mid H, \text{Encoder}(A_t)) \\ & P_\theta(R_{response} \mid R_{reason}, H, \text{Encoder}(A_t)) \end{aligned}$$

where T denotes the target answer with both reasoning and response. θ denotes the trainable parameters of LLM. The parameters of the audio encoder are frozen to preserve the LLM's speech understanding capability. More specifically, the loss of predicting the reasoning trace can be formalized as:

$$\mathcal{L}_{\text{reason}} = -\sum_{t=1}^{T_r} \log P_\theta(r_t \mid r_{<t}, H, \text{Encoder}(A_t))$$

where r_t is the t-th token in the reasoning trace, and T_r is the total number of tokens in the reasoning sequence. Similarly, the loss of predicting the final response can be written as:

$$\mathcal{L}_{\text{response}} = -\sum_{t=1}^{T_y} \log P_\theta(y_t \mid y_{<t}, R_{reason}, H, \text{Encoder}(A_t))$$

where y_t is the t-th token in the final response, and T_y is the number of tokens in the response. The model conditions on the predicted or gold reasoning trace R_{reason} during inference and training, respectively.

The total loss is the weighted sum of the two objectives:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{reason}} + \lambda_2 \cdot \mathcal{L}_{\text{response}}$$

We treat both components as equally important by setting $\lambda_1 = \lambda_2 = 1$, and performing a one-stage training.

5 Experiments

5.1 Experimental Setups

Training setup To validate the effectiveness of reasoning in speech-emotion dialogue, EmoDial-Reason is used to finetune Qwen2-Audio[5], due to its capability of handling both audio QA and dialogue generation tasks. We randomly sample 10% from EmoDial-Reason for validation during training. For evaluation, we prepare two test sets: one that shares the same distribution as the training data, and another that is out-of-distribution. we sample 100 instances from DailyTalk and StyleTalk for the former one and 151 from IEMOCAP for the other.

During training, we adopt LoRA with a rank of 8. The learning rate is set to 3e-5. The batch size is set to 4 due to resource constraints. The checkpoint with the lowest validation loss is selected as the final model for evaluating. Speech encoder is frozen during training.

Baselines Qwen2-Audio[5], BLSP-Emo[24] and LLaMA-Omni[10] are chosen as our baselines.

Qwen2-Audio. An audio-language model excelling in voice chat and audio analysis with instruction following ability, on which we finetune our model.

LLaMA-Omni. A multimodal model capable of speech interaction by directly generating text and speech from speech instructions. Only the textual outputs are used for comparison in our experiments.

BLSP-Emo. Built on Whisper and Qwen, the model focuses on empathetic speech understanding and generation by aligning semantic and emotional cues with LLM, demonstrating remarkable performance on empathetic response.

Evaluation Metrics To evaluate the model's ability, three objective metrics are adopted: DIST-1[14], EMOScore and BERTScore[30]. Specifically, DIST-1 measures response diversity by calculating the ratio of distinct uni-grams to the total number of generated tokens across all responses. EMOScore measures emotion similarity between the model's response and the target response, using a BERT[9] model trained on the GoEmotions[8] dataset. Similarly, BERTScore is adopted to evaluate the semantic similarity between the generated and reference responses. We choose the response from raw datasets rather than generated by reasoning model as target response here.

Following previous work[24], we introduce the LLM evaluation due to the open-ended property of dialogue generation tasks. GPT-4 is used to assess the generated responses across three dimensions: content relevance, logical coherence, and emotional appropriateness. Specifically, content relevance evaluates how well a response aligns with the main topic or user intent; logical coherence evaluate whether the information in the dialogue is logically clear and consistent; emotional appropriateness measures whether the emotional expression in the response fits the context and user's mood. It is prompted to score each

Model	DIST-1	EMOScore	BERTScore	Format	LLM Evaluation		
					Content	Logic	Emotion
In-Distribution Dataset							
Qwen2-Audio	0.34	60.0	64.6	-	2.91	2.77	2.40
LLaMA-Omni	0.32	58.5	65.3	-	3.06	2.65	2.52
BLSP-Emo	0.28	50.7	66.4	-	3.88	3.34	3.19
Raw-Res	0.37	69.9	70.9	-	4.84	4.66	4.42
Free-Reason	0.46	62.9	**67.1**	0.98	4.72	4.63	4.48
Guided-Reason	**0.47**	**64.1**	66.7	0.97	**4.83**	**4.67**	**4.54**
Out-of-Distribution Dataset							
Qwen2-Audio	0.25	50.9	58.8	-	3.30	3.17	2.65
LLaMA-Omni	0.23	50.7	58.6	-	3.32	3.18	2.77
BLSP-Emo	0.16	40.8	58.9	-	3.46	3.07	2.84
Raw-Res	0.05	60.3	**62.7**	-	4.13	3.78	3.46
Free-Reason	**0.34**	**61.2**	61.8	0.99	**4.78**	**4.60**	**4.39**
Guided-Reason	0.33	58.0	60.2	0.97	4.66	4.45	4.32

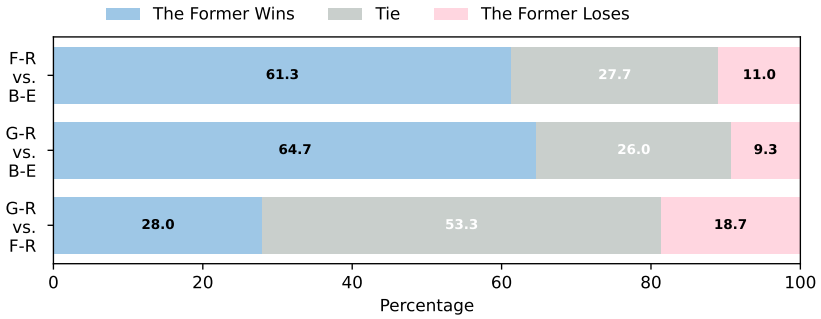
Table 2: Performance of reasoning-enhanced dialogue generation. *Format* reports the percentage of outputs that follow the required prompt format. *Raw-Res* refers to the model trained only on the original dataset. *Guided-Reason* is fine-tuned with template-guided reasoning paths, and *Free-Reason* is fine-tuned with free-form reasoning paths. *ID* and *OOD* denote the in-distribution and out-of-distribution test splits, respectively

response from 1 to 5. A human evaluation is further conducted for a more comprehensive assessment, involving comparisons of outputs from different models and manual scoring the responses from 1-5 across 3 dimensions, consistent with the LLM-as-a-judge evaluation.

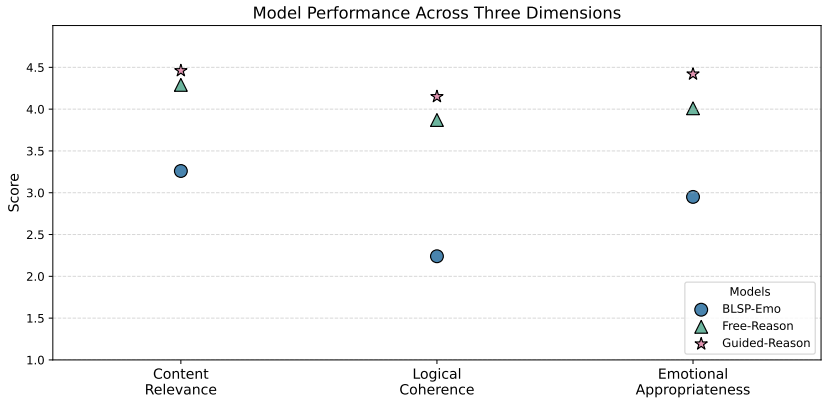
5.2 Experimental Results

Table 2 summarizes automatic scores and LLM-as-a-judge ratings for three baselines (Qwen2-Audio, LLaMA-Omni and BLSP-Emo) and our three fine-tuned variants, Raw-Res (no reasoning), Guided-Reason (template-guided reasoning) and Free-Reason (free-form reasoning). Evaluation is performed on two test splits: an in-distribution (ID) set rendered with studio-quality TTS and an out-of-distribution (OOD) set recorded in real acoustic conditions which is sampled from IEMOCAP[3]. Compared to the ID set, the OOD test set contains more filler words, repetitions, and casual expressions with loose logic, making it more challenging for the model.

Reasoning improves overall quality. Compared to the baselines, the models with reasoning show significant performance improvements across automatic metrics and LLM-based evaluations. For ID test set, the reasoning approach with template achieves better performance on most metrics. Although



(a) Evaluation of response quality in pairwise model comparisons. *G-R*, *F-R*, and *B-E* denote Guided-Reason, Free-Reason, and BLSP-Emo, respectively.



(b) Human evaluation scores on three dimensions.

Fig. 3: Human evaluation results.

the performance of the free-form reasoning model slightly declines, it still significantly outperforms the baseline models. The model Raw-Res in Table 2 is trained using original responses from raw dataset Dailytalk and Styletalk(instead of those generated by reasoning model) as ground truth. This model shows strong performance on metrics based on similarity calculation using BERT-based models, due to the similar distribution of its training data and test data. Our models achieve comparable or even better results than those of Raw-Res in LLM evaluations, clearly demonstrating the effectiveness of our dataset and model.

Generalization Analysis Although the reference model Raw-Res performs well on ID dataset, its performance drops significantly on OOD set. By contrast, our reasoning-enhanced models demonstrate strong robustness across different datasets. This indicates that incorporating reasoning processes into emotional dialogue generation not only improves the model's response quality but also enhances its adaptability to unseen scenarios, highlighting its strong potential for real-world applications. By comparing the ID and OOD results, it can also be uncovered that *Guided-Reason achieves superior performance on synthetic*

speech scenario, whereas Free-Reason outperforms it on real-world situations. We attribute this to the fine-grained reasoning involves in free-form reasoning, which better adapts to the diversity of real acoustic conditions.

Human Evaluation Human evaluation is also conducted for better assessment. For each dialogue context, three graduate student evaluators were asked to compare the responses generated by different models using a win/tie/lose scheme. The comparison was carried out among the best-performing baseline model BLSP-Emo and our proposed models, Free-Reason and Guided-Reason. The evaluation results are presented in Figure 3a. It can be noticed that the responses generated by the reasoning-enhanced models are consistently superior to those of the baseline model. Additionally, Guided-Reason exhibits a slight edge over Free-Reason in response quality.

Following the same protocol of LLM evaluation, we also manually rated the responses of each model on the ID test split across three dimensions, as the result can be seen at Figure 3b, which keeps the same as the comparison evaluation. Guided-Reason achieved the highest scores, followed closely by Free-Reason. In contrast, BLSP-Emo underperformed across all dimensions compared to the models fine-tuned on the EmoDial-Reason. These results are consistent with the outcomes of the LLM evaluation.

We further performed a manual quality inspection of the reasoning process. Given resource constraints, we randomly selected 70 outputs from the Free-Reason and Guided-Reason models. Among these, 85.7% outputs effectively utilize the paralinguistic information presented in the speech, 80.0% exhibit coherent reasoning and produce appropriate responses accordingly. In addition, across all samples, the final responses are closely aligned with their reasoning processes.

The results above provide additional evidence for the effectiveness of incorporating reasoning into empathetic dialogue and demonstrating the value of our proposed dataset, as well as the reliability of the LLM-as-a-judge.

5.3 Template Guide Incorporation

To combine the advantages of both reasoning strategies, a mix training is performed using both template-guided and free-form reasoning paths. Specifically, we combine samples corresponding to different reasoning paths and use them jointly to train the model. This approach encourages the model to adaptively select reasoning strategy itself depending on the scenario. The experiment results are shown in Table 3, where the mix-path reasoning model is referred to as Mix-Reason. As illustrated in the table, the model trained with mixed reasoning achieves comparable or even superior performance to the previous best models across both simple and complex scenarios.

By analyzing the reasoning paths selected by the model in different scenarios, as shown in Figure 4, a clear pattern can also be observed that in simpler scenarios, the model chooses to follow template. However, as the scenario becomes more complex, the model tends to incorporate a more fine-grained and detailed reasoning, with a higher ratio to choose free-form reasoning.

Model	LLM Evaluation		
	Content	Logic	Emotion
In-Distribution Dataset			
Mix-Reason	4.85	4.75	4.53
- w/o G-R	4.72 -0.13	4.63 -0.12	4.48 -0.05
- w/o F-R (Prev-best)	4.83 -0.02	4.67 -0.08	4.54 +0.01
- w/o Reason	4.69 -0.16	4.48 -0.27	4.43 -0.10
Out-of-Distribution Dataset			
Mix-Reason	4.80	4.65	4.32
- w/o G-R (Prev-best)	4.78 -0.02	4.60 -0.05	4.39 +0.07
- w/o F-R	4.66 -0.14	4.45 -0.20	4.32 -0.00
- w/o Reason	4.58 -0.22	4.35 -0.30	4.17 -0.15

Table 3: Ablation study evaluating the effectiveness of reasoning process. *G-R* and *F-R* denote Guided-Reason and Free-Reason, respectively. *Prev-best* refers to the best-performing single-reason-path model on corresponding dataset.

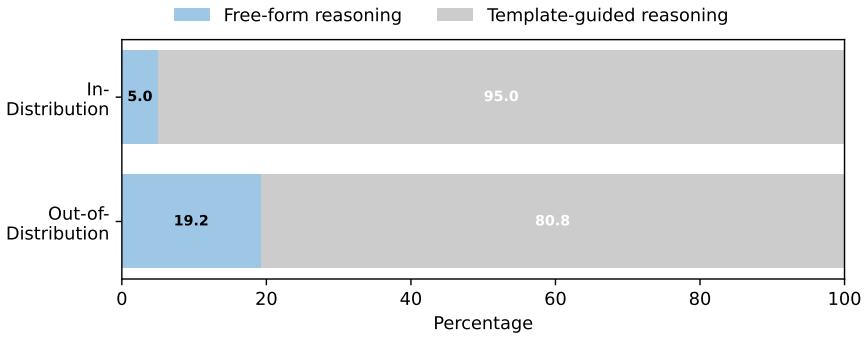


Fig. 4: Distribution Shift in Reasoning Strategies.

5.4 Ablation Study

Ablation experiments are conducted to investigate the impact of reasoning processes on response quality, with the results shown in Table 3. As previously mentioned, compared to models trained using a single reasoning path, model trained with a mixture of two reasoning paths demonstrates superior adaptability across various dialogue scenarios.

Furthermore, when using only the responses generated by the reasoning model DeepSeek-R1 without corresponding reasoning processes during training, a significant performance drop can be observed across all three evaluation dimensions. This indicates that the reasoning process substantially enhances the model's ability to generate emotionally and semantically appropriate responses, demonstrating the effectiveness of our proposed dataset and approach.

6 Conclusion

In this work, we introduce a speech-emotion dialogue dataset with two types of reasoning chains. Our extensive experiments affirm the value of reasoning in speech-emotion dialogue. Moreover, exposing the model to mixed reasoning paths helps it adapt to diverse scenarios, paving the way for more trustworthy and effective empathetic voice agents.

Acknowledgements. This research is supported by the General Office of National Language Commission Research Planning Committee (No. ZDA145-18).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
4. Chae, H., Song, Y., Ong, K., Kwon, T., Kim, M., Yu, Y., Lee, D., Kang, D., Yeo, J.: Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5606–5632 (2023)
5. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al.: Qwen2-audio technical report. arXiv e-prints pp. arXiv–2407 (2024)
6. Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., Liu, T.: Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1173–1203 (2024)
7. D'efossez, A., Mazar'e, L., Orsini, M., Royer, A., P'erez, P., J'egou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue. ArXiv **abs/2410.00037** (2024), `https://api.semanticscholar.org/CorpusID:273022979`
8. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A dataset of fine-grained emotions. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4040–4054. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.372, `https://aclanthology.org/2020.acl-main.372/`
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. As-

sociation for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
10. Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y.: Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666 (2024)
11. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Lee, K., Park, K., Kim, D.: Dailytalk: Spoken dialogue dataset for conversational text-to-speech. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1–5 (2022), https://api.semanticscholar.org/CorpusID:250265010
13. Li, G., Liu, J., Dinkel, H., Niu, Y., Zhang, J., Luan, J.: Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. ArXiv **abs/2503.11197** (2025), https://api.semanticscholar.org/CorpusID:277043724
14. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: 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. pp. 110–119 (2016)
15. Li, Z.Z., Zhang, D., Zhang, M.L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.J., Chen, X., et al.: From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419 (2025)
16. Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., Liu, K.: A survey of multimodel large language models. In: Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. pp. 405–409 (2024)
17. Lin, G.T., Chiang, C.H., Lee, H.Y.: Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6626–6642 (2024)
18. Lin, G.T., Shivakumar, P.G., Gandhe, A., Yang, C.H.H., Gu, Y., Ghosh, S., Stolcke, A., Lee, H.y., Bulyko, I.: Paralinguistics-enhanced large language modeling of spoken dialogue. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10316–10320. IEEE (2024)
19. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M.A., Socher, R., Amatriain, X., Gao, J.: Large language models: A survey. ArXiv **abs/2402.06196** (2024), https://api.semanticscholar.org/CorpusID:267617032
20. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: I know the feeling: Learning to converse with empathy. ArXiv **abs/1811.00207** (2018), https://api.semanticscholar.org/CorpusID:53153815
21. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Zhang, C.: Salmonn: Towards generic hearing abilities for large language models. ArXiv **abs/2310.13289** (2023), https://api.semanticscholar.org/CorpusID:264406171
22. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
23. Wang, C., Liao, M., Huang, Z., Lu, J., Wu, J., Liu, Y., Zong, C., Zhang, J.: Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. arXiv preprint arXiv:2309.00916 (2023)

24. Wang, C., Liao, M., Huang, Z., Wu, J., Zong, C., Zhang, J.: Blsp-emo: Towards empathetic large speech-language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 19186–19199 (2024)
25. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
26. Xie, J., Lei, S., Yu, Y., Xiang, Y., Wang, H., Wu, X., Wu, Z.: Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. arXiv preprint arXiv:2501.10937 (2025)
27. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
28. Xue, H., Liang, Y., Mu, B., Zhang, S., Chen, M., Chen, Q., Xie, L.: E-chat: Emotion-sensitive spoken dialogue system with large language models. In: 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 586–590 (2024). https://doi.org/10.1109/ISCSLP63861.2024.10800447
29. Ye, J., Xiang, L., Zhang, Y., Zong, C.: Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 4646–4669 (2025)
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
31. Zhao, J., Wei, X., Bo, L.: R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv preprint arXiv:2503.05379 (2025)