Flexible Thinking for Multimodal Emotional Support Conversation via Reinforcement Learning

Fanfan Wang[1], **Xiangqing Shen**[2,1], **Jianfei Yu**[1,2,3*], **Rui Xia**[2,3*],
[1]School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
[2]School of Intelligence Science and Technology, Nanjing University, China
[3]University of Chinese Academy of Sciences, Nanjing, China
{ffwang, jfyu}@njust.edu.cn
{siangching.shen, rxia.cn}@gmail.com

Abstract

Emotional Support Conversation (ESC) systems aim to alleviate user distress. However, current Chain-of-Thought based ESC methods often employ rigid, text-only reasoning, limiting adaptability in dynamic, multimodal interactions and introducing reasoning noise that degrades support quality. To address this, we introduce "Flexible Thinking" for multimodal ESC, enabling models to adaptively select contextually relevant thinking aspects: Visual Scene, Emotion, Situation, and Response Strategy. We first construct training data by manually curating flexible thinking demonstrations on the MESC dataset, then using a Multimodal Large Language Model to synthesize these processes for the full training set. Then, we propose FIRES, a framework integrating Supervised Fine-Tuning (SFT) for initial learning with Reinforcement Learning for refinement. This two-stage approach helps FIRES transcend SFT's generalization limits and, crucially, directly links thinking processes to response quality via tailored rewards, moving beyond imitating potentially imperfect synthetic data. Experiments on MESC and EMOTyDA datasets demonstrate FIRES's effectiveness and generalizability in fostering higher-quality emotional support responses through adaptive reasoning.

1 Introduction

Emotional support and mental well-being are increasingly crucial in today's fast-paced society, where individuals commonly face pressures from both work and life. With the rapid advancements in artificial intelligence, providing emotional support through automated dialogue systems, known as Emotional Support Conversation (ESC) task, has emerged as a promising research topic of significant interest. The goal of ESC is to alleviate users' emotional distress and help them cope with encountered challenges (Burleson, 2003; Liu et al., 2021).

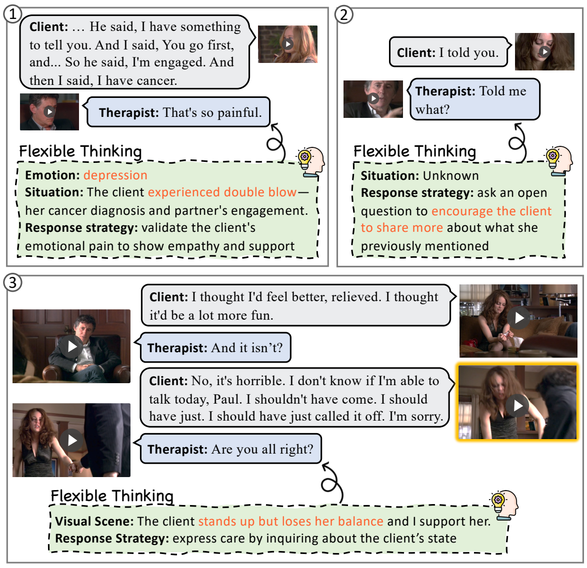

Figure 1: Some examples for the multimodal ESC task from the MESC dataset. The green box contains the flexible thinking process before the final response.

It holds broad application prospects across various scenarios, such as daily interactions, social companionship, mental health support, and customer service, by improving user experiences through timely emotional support.

Early research on ESC mostly focused on controlling response strategies (Tu et al., 2022; Cheng et al., 2022; Li et al., 2024b; Liu et al., 2024) or incorporating external knowledge (Deng et al., 2023; Cheng et al., 2023; Bao et al., 2024; Hao and Kong, 2025) to strengthen context understanding and improve response quality. In recent years, the ascent of Chain-of-Thought (CoT) reasoning has inspired studies that simulate human-like step-by-step thinking before generating a response (Zhang et al., 2024b; Cao et al., 2024), aiming to enhance the interpretability and reliability.

However, existing CoT-based methods for ESC mostly focus on text-only conversations and typically adopt fixed thinking steps (Zhang et al.,

* Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1341–1356
November 4-9, 2025

2024b; Cao et al., 2024; Chen et al., 2025; Wu et al., 2024). This design results in a rigid "one-size-fits-all" reasoning paradigm that struggles to adapt to the dynamic contexts of real-world conversations. In contrast, human supportive interactions are inherently multimodal (e.g., involving both textual and visual information), and human supporters do not mechanically deliberate on a fixed set of aspects; instead, they flexibly adjust their cognitive focus based on the specific context and the interlocutor's state. This flexibility is crucial for providing effective support. As shown in Figure 1, an empathetic response may require a deep analysis of the interlocutor's emotion cause and specific situation (example 1). When the available information in the limited conversational history is insufficient, quickly initiating further inquiry to gather more details is more effective than complex deep thinking (example 2). Sometimes an appropriate response may critically hinge on the perception of events within the visual scene (example 3). A rigid thinking mode with fixed steps overlooks the actual utility of different reasoning aspects in specific contexts. When predefined thinking aspects mismatch current contextual requirements, such a mode not only fails to provide constructive guidance but may also introduce "reasoning noise", thereby undermining the quality of support.

In this work, we introduce Flexible Thinking for the multimodal ESC task, which aims to adaptively select and integrate those thinking aspects beneficial for generating high-quality emotional support responses based on the specific conversational context. Specifically, we predefine four potential thinking aspects: perception of the current visual scene (Visual Scene), recognition of the interlocutor's emotions (Emotion), analysis of the specific causes of emotions or the problems faced by the interlocutor (Situation), and planning of the response strategy (Response Strategy). To construct training data, we first manually annotated demonstrations of flexible thinking on the public MESC dataset (Chu et al., 2024). Subsequently, we leverage the powerful multimodal large language model (MLLM) to synthesize the flexible thinking processes for the entire training set based on these demonstrations.

We propose `FIRES` (**F**lexible th**I**nking via **R**einforcement l**E**arning for Multimodal Emotional **S**upport Conversation), a framework that synergizes Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). Initially, SFT on synthesized data teaches the model foundational thinking patterns. However, SFT alone has limitations, as models may memorize flaws from the synthetic data (like inaccurate inferences or redundant steps) and struggle to generalize by imitating these potentially imperfect examples. To overcome this, employs the latest RL algorithm, Group Relative Policy Optimization (GRPO), which is derived from recent work on reasoning models (Shao et al., 2024; Guo et al., 2025). This RL stage allows the model to explore diverse reasoning paths and learn from carefully designed reward function that directly links the thinking process to response quality, promoting effective reasoning beyond mere imitation of synthetic data.

We conduct extensive experiments using both open-source and closed-source MLLMs on the MESC dataset. The results demonstrate the advantages of flexible thinking and the superiority of the proposed `FIRES`. Further experiments on another multimodal dialogue dataset EMOTyDA (Saha et al., 2020) also validate the generalizability of our framework.

Our contributions are summarized as follows: (1) We introduce the concept of flexible thinking to the multimodal ESC task, encouraging the model to, based on the specific conversational context, adaptively integrate aspects that facilitate effective response generation. (2) We propose the `FIRES` framework, which involves an initial cold start via SFT and a subsequen RL optimization through GRPO coupled with a customized reward function for ESC, to stimulate the model towards effective flexible thinking and the generation of high-quality emotional support responses. (3) Experimental results demonstrate the effectiveness and generalizability of our framework.

2 Related Work

2.1 Emotional Support Conversation

Emotional Support Conversation (ESC) is a dialogue generation task aiming to alleviate emotional distress for the user. Liu et al. (2021) first introduced this task and manually constructed the ESConv dataset. Since then, numerous studies focusing on this task have emerged. Early research on ESC focused on improving model performance by refining dialogue modeling (Peng et al., 2022; Zhao et al., 2023; Li et al., 2024a), modeling support strategy (Tu et al., 2022; Cheng et al., 2022; Li et al., 2024b; Liu et al., 2024; Hu et al., 2024; Peng et al., 2023) or integrating external knowledge, which commonly included leveraging commonsense reasoning (Tu et al., 2022; Deng et al., 2023; Bao et al., 2024; Xu et al., 2024; Liu et al., 2024) to enhance model understanding of the user's state, as well as incorporating persona information (Cheng et al., 2023; Han et al., 2024; Hao and Kong, 2025) to generate more appropriate supportive responses.

Recently, the advent of Large Language Models (LLMs) has notably influenced ESC research (Kang et al., 2024; Lissak et al., 2024; Cheng et al., 2024; Wang et al., 2024).Many studies leveraged LLMs for dialogue augmentation (Zheng et al., 2023a; Ye et al., 2025; Chen et al., 2025; Zheng et al., 2024). Some researchers combined LLMs with Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Lyu et al., 2023), and designed fixed thinking steps to mimic the human reasoning process and then guide response generation (Zhang et al., 2024b; Cao et al., 2024; Wu et al., 2024; Chen et al., 2024). The aforementioned studies all focus on text-only conversations. In contrast, the multimodal ESC task remains underexplored (Wu et al., 2025; Fei et al., 2024; Zhang et al., 2024c), and there is a scarcity of suitable datasets, particularly for real-world scenarios involving video and audio (Chu et al., 2024).

2.2 Reinforcement Learning

Reinforcement Learning (RL) aligns language models with human preferences by optimizing them against a reward model, frequently using the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) due to its empirical stability and performance (Liu et al., 2025a; Zhang et al., 2024a; Zheng et al., 2023b). The recent proposed Group Relative Policy Optimization (GRPO) Shao et al. (2024) streamlines PPO by discarding the value network, instead estimating the relative advantage within a group of sampled completions to achieve similar performance with lower computational cost. Following this, a number of studies have emerged to either improve the GRPO algorithm (Sane, 2025; Liu et al., 2025b; Chu et al., 2025b), or apply it to various task (Huang et al., 2025; Li et al., 2025; Jin et al., 2025; Pan et al., 2025). To the best of our knowledge, limited prior work has utilized traditional RL algorithms to optimize emotional support systems (Zhou et al., 2023; Li et al., 2024c). Our work, however, targets the multimodal ESC task and integrates MLLMs with GRPO to promote effective reasoning and response generation.

3 Methodology

3.1 Task Formalization

In this paper, we focus on multimodal emotional support conversations involving both text and video, which are close to real-world scenarios. Given a multimodal conversation history $C = \{U_1, U_2, \ldots, U_{t-1}\}$ between a seeker and a supporter, where each utterance contains textual and visual data, the goal of the Multimodal Emotional Support Conversation (MESC) task is to generate a contextually appropriate response R_t as the supporter to alleviate the seeker's emotional distress.

3.2 Aspects of Flexible Thinking

Although humans may appear to respond quickly in face-to-face communication, they typically undergo a rapid and internal thinking process: visual observations are integrated with the interlocutor's utterances to understand their emotions and situation, followed by deliberation on how to reply. Moreover, the emphasis of human thinking shifts according to the specific context.

To emulate this nuanced human capability for pre-response flexible thinking, our framework requires the model to provide both its thinking process and the response. We define four potential thinking aspects, from which the model adaptively considers the relevant ones to construct its thinking, thereby facilitating the generation of more helpful and appropriate responses. The specific content of each aspects is as follows:

- **Visual Scene**: Decribe the key elements observed in the video, including human behaviors, facial expressions, posture or gestures that could provide context to the interlocutor's emotional state or the ongoing interaction.
- **Emotion**: Recognize the interlocutor's emotions based on the textual utterances and visual cues.
- **Situation**: Analyze the causes that trigger the interlocutor's emotions or summarize the issues they are facing, to inform a more targeted supportive response.
- **Response Strategy**: Outline the intended response strategy and the communicative goal to guide the response generation.

The depth and breadth of an ideal flexible thinking process will dynamically change with the conversational context: sometimes, the model may only need to focus on one or two certain aspects to generate an appropriate response; sometimes it re-

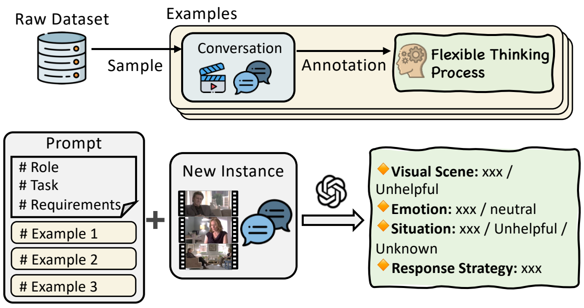

Figure 2: The synthesis of flexible thinking data.

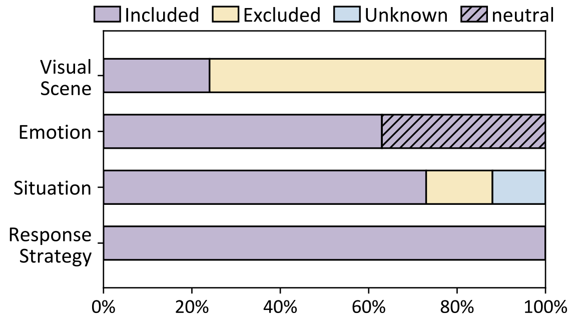

Figure 3: Proportional inclusion of thinking aspects in synthesized training data. The purple segment represents instances where the aspect is deemed relevant and included in the thinking process. The yellow segment indicates instances where the aspect is deemed not helpful for the response and omitted. The striped shading indicates the interlocutor's emotion is neutral. The blue segment marks instances where the limited dialogue history is insufficient to know the interlocutor's situation.

quires a thorough and comprehensive consideration of all the aspects mentioned above.

3.3 Synthesis of Flexible Thinking Data

To obtain supervision data for SFT, we leverage MLLMs to synthesize flexible thinking processes for conversation responses. The synthesis process is shown in Figure 2.

Specifically, based on data observation, we first manually select three classic instances of emotional support conversations and annotate their corresponding flexible thinking processes. They emphasize different thinking aspects (similar to the examples illustrated in Figure 1), and are used as few-shot examples to discourage the model from producing fixed-pattern thinking. We then craft a detailed prompt (Appendix A) introducing the task and its requirements, which is subsequently provided to GPT-4o[1] along with these examples and new conversation instances. For each instance, in addition to the conversation history text, key frames extracted using FFmpeg from the corresponding video are also provided as input context, and the number of keyframes is capped at 16 to limit the input length.

For each response generation instance in the training set of the dataset MESC (Chu et al., 2024), we synthesize its corresponding flexible thinking process. Figure 3 uses stacked bars to show the percentage of training instances incorporating each of the four defined aspects into their synthesized thinking processes. Notably, considering the crucial guiding role of "Response Strategy" in formulating supportive responses, this aspect is consistently included in the thinking process for all instances. Interlocutors do not always exhibit strong or clearly discernible emotions, leading to a small proportion of instances where "Emotion" is identified as neutral. For both the "Visual Scene" and "Situation" aspects, there are cases where analysis of them is omitted from the thinking process as they may be deemed "not helpful" for the subsequent response. Furthermore, in some instances, "Situation" might be marked as "Unknown" due to insufficient contextual information within the conversation history, suggesting a potential need for further inquiry to explore and ascertain the necessary details.

3.4 Response Supervision RL with GRPO

Due to potential imperfections that can sometimes arise in synthetic data, conducting SFT alone may lead to the model being influenced by minor errors or noise present in the data, thereby subtly impacting its performance. In our framework, after an initial cold start through SFT, a response supervision RL with GRPO is employed to further optimize the model. We design a reward function specifically tailored to the ESC task, which supervises the model's final responses after its thinking process, encouraging the model to engage in effective flexible thinking that results in more appropriate and supportive responses.

Notably, the dynamic integration of thinking aspects is learned implicitly through our two-stage framework. Initially, the SFT cold start teaches the model the form of flexible thinking; it learns from diverse examples among our synthesized data that generating a subset of thinking aspects is a valid pattern. Subsequently, the RL process optimizes the ability of flexible thinking by connecting thinking patterns to the quality of the final response.

[1] https://platform.openai.com/docs/guides/images-vision (gpt-4o-2024-11-20)

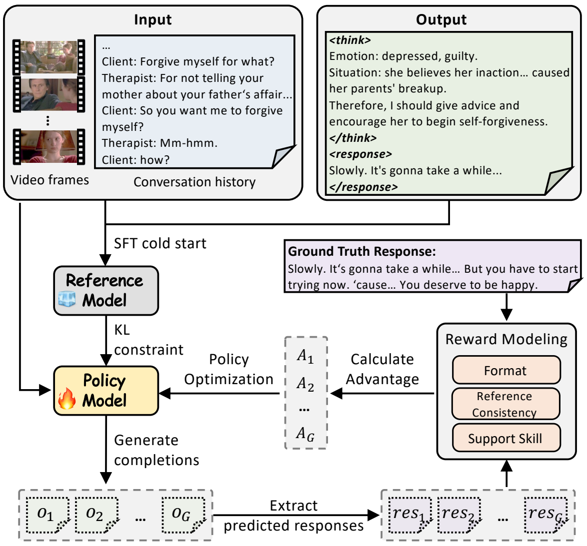

Figure 4: Overview of our framework FIRES.

Since the reward is applied to the final response, patterns that include noisy or irrelevant aspects lead to poor responses and are penalized with lower rewards. Conversely, effective thinking patterns are positively reinforced, allowing the model to implicitly optimize its policy to select the most rewarding combination of aspects for any given context.

3.4.1 Cold Start via SFT

Based on the powerful vision-language LLM, Qwen2.5-VL (Bai et al., 2025), we initially conduct SFT with synthesized thinking data, to equip the model with a foundational capability for flexible thinking. The input system prompt is set to: "*You are an expert in Emotional Support Conversation. Based on the conversation history, please first describe the thinking process and then provide an appropriate response.*" When provided with the dialogue history text and corresponding video frames as input, the model is expected to output both its internal thinking process and the final response.

3.4.2 Group Relative Policy Optimization

As shown in Figure 4, the model obtained from the SFT cold start stage is utilized to initialize the policy model $\pi_{\theta_{old}}$ and concurrently serves as the reference model $\pi_{\theta_{ref}}$. For a given instance q comprising dialogue history text and video frames, the model first generates a group of G distinct candidate outputs (termed "completions") $o = \{o_1, \ldots, o_i \ldots, o_G\}$ via sampling. Each of these completions encompasses both the thinking process and the response. We then extract the predicted responses from these completions and calculate a reward score r_i for each response res_i using the designed reward function. Subsequently, the advantage A_i for each completion is calculated based on its reward relative to the average reward of G completions:

$$A_i = \frac{r_i - \text{mean}\left(\{r_i\}_{i=1}^{G}\right)}{\text{std}\left(\{r_i\}_{i=1}^{G}\right)}, \tag{1}$$

which reflects the relative response quality.

GRPO encourages the model to generate outputs with higher advantages within each group. The optimization objective is to maximize the cumulative reward, with the objective function formulated as follows:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{old}}(q)} \left[\frac{1}{G} \sum_{i=1}^{G} p_i A_i - \beta \mathrm{D}_{\mathrm{KL}} \left(\pi_\theta \| \pi_{ref} \right) \right], \tag{2}$$

$$p_i = \frac{\pi_\theta(o_i)}{\pi_{\theta_{old}}(o_i)}, \tag{3}$$

where β is a regularization coefficient and $\mathrm{D}_{\mathrm{KL}}(\pi_\theta \| \pi_{ref})$ represents the KL divergence between the current policy model π_θ and the original reference model $\pi_{\theta_{ref}}$, which is incorporated to prevent excessive deviation from the reference model. p_i is the probability ratio of output o_i being generated by the current policy model π_θ to that of the old policy model $\pi_{\theta_{old}}$.

3.4.3 Reward Modeling for ESC

The reward function is pivotal to the GRPO algorithm as it provides guidance for the model's optimization. Our reward function, r_i, is a composite metric specifically designed for the ESC task, comprises the following three components:

Format Reward. To ensure the model learns the "think-then-respond" output paradigm, we assign a basic reward of $r_i^F = 1$ if the output o_i strictly adheres to the form of "<think> ... </think><response> ... </response>".

Reference Consistency Reward. Conversational responses, unlike mathematical problems, are relatively subjective and lack definitive standard answers. To encourage the model to emulate the human therapist's conversational patterns, we regard the ground-truth responses from the dataset as references, and reward the generated response res_i for capturing salient information from the reference response res_gt. For this, the ROUGE-L (Lin, 2004) score, a common metric for text generation, is employed as a reward component, i.e., $r_i^R = F_{\text{LCS}}(res^{gt}, res_i)$, which is calculated as an

F_1-score based on the Longest Common Subsequence (LCS) to measure the structural and lexical overlap, thereby promoting consist and topically relevant generations.

Support Skill Reward. ESC system should reduce the user's emotional distress through the utilization of proper support skills (Liu et al., 2021), e.g., appropriately exploring the interlocutor's situation, understanding and acknowledging their emotions, or expressing empathy and concern. To align the model with effective emotional support skills, we leverage the EPITOME framework (Sharma et al., 2020) which defines three communication mechanisms: Explorations (EX), Interpretations (IP), and Emotional Reactions (ER). Specifically, we use pre-trained RoBERTa (Lee et al., 2022) to predict the intensity of each mechanism (on a scale of 0, 1, or 2) for both the generated response res_i and the ground-truth response res^{gt}, i.e., $\mathrm{EP}_m(y) = \mathrm{RoBERTa}_m(C, y)$, where $m \in \{\mathrm{EX}, \mathrm{IP}, \mathrm{ER}\}$, C and y denote the conversation context and response, respectively. The mean squared error between the predicted intensity scores is then calculated as the DIFF-EPITOME score:

$$\mathrm{DE}_i = \frac{1}{3} \sum_{m=1}^{3} \left(\mathrm{EP}_m\left(res^{gt}\right) - \mathrm{EP}_m\left(res_i\right)\right)^2. \tag{4}$$

When the DIFF-EPITOME score is less than or equal to 1, we add a reward component $r_i^{DE} = 1$ to encourage the model to emulate the communication skills of real therapists.

If the output format of the completion o_i is correct, its total reward is the weighted sum of the three components:

$$r_i = w^F r_i^F + w^R r_i^R + w^{DE} r_i^{DE}. \tag{5}$$

Otherwise, its total reward $r_i = 0$.

4 Experiments

4.1 Experimental Settings

Datasets. `FIRES` is fine-tuned on the MESC dataset (Chu et al., 2024), which is a multimodal conversation dataset dedicated to ESC and involving real-world scenarios. It is sourced from the TV series *In Treatment* and contains psychotherapy sessions between therapists and clients. To evaluate the generalizability of our framework, we also perform inference on a subset drawn from another multimodal dialogue dataset EMOTyDA (Saha et al., 2020), named EMOTyDA-ESC. Further details about the datasets are provided in Appendix B.

Evaluation Metrics. To comprehensively evaluate the quality the generated responses, we employed: (1) Automatic Evaluation, including general text generation metrics such as BLEU-2 (B-2) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), METEOR (M.) (Banerjee and Lavie, 2005), Distinct-3 (Dist-3) (Li et al., 2016) and BERTScore (F_BERT) (Zhang et al., 2020) , as well as DIFF-EPITOME (DIFF-E.) (Lee et al., 2022) specialized for dialogue systems. (2) Human Evaluation, where two expert human evaluators are employed to compare the responses generated by our model and the baselines. More details about evaluation metrics are provided in Appendix C.

Baselines. We compared the following models to validate the effectiveness of our framework:

- Previous methods based on conversational agent BlenderBot: BlenderBot-Joint (Liu et al., 2021) and BBMHR (Zhang et al., 2023);
- Closed-source LLMs: we perform few-shot learning with GPT-3.5, GPT-4, GPT-4o, and the recent reasoning model DeepSeek-R1[2]. We also conduct experiments using different CoT prompt templates (Fixed CoT or Flexible CoT) with GPT-4o.
- Open-source LLMs: we choose Qwen2.5-VL-7B (Bai et al., 2025), a recent powerful MLLM that supports visual and text input, as the baseline and backbone of our framework.
- The baseline model SMES (Chu et al., 2024), proposed by the creators of the MESC dataset, which first utilizes MLLM to extract emotion cues from videos, and then fine-tunes a small model to sequentially generate user emotion category, strategy type, system emotion category and response. We also reproduce the SMES framework based on Qwen2.5-7B, denoted as SMES_Qwen2.5.

Implementation Details. Base on experimental observations, the value of reference consistency reward r^R is relatively small (around 0.3), so we set its weight $w^R = 1$ and the weight of support skill reward $w^{DF} = 0.1$. Since the model can generally follow the required format after the SFT cold start, we set $w^F = 0.1$ to better highlight the differences in response quality among completions. Due to the scarcity and high computational cost of MLLMs that supports tri-modal fine-tuning, we only utilize

[2]https://api-docs.deepseek.com/

Methods	B-2 ↑	R-L ↑	M. ↑	F_BERT ↑	Dist-3 ↑	DIFF-E. ↓
			Few-shot			
GPT-3.5*	1.01	4.60	-	84.50	-	-
GPT-4*	4.98	9.96	-	84.60	-	-
Deepseek-R1 (3-shot)	4.58	13.69	15.28	84.96	99.87	0.8774
GPT-4o (3-shot)	4.87	14.83	**18.06**	85.17	99.79	0.9001
GPT-4o (3-shot) + Fixed CoT	5.52	15.14	14.26	85.51	99.86	0.8667
GPT-4o (3-shot) + Flexible CoT	**6.07**	**16.28**	15.42	**85.65**	**99.90**	**0.8260**
Qwen2.5-VL (0-shot)	4.23	12.16	12.56	85.08	99.89	1.0007
			Fine-tune			
BlenderBot-Joint*	4.85	15.25	-	85.50	-	-
BBMHR*	1.31	15.38	-	86.60	-	-
SMES*	5.13	15.42	-	**86.80**	-	-
Qwen2.5-VL	9.77	20.30	14.05	86.35	84.01	0.8744
SMES_Qwen2.5	9.24	19.88	13.16	85.62	92.61	0.7577
FIRES	**10.18**	**22.76**	**19.54**	86.26	**98.93**	**0.7209**

Table 1: Performance comparison of different methods on MESC. The superscript * denotes results copied from Chu et al. (2024). "Fixed CoT" prompts GPT-4o to think about all four predefined aspects step by step, while "Flexible CoT" prompts it to flexibly consider only the helpful aspects.

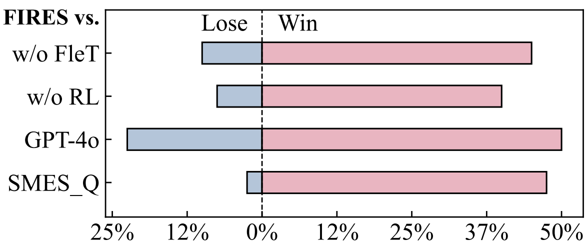

Figure 5: Human evaluation of responses. "SMES_Q" refers to SMES_Qwen2.5, "FleT" refers to Flexible Thinking, and "RL" refers to the RL algorithm GRPO.

the texts and videos of the conversations. The recently released powerful open-source large vision-language model, `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025), is employed as our backbone, and is fine-tuned using LoRA (Hu et al., 2022) on 4 A6000 GPUs. To reduce the resource consumption and training duration, we extract key video frames using FFmpeg as visual input to the model. The hyperparameters G and β for GRPO are set to 4 and 0.001, respectively. All codes are implemented with PyTorch. More hyperparameter settings are detailed in the Appendix E.

4.2 Main Results

The automatic evaluation results, presented in Table 1, demonstrate that our proposed `FIRES` framework significantly outperforms other baseline methods in both the fine-tuning and few-shot settings. In the fine-tune comparison, `FIRES` shows clear advantages over strong baselines like "Qwen2.5-VL" and "SMES_Qwen2.5". While SMES predicts emotion and strategy categories before generating responses, the proposed flexible thinking in the form of open natural language proves more beneficial for generating high-quality supportive responses. Under the few-shot learning setting, LLMs show high F_BERT scores and excellent diversity, but their overall performance still falls short of fine-tuned models and our framework `FIRES`. DeepSeek-R1, despite being a reasoning model that generates both thinking process and response, performs slightly poorly. GPT-4o achieves better results its performance is further improved by CoT prompting. More importantly, the fact that a flexible CoT prompt outperforms a fixed one further corroborates our core hypothesis: adaptive, flexible thinking is more effective than a rigid, fixed process for the MESC task.

Complementing automatic metrics, the results of human evaluation through A/B test are illustrated in Figure 5. Compared with the baselines GPT-4o and SMES_Qwen2.5, `FIRES` achieves significantly higher win rates, indicating that its generated responses are of higher quality and thereby more preferred by humans. This further substantiates the advantages of our framework.

4.3 Ablation Study

Impact of different components. As shown in Table 2, by prompting the MLLM to engage in

Methods	B-2 ↑	R-L ↑	M. ↑	DIFF-E. ↓
FIRES	10.18	**22.76**	**19.54**	0.7209
w/o FleT	9.17	21.62	18.79	**0.7167**
w/o SFT	9.73	21.56	15.79	0.8480
w/o RL	8.95	19.58	13.58	0.8972
w/o SSR	**10.32**	22.46	18.38	0.7651
w/o Video	9.99	21.08	15.82	0.7702

Table 2: Impact of different components of our model FIRES. "SSR" refers to the support skill reward. "w/o Video" represents the removal of visual input.

FIRES vs.	FIRES w/o RL	FIRES w/o FleT
Coherence	14% / 2% / 84%	6% / 5% / 89%
Usefulness	37% / 8% / 55%	23% / 16% / 61%

Table 3: Human evaluation of thinking processes. The percentages represent the Win/Lose/Tie rates.

Methods	B-2 ↑	R-L ↑	M. ↑	DIFF-E. ↓
FIRES	10.18	**22.76**	**19.54**	0.7209
w/o Scene	**10.41**	22.62	18.23	0.7504
w/o Emotion	10.00	22.26	19.34	**0.7107**
w/o Situation	9.40	21.18	16.7	0.7349
w/o Strategy	9.42	21.46	17.72	0.7394
w/o All	8.66	20.15	15.57	0.8333

Table 4: Impact of thinking aspects. "w/o All" means that the model freely explore its thinking process without any predefined aspects.

Methods	B-2 ↑	R-L ↑	M. ↑	DIFF-E. ↓
GPT-4o	3.77	10.22	10.35	**0.6441**
Qwen2.5-VL	5.05	12.67	10.81	0.8720
FIRES w/o RL	6.32	11.79	7.48	1.04517
FIRES	**7.47**	**16.29**	**12.26**	0.8817

Table 5: Cross-dataset generalization. All models directly make inferences on EMOTyDA-ESC.

comprehensive, fixed thinking for each aspect during data synthesis, and then applying the same SFT cold start and GRPO optimization, the resulting model (w/o FleT) exhibits a decline in performance, illustrating the necessity and effectiveness of flexible thinking. Removing the GRPO stage (w/o RL) leads to the most significant performance degradation, underscoring its critical role in eliciting more effective thinking patterns through the supervision of responses. Figure 5, where FIRES demonstrates a clear win rate against "w/o FleT" and "w/o RL", also corroborate these findings. The cold-start stage is equally important and if removed (w/o SFT), the model lacks a reference from the synthetic data and struggles to learn effective flexible thinking patterns, thus resulting in suboptimal performance. Furthermore, ESC-related support skill reward offer beneficial guidance during the GRPO stage. Additionally, removing visual input (w/o Video) markedly weakens the performance, demonstrating its importance for the MESC task. To further validate the critical role of the flexible thinking and RL stage, we also conduct human evaluation of the generated thinking processes (see Appendix C for details). As shown in Table 3, the thinking processes from FIRES are judged as more coherent and useful than those from the ablated versions, providing direct evidence that our framework generates more effective reasoning.

Impact of thinking aspects. To further validate the effectiveness of each of the four proposed thinking aspects, we conduct a detailed ablation study, with the results presented in Table 4. The results indicate that our FIRES framework, which adaptively leverages four thinking aspects, achieves the best or near-best performance across several key metrics. Specifically, removing any single thinking aspect from our synthesized data leads to a performance degradation on certain core metrics. The most significant performance drop occurs when removing the Situation or Response Strategy aspects. This suggests that analyzing the user's underlying problem and planning the response strategy are two of the most critical steps for generating high-quality supportive responses. Removing the Visual Scene aspect, while resulting in a slight increase in the B-2 score, degrades performance on other metrics, particularly on DIFF-EPITOME. This indicates that incorporating visual information helps the model better understand the user's state and generate higher-quality responses. Finally, the "w/o All" variant, which removes predefined aspects and allows the model to think freely, performs the worst across all metrics. This result confirms that unguided free thinking may lead to divergent and less helpful responses, thereby validating the necessity of our multi-aspect flexible thinking mechanism.

4.4 Further Analysis

Cross-Dataset Generalization. Compared to powerful MLLMs such as GPT-4o and Qwen2.5-VL, FIRES achieves better results on most metrics, demonstrating the strong generalizability. The performance degradation of the variant "FIRES w/o RL" suggests that solely performing SFT on flexible thinking data limits its generalization on unseen data. This also corroborates recent research find-

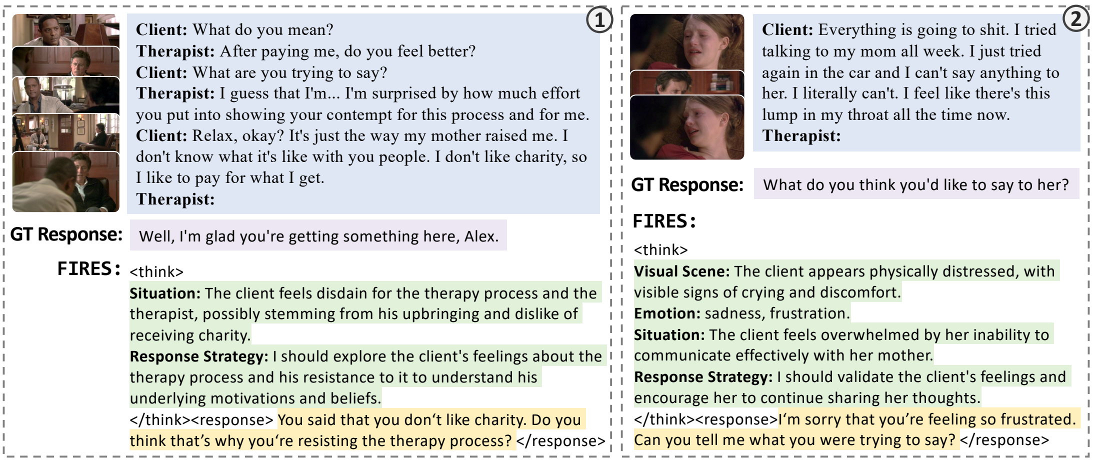

Figure 6: Model outputs for two instances from the test set. "GT" means Ground Truth.

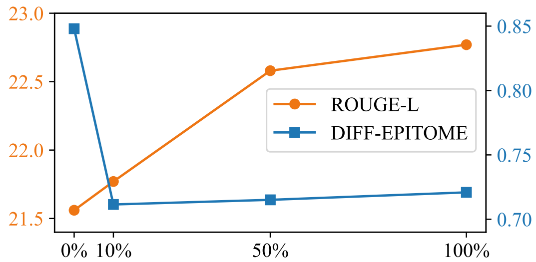

Figure 7: Impact of SFT cold start data size.

ings of "SFT for memorization, RL for generalization" (Chu et al., 2025a). Notably, due to the inherent discrepancies between MESC (targeting psychotherapy scenarios) and EMOTyDA-ESC (containing diverse daily dialogues across various scenarios), our model FIRES, fine-tuned on the MESC dataset, might be more inclined to explore the interlocutor's situation, leading to a higher EX score and then a lower DIFF-EPITOME score compared to the general large model GPT-4o.

Impact of Cold Start Data Size. As shown in Figure 7, during the SFT cold-start stage, FIRES requires only 50% of the supervised data to achieve performance comparable to that achieved with the full data. Given the limited size of the MESC dataset (the training set contains only 3.7k instances) and our observation that increasing the data size helps mitigate the model's tendency towards a fixed thinking pattern, the results we report above all utilize synthetic data from the entire training set for cold start.

4.5 Case Study

The case 1 in Figure 6 illustrates the FIRES's ability to flexibly think about the helpful aspects based on the actual context. In the video, the client is conversing calmly with the therapist and exhibits no overt emotions; thus, Visual Scene and Emotion are excluded from the thinking process. FIRES analyzes the client's current situation and the corresponding response strategy, subsequently generating a response to actively explore the reasons for the client showing contempt for therapy process, which is consistently preferred by human evaluators even compared to the ground truth response. For the case 2, the generated thinking process, which comprehensively encompasses the visual scene of the client crying grievously and the analysis of other aspects, guides the generated response to be more empathetic and supportive. Comparison of outputs from other models is shown in Figure 9 in the Appendix.

5 Conclusion

In this paper, we introduce "Flexible Thinking" to address the limitations of rigid, fixed-step, and text-only reasoning. For the multimodal ESC task, we propose a framework named FIRES to cultivate helpful flexible thinking and generate effective response via a two-stage process: an initial SFT cold start on flexible thinking data synthesized by a MLLM, followed by refinement using an RL algorithm with a meticulously designed reward function. Extensive experiments on the MESC and EMOTyDA datasets have demonstrated the effectiveness and generalizability of our framework. We believe that enabling models to think flexibly like humans, i.e., adaptively adjusting their thinking focus based on the evolving multimodal context, is a promising path toward developing more human-like conversational AI capable of providing genuine and helpful emotional support.

Limitations

Despite the promising results achieved by our proposed framework, we acknowledge several limitations that warrant future exploration. Firstly, our design of reward function still has room for improvement. It relies on ROUGE-L to measure lexical overlap and DIFF-EPITOME to maintain support skill alignment, future work could explore more comprehensive rewards, such as those from an LLM-as-a-Judge, to better guide and evaluate response quality. Secondly, the MESC dataset we use for fine-tuning has constraints in terms of size and scenario diversity, which could influence the breadth of flexible thinking patterns learned and limit the model's generalizability to other diverse scenarios. More varied data could enhance the robustness and real-world applicability of our framework. Finally, to mitigate resource consumption and potential noise from full video processing, we simplifies visual input by extracting keyframes using FFmpeg Toolkit, which, while efficient, may not always precisely capture the most critical information. Future work could incorporate a dedicated keyframe selection module to more effectively identify and utilize the most salient visual cues for enhanced emotional support.

Acknowledgments

This work was supported by the National Natural Science Fouandation of China under Grant No. 62476132 and No. 62476134.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yinan Bao, Dou Hu, Lingwei Wei, Shuchong Wei, Wei Zhou, and Songlin Hu. 2024. Multi-stream information fusion framework for emotional support conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11981–11992.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2024. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 175–187. Springer.

Siyuan Chen, Cong Ming, Zhiling Zhang, Yanyi Chen, Kenny Q Zhu, and Mengyue Wu. 2024. Mixed chain-of-psychotherapies for emotional support chatbot. *arXiv preprint arXiv:2409.19533*.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. 2025. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1274–1282.

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

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17853–17861.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025a. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. 2025b. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2024. Towards multimodal emotional support conversation systems. *arXiv preprint arXiv:2408.03650*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 4079–4095.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. Empathyear: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Seunghee Han, Se Jin Park, Chae Won Kim, and Yong Man Ro. 2024. Persona extraction through semantic similarity for emotional support conversation generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11321–11325. IEEE.

Jiawang Hao and Fang Kong. 2025. Enhancing emotional support conversations: A framework for dynamic knowledge filtering and persona extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3193–3202.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 900–909.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Dongjin Kang, Sunghwan Mac Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261.

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

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu. 2024a. Dq-hgan: A heterogeneous graph attention network based deep q-learning for emotional support conversation generation. *Knowledge-Based Systems*, 283:111201.

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

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024b. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5794–5805.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024c. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1976–1988.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2040–2079.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, and 1 others. 2025a. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. 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 (Volume 1: Long Papers)*, pages 3469–3483.

Yiting Liu, Liang Li, Yunbin Tu, Beichen Zhang, Zheng-Jun Zha, and Qingming Huang. 2024. Dynamic strategy prompt reasoning for emotional support conversation. *IEEE Transactions on Multimedia*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

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

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, 264:110340.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Soham Sane. 2025. Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization. *arXiv preprint arXiv:2502.01652*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted ai feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 567–585.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. Coke: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15984–16007.

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 2024. Dynamic demonstration retrieval and cognitive understanding for emotional support conversation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 774–784.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694.

Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. 2024a. Improving reinforcement learning from human feedback with efficient reward model ensemble, 2024. *URL https://arxiv. org/abs/2401.16635*.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024c. Stickerconv: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733.

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

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, and 1 others. 2023b. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729.

A Prompt Templates

The prompt template, illustrated in Figure A, guides GPT-4o in synthesizing the flexible thinking process for each instance. It details the expected role, the task definition, and necessary requirements, along with input-output pairs that serve as few-shot demonstrations for in-context learning.

B Details of Datasets

The MESC dataset[3] (Chu et al., 2024) also provides annotations for emotion categories and strategy types, which served as auxiliary information for our data synthesis process. Based on the released timestamps, we utilize the FFmpeg tool to segment the video clips from the raw episodes of *In Treatment*. During the experiments, we concatenate consecutive utterances from the same speaker to consolidate their conversational turns, and designate each therapist's turn as the target response for a instance, with preceding utterances serving as the conversation history. The training, validation, and test sets of the MESC dataset comprise 3,721, 445, and 454 instances, respectively.

The EMOTyDA dataset[4] (Saha et al., 2020) is derived from the TV series *Friends* and is constructed for emotion recognition and dialogue act classification. Based on the emotion and dialogue act annotations provided in its test set, we automatically curate a subset of 310 instances suitable for the ESC task, where the target response's emotion must not be negative, and the dialogue act should not be trivial "Backchannel" or "Greeting".

C Details of Evaluation Metrics

To comprehensively evaluate the quality the generated responses, we adopt the following general text generation metrics: BLEU-2 (B-2) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and METEOR (M.) (Banerjee and Lavie, 2005) for lexical overlap; Distinct-3 (Dist-3) (Li et al., 2016) for text diversity; BERTScore (Zhang et al., 2020) for semantic similarity. The pre-trained model `roberta-large` is used to obtain embeddings and compute the F_1 of BERTScore (F_BERT). Other metrics are calculated using the NLTK package, following the implementation[5] by Liu et al.

[3]https://github.com/chuyq/MESC

[4]https://github.com/sahatulika15/EMOTyDA

[5]https://github.com/thu-coai/Emotional-Support-Conversation/blob/main/codes_zcj/metric/myMetrics.py

Role
You are an expert in thinking process generation.
Task
Given a multimodal client-therapist conversation, including conversation history, video frames, and the therapist's final response, please infer the therapist's latent thinking process preceding his response. Flexibly include the thinking aspects below:
1. Visual Scene: Decribe the key elements observed in the video. This includes …
2. Emotion: Recognize the client's emotions based on the textual utterances and visual cues.
3. Situation: If the client's emotion is not neutral, describe the causes that triggered the emotion. If neutral, summarize the issues he/she is facing.
4. Response Strategy: Outline the intended response strategy and the communicative goal.
Requirements
1. List each aspect strictly following the format above.
2. Depending on the specific conversation context, describe only aspects that is helpful to the response; otherwise, write "Not helpfu". Write "Unknown" if there is insufficient information.
3. Do not mention the therapist's final response.
4. Be concise and to the point.
Example 1
* **Conversation history: …** \n*<image> <image> <image> <image>* **Therapist**: For not telling your mother about your father's affair…You are not responsible for your parents' actions. The only actions that we can control are our own... \n*<image>* **Client**: So you want me to forgive myself?\n*<image>* **Therapist**: Mm-hmm.\n<image>**Client**: how?
* **Therapist's Response:** Slowly. It's gonna take a while. It's taken you a few years to fall into this and it's probably gonna take you a while to get out of it. But you have to start trying now. 'cause you know something? You deserve to be happy.
* **Thinking process:** 1. **Visual Scene**: None\n2. **Emotion**: depressed, guilty\n3. **Situation**: she believes her inaction (not telling her mother about her father's affair) caused her parents' breakup. \n4. **Response Goal**: Therefore, the therapist should give advice and encourage the client to begin the process of self-forgiveness.

Example 2
…
Example 3
…
New Instance
…

Figure 8: The prompt template used for the synthesis of flexible thinking data.

(2021). Furthermore, the DIFF-EPITOME[6] (DIFF-E.) score (Lee et al., 2022) that is specialized for dialogue systems and utilized in our reward modeling (Section 3.4.3), is also employed as an automatic evaluation metric.

In addition to these automatic metrics, we conduct a human evaluation for a more nuanced assessment of response quality. For this, two expert human evaluators are employed to perform A/B tests, comparing the responses generated by our framework `FIRES` against those from the baselines. In line with common practice in prior work (Liu et al., 2021; Chu et al., 2024), we randomly sample 100 instances from the test set of the MESC dataset. This sample size is sufficient to indicate general performance trends, as it constitutes a significant portion (approximately 22%) of the whole test set. During each A/B test, evaluators need to rate `FIRES`'s response as "win", "lose", or "tie" based on whether it is more supportive, effective, and human-preferred. Subsequently, we calculate the average win and loss rates from these ratings. The results are illustrated in Figure 5.

We also conduct a human evaluation to directly assess the quality of the generated thinking processs. The human evaluators are employed to perform A/B comparison on 100 sampled test instances, comparing the thinking processs from our full framework `FIRES` against those from ablation variants. Each comparison was judged on two dimensions:

- Coherence: Is the thinking process logical and relevant to the conversational context?
- Usefulness: Is the thinking process helpful for the final response?

The results are reported in Table 3.

[6] `https://github.com/passing2961/EmpGPT-3/blob/main/modules/empathy_scorer.py`

D Details of Case Study

In addition to our model's output, Figure 9 also displays the outputs of two variants and GPT-4o.

E Hyperparameter Settings

Table 6 illustrates the basic hyperparameter settings during our SFT and GRPO fine-tuning stages.

The parameter β in GRPO is a regularization coefficient that controls the KL-divergence between the current policy and the reference SFT model. We set $\beta = 0.001$ empirically during our initial experimental exploration, based on observations of training stability and final performance. For instance:

- Setting β to a high value (e.g., 0.04) overly constrains the policy optimization. The strong KL-divergence penalty prevents the model from learning effectively from the reward signal. This results in a noisy reward curve that fails to converge, and the model's performance does not improve.
- Conversely, setting β to a very low value or even zero, i.e., removing the crucial regularization against the reference model, leads to the problem of Reward Hacking, which in our case manifests as policy collapse. For example, the model would frequently select the thinking pattern like "Situation: Unknown" and then tend to generate a generic and low-effort response like "What do you mean?", ultimately degrading the overall performance. Our chosen value of $\beta = 0.001$ provides a trade-off, allowing for effective policy optimization while maintaining stable training.

Due to the significant computational cost associated with training, all experiments in this work were conducted once with a fixed random seed of 42 to ensure reproducibility. The reported scores reflect the performance from this single run.

Hyperparameter	SFT Cold Start	GRPO
lora_rank	8	8
lora_alpha	16	32
batch size	32	32
learning rate	1e-5	1e-5
epoch	3	3
warmup ratio	0.1	0.05
deepspeed	-	zero2
G	-	4
β	-	0.001

Table 6: Hyper-Parameter Settings.

F License Discussion

The scientific artifacts used and developed in this work include pre-trained model and datasets derived for experimental purposes. We detail their licensing and usage terms below:

- **Model**: The MLLM utilized for fine-tuning, `Qwen2.5-VL-7B-Instruct`[7] (Bai et al., 2025), is made available under the Apache 2.0 license. We have adhered to the terms of this license in our use of the model.
- **Dataset**: The MESC dataset was publicly released by Chu et al. (2024). The visual data for our multimodal experiments consists of video clips derived from the TV series *In Treatment*. The specific video segments were identified using timestamps originally provided by Chu et al. (2024). The process of downloading the videos of the TV series and segmenting them into clips based on the aforementioned timestamps was undertaken by us strictly for internal, non-commercial academic research purposes, specifically for the analysis and generation of multimodal emotional support conversations. Due to the copyrighted nature of the source data, we will not be distribute, publish, or share any of the raw video clips or derived visual content.

[7] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

	①	②
Conversation History	 **Client:** What do you mean? **Therapist:** After paying me, do you feel better? **Client:** What are you trying to say? **Therapist:** I guess that I'm... I'm surprised by how much effort you put into showing your contempt for this process and for me. **Client:** Relax, okay? It's just the way my mother raised me. I don't know what it's like with you people. I don't like charity, so I like to pay for what I get. **Therapist:**	 **Client:** Everything is going to shit. I tried talking to my mom all week. I just tried again in the car and I can't say anything to her. I literally can't. I feel like there's this lump in my throat all the time now. **Therapist:**
GT Response	Well, I'm glad you're getting something here, Alex.	What do you think you'd like to say to her?
FIRES	<think> **Situation:** The client feels disdain for the therapy process and the therapist, possibly stemming from his upbringing and dislike of receiving charity. **Response Strategy:** I should explore the client's feelings about the therapy process and his resistance to it to understand his underlying motivations and beliefs. </think><response> You said that you don't like charity. Do you think that's why you're resisting the therapy process? </response>	<think> **Visual Scene:** The client appears physically distressed, with visible signs of crying and discomfort. **Emotion:** sadness, frustration. **Situation:** The client feels overwhelmed by her inability to communicate effectively with her mother. **Response Strategy:** I should validate the client's feelings and encourage her to continue sharing her thoughts. </think><response>I'm sorry that you're feeling so frustrated. Can you tell me what you were trying to say? </response>
FIRES w/o RL	<think> **Visual Scene:** The client appears relaxed, sitting back in his chair with a casual posture. **Situation:** The client feels uncomfortable with the therapeutic process and perceives it as a form of charity. **Response Strategy:** I should explore the client's feelings about the therapeutic relationship and address his resistance to the process. </think><response>So you're not comfortable with the therapeutic relationship?</response>	<think> **Visual Scene:** The client appears physically distressed, with visible signs of crying (red eyes, tear-streaked face). **Emotion:** sadness. **Situation:** The client expresses distress over communication difficulties with her mother. **Response Strategy:** I should ask an open-ended question to explore the client's feelings about her mother and the difficulty of expressing herself. </think><response>And you said you tried talking to your mom. What did you try to say? </response>
FIRES w/o FleT	<think> **Visual Scene:** The client is sitting comfortably, gesturing with his hands while speaking. The therapist is seated across from him, listening attentively. Both appear relaxed but engaged in the conversation. **Emotion:** neutral. **Situation:** The client is discussing his approach to therapy, emphasizing his preference for paying for services rather than receiving them for free. **Response Strategy:** the therapist should use 'restatement' to reflect back what the client has said, encouraging further discussion. </think><response> You said that you don't like charity. Do you mean that you don't like being given something for free? </response>	<think> **Visual Scene:** The client appears distressed, with tears in her eyes and a tense expression. She is lying down, indicating she might be feeling overwhelmed or exhausted. **Emotion:** sad. **Situation:** She is unable to communicate effectively with her mother and feels frustrated and emotionally stuck. **Response Strategy:** I should use an 'open question' strategy to facilitate further discussion and understanding of the client's emotions and thoughts. </think><response> You said you couldn't talk to your mother. What do you mean by that? </response>
GPT-4o	It sounds like paying for this session is a way for you to maintain a sense of control and independence, perhaps even to protect yourself from feeling vulnerable. Do you think that might be connected to how your mother raised you?	It sounds like you're feeling really overwhelmed and stuck right now. That lump in your throat might be your body's way of expressing how difficult it is to communicate with your mom. What do you think is stopping you from saying what you want to say to her?

Figure 9: Comparison of model outputs for two instances from the test set.