

Beyond Semantic Similarity: Appraisal-Guided Chain-of-Thought Reasoning and Retrieval for Multimodal Emotional Support Conversations

Yuqi Chu^{1,*}, Lizi Liao², Jinggui Liang², Boyang Li³, Richang Hong^{1,†}

¹Hefei University of Technology, Hefei, China

²Singapore Management University, Singapore

³Nanyang Technological University, Singapore

{chuyuqi127, hongrc.hfut}@gmail.com, lzliao@smu.edu.sg,
jg.liang.2023@phdcs.smu.edu.sg, boyang.li@ntu.edu.sg

Abstract

Emotional support conversation systems strive to emulate the empathetic depth of human therapists, yet current approaches often fail due to the "*Cognitive Gap*"—the inability to discern the latent psychological evaluations driving a user’s distress. Existing retrieval-augmented generation paradigms exacerbate this by relying on semantic similarity, frequently retrieving historical dialogues that are surface analogous but therapeutically incongruent. To bridge this gap, we introduce Appraisal-Guided Chain-of-Thought Reasoning & Retrieval (**AG-CTR**²) for better emotional support. Specifically, we bootstrap the MLLM to generate appraisal-guided reasoning chains and apply a dual-signal verification mechanism using ground-truth emotion labels and a teacher model to verify and refine them. Under such instance-level guidance, we finetune the MLLM to internalize such reasoning capability. At inference, the model utilizes its generated appraisal chain as a structured query to help retrieve historical therapeutic responses based on psychological situation similarity rather than content surface proximity. Extensive experiments and analyses on two ESC benchmarks demonstrate that AG-CTR² significantly outperforms state-of-the-art baselines. Our data and codes are available.¹

1 Introduction

The surging demand for accessible mental health care has catalyzed the development of Emotional Support Conversation (ESC) systems designed to alleviate distress (Zheng et al., 2024b; Chu et al., 2025). While the evolution toward multimodal frameworks that integrate text, acoustic, and visual cues promises richer emotion detection, bridging the gap between recognition and the delivery of genuine, cognitively informed support remains a challenge requiring deep psychological insight.

*Work was done during an internship at SMU.

†Corresponding author.

¹https://github.com/chuyq/Appraisal_esc

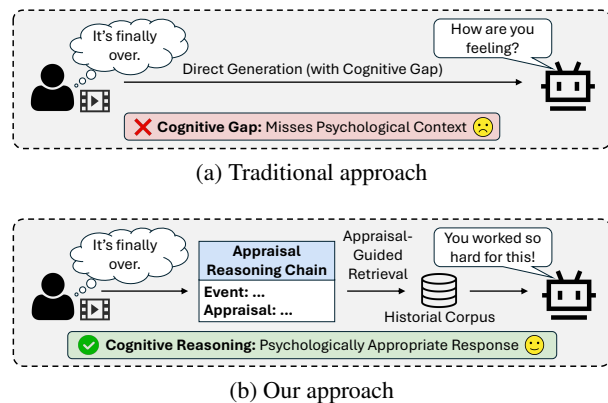


Figure 1: Compared to direct generation in traditional approaches, our AG-CTR² generates explicit appraisal-grounded reasoning chains to serve as structured retrieval queries, guiding the system toward psychologically resonant retrieval and generation.

Despite recent advances, existing ESC methods still suffer from a critical *Cognitive Gap*, as shown in Figure 1. Consider a user saying "*It's finally over*": without understanding whether they appraise this as relief from burden (warranting celebration) or grief over loss (requiring consolation), any support system risks providing therapeutically misaligned responses. Prior methods either rely on strategy prediction followed by response generation (Liu et al., 2021; Cheng et al., 2022) or incorporate commonsense knowledge to augment generation (Tu et al., 2022; Zhou et al., 2023), but neither explicitly models the appraisal process underlying emotions.

Recent work has incorporated Retrieval-Augmented Generation (RAG) (Chen et al., 2022; Deng et al., 2023) from external knowledge data, aiming to improve response quality. **However, the Cognitive Gap not only persists in RAG systems but becomes even more detrimental:** by relying on semantic similarity to match user inputs with historical dialogues, RAG retrieves based on surface text overlap rather than psychological congruence (Lewis et al., 2020; Jiang et al., 2023).

Returning to the earlier example, a semantic retriever may fetch grief-oriented support for a relief scenario simply because both contain "finally over", yielding retrieved exemplars that are superficially relevant but therapeutically inappropriate. This surface-level matching amplifies the cognitive gap's consequences, as misaligned retrievals directly condition on response generation.

To bridge this gap, we draw upon Cognitive Appraisal Theory (Lazarus, 1991; Folkman et al., 1986), a foundational psychological framework positing that emotions are not direct reactions to events, but products of an individual's evaluation of those events. According to this theory, the specific emotion felt is determined by how one assesses the event's significance (Primary Appraisal) and one's ability to cope with it (Secondary Appraisal). We argue that to close the cognitive gap, ESC systems could better emulate this cognitive process. By explicitly modeling the reasoning chain from Event \rightarrow Appraisal \rightarrow Coping \rightarrow Emotion, an MLLM chatbot can understand the "psychological situation" of the user more accurately. This structured reasoning can then serve as a sophisticated retrieval query within a RAG framework, allowing the system to find historical cases where users exhibited the same cognitive patterns, regardless of whether the specific topics (e.g., job loss vs. breakup) differ.

In this work, we propose AG-CTR² (Appraisal-Guided Chain-of-Thought Reasoning & Retrieval), a novel framework that realizes this theory through a "Generate-Verify-Internalize" pipeline. First, we bootstrap a Multimodal Large Language Model (MLLM) to generate diverse candidate reasoning chains grounded in appraisal theory. Recognizing that unverified reasoning chains may contain inconsistencies or misalignments with the user's emotional state, we employ a dual-signal verification mechanism for reasoning learning: ground-truth emotion labels provide initial filtering based on emotion alignment, while a strong teacher model evaluates the appraisal chains across multiple criteria and generates refined versions when quality is insufficient. This process constructs high-quality and low-quality reasoning repositories. Under this instance-level guidance, we finetune the MLLM to internalize this reasoning capability through supervised learning on high-quality chains and negative preference optimization against low-quality chains, learning to produce appraisal chains that better align with verified chains rather than noisy or weakly grounded reasoning patterns. Finally, at in-

ference, the model generates an appraisal chain that serves as a structured query, guiding the retrieval module to fetch therapeutic responses based on appraisal-level similarity rather than surface-level content proximity.

Our contributions are summarized as follows:

- We propose AG-CTR² to explicitly integrate Cognitive Appraisal Theory into both the reasoning and retrieval stages, shifting the paradigm from semantic matching to "cognitive resonance".
- We design a bootstrapping and fine-tuning method that leverages ground-truth emotion labels and a teacher model to verify and internalize high-quality appraisal reasoning, leading to more stable and supervision-aligned reasoning chains.
- Extensive experiments demonstrate that AG-CTR² significantly outperforms all baselines on two emotional support datasets.

2 Related Work

2.1 Emotional Support Conversation

With the growing demand for mental health support, ESC has emerged as a promising digital mental health solution. Early research focused on strategy-driven generation (Liu et al., 2021; Zheng et al., 2024a) or enriching context via knowledge-augmented (Bao et al., 2024; Liang et al., 2025) or persona modeling (Cheng et al., 2023; Hao and Kong, 2025). Recently, some studies have focused on multimodal ESC to capture non-verbal cues, such as acoustic and visual signals, which are essential for understanding the user's emotional state (Chu et al., 2025). However, a persistent "Cognitive Gap" remains. While Wang et al. (2025) utilizes advanced optimization techniques like GRPO to improve response quality, they do not explicitly model the underlying causal mechanisms of emotions. Unlike prior methods that infer user states without explicit psychological modeling, we model the appraisal chain to capture the latent psychological evaluations underlying distress.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is used in dialogue systems to mitigate hallucinations and provide grounded responses (Chen et al., 2022; Liu et al., 2024). Existing retrievers typically rely on surface-level semantic similarity (Jiang et al., 2023; Zhou et al., 2025; Hu et al., 2025), which may fail in therapeutic contexts where identical utterances

(e.g., "It is over") signify vastly different psychological states (e.g., *depression* vs. *neutral*). In the context of ESC, prior methods have explored RAG by retrieving information from mental health knowledge graphs to assist response generation (Deng et al., 2023). Meanwhile, some methods incorporate Chain-of-Thought (CoT) reasoning to improve the interpretability and reasoning quality of ESC (Wei et al., 2022; Zhang et al., 2024). Despite these advances, CoT reasoning has rarely been integrated into the retrieval process itself. Our work bridges this gap by leveraging appraisal-grounded reasoning chains as structured retrieval queries, enabling the system to retrieve historical cases based on psychological resonance rather than superficial dialogue similarity.

3 Methodology

Figure 2 illustrates our AG-CTR² framework, which mainly consists of three components: **(1) Appraisal-Grounded Reasoning Generation:** To capture underlying cognitive evaluations rather than surface-level content, we first bootstrap an MLLM to generate structured appraisal reasoning chains that decompose user emotions into appraisal-inspired dimensions reflecting inferred event, concerns, and coping assessments. **(2) Dual-Signal Verification Reasoning:** To ensure the reliability and psychological validity of these chains, we introduce a dual-signal verification mechanism that filters reasoning using both ground-truth emotion labels and teacher model assessment, followed by training the MLLM via SFT and negative preference optimization (NPO) to internalize accurate reasoning patterns. **(3) Appraisal-Guided Retrieval and Response Generation:** To better utilize the generated appraisal reasoning, we leverage these chains as queries to retrieve therapeutically aligned dialogues for response generation.

3.1 Problem Formulation

Given a multimodal user query $U_t = \{u_t, v_t\}$ where u_t represents the textual user utterance and v_t denotes the associated video, along with the dialogue history $H_t = \{U_1, \dots, U_{t-1}\}$, the model M_{θ_r} first generates a textual appraisal-based reasoning chain $A_t \sim M_{\theta_r}(A_t | H_t, U_t)$ to capture the underlying cognitive evaluations of the user. This chain A_t is then employed as a query to retrieve psychologically aligned responses \mathcal{T}_t from the historical corpus, ensuring logical rather than

merely surface-level alignment. The goal is to optimize a generation function $P(r_t | H_t, U_t, A_t, \mathcal{T}_t)$ that produces a textual response r_t .

3.2 Appraisal-Guided Reasoning Generation

To enable accurate interpretation of user distress beyond surface-level emotional cues, we model the latent cognitive evaluations that underpin a user’s emotional state. Inspired by Cognitive Appraisal Theory (Lazarus, 1991; Folkman et al., 1986), this step enables the model to discern that emotions do not arise directly from triggering events (e.g., a breakup or illness, which can be easily observed from the dialogue context), but from an individual’s specific cognitive evaluation of those events. We define reasoning as a structured, appraisal-guided process in which each step corresponds to an explicit, appraisal-inspired component, forming an ordered reasoning process from multimodal perception to emotional inference.

Specifically, we decompose the appraisal process into a reasoning structure comprising following key dimensions: *Event Understanding* (A_e), *Negative Appraisal* (A_n), and *Coping Assessment* (A_c), which collectively determine the *Inferred Emotion* (\hat{L}_e). Formally, the appraisal reasoning chain is defined as a structured sequence:

$$A_t = (A_e, A_n, A_c) \rightarrow \hat{L}_e. \quad (1)$$

This notation explicitly demonstrates the reasoning path from latent psychological attributes to the emotion label prediction, which would be verified in the next section.

To extract these reasoning chains from our multimodal dataset, given the multimodal query $U_t = \{u_t, v_t\}$ and dialogue history H_t , we utilize the M_{θ_r} to generate the reasoning. We design an Appraisal Reasoning Prompt (ARP) containing task descriptions and human-curated examples of multimodal context-appraisal pairs, see Appendix B. For each instance in the dataset, we elicit the appraisal reasoning as:

$$A_t \sim M_{\theta_r}(\text{ARP} \oplus H_t \oplus U_t), \quad (2)$$

where \oplus denotes the concatenation operation. We get the model to generate the reasoning paths. This generated set A serves as the cognitive reasoning base for subsequent verification and refinement.

3.3 Dual-Signal Verification Reasoning

To address the limitations of prompt-based generation (§3.2), where directly generated reason-

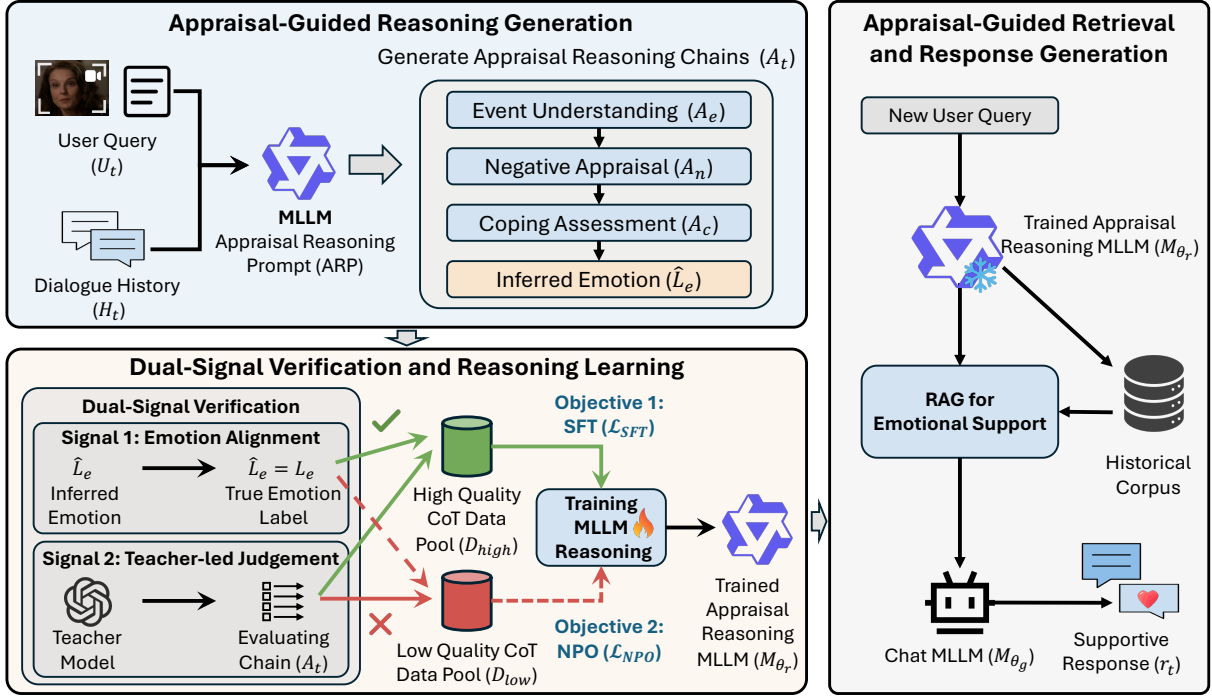


Figure 2: Overview of AG-CTR². Given a multimodal user query, an MLLM generates appraisal-grounded reasoning chains that model cognitive evaluations along Event \rightarrow Appraisal \rightarrow Coping \rightarrow Emotion. A dual-signal verification and reasoning mechanism filters and refines these chains using emotion supervision and teacher assessment, and trains the reasoning MLLM via supervised fine-tuning and negative preference optimization. Finally, the generated appraisal chain serves as a structured retrieval query to fetch psychologically aligned support responses and guide emotional support generation.

ing chains often lack reliability and psychological validity, we propose a dual-signal filtered reasoning learning approach. This method first employs a dual-signal filtering mechanism that leverages both ground-truth emotion labels and a high-capacity teacher model to construct high-quality and low-quality reasoning repositories ($\mathcal{D}_{\text{high}}$ and \mathcal{D}_{low}), then trains the MLLM through supervised fine-tuning and negative preference optimization to internalize accurate reasoning patterns.

Dual-Signal Verification The filtering process operates through two complementary signals. First, for each candidate chain $A_t \in \mathcal{A}$, we compare the inferred emotion \hat{L}_e with the ground-truth label L_e :

$$\mathcal{V}_{\text{align}}(A_t) = \begin{cases} \mathcal{D}_{\text{high}} \leftarrow A_t, & \text{if } \hat{L}_e = L_e \\ \mathcal{D}_{\text{low}} \leftarrow A_t, & \text{if } \hat{L}_e \neq L_e \end{cases}. \quad (3)$$

Beyond automated alignment, we utilize a high-capacity teacher model M_T to evaluate and rectify the reasoning. The teacher model assesses each chain A_t across four psychological and functional metrics $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$, scored on a five-point scale from 1 (poor) to 5 (excellent). Details are provided in Appendix A.

We define a composite quality score $S_{\text{total}} = \frac{1}{4} \sum_{i=1}^4 s_i$. If S_{total} falls below threshold τ , the teacher model generates a refined reasoning chain A_t^* . To support this process, we design an Appraisal Teaching Prompt (ATP) that includes task descriptions and human-curated examples of multimodal context–appraisal pairs (details are provided in the appendix A).

$$A_t^* = M_T(\text{ATP} \oplus H_t \oplus U_t \oplus L_e). \quad (4)$$

The original flawed chain is archived in \mathcal{D}_{low} as a negative signal, while A_t^* is added to $\mathcal{D}_{\text{high}}$. This dual-pool construction provides both positive demonstrations and contrastive negative examples for subsequent learning.

Reasoning Learning We train the M_{θ_r} to generate appraisal-based reasoning chains. The training follows a multi-stage schedule. Specifically, we first perform supervised fine-tuning (SFT) on high-quality chains from $\mathcal{D}_{\text{high}}$:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(H_t, U_t, A_t) \sim \mathcal{D}_{\text{high}}} \sum_{i=1}^{|A_t|} \log P_{\theta_r}(a_i | H_t, U_t, a_{<i}).$$

To enhance the model’s ability to discriminate between high-quality and flawed appraisal reasoning, we then apply negative preference optimization (NPO), which contrasts high-quality chains from $\mathcal{D}_{\text{high}}$ against flawed chains from \mathcal{D}_{low} :

$$\mathcal{L}_{\text{NPO}} = \underbrace{-\mathbb{E}_{A_t \sim \mathcal{D}_{\text{high}}} \sum_{i=1}^{|A_t|} \log P_{\theta_r}(a_i^+ | H_t, U_t, a_{<i}^+)}_{\text{positive reasoning reinforcement}} + \underbrace{\mathbb{E}_{A_t \sim \mathcal{D}_{\text{low}}} \sum_{i=1}^{|A_t|} \log P_{\theta_r}(a_i^- | H_t, U_t, a_{<i}^-)}_{\text{negative reasoning penalization}}.$$

This contrastive learning ensures the model unlearns flawed psychological attributes while reinforcing correct appraisal reasoning.

3.4 Appraisal-Guided Retrieval and Response Generation

To retrieve information based on appraisal reasoning patterns rather than surface-level textual similarity, our method leverages the trained MLLM to generate structured appraisal reasoning chains A_t as high-level queries, enabling the retrieval of historical dialogues with aligned psychological evaluations and supportive responses.

Retrieval Setup We retrieve only from the training split. The retrieval corpus \mathcal{D} consists of turn-level units, each comprising a supportive response with its dialogue context. For each unit, we store both the raw text and a precomputed appraisal chain generated offline. At inference, the model uses the generated appraisal chain A_t as the query. Queries and indexed units are encoded with the same frozen text encoder, and similarity is computed via cosine distance in the embedding space. For efficiency, we employ an approximate nearest-neighbor index to retrieve the top- k candidates. To prevent data leakage, all test instances are excluded from \mathcal{D} .

During inference, the trained M_{θ_r} first generates the appraisal reasoning chain A_t , we then retrieve a set of therapeutically relevant examples $\mathcal{T}_t = \{d_1, d_2, \dots, d_k\}$ from a historical corpus \mathcal{D} . This mechanism encourages the retrieval of therapeutic responses that align with the inferred appraisal structure underlying the user’s distress, rather than relying solely on keyword overlap or surface-level semantic similarity.

The final response is generated by a chat MLLM, M_{θ_g} . This model incorporates the multimodal con-

text, the generated appraisal reasoning chain, and the retrieved therapeutic insights as inputs. The training objective is defined as follows:

$$\mathcal{L}_{\text{gen}} = -\log P(r_t | H_t, U_t, A_t, \mathcal{T}_t).$$

4 Experiments

To verify the effectiveness of our proposed AG-CTR² framework and each of its components, we conduct experiments on two emotional support conversation benchmarks.

4.1 Experimental Setups

Datasets To evaluate the effectiveness of our framework, we conduct experiments on two ESC datasets: (i) **MESC** (Chu et al., 2025), a multimodal ESC dataset consisting of 1019 dialogues (28762 utterances). It provides rich annotations for seven emotion types across text, audio, and video modalities. Due to the lack of other open-source multimodal ESC datasets, we also select ESConv. (ii) **ESConv** (Liu et al., 2021), a text-based ESC dataset comprising 1053 dialogues (31410 utterances) with seven emotion types. We follow the original data splits for all experiments.

Evaluation Metrics For automatic evaluation of response generation, we adopt BLEU-2/4 (B-2, B-4) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and BERTScore-F1 (BERT-F1) (Zhang et al., 2020) to assess n-gram overlap, sequence-level similarity, and semantic consistency between generated responses and golden responses.

Baselines We compare our method against previous ESC models, general MLLM models, and fine-tuned MLLM ESC models. Specifically, the baselines include four text-based ESC methods: BlenderBot (Roller et al., 2021), BlenderBot-Joint (Liu et al., 2021), MISC (Tu et al., 2022) and KEMI (Deng et al., 2023). Four general MLLM methods: Video-LLaMA (Zhang et al., 2023), GPT-4o mini (Hurst et al., 2024), Qwen2-VL (Wang et al., 2024), and Qwen2.5-VL (Bai et al., 2025). Four fine-tuned MLLM ESC methods: Emotion-LLaMA (Cheng et al., 2024), SMES (Chu et al., 2025), Qwen2.5-VL(SFT) (Bai et al., 2025), and FIRES (Wang et al., 2025). More details about the baselines can be found in Appendix C.

Implementation Details We employ two separate multimodal large language models in our framework. The reasoning M_{θ_r} is Qwen2.5-VL-7B-Instruct (Bai et al., 2025), which is optimized with \mathcal{L}_{SFT} and \mathcal{L}_{NPO} . The chat M_{θ_g} is MiniGPT-v2,

Methods	MESC				ESConv			
	B-2 \uparrow	B-4 \uparrow	R-L \uparrow	BERT-F1 \uparrow	B-2 \uparrow	B-4 \uparrow	R-L \uparrow	BERT-F1 \uparrow
BlenderBot (Roller et al., 2021)	4.36	0.80	14.20	84.80	5.45	1.05	15.43	85.07
BlenderBot-Joint (Liu et al., 2021)	4.85	1.05	15.25	85.12	5.52	1.29	15.51	85.50
MISC (Tu et al., 2022)	5.43	1.50	15.61	85.19	7.62	2.19	16.40	85.79
KEMI (Deng et al., 2023)	5.44	1.41	15.84	85.24	8.31	2.51	17.05	85.88
Video-LLaMA (Zhang et al., 2023)	5.02	0.74	15.90	84.76	1.82	0.51	11.16	83.73
Gpt-4o mini (Hurst et al., 2024)	5.98	1.20	15.21	85.12	3.58	1.19	12.33	84.93
Qwen2-VL (Wang et al., 2024)	5.61	0.96	17.49	84.80	3.05	1.02	13.54	85.19
Qwen2.5-VL (Bai et al., 2025)	5.93	1.39	15.19	84.86	2.64	0.86	12.31	85.32
Emotion-LLaMA (Cheng et al., 2024)	8.76	4.55	21.81	85.52	5.31	2.84	18.18	85.74
SMES (Chu et al., 2025)	5.13	1.37	15.42	86.80	6.61	2.41	17.19	86.16
Qwen2.5-VL (SFT) (Bai et al., 2025)	8.91	4.02	19.68	86.48	6.02	2.78	20.03	86.00
FIRES (Wang et al., 2025)	10.18	4.54	22.76	86.26	6.78	2.32	18.12	85.91
AG-CTR²	13.93	5.85	23.09	86.87	8.98	3.92	21.47	86.36

Table 1: Comparison of response generation task on the MESC and ESConv datasets.

which is trained with \mathcal{L}_{gen} to generate responses. For the teacher model, we utilize Gemini-3-Pro-Preview (Team et al., 2023) to verify and refine the appraisal-guided reasoning chains. The verification threshold τ is set to 3.5. We adopt LoRA (Hu et al., 2022) for parameter-efficient fine-tuning with a learning rate of $5e-5$ and a batch size of 32; the negative penalization coefficient is set to 0.1. For the retrieval, the number of retrieved therapeutic responses k is set to 3. All experiments are carried out on a server with $4 \times \text{L40S GPUs}$.

4.2 Main Results

Table 1 reports the main results on MESC and ESConv, we compare AG-CTR² with representative baselines and draw the following observations.

Traditional ESC models such as MISC and KEMI outperform vanilla dialogue systems (e.g., BlenderBot) by incorporating task-specific supervision and external knowledge. However, when compared with advanced general MLLMs (e.g., GPT-4o mini and Qwen2.5-VL), their advantages are inconsistent, with performance alternating across different metrics and datasets. This indicates that effective emotional support generation requires not only ESC-specific modeling, but also the generation capabilities provided by modern MLLMs.

Fine-tuned ESC MLLMs, such as Emotion-LLaMA and FIRES, consistently outperform both traditional ESC methods and general-purpose MLLMs. These approaches achieve stronger performance by analyzing users’ emotional cues from multimodal inputs, enabling a more accurate under-

standing of users’ affective states. FIRES further improves performance by incorporating structured thinking mechanisms to guide response generation. These results indicate that integrating fine-grained emotional understanding is critical for high-quality emotional support generation.

AG-CTR² consistently achieves the best performance across all evaluation metrics on both datasets by leveraging appraisal-guided reasoning and CoT-driven retrieval. In contrast to prior methods that model users’ psychological states implicitly, AG-CTR² explicitly incorporates cognitive appraisal theory into both the reasoning and retrieval stages, shifting the paradigm from surface-level semantic matching to cognitive resonance. By modeling appraisal-guided reasoning chains, the model interprets user distress through underlying cognitive evaluations rather than surface-level content and retrieves historically relevant therapeutic responses based on shared appraisal patterns, thereby enabling more meaningful and effective emotional support.

4.3 Ablation Study

To investigate the contribution of each component in AG-CTR², we conduct ablation experiments on the two datasets, as shown in Figure 3 and Table 2.

Impact of Dual-Signal Verification. To evaluate the effectiveness of dual-signal verification, we introduce an ablated variant, AG-CTR² *w/o* Verification, which directly uses generated appraisal reasoning without dual-signal verification during training. As shown in Figure 3, this variant con-

sistently exhibits performance degradation across both datasets. These results indicate that verification based on emotion supervision and teacher quality judgment is essential for filtering hallucinated or low-quality appraisal chains, reducing training noise, and enabling the model to internalize reliable and psychologically grounded appraisal reasoning.

Impact of Appraisal-Guided Reasoning. This reasoning mechanism enables the model to interpret user distress through cognitive evaluations rather than surface-level content. In the variant AG-CTR² w/o Reasoning, retrieval is performed directly using the user query. As shown in Figure 3, this variant results in the largest performance degradation across both datasets. This finding demonstrates that explicit reasoning over users’ cognitive evaluations is essential not only for accurately understanding users’ distress, but also for guiding retrieval toward more relevant and therapeutically useful information, rather than relying on surface-level textual similarity.

Impact of Appraisal-Guided Retrieval. To evaluate the contribution of appraisal-guided retrieval, we introduce an ablated variant, AG-CTR² w/o CoT Retrieval, in which response generation relies on the generated appraisal reasoning without retrieving historical therapeutic responses. As shown in Figure 3, performance drops on both datasets, particularly on ESConv. This indicates that retrieving historically relevant therapeutic responses could provide crucial contextual grounding, enabling the model to incorporate relevant responses, thereby improving the effectiveness of emotional support generation.

Impact of Sparse Dialogue History. To assess the robustness of AG-CTR² under incomplete dialogue context, we compare performance with full dialogue history and with incomplete history containing only the most recent two turns. As shown in Table 2, AG-CTR² remains robust under incomplete history, with only minor performance drops across all metrics. This suggests that the appraisal chain mainly depends on recent multimodal input, rather than relying heavily on long-range dialogue context. In addition, the retrieval module further compensates for limited history by introducing therapeutically relevant examples with similar psychological appraisal patterns.

4.4 Human Evaluation

Following previous ESC studies (Liu et al., 2021; Chu et al., 2025), we conduct a human evaluation

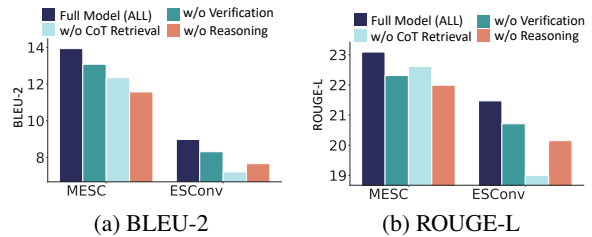


Figure 3: Ablation results on two datasets. Full denotes the AG-CTR² framework, w/o Verification removes the dual-signal verification module, w/o CoT Retrieval disables appraisal-guided retrieval, w/o Reasoning eliminates appraisal reasoning while retaining retrieval.

History Setting	B-2	B-4	R-L	BERT-F1
Full History	13.93	5.85	23.09	86.87
Incomplete	13.85	5.83	22.23	86.81

Table 2: Ablation study on dialogue history.

to compare the responses generated by two models across five dimensions: *Fluency*, *Identification*, *Comforting*, *Suggestion*, *Overall*, and we invite three annotators with master’s degrees, who carefully study the evaluation guidelines and assess each pairwise comparison using a Win/Tie/Lose (details are provided in Appendix F). We randomly sample 100 dialogues from two datasets. Figure 4 reports the human evaluation results on the MESC dataset, while results on ESConv are presented in Appendix D. As shown in Figure 4, AG-CTR² consistently outperforms FIRES across all five evaluation dimensions, with particularly strong gains in Overall (65% Win), Identification (62%), and Fluency (70%). The improvement in *Identification* indicates that appraisal-guided reasoning enables accurate recognition of users’ underlying concerns, while the superior performance in *Suggestion* highlights the advantage of reasoning-aligned retrieval over surface-level semantic matching. Overall, these results demonstrate that appraisal-guided reasoning improves the quality of supportive responses and promotes a better understanding of users’ concerns, leading to more contextually appropriate and helpful guidance.

4.5 Effect of Reasoning Quality

We analyze the effect of different reasoning learning mechanisms on the generated appraisal reasoning and its impact on response generation. As shown in Table 3, we evaluate reasoning quality using both automatic and human evaluations, and

Methods	Automatic Evaluation (Agent)				Human Evaluation				Impact of Response		
	E-Cons.	C-Acc.	L-Coh.	R-Use.	E-Cons.	C-Acc.	L-Coh.	R-Use.	B-2	R-L	BERT-F1
Base Reasoning	2.72	3.13	3.47	3.17	2.15	2.95	3.16	3.03	12.28	21.11	85.21
+ Dual-Signal Verification	3.36	3.54	3.62	3.42	3.20	3.90	3.55	3.25	13.08	22.31	86.36
+ Negative Preference Optimization	3.80	3.67	3.68	3.50	3.45	4.05	4.10	3.65	13.93	23.09	86.87

Table 3: Reasoning quality evaluation and its impact on response generation. Automatic evaluation is conducted by an agent, while human evaluation assesses the perceived quality of reasoning. And reports the impact on response generation performance.

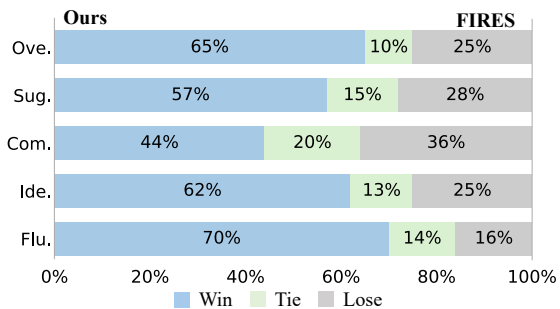


Figure 4: Human evaluation. Inter-annotator agreement measured by Fleiss’s kappa yields κ values of 0.52 (Overall), 0.43 (Suggestion), 0.51 (Comforting), 0.42 (Identification), and 0.45 (Fluency).

further examine how reasoning quality influences response generation performance. The generated appraisal reasoning chains are evaluated on a five-point scale (1 = poor, 5 = excellent) across four dimensions: *Emotion Consistency* (E-Cons.), *Concern Inference Accuracy* (C-Acc.), *Logical Coherence* (L-Coh.), and *Retrieval Usefulness* (R-Use.). For automatic evaluation, reasoning quality is scored by GPT-4o mini. For human evaluation, we invite three annotators to independently assess the generated appraisal reasoning chains.

As shown in Table 3, compared with base reasoning, incorporating dual-signal verification consistently improves reasoning quality across all four dimensions under both automatic and human evaluations. This indicates that emotion supervision and teacher-based quality judgment help filter noisy or weakly aligned appraisal chains, enabling more consistent and appraisal-aligned intermediate reasoning, which in turn improves response generation performance. Building upon this, negative preference optimization further refines appraisal reasoning, demonstrating that discouraging implausible or therapeutically irrelevant reasoning chains helps the model learn more structured and practically useful appraisal reasoning patterns, which in turn consistently enhance response generation.

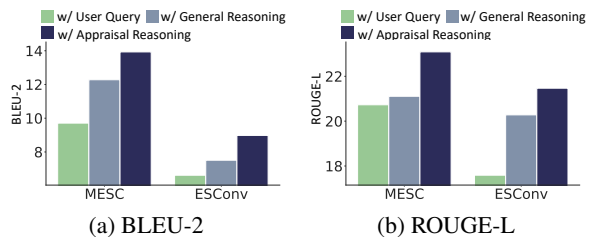


Figure 5: Using different retrieval queries on the impact of response generation on MESC and ESConv.

Retrieval Query	Cons.	Info.	R-Use.	κ
User Utterance	2.8	2.6	3.2	0.43
General Reasoning	3.1	3.3	3.7	0.55
Appraisal Reasoning	3.4	3.5	4.1	0.45

Table 4: Human evaluation of different retrieval queries. Here, κ denotes the inter-annotator agreement measured using Fleiss’s kappa (Fleiss, 1971).

4.6 Effect of Different Retrieval Queries

To examine how different retrieval queries affect retrieval quality and downstream response generation, we evaluate three retrieval methods: using the raw user utterance, general emotion reasoning, and appraisal-guided reasoning as retrieval queries. Table 4 reports human evaluation results on *Consistency*, *Informativeness*, and *Retrieval Usefulness* (details see Appendix G), while Figure 5 illustrates their impact on response generation quality. Using the raw user query yields the weakest performance, indicating that surface-level semantic matching often retrieves therapeutically misaligned responses. Replacing it with general reasoning improves all metrics, suggesting that incorporating reasoning helps capture users’ emotional cues. Notably, appraisal-guided reasoning achieves the best performance across all evaluation dimensions and consistently leads to higher-quality generated responses. This demonstrates that modeling users’ cognitive appraisals enables retrieval based on shared psychological situations

rather than superficial similarity, resulting in more therapeutically relevant responses.

4.7 Case Study

We present a representative case study in Appendix E to illustrate the effectiveness of AG-CTR². While baseline methods generate generic empathetic responses that overlook the user’s latent psychological trigger, AG-CTR² captures the underlying cognitive appraisal, retrieves more relevant responses, and generates responses that directly address the implicit cause, offering more meaningful supportive responses.

5 Conclusion

In this paper, we address the cognitive gap in ESC systems by proposing AG-CTR², an appraisal-guided Chain-of-Thought reasoning and retrieval framework that moves beyond surface-level matching toward cognitively grounded understanding. By integrating Cognitive Appraisal Theory into both reasoning and retrieval, AG-CTR² generates appraisal-based reasoning chains and uses them as structured queries to retrieve therapeutically relevant responses based on shared psychological evaluations. Experiments and human evaluations on two ESC benchmarks show that AG-CTR² consistently outperforms strong baselines.

Limitations

Our framework demonstrates promising performance in enhancing the quality of emotional support; however, some limitations remain. The current evaluation does not include longitudinal analysis, and the long-term effectiveness of the proposed framework has yet to be empirically validated. Future work could incorporate extended timeframes or follow-up interactions to assess sustained impacts on users’ emotional well-being and to guide further refinements. Additionally, although our model achieves strong performance by leveraging appraisal-guided reasoning, it primarily relies on conventional multimodal inputs. Incorporating richer multimodal signals, such as eye-tracking data or electroencephalogram (EEG) signals, may enable more fine-grained psychological analysis and facilitate more nuanced modeling of users’ evolving emotional states.

Ethical Considerations

The datasets used in this work are publicly available and widely studied benchmarks, collected and annotated by professionals or well-trained annotators. All personally identifiable and sensitive information has been removed to protect user privacy. Our method does not provide medical diagnoses or treatment recommendations. Instead, it is designed to offer supportive and empathetic responses. Compared with existing emotional support conversation systems, our approach may contribute to safer and more reliable ESC systems by grounding responses in cognitively aligned reasoning.

Acknowledgements

This research was supported by the National Science Foundation of China (92467302). This work was also supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (24-SIS-SMU-083), the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP)(I2301E0026), the Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), and the National Research Foundation Fellowship (NRFF13-2021-0006), Singapore. We are sincerely grateful to all reviewers for their insightful feedback.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- 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)*.
- Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- 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, Toronto, Canada, July 9-14, 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15*.
- Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE TMM*.
- 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4079–4095.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Susan Folkman, Richard S Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J Gruen. 1986. Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes. *Journal of personality and social psychology*, 50(5):992.
- 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*.
- Edward J Hu, Yelong Shen, 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. *ICLR*, 1(2):3.
- Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. Synnergizing llms with global label propagation for multimodal fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1426–1440. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- Richard S Lazarus. 1991. *Emotion and adaptation*. Oxford University Press.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Jinggui Liang, Dung Vo, and Lizi Liao. 2025. Intentionframe: A semi-structured, multi-aspect framework for fine-grained conversational intention understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 28108–28125. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 3001–3004. ACM.
- 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483.
- 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.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL 2021*, pages 300–325.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan

- Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- 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 *ACL 2022, Dublin, Ireland, May 22-27*, pages 308–319.
- Fanfan Wang, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2025. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- 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.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP 2023 - System Demonstrations, Singapore, December 6-10*, pages 543–553.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 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*. OpenReview.net.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Ee-Peng Lim, Minlie Huang, and Liqiang Nie. 2024a. Thoughts to target: Enhance planning for target-driven conversation. In *EMNLP 2024, Miami, FL, USA, November 12-16*, pages 21108–21124.
- Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024b. 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)*.
- 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)*.
- Xiaofeng Zhou, Heyan Huang, and Lizi Liao. 2025. Debate, reflect, and distill: Multi-agent feedback with tree-structured preference optimization for efficient language model enhancement. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 9122–9137. Association for Computational Linguistics.

Appendix

A Details of Reasoning Verification

This is the detail of the psychological metrics.

- **Emotion-Consistency (s_1):** Evaluates the logical alignment between the appraisal steps and the final emotion label.
- **Concern Inference Accuracy (s_2):** Measures the precision in identifying the help-seeker’s underlying stressors.
- **Logical Coherence (s_3):** Assesses the structural integrity of the causal chain from event perception to coping.
- **Utility for Retrieval (s_4):** Determines the potential of the reasoning chain to serve as an effective situational query for RAG.

The prompt used for the teacher model verification reasoning chain is illustrated in Figure 6, and the prompt used for the teacher model to refine reasoning chains is illustrated in Figure 7.

B Details of Appraisal Reasoning

The prompt used to generate appraisal reasoning is illustrated in Figure 8.

C Details of Baseline Methods

We compare our proposed methods against a diverse set of competitive baselines, categorized into specialized ESC models and multimodal large language models:

- **BlenderBot (Roller et al., 2021):** A Transformer-based open-domain dialogue model pre-trained on diverse conversational skills, including empathy and personification.

Prompt for Teacher Model Verification

Instruction: You are a professional therapist acting as an expert evaluator. You should assess the quality of a given CoT reasoning for emotional support systems, specifically whether it is psychologically grounded, logically coherent, and useful for RAG-oriented retrieval. Score the CoT reasoning on four dimensions using a 1–5 scale (1=Poor, 2=Weak, 3=Fair, 4=Good, 5=Excellent):

1. Emotion Consistency: Does the reasoning logically and psychologically support the target emotion {emotion}?
2. Concern Inference Accuracy: Does it accurately infer the user’s underlying concerns/needs/stressors?
3. Logical Coherence: Is the reasoning clear, structured, and internally consistent?
4. RAG Usefulness: Is it sufficiently structured and informative to guide downstream retrieval?

Examples:

Emotion Consistency : 3
Concern Inference Accuracy : 4
Logical Coherence :3
RAG Usefulness : 4

Figure 6: Prompt used for teacher model verification.

- **BlenderBot-Joint** (Liu et al., 2021): An adaptation of BlenderBot that utilizes a multi-task learning framework to jointly predict support strategies and generate responses.
- **MISC** (Tu et al., 2022): A text-based framework that integrates commonsense knowledge from COMET to better understand user states and apply mixed support strategies.
- **KEMI** (Deng et al., 2023): A model that enhances emotional support by integrating both commonsense knowledge and mental health knowledge retrieved from the graph into the dialogue generation process.
- **Video-LLaMA** (Zhang et al., 2023): A multi-instructionally tuned model that enables LLMs to perceive and understand video content by connecting visual and audio encoders to a language backbone.
- **Gpt-4o mini** (Hurst et al., 2024): A small state-of-the-art, natively multimodal large language model designed to seamlessly process and generate across text and visual modalities in real-time.
- **Qwen2-VL** (Wang et al., 2024): Leading

Prompt for Appraisal Reasoning Refinement

Instruction: You are acting as an expert psychological analyst specializing in multimodal emotion support. Your objective is to perform a deep-layer analysis of the user's emotional state using Cognitive Appraisal Theory. Using Cognitive Appraisal Theory, synthesize the verbal dialogue and visual/behavioral cues from the video to refine the "Draft Reasoning." You must provide a psychologically grounded explanation that bridges the gap between the environmental stimulus and the resulting emotional state.

Inputs

Video Interaction: {video}
Dialogue History: {chat_history}
Current Utterance: {user_input}
Draft Reasoning (For Refinement):
{cot_reasoning}
Emotion: {emotion}

Examples:

Reasoning: {reasoning}
Emotion: {emotion}

Figure 7: Prompt used for teacher model refinement.

multimodal models that integrate strong vision–language understanding and support video inputs of varying lengths and resolutions.

- **Qwen2.5-VL** (Bai et al., 2025): An improved version of Qwen2-VL with advanced vision–language understanding capabilities, capable of processing videos of varying lengths and resolutions.
- **SMES** (Chu et al., 2025): A specialized multimodal ESC model that utilizes cross-modal emotional cues and is optimized via four auxiliary emotion-recognition tasks.
- **Emotion-LLaMA** (Cheng et al., 2024): An emotion-aligned MLLM specifically fine-tuned to capture nuanced affective states and provide emotionally resonant responses in complex social scenarios.
- **FIRES** (Wang et al., 2025): A multimodal model that leverages Group Relative Policy Optimization (GRPO) to align model responses with human-centric emotional support.

Prompt for Appraisal Reasoning Generation

Instruction: You are a professional therapist tasked with analyzing the user’s current emotional state by reasoning about the underlying causes of distress. Given the {dialogue history}, {current user utterance}, and {video}, integrate verbal content with visual cues (e.g., facial expressions, gaze, posture, and body movements) to support your analysis. Your reasoning must follow Cognitive Appraisal Theory and should:

1. Identify the psychologically relevant event or situation the user is experiencing based on the dialogue and video;
2. Explain why this situation is appraised as negative or distressing, integrating verbal and visual evidence;
3. Assess the user’s perceived coping ability (able vs. unable) and show how this appraisal leads to the inferred emotional state.
4. Based strictly on the above reasoning, predict the user’s emotion label chosen from the following set:
["anger", "sadness", "disgust", "depression", "neutral", "joy", "fear"].

Examples:

Dialogue History: {history}

User Input: {text}

Reasoning: {reasoning}

Emotion: {emotion}

Figure 8: Prompt used for generating appraisal reasoning.

D Human Evaluation Results on the ESConv Dataset

The human evaluation results on ESConv are shown in Figure 9.

E Case Study

Case study for the generating reasoning chain and response generation, as shown in Figure 10.

F Details of Human Evaluation

The metric of the human evaluation:

- *Fluency:* Which model produces more fluent responses?
- *Identification:* Which model is more adept at identifying the user’s problem?
- *Comforting:* Which model better comforts the user?

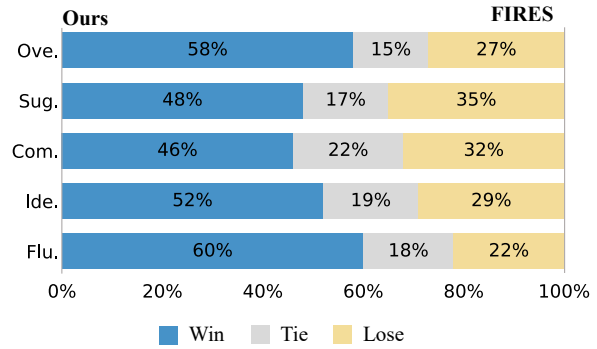


Figure 9: Human evaluation results. We assess inter-annotator agreement using Fleiss’s kappa (Fleiss, 1971), obtaining κ values of 0.45 (Overall), 0.51 (Suggestion), 0.46 (Comforting), 0.43 (Identification), 0.52 (Fluency).

- *Suggestion:* Which model provides more helpful and informative suggestions?
- *Overall:* Which model generates generally better responses?

We provide detailed guidelines for human evaluation of appraisal reasoning quality in Table 5 and response generation in Table 6. Annotators are asked to assess the quality of the generated appraisal-grounded reasoning by the trained MLLM model.

G Details of Retrieval Queries

To assess the quality of retrieved results and their suitability for supporting response generation, we conduct a human evaluation along three dimensions that reflect both semantic alignment and therapeutic relevance:

- *Consistency:* Whether the retrieved content is logically consistent with the user’s expressed situation and underlying emotional context.
- *Informativeness:* Whether the retrieved results provide sufficient and meaningful information to support the generation of an appropriate emotional support response.
- *Retrieval Usefulness:* Whether the retrieved results are helpful and actionable for guiding the downstream response generation process.

H LLM Usage

Large Language Models (LLMs) were utilized in this work solely as auxiliary tools for linguistic refinement. Their function was restricted to enhancing the grammar, clarity, and stylistic consistency of text that had been originally drafted by

Guidelines for Human Evaluation of Appraisal Reasoning	
You are asked to evaluate the quality of appraisal-based reasoning chains generated by a model for emotional support conversations. Each reasoning chain should be assessed independently based on how well it explains the user’s emotional state and underlying concerns, and how useful it would be for guiding retrieval in a Retrieval-Augmented Generation (RAG) system. Please score each dimension on a 1–5 Likert scale , following the definitions below.	
Evaluation Scale 1 = Poor, 2 = Weak, 3 = Fair, 4 = Good, 5 = Excellent	
(1) Emotion Consistency	
Definition	Evaluate whether the reasoning logically and psychologically supports the target emotion label, and whether the inferred emotion is consistent with the user’s dialogue and context.
Example	High (5): Explains the user’s anxiety through appraisal of the cognitive process. Low (1): The reasoning contradicts the mentioned emotions unsupported by the dialogue.
(2) Concern Inference Accuracy	
Definition	Evaluate how accurately the reasoning identifies the user’s underlying concerns, needs, stressors, or cognitive evaluations driving the emotion.
Example	High (5): Correctly infers fear of abandonment underlying the distress. Low (1): The reasoning misidentifies or ignores the user’s core concern.
(3) Logical Coherence	
Definition	Evaluate the clarity, structure, and internal consistency of the reasoning, including whether it follows a logical progression without contradictions.
Example	High (5): The reasoning follows a clear progression from situation to appraisal to emotional outcome. Low (1): The reasoning is fragmented, repetitive, or logically inconsistent.
(4) Usefulness for Retrieval (RAG-Oriented)	
Definition	Evaluate whether the reasoning is structured and informative enough to guide retrieval of psychologically relevant historical cases or therapeutic strategies in a RAG system.
Example	High (5): Explicitly captures appraisal patterns to guide retrieval. Low (1): The reasoning is too vague or generic to support meaningful retrieval.

Table 5: Guidelines for Human Evaluation of Appraisal-Based Reasoning Quality

the authors. At no stage did LLMs contribute to research ideation, methodological design, data collection, analysis, or interpretation of results. All intellectual contributions, scientific content, and conclusions presented in this paper are entirely attributable to the authors. The authors accept full responsibility for the accuracy, originality, and integrity of the submission, including sections of text that may have been refined with the assistance of LLMs.

Guideline for Human Evaluation	
<p>You need to evaluate the performance of two different models in the emotional support conversation task, specifically assessing whether they effectively help users alleviate emotional problems. For each model, you will receive generated responses. Your task is to determine which model performs better across five specific aspects. Please refer to the definitions below to understand the aspects we need to focus on. The provided examples will demonstrate how to assess each metric.</p>	
Results	1 (First method wins), 0 (Equal), -1 (First method loses)
(1) Fluency	
Definition	Which model produces more fluent and grammatically correct responses without awkward phrasing or errors?
Example	<p>High Fluency: "I'm sure you're doing your best, and it's okay to seek support when needed. How can I assist you today?"</p> <p>Low Fluency: "I'm not sure... I'm just... I'm... I'm..."</p>
(2) Identification	
Definition	Which model is more adept at accurately identifying and understanding the user's emotional problem or concern?
Example	<p>High Identification: "It sounds like you're feeling overwhelmed with work responsibilities. Let's explore ways to manage your stress."</p> <p>Low Identification: "I'm sorry. i'm not sure what you mean."</p>
(3) Comforting	
Definition	Which model better provides comfort and reassurance to the user in a compassionate and empathetic manner?
Example	<p>High Comforting: "I'm really sorry you're going through this. Tell me a little bit more about that, about that girl."</p> <p>Low Comforting: "I understand you're upset. Let's move on to another topic."</p>
(4) Suggestion	
Definition	Which model provides more helpful, relevant, and actionable suggestions or advice to the user?
Example	<p>High Suggestion: "Do you really accept that? That life is better lived unexamined."</p> <p>Low Suggestion: "Maybe you should try to relax more."</p>
(5) Overall	
Definition	Which model generally generates better, more cohesive, and supportive responses that effectively address the user's needs?
Example	<p>High Overall: "I'm really sorry you're feeling this way. Can I ask you if anything else happened in those sessions that might have? "</p> <p>Low Overall: "I'm here if you need to talk."</p>

Table 6: Guidelines for Human Evaluation of Emotional Support Conversations


<p>Case Id: Sophie, Problem: The relationship with friends and family, Situation: The user was eating pizza but stopped after hearing the word “appetite.”</p>

Dialogue history
<p>User: (Anger) I'm stressed about the exam, I think I have a cavity, and I have a stomachache from that fucking pizza. What a mistake that was. Therapist: I have a theory about that pizza. User: (Depression) Surprise, surprise. Therapist: Would you like to hear it? User: (Depression) What's your theory?</p>
Generated Reasoning Chain
<p>Emotional Reasoning: The user appears to be sitting on a couch with a somewhat concerned facial expression, suggesting they might be reflecting on something troubling. These issues indicate a sense of worry and anxiety. The user's tone suggests they are seeking understanding or validation, which could imply a need for emotional support. Appraisal Reasoning: The user is discussing their concerns about an exam, a potential cavity, and a stomachache from pizza, indicating a range of physical and academic stressors. The user's tone is matter-of-fact and slightly dismissive, suggesting they may feel overwhelmed but are trying to manage their emotions. The visual cues show a calm demeanor, but the underlying dialogue reveals a complex emotional state with multiple sources of distress.</p>
Retrieval Results
<p>Raw User Query: Your experience matters. Appraisal Reasoning Query: You make it sound like you don't think that's going to happen.</p>
Generated Response
<p>FIRES: I want you to know that I'm not attacking you or anything like that. I'm just trying to understand more about yourself. AG-CTR²: It's okay if you're not sure about your theory. When I used the word gusto, it may have echoed something your father's girlfriend used to say to you, and that association quietly shifted how you felt in the moment. Ground-truth: I think that you were enjoying that pizza. Until I said something that reminded you of something your father's girlfriend used to say to you. I used the word gusto.</p>

Figure 10: Case Study