Causal-ESC: Reliable Policy Learning for Emotional Support Conversation via Causal Inference

Xv Wang[1] **Zhenyu Wang**[1*] **Guanyu Zheng**[1] **Rui Zhang**[2]
[1]South China University of Technology
[2]Huya Inc.
sexvwang@mail.scut.edu.cn, wangzy@scut.edu.cn, sezgyhtt@mail.scut.edu.cn,
zhang1rui4@outlook.com

Abstract

While Large Language Models (LLMs) have significantly advanced the fluency of Emotional Support Conversation (ESC) systems, current research predominantly focuses on engineering increasingly complex architectures—from intricate reasoning chains to multi-agent collaborations. While these advancements (e.g., CoT) offer semantic traces of reasoning, they remain mechanistically opaque, obscuring the fundamental causal mechanisms between dialogue features and effective empathic strategies, leading to poor interpretability and susceptibility to distribution shifts in offline learning. To address these limitations, we propose a novel framework **Causal-ESC**. Departing from conventional paradigms that directly utilize raw dialogue history as input, our approach introduces Doubly Robust (DR) learning to explicitly model the causal effect of utterance features on strategy selection, effectively mitigating the biases and counterfactual unobservability inherent in offline datasets. We further integrate an LLM-based stylized rewriting mechanism to translate these rigorously learned causal strategies into natural, context-consistent responses. Comprehensive experiments, supported by statistical verification (e.g., Outcome R^2) and human-like evaluation, demonstrate that our framework not only significantly outperforms state-of-the-art baselines in empathy and helpfulness but also provides a theoretically grounded, interpretable solution to the mechanistic interpretability dilemma in affective computing.

1 Introduction

Empathy, defined as the ability to understand and share the feelings of others, serves as the cornerstone of effective human communication (Decety and Jackson, 2004; Cohen and Strayer, 1996; Yuan et al., 2024). In the realm of Human-Computer Interaction (HCI) (Paiva et al., 2017; Zhang et al., 2017, 2021) Emotional Support Conversation (ESC) represents a more advanced and challenging task. Pioneered by the release of the ESConv dataset (Liu et al., 2021), ESC aims not only to express empathy but also to reduce the user's emotional distress through specific counseling strategies (e.g., questioning, reflection, and suggestion) within a multi-turn interaction.

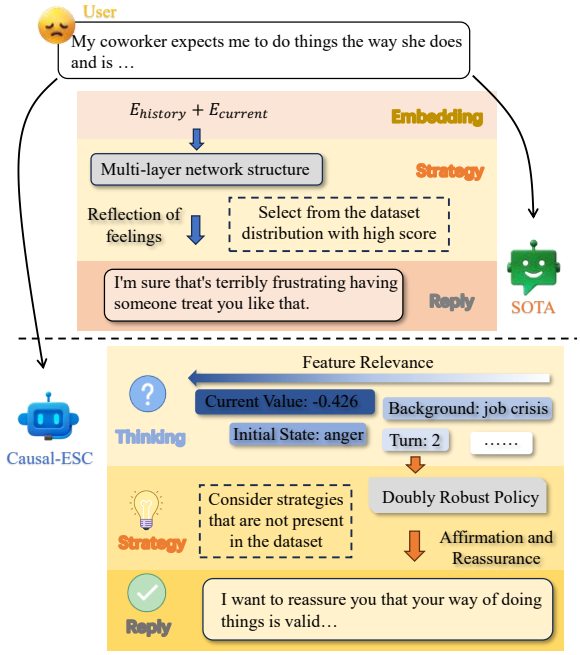

Figure 1: Causal-ESC learns strategies by analyzing the causal relationship between strategies and specific features, and generates corresponding responses.

Early research in ESC primarily relied on pipeline approaches (Liu et al., 2021) or joint learning frameworks (Tu et al., 2022; Peng et al., 2022) based on Pre-trained Language Models (PLMs) like BART (Lewis et al., 2020) or BlenderBot (Roller et al., 2021). These methods typically treated strategy prediction and response generation as auxiliary tasks. Recently, the paradigm has shifted towards utilizing Large Language Models (LLMs) for ESC (Deng et al., 2023; Zheng

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19386–19401

et al., 2022; Hua et al., 2025), leveraging their emergent reasoning and instruction-following capabilities. Researchers have explored Chain-of-Thought (CoT) (Wei et al., 2022) prompting to explicitly guide LLMs through the cognitive process of psychological counseling.

Nevertheless, a critical methodological limitation persists: current approaches are predominantly oriented towards engineering increasingly complex architectures that, while sometimes offering semantic reasoning traces, remain mechanistically opaque, rather than attempting to deconstruct or minimize the causal ambiguity of the generation process. Instead of isolating the causal drivers of empathy to simplify model inference, recent works continue to stack intricate reasoning modules and multi-agent interactions to force performance gains. This trajectory exacerbates the disconnection between input features and the final empathic outcome, leaving the fundamental question of why a specific strategy yields a superior result unanswered. Such a lack of interpretability not only hinders effective error analysis but also raises serious concerns regarding system reliability in sensitive mental health contexts. Furthermore, the practical deployment of these systems faces significant hurdles: researchers are often caught in a dilemma between the susceptibility of offline learning to severe overfitting (Fujimoto et al., 2019; Levine et al., 2020) and the prohibitive costs associated with online training—particularly when employing RLHF (Ouyang et al., 2022), while the inherent subjectivity of the empathy task makes it intrinsically challenging to define a universal "gold standard" for optimization (Yang et al., 2024; Liu et al., 2016).

To address these issues, we propose a novel framework Causal-ESC. Unlike previous methods that treat strategy selection as an opaque mapping task driven by surface-level correlations, our work introduces Doubly Robust (DR) learning (Bang and Robins, 2005) to explicitly model the causal mechanism between dialogue utterance features and the deployment of emotional strategies. As shown in Figure 1, this approach enhances the interpretability of the system by isolating the true causal drivers of empathy from spurious correlations. Furthermore, to bridge the gap between abstract strategy formulation and natural language generation, we employ an LLM-based stylized rewriting mechanism. This module generates responses that are strictly aligned with the identified causal strategies while maintaining high content consistency. To rigorously validate the credibility of our learned strategies, we pose a fundamental research question:

RQ) *Are these features genuinely causally related to empathy strategies, and what is the strength of this relationship?*

We answer this through comprehensive statistical analyses—including Outcome R^2 and propensity score distribution examinations—thereby verifying the validity of the causal links.

In summary, the contributions of our work are as follows:

(1) We pioneer the application of causal analysis to empathic policy learning, effectively reducing the mechanistic opacity of the model and enhancing causal interpretability. By statistically quantifying the impact of specific features on strategy selection, we provide a mathematical and theoretical foundation for the mechanics of empathic dialogue.

(2) We propose a pipeline framework combining Doubly Robust learning with LLM-based stylized generation, which effectively mitigates common challenges in offline policy learning, such as distribution shift and unobservable counterfactuals. Additionally, the two-stage stylized generation framework can alleviate part of the subjectivity in generative process.

(3) Extensive experiments demonstrate that our causality-grounded policy model is statistically more reliable and achieves significant performance improvements over both traditional and SOTA LLM-based empathic dialogue methods. Human evaluation results further confirm that our method generates more appropriate and supportive responses compared to existing baselines.

2 Related Work

Existing research on ESC can be broadly categorized into two paradigms based on how they handle counseling strategies: Explicit Policy Learning, which treats strategy selection as a distinct classification task prior to generation, and Implicit Policy Learning, which models empathy and strategy latently within an end-to-end generation process.

2.1 Explicit Policy Learning Methods

Explicit methods operate on a pipeline where the model first predicts a specific support strategy (e.g., Question, Reflection) to guide response generation. Traditional approaches like the ESConv baseline (Liu et al., 2021) and graph-based models such

as MISC (Tu et al., 2022) and KEMP (Li et al., 2022) utilized commonsense knowledge and auxiliary classification tasks to predict strategy labels, though they often suffered from error propagation. In the era of Large Language Models (LLMs), this paradigm has evolved into "Planning-based" generation using Chain-of-Thought (CoT). Recent works like MultiESC (Cheng et al., 2022) and Cognitive Prompting (Majumder et al., 2022) leverage the instruction-following capabilities of LLMs to explicitly generate a "thought" process or plan the strategy before producing the final utterance, thereby making the decision-making step more interpretable.

2.2 Implicit Policy Learning Methods

Implicit methods aim to internalize the logic of emotional support without relying on hard strategy labels during inference. Early neural models such as MIME (Majumder et al., 2020) and GLHG (Peng et al., 2022) focused on learning latent representations of emotion and hierarchical dialogue structures to naturally weave empathy into responses. With the shift to LLMs, implicit learning has manifested through data augmentation and multi-agent collaboration. For instance, AugESC (Zheng et al., 2022) fine-tunes models on LLM-augmented data to learn support patterns implicitly, while Chatcounselor (Liu et al., 2023) has fine-tuned its prompts based on the actual conversations between clients and professional psychologists. These approaches rely on the emergent reasoning of LLMs to handle empathetic nuances as a latent byproduct of high-quality generation, rather than through rigid classification.

3 Problem Formulation

The empathetic dialogue task is defined as a conditional text generation problem within a specific scenario. It aims not only to understand the semantic content of the conversation but also to accurately perceive the user's emotional state and generate responses that are emotionally appropriate (e.g., comforting, encouraging, empathetic) and contextually coherent (Rashkin et al., 2019; Zhou et al., 2018). Formally, a dialogue session consists of T turns. In the t-th turn, the Speaker (user) initiates or continues the conversation describing their situation, and the Listener (system) responds. Let $C_t = \{u_1, r_1, u_2, r_2, ..., u_t\}$ denote the dialogue context history up to the current turn, where u represents the user's utterance and r represents the system's response. The goal of the task is to learn a generation probability distribution $p(r|C_t)$. To enhance interpretability and emotional control, we explicitly decompose the traditional end-to-end generation task into two stages: Policy Decision and Response Generation. The generation probability is formulated as:

$$p(r|C_t) = \sum_{d \in \mathcal{D}} p(r|d, C_t) \cdot p(d|C_t) \quad (1)$$

where d represents a latent action selected from a predefined space of empathetic strategies D.

4 Method

Our method consists of two sub-tasks: Empathetic Policy Learning and Strategy-styled Response Generation. First, the Policy Learning Module leverages Doubly Robust (DR) learning to infer the optimal empathetic strategy d^* from offline logs based on the dialogue state. Second, the Response Generation Module generates a response r conditioned on both the selected strategy d^* and the context C_t. Finally, the two modules are combined to complete the empathetic dialogue generation. Figure 2 illustrates the overall workflow of our proposed method. This section focuses on the DR-based policy learning framework.

4.1 Empathetic Policy Learning

We employ a Doubly Robust learning approach to train the policy network offline. The core objective of this sub-task is to learn a Target Policy $\pi_\theta(d|x)$ that maximizes the Expected Cumulative Sentiment Gain (Y) across the user population (Dudík et al., 2011). The objective function is defined as:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{x \sim P(x), d \sim \pi_\theta(\cdot|x)}[r(x, d)] \quad (2)$$

where x represents the dialogue context features, d is the empathetic strategy adopted, and $r(x, d)$ denotes the observed reward generated by the strategy.

4.1.1 Feature Extraction and State Representation

Traditional empathetic dialogue models typically encode raw dialogue text directly via Transformer encoders as input embeddings. However, since the noise processing within attention mechanisms is implicit and difficult to control, using raw embeddings alone can lead to unstable policy learning.

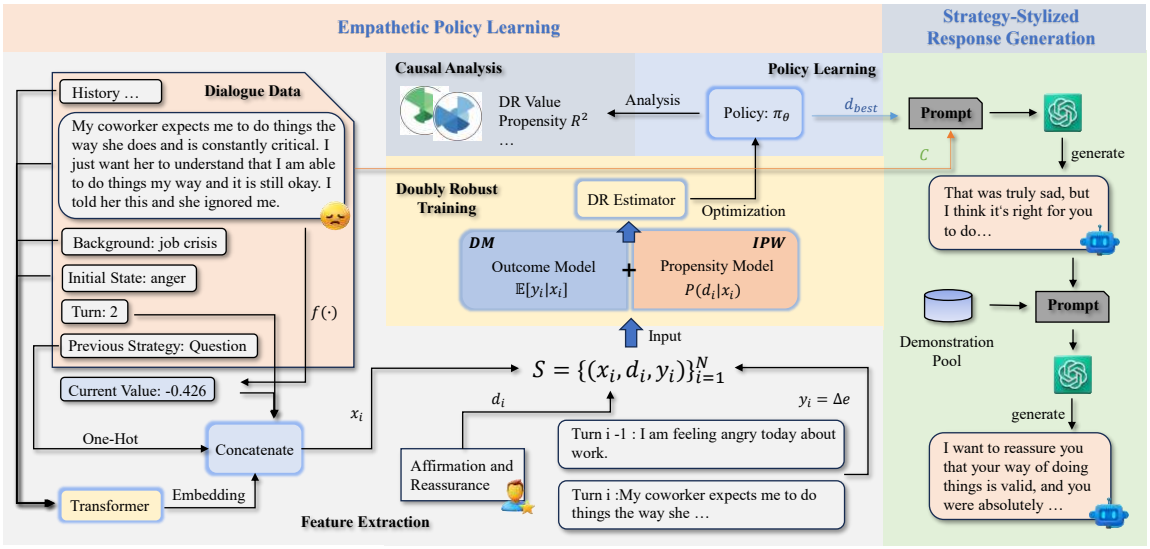

Figure 2: The overall framework of Causal-ESC. The left part is Empathetic Policy Learning, which generates the optimal strategy d_{best}. The right part is Strategy-Stylized Response Generation, which generates the final response based on d_{best}.

Following (Sharma et al., 2020), we explicitly extract features strongly correlated with empathy to construct a robust state vector x. The state x is constructed by concatenating the historical semantic embedding with five explicit feature components:

$$x = [h_{context} \oplus h_{desc} \oplus h_{init} \oplus h_{turn} \oplus h_{cuur} \oplus h_{prev}] \quad (3)$$

The components are defined as follows: **Context Embedding** ($h_{context}$): The semantic vector of the dialogue history encoded by a pre-trained language model (e.g., Transformer). **Problem Description Embedding** (h_{desc}): Represents the background and topic of the conversation (e.g., "job crisis", "sleep problems"). **Initial Emotion State** (h_{init}): The user's emotion vector at the beginning of the session, serving as a baseline for emotional change. **Turn Information** (h_{turn}): The normalized dialogue turn index t/T_{max}, indicating the stage of the conversation (e.g., greeting strategies are frequent in early stages, while suggestions appear later). **Current Sentiment** (h_{cuur}): A vector quantifying the sentiment intensity of the current utterance, extracted using a sentiment analysis tool (Zhou et al., 2023). **Previous Strategy** (h_{prev}): The one-hot encoding of the system's strategy d_{t-1} from the previous turn, used to maintain policy continuity.

4.1.2 Action Space and Reward Function

Empathetic Strategy Space ($\mathcal{D}$): Following the taxonomy in the ESConv dataset (Liu et al., 2021), we categorize the listener's strategies into $K = 8$ distinct types, denoted as $\mathcal{D} = \{d_1, d_2, ..., d_K\}$ (e.g., Affirmation, Reflection; see Appendix for details). Observed Reward (y): Based on the theory of emotional regulation in psychology (Gross, 1998), we quantify the effectiveness of a strategy by calculating the change in user sentiment before and after the response. Let $f(\cdot)$ be a sentiment scoring function, which is implemented by the scorer according to (Hartmann, 2022). the reward is defined as:

$$y = \Delta e = f(u_{t+1}) - f(u_t) \quad (4)$$

The rationale behind the setting of y is detailed in the Experiments section.

4.1.3 Doubly Robust Policy Optimization

We learn from an observational dataset $\mathcal{S} = \{(x_i, d_i, y_i)\}_{i=1}^N$, where actions d_i were generated by a Behavior Policy (i.e., human counselors), denoted as $\pi_b(d|x)$. This introduces selection bias. To address this while maintaining low variance, we adopt the Doubly Robust (DR) estimator to evaluate and optimize the new policy π_θ.

Direct Method (DM) Trains a regression model (Reward Model) $\hat{Q}_\phi(x, d)$ to fit the observed rewards, predicting the expected sentiment change for a given state-action pair:

$$\hat{V}_{DM}(\pi_\theta) = \mathbb{E}_{\mathcal{S}}\left[\sum_{d \in \mathcal{D}} \pi_\theta(d \mid x)\hat{Q}_\phi(x, d)\right] \quad (5)$$

Inverse Propensity Weighting (IPW) Trains a classification model to estimate the propensity

score $\hat{\pi}_b(d|x)$ of the behavior policy and weights the observed data to simulate a randomized experiment.

$$\hat{V}_{IPW}(\pi_\theta) = \mathbb{E}_{\mathcal{S}}\left[\frac{\pi_\theta(d_i \mid x_i)}{\hat{\pi}_b(d_i \mid x_i)} y_i\right] \tag{6}$$

The Doubly Robust (DR) Estimator uses the regression model $\hat{Q}$ as a baseline to reduce the variance of IPW, while using the IPW term to correct the bias of the regression model. For each sample (x, d, y), the estimated DR reward vector for the target policy π_θ across the full action space is defined as:

$$\hat{r}_{DR}(x, d=k) = \hat{Q}_\phi(x,k) + \frac{\mathbb{I}(d_{obs}=k)}{\hat{\pi}_b(k \mid x)} \left(y_{obs} - \hat{Q}_\phi(x,k)\right) \tag{7}$$

Where $\mathbb{I}(d_{obs}=k)$ is an indicator function, meaning that if the actual action d_i in the historical data exactly matches the current action, the value will be 1. To efficiently learn the empathetic policy, we first pre-train the propensity network π_b and the reward network $\hat{Q}_\phi$ via supervised learning. During the policy learning phase, we freeze these two networks and optimize the policy network π_θ by minimizing the negative expected DR reward:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \pi_\theta(k \mid x_i)\left(\hat{Q}_\phi(x_i,k) + \frac{\mathbb{I}(d_i=k)}{\hat{\pi}_b(k \mid x_i)}\left(y_i - \hat{Q}_\phi(x_i,k)\right)\right) \tag{8}$$

To improve numerical stability, following (Swaminathan and Joachims, 2015), we apply Propensity Clipping. We set a threshold $\delta > 0$ and use max $(\hat{\pi}_b(d|x), \delta)$ in the denominator to prevent gradient explosion caused by extremely small probabilities. (See Appendix for the detailed algorithm).

4.1.4 Causal Analysis

This section introduces several metrics for causal analysis to validate the effectiveness and suitability of the proposed features. Based on the Inverse Probability Weighting (IPW) equation 6, we define the importance weight as $\rho_i = \frac{\pi_\theta(d_i|x_i)}{\hat{\pi}_b(d_i|x_i)}$. Using this defined weights, the following diagnostic metrics are reported:

Outcome R^2 and Propensity R^2: These quantify the goodness-of-fit for the reward model and the behavior policy, respectively. While Outcome R^2 measures predictive accuracy, Propensity R^2 is crucial for checking if the model captures the underlying treatment assignment mechanism (Voloshin et al., 2019).

Effective Sample Size (ESS): Adapted from Kish's approximation (Freeman, 1966), ESS measures the effective number of samples supporting the counterfactual estimate, defined as:

$$\text{ESS} \approx \frac{\left(\sum_{i=1}^{N}\rho_i\right)^2}{\sum_{i=1}^{N}\rho_i^2} \tag{9}$$

A low ESS implies that the estimator is dominated by a few data points with extreme weights, signaling severe positivity violations.

IPW Variance: Calculated as $Var(\rho)$, high variance indicates instability in the weighting process and susceptibility to distribution shifts.

4.2 Strategy-Stylized Response Generation

Upon determining the optimal strategy d_{best}, we formulate the response generation as a strategy-driven stylized rewriting task. To endow the response with distinct strategic stylistic features while maintaining conversational logical consistency, we propose a two-stage framework based on the "generate-then-rewrite" paradigm.

4.2.1 Construction of Style Demonstration Pool

To provide high-quality stylistic references during prompt construction while mitigating the risk of overfitting caused by excessive reliance on redundant data, we adopt a sparse sampling strategy to construct the demonstration pool. Specifically, we randomly sample a subset containing $\gamma = 20\%$ of the original training data to construct a set of strategy-response pairs, denoted as the demonstration pool $\Omega_{pool} = \{(d_i, r_i)\}_{i=1}^{N}$. This compact pool Ω_{pool} serves as the sole source for retrieving few-shot examples in the subsequent rewriting stage. By limiting the scale of the demonstration pool, we aim to encourage the model to capture the general distributional features of the strategic styles rather than mechanically memorizing specific training samples.

4.2.2 Stage I: Context-Aware Content Generation

The objective of this stage is to generate a content draft that is semantically accurate and highly relevant to the context. Given the dialogue history C and the predicted optimal strategy d_{best}, we first utilize the LLM to generate an intermediate response

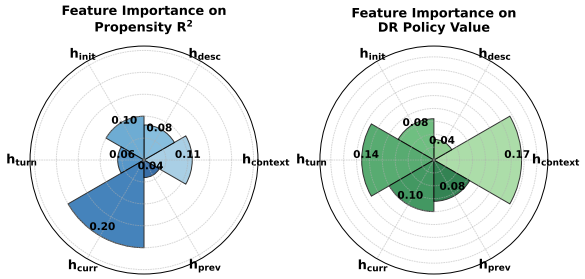

Figure 3: Feature Importance Analysis on Propensity R^2 and DR Policy Value.

$\hat{r}_{content}$ that satisfies the strategic intent but is not explicitly constrained by stylistic features.

This generation process is modeled as maximizing the following conditional probability:

$$\hat{r}_{content} = \arg\max_{r} \; P_\theta(r|C, d_{best}) \tag{10}$$

Here, $\hat{r}_{content}$ ensures the coherence of the response with respect to the context C at the logical level, providing a solid semantic foundation for the subsequent stylization.

4.2.3 Stage II: Prompt-Based Stylized Rewriting

In the second stage, the focus shifts to the alignment of linguistic style. We design a rewriting prompt to guide the model in mapping the intermediate response $\hat{r}_{content}$ into the target strategic style space. Based on the current strategy d_{best}, we retrieve the l most relevant samples from the demonstration pool Ω_{pool} as in-context exemplars. The final stylized response r_{final} is generated through the following rewriting process:

$$\hat{r}_{content} = \arg\max_{r} \; P_\phi(r|\hat{r}_{content}, \mathcal{I}_{rewrite}, TopK(d_{best}, \Omega_{pool})) \tag{11}$$

where $\mathcal{I}_{rewrite}$ represents the style transfer instruction, and $TopK(\cdot)$ denotes the retrieval of stylistic exemplars based on semantic similarity. Through this process, the model reconstructs the syntax, tone, and wording of $\hat{r}_{content}$ to align with the target style while preserving its core semantics. [1]

5 Experimentals

5.1 Datasets

We conduct experiments on EmpatheticDialogues (Rashkin et al., 2019) and ESConv (Liu et al., 2021).

[1]The detailed prompt is shown in the appendix.

Metric	instructor-larg		e5-large	
	History Only	Causal ESC	History Only	Causal ESC
DR Policy Value	0.796	0.814	0.759	0.786
Outcome Model R^2	0.851	0.865	0.842	0.856
Propensity Model R^2	0.126	0.285	0.157	0.261
ESS	0.362	1.653	0.103	1.247
IPW Variance	19.11	4.75	25.11	6.59

Table 1: Comparison of metrics after using different model embeddings: only historical versus added features.

EmpatheticDialogues contains 25k open-domain conversations grounded in 32 fine-grained emotions. ESConv focuses on Emotional Support Conversation, consisting of long-range dialogues incorporating specific support strategies to alleviate user distress.

5.2 Baseline Methods & Automatic Metrics

We compare the Causal-ESC against several state-of-the-art baselines, categorized into explicit and implicit strategy methods.

Explicit Strategy Methods: MISC (Tu et al., 2022) incorporates fine-grained emotion detection and strategy selection within a unified framework. **KEMP** (Li et al., 2022) enhances context understanding by leveraging ConceptNet to model emotional dependencies explicitly. **Multi-ESC** (Cheng et al., 2022) utilizes multi-turn context to predict specific emotional strategies, steering the model to generate strategy-aware responses.

Implicit Strategy Methods: MIME (Majumder et al., 2020) mimics human empathetic behavior by clustering emotions into positive and negative groups to influence generation. **GLHG** (Peng et al., 2022) constructs a global-local hierarchical graph to capture subtle interaction dynamics and context implicitly. **Sibyl** (Wang et al., 2025) focuses on capturing latent conversational topics and sentiment shifts to provide diverse and contextually relevant responses without explicit strategy constraints. **CFEG** (Chen et al., 2024) employs Chain-of-Thought (CoT) fine-tuning to explicitly infer the underlying causes of users' emotions prior to generating empathetic responses. **EmpCRL** (Cai et al., 2024) utilizes in-context commonsense reasoning combined with reinforcement learning to achieve controllable empathetic generation. **TOOL-ED**

Method		ACC.	BLEU-2	BLEU-4	Dist-1	Dist-2	Rou-L	MET.
Explicit Strategy Methods	**MISC**	31.61	7.31	2.20	4.41	19.71	17.91	5.16
	KEMP	39.31	7.82	2.34	4.87	22.55	18.69	6.42
	Muti-ESC	42.01	9.18	3.09	5.34	25.90	20.41	8.84
Implicit Strategy Methods	**MIME**	-	5.23	1.17	2.11	10.94	14.74	6.43
	GLHG	-	7.57	2.13	3.50	21.61	16.37	7.52
	GPT-4o	-	8.45	3.93	6.43	31.39	18.86	8.5
	Sibyl	-	9.02	**4.10**	6.52	32.09	19.20	9.65
	Causal-ESC	**53.53**	**9.54**	4.07	**8.25**	**40.86**	**22.15**	**10.23**

Table 2: Automatic Evaluation Results on ESConv.

Method		BLEU-2	BLEU-4	Dist-1	Dist-2	Rou-L	MET.
Explicit Strategy Methods	**MISC**	7.63	2.56	4.57	20.18	17.65	6.21
	KEMP	8.05	2.79	5.04	23.50	19.17	7.08
Implicit Strategy Methods	**MIME**	6.42	2.36	2.94	12.58	15.21	6.80
	GLHG	8.51	3.62	4.91	25.75	15.46	8.63
	TOOL-ED	9.61	3.60	2.89	13.95	17.93	-
	EmpCRL	-	-	4.27	16.11	-	-
	GPT-4o	9.73	5.62	14.23	38.97	14.01	10.66
	MultiAgentESC	9.82	5.73	9.84	36.75	17.56	10.26
	Sibyl	10.25	**7.57**	9.70	39.86	21.20	10.09
	CFEG	10.54	5.17	2.96	19.52	-	-
	Causal-ESC	**10.74**	7.10	**10.25**	**42.69**	**25.74**	**11.48**

Table 3: Automatic Evaluation Results on Empathetic Dialogues.

(Cao et al., 2025) leverages the tool-calling capabilities of LLMs to dynamically integrate external knowledge and supportive strategies. Finally, **MultiAgentESC** (Xu et al., 2025) constructs a multi-agent collaboration framework, simulating interactions among different specialized roles to iteratively refine the emotional support conversation.

For the evaluation metrics, we utilize Accuracy (**Acc.**) to measure the correctness of explicit strategy selection. For the generated responses, we employ BLEU (Papineni et al., 2002), ROUGE-L (**ROU-L.**) (Lin, 2004), METEOR (**MET.**) (Banerjee and Lavie, 2004), and Distinct-n (**Dist-n**) (Li et al., 2016) to assess the generation quality.

5.3 DR Estimator Validation and Stability Analysis

Table 1 compares our feature selection method (Sec.4.1.1) with raw embedding baselines (instructor-large (Su et al., 2023) and e5-large (Wang et al., 2022)). Reported values are averaged over multiple runs using 20% data samples with varying random seeds. It can be observed that using e5-large as the encoder yields overall performance inferior to instructor-large, which can be attributed to the encoder size, as the T5-based model demonstrates a stronger ability to capture textual features. Under the same encoder setting, although History Only achieves high Policy Value and Outcome R^2, both the Propensity and ESS remain very low. This indicates that the raw history embeddings function as a "black box"—they can overfit the reward signal (high Outcome R^2) but fail to disentangle the causal factors driving strategy selection (low Propensity R^2). The near-zero ESS confirms that the baseline suffers from severe positivity violations, meaning the learned policy relies almost entirely on extrapolation rather than valid causal overlap.

5.4 Necessity of Feature Representation

While Section 4.1.1 theoretically grounded the construction of our feature set $(h_{desc} \oplus ... \oplus h_{prev})$ in psychological principles, we now provide empirical statistical validation to verify their necessity and quantify their individual contributions. To empirically measure the influence and "weight" of each component, we conduct an ablation study by individually removing specific features from the full input vector and observing the subsequent performance degradation. Figure 3 shows the impact of these exclusions on two critical dimensions(Propensity R^2 and DR Policy Value), We observe distinct functional roles across features: emotional intensity (h_{curr}) acts as the primary causal anchor for strategy selection (dominant Propensity

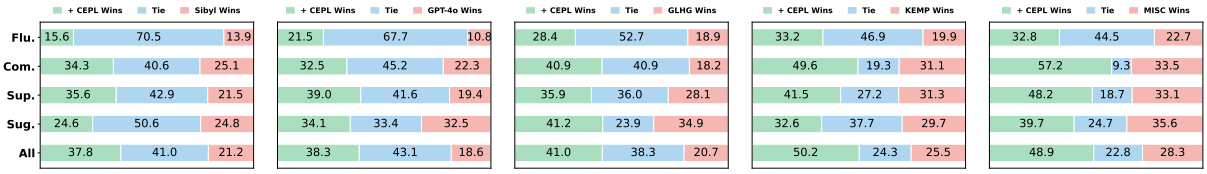

Figure 4: Human A/B test on ESConv (%) dataset.

	Method	BLEU-2	BLEU-4	Dist-1	Dist-2	Rou-L	MET.
GPT-4o	**Causal-ESC**	**9.54**	**4.07**	**8.25**	**40.86**	**22.15**	**10.23**
	w/o **Empathetic Policy**	8.34	3.83	6.43	32.37	17.45	8.3
	w/o **Stylized Rewriting**	9.27	3.95	6.45	31.75	19.26	9.2
DeepSeek-R1	**Causal-ESC**	**9.26**	**3.89**	**7.94**	**37.25**	**21.67**	**9.4**
	w/o **Empathetic Policy**	8.32	3.71	6.16	28.67	17.16	8.1
	w/o **Stylized Rewriting**	8.61	3.85	6.15	28.82	18.44	8.9

Table 4: The ablation experiment results on GPT-4o and DeepSeek.

drop, $\Delta = 0.25$), and simultaneously, semantic content ($h_{context}$) and temporal dynamics (h_{turn}) serve as the decisive drivers for response quality (major DR Value drops, $\Delta = 0.17$ and 0.14). This confirms that emotion triggers the intent, and concurrently, precise context and timing dictate the outcome. Concurrently, the significant degradation in these metrics upon feature removal also validates the rationality of our reward signal y in Equation4, confirming its sensitivity to contextually vital information.

5.5 Main Result

We evaluated the Causal-ESC against state-of-the-art baselines on both the ESConv and ED datasets, with the comparative results summarized in Table 2 and 3. The generally higher scores on ED are due to its shorter dialogue turns and lower complexity relative to ESConv. Regarding generation quality, while our BLEU-4 score is marginally lower than the optimal baseline, the BLEU-2 score is higher. This trade-off implies that our model aligns well with ground-truth keywords while avoiding overfitting, thereby generating more diverse responses. This improved richness is also evidenced by the significant gains in the Distinct (Dist) metrics.

Figure 4 presents the human evaluation results on the ESConv dataset. We compared Causal-ESC with SOTA baselines using five metrics adapted from (Peng et al., 2022): Fluency (**Flu.**), Comforting (**Com.**), Supportive (**Sup.**), Suggestion (**Sug.**), and Overall (**All.**). For this evaluation, we randomly selected 100 dialogue samples, which were assessed by five professional annotators. The results show statistically significant improvements ($p = 0.031 < 0.05$) with a Kappa coefficient of 0.43, reflecting moderate inter-annotator agreement. The results indicate a substantial proportion of ties when comparing our model with large models like Sibyl and GPT-4o. Despite this, Causal-ESC still outperforms them overall by margins of 7.7% and 9%, respectively. In contrast, against GLHG, KEMP, and MISC, the frequency of ties notably diminishes, translating into a more distinct winning advantage. Furthermore, Causal-ESC demonstrates consistent superiority across all evaluated metrics.

Method	ACC
DR	**53.53**
PPO	41.26
Dueling	38.54
Q-Learning	34.21

Table 5: Comparison between DR and classic reinforcement learning methods.

5.6 Policy Learning Analysis

Figure 5 illustrates the comparison between the predicted and ground-truth strategy distributions. It can be observed that the method optimized via Doubly Robust learning aligns much more closely with the true distribution. Furthermore, while the history-only baseline leans towards the "Others" category—opting for safety over efficacy compared to the ground truth—Causal-ESC demonstrates a distinct shift from "Reflection of feelings" to "Self-disclosure." This indicates that our model adopts a more proactive approach to empathetic engagement.

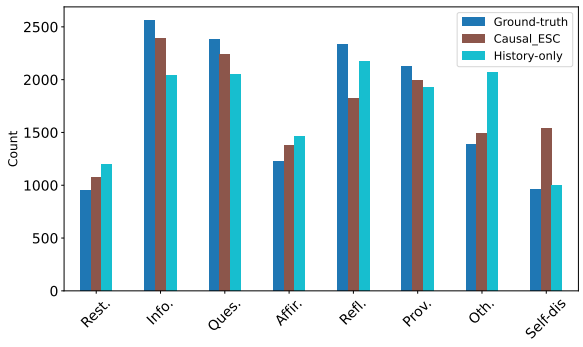

Figure 5: Comparison between the predicted and ground-truth strategy distributions.

Furthermore, we compared DR against several classic reinforcement learning methods. As shown in Table 5, DR consistently outperforms these traditional RL baselines.

5.7 Ablation Experiment

To investigate the effectiveness of individual components within Causal-ESC, we conducted an ablation study by removing the Empathetic Policy and Stylized Rewriting modules, respectively. We utilized both GPT-4o and DeepSeek as backbones to verify the generalizability of our results. As shown in Table 4, the variant without the Empathetic Policy exhibits a significant performance degradation across all metrics, demonstrating that explicit strategy inference is crucial for effective empathetic dialogue. Furthermore, removing the Stylized Rewriting module leads to a notable decline in diversity metrics (Dist-1/2). This indicates that relying solely on policy guidance tends to result in monotonous responses, highlighting the importance of stylization for generating diverse replies.

In Table 6, we present the ablation study concerning the DM and IPW components within the DR framework. The results indicate that while IPW achieves a comparable average policy value, it suffers from significantly higher variance and a markedly lower Effective Sample Size (ESS). Conversely, although the DM method exhibits low variance, its average policy value is poor. Consequently, the Accuracy (ACC) for both isolated methods is lower than that of the full DR approach.

6 Conclusion

In this paper, we presented Causal-ESC, a novel framework designed to address the interpretability and robustness challenges inherent in current

Method	Policy Value	Std	ESS	ACC
DR	**0.814**	**1.47**	**1.653**	**53.53**
DM	0.765	1.87	-	49.78
IPW	0.809	4.75	1.265	51.03

Table 6: Comparison of ablation experiment results between DM and IPW

Emotional Support Conversation systems. Moving beyond the mechanistically opaque paradigm of increasing architectural complexity, we introduced a causal inference perspective to the field. By employing Doubly Robust learning, we explicitly modeled the causal mechanisms between dialogue features and strategy selection, effectively mitigating the biases and distribution shifts common in offline learning scenarios. Furthermore, we bridged the gap between abstract causal strategies and natural conversation through an LLM-based stylized rewriting mechanism.

Our comprehensive experiments, supported by rigorous statistical verification (e.g., Outcome R^2) and human evaluation, confirm that Causal-ESC not only outperforms state-of-the-art baselines in terms of empathy and helpfulness but also ensures theoretical validity. By quantifying the causal drivers of effective support, this work provides a transparent, mathematically grounded solution to affective computing, marking a significant step towards achieving causal and mechanistic transparency in empathic dialogue systems.

Limitations

Despite the contributions of this work, several limitations remain. First, while our method aims to enhance the interpretability of empathic dialogue through causal inference, the intrinsic subjectivity of human conversation implies that statistical causal metrics cannot achieve absolute deterministic correlations. Consequently, our results should be interpreted as a relative enhancement in causal grounding compared to prior methods, rather than establishing a perfect causal link.

Second, the heterogeneity of user needs poses a significant challenge. In emotional support scenarios, different users may require distinct comforting strategies even when facing identical problems. A finite offline dataset cannot exhaustively represent the true distribution of real-world help-seekers or cover every nuanced emotional context. Addressing this gap between limited training data and the

infinite diversity of human emotion remains a persistent challenge for the entire field, which we aim to explore further in future work.

7 Ethics Statement

The datasets utilized in our experiments are exclusively derived from the open-source datasets ESConv (Liu et al., 2021) and ED (Rashkin et al., 2019). As these are publicly available datasets, there are no issues regarding personal privacy or personally identifiable information. For our human evaluation, we recruited crowdsourced workers who possess prior experience in assessing the quality of responses generated by empathetic dialogue systems. All participants were fully informed about the nature and content of the evaluation tasks. Furthermore, all participants received fair and appropriate compensation for their annotation work.

References

Satanjeev Banerjee and Alon Lavie. 2004. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72.

Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746.

Huiying Cao, Yiqun Zhang, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. 2025. Tool-ed: Enhancing empathetic response generation with the tool calling capability of llm. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5305–5320.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

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. *arXiv preprint arXiv:2210.04242*.

Douglas Cohen and Janet Strayer. 1996. Empathy in conduct-disordered and comparison youth. *Developmental psychology*, 32(6):988.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Linton C Freeman. 1966. Kish: Survey sampling (book review). *Social Forces*, 45(1):132.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

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*.

Jochen Hartmann. 2022. Emotion english distilroberta-base. *See https://huggingface. co/j-hartmann/emotion-english-distilroberta-base*.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, David A Clifton, and 1 others. 2025. Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry*, 12(1):27.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

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.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

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

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2122–2132.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

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 (ACL)*, pages 3469–3483.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ana Paiva, Iolanda Leite, and Tiago Ribeiro. 2017. Emotion modeling for social robots. In *The Oxford Handbook of Affective Computing*, pages 296–308. Oxford University Press.

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 31st International Joint Conference on Artificial Intelligence (IJCAI)*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and 1 others. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*, pages 300–325.

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.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Donghong Ji, and Yansong Zhou. 2022. Misc: A mixed strategy-aware model integrating commonsense for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 308–319.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. 2019. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Li Yuan, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2024. Encoding syntactic information into transformers for aspect-based sentiment triplet extraction. *IEEE Transactions on Affective Computing*, 15(2):722–735.

Rui Zhang, Zhenyu Wang, and Dongcheng Mai. 2017. Building emotional conversation systems using multitask seq2seq learning. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 612–621. Springer.

Rui Zhang, Zhenyu Wang, Mengdan Zheng, Yangyang Zhao, and Zhenhua Huang. 2021. Emotion-sensitive deep dyna-q learning for task-completion dialogue policy learning. *Neurocomputing*, 459:122–130.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *arXiv preprint arXiv:2202.13047*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

A Doubly Robust Empathetic Policy Learning

This section presents the algorithm flow of Doubly Robust Policy Optimization as shown in Algorithm 1 and Algorithm 2. The overall algorithm consists of 2 steps: 1) Pre-training the propensity network and reward network; 2) Policy learning using doubly robust estimation.

B Prompt Template Design

This section presents the prompts designed in the Context-Aware Content Generation and Prompt-Based Stylized Rewriting parts of Strategy-Stylized Response Generation, as shown in Tables 7 and 8.

C Implementation Details

To construct the state representation x, we utilize Instructor-Large (Su et al., 2023) and E5-Large (Wang et al., 2022) to generate high-quality semantic embeddings for the dialogue context and problem description. For the Doubly Robust Policy Learning phase, we optimize the networks using the Adam optimizer with a batch size of 68. Distinct learning rates are assigned to ensure stability: $\alpha_\beta = 1e-4$, $\alpha_\phi = 1e-4$, $\alpha_\theta = 1e-5$. To mitigate high variance in IPW, we apply propensity clipping within the range $[0.05, 0.95]$. Finally, for the Strategy-Stylized Response Generation, we leverage the APIs of state-of-the-art LLMs, specifically GPT-4o (OpenAI, 2023) and DeepSeek (Guo et al., 2025), to generate empathetic responses conditioned on the strategies selected by our policy model.

For the ESConv dataset, we randomly selected 20% for policy training and the construction of the Style Demonstration Pool, while the ED dataset was entirely used for testing.

D More Experiment

Figure 6 supplements the human evaluation experiment on the ED dataset. It can be observed that compared to ESConv, the probability of ties has increased. We believe this is due to the shorter conversation, which result in less clarity in the direction of the conversation.

Figure 9 shows the impact of hyperparameter γ. The Dist and ROUGE-L scores increase with γ but eventually exhibit diminishing returns, suggesting that a larger style pool helps the model reach optimal diversity faster. Conversely, BLEU does not

Prompt

Instruction:
You are an intelligent dialogue planning agent.

Your task is to generate a response draft based on the dialogue history and the target strategy provided below.

Requirements:
1. Ensure the response is logically consistent with the context.
2. Strictly reflect the intent of the target strategy.
3. Focus on semantics (what to say) rather than style.

Input:

Dialogue History: {{Dialogue_History_C}}
Target Strategy: {{Best_Strategy_d_best}}

Output:

Draft Response: {{draft_response}}

Table 7: Prompt template for Stage I: Context-Aware Content Generation.

Prompt

Instruction:
You are an expert text style editor.

Your task is to rewrite the provided draft response to match a specific strategic style, without changing the original semantic meaning.

Refer to the style demonstrations below to understand the tone and sentence structure.

 # Style Demonstrations:

Strategy: {{Best_Strategy}}
{{Demo_Text}}
(Top-K retrieved examples)

Input:

Target Strategy: {{Best_Strategy_d_best}}
Draft Response: {{Base_Response_r_content}}

Output:

Stylized Response: {{final_response}}

Table 8: Prompt template for Stage II: Prompt-Based Stylized Rewriting using retrieved exemplars.

Algorithm 1: Supervised Pre-training for Propensity and Reward Models

input : Logged dataset $\mathcal{S} = \{(x_i, d_i, y_i)\}_{i=1}^{N}$, Strategy space $\mathcal{D} = \{1, ..., K\}$ where $K = 8$, Propensity Network f_β, Reward Network f_ϕ, Learning rates α_β, α_ϕ, Batch size B

output : Pre-trained parameters β^* and ϕ^*

// Train Propensity Network f_β

Initialize parameters β;

while *loss convergence* **do**

  Sample a mini-batch $\{(x_j, d_j)\}_{j=1}^{B}$ from $\mathcal{S}$;

  Compute predicted probability distribution $\hat{\pi}_b(x_j) = \text{Softmax}(f_\beta(x_j))$;

  Calculate Cross-Entropy Loss: $\mathcal{L}_{prop}(\beta) = -\frac{1}{B}\sum_{j=1}^{B} \log\left(\hat{\pi}_b(d_j|x_j)\right)$;

  Update $\beta \leftarrow \beta - \alpha_\beta \nabla_\beta \mathcal{L}_{prop}(\beta)$;

end

$\beta^* \leftarrow \beta$;

// Train Reward Network f_ϕ

Initialize parameters ϕ;

while *loss convergence* **do**

  Sample a mini-batch $\{(x_j, d_j, y_j)\}_{j=1}^{B}$ from $\mathcal{S}$;

  Compute predicted reward $\hat{Q}(x_j, d_j) = f_\phi(x_j, d_j)$;

  Calculate MSE Loss: $\mathcal{L}_{reward}(\phi) = \frac{1}{B}\sum_{j=1}^{B}\left(y_j - \hat{Q}(x_j, d_j)\right)^2$;

  Update $\phi \leftarrow \phi - \alpha_\phi \nabla_\phi \mathcal{L}_{reward}(\phi)$;

end

$\phi^* \leftarrow \phi$;

return β^*, ϕ^*;

scale monotonically with pool size and peaks at $\gamma = 20$. To strike a balance across all metrics, we choose $\gamma = 20$.

E Annotation Guidelines

Table 10 presents the annotation guidelines that we provided to the assessors during the human evaluation stage.

F Case Study

We randomly selected an annotator and asked him to familiarize themselves with the empathetic responses in the dataset. We then constructed a specific scenario and required both the annotator and Causal-ESC to generate a response. The comparative results are shown in Table 11. As observed, compared to the human annotator's choice, our model selected a more contextually appropriate strategy.

Algorithm 2: Doubly Robust Empathetic Policy Learning

input : Logged dataset $\mathcal{S}$, Strategy space $\mathcal{D}$ size K, Pre-trained parameters β^*, ϕ^* (Fixed), Policy Network π_θ, Clipping threshold δ, Learning rate α_θ

output : Optimized Policy Network π_{θ^*}

Initialize policy parameters θ;
for M *epochs* **do**
  for *each mini-batch* $\mathcal{B} = \{(x_j, d_j, y_j)\}_{j=1}^{B}$ *in* $\mathcal{S}$ **do**
    Initialize batch gradient $\nabla_\theta \mathcal{J} = 0$;
    `// Pre-calculate nuisance parameters`
    Compute propensity scores $\hat{\pi}_b(k|x_j)$ for all strategies $k \in \mathcal{D}$ using f_{β^*};
    Apply clipping: $\hat{\pi}_{clip}(k|x_j) = \max(\hat{\pi}_b(k|x_j), \delta)$;
    Compute reward estimates $\hat{Q}(x_j, k)$ for all strategies $k \in \mathcal{D}$ using f_{ϕ^*};
    `// Construct Doubly Robust Reward Vector`
    for *each sample* $j \in \{1..B\}$ **do**
      for *each possible strategy* $k \in \mathcal{D}$ **do**
        $\hat{r}_{DR}(x_j, k) \leftarrow \hat{Q}(x_j, k) + \frac{\mathbb{I}(d_j=k)}{\hat{\pi}_{clip}(k|x_j)} \cdot (y_j - \hat{Q}(x_j, k))$;
      end
    end
    `// Policy Optimization`
    Compute target policy distribution $\pi_\theta(x_j) = \text{Softmax}(\pi_\theta(x_j))$;
    Calculate Policy Objective (Maximize Expected Reward):
    $\mathcal{J}(\theta) = \frac{1}{B} \sum_{j=1}^{B} \sum_{k=1}^{K} \pi_\theta(k|x_j) \cdot \hat{r}_{DR}(x_j, k)$;
    $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta \mathcal{J}(\theta)$;
  end
end
return π_{θ^*}

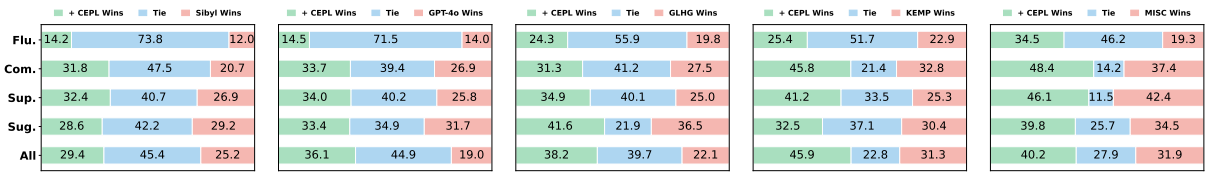

Figure 6: Human A/B test on ED (%) dataset.

γ Value(%)	BLEU-2	BLEU-4	Dist-1	Dist-2	Rou-L	MET.
0	9.27	3.95	6.45	31.75	19.26	9.2
10	9.38	4.05	7.34	36.28	20.92	10.05
20	**9.54**	**4.07**	8.25	40.86	22.15	**10.23**
30	9.52	4.05	8.28	41.37	22.23	10.19
40	9.47	3.98	8.31	41.75	22.30	10.15
50	9.32	3.87	8.33	41.82	22.32	10.08

Table 9: The impact of the demonstration pool sampled rate $\gamma(\%)$ on the experimental results.

Annotation Guidelines for Emotional Support Dialogue Systems

Task Overview:

You will be presented with a Dialogue History (a conversation between a user seeking help and a supporter) and a Candidate Response generated by an AI system. Your task is to evaluate the quality of the response based on 5 specific metrics. For each metric, please rate the response on a scale of 1 to 5.

Metric Definitions:

1. Fluency (Flu.)
- Definition: Evaluating the model based on the fluency of its response.
- 1 = Unreadable / Broken English.
- 5 = Perfect grammar and natural flow.

2. Comforting (Com.)
- Definition: Assessing the model's skill in providing comfort.
- 1 = Cold, indifferent, or makes the user feel worse.
- 5 = Very warm, empathetic, and makes the user feel deeply understood.

3. Supportive (Sup.)
- Definition: Determining to what extent the model offers supportive or helpful responses.
- 1 = Dismissive or unhelpful.
- 5 = Highly encouraging and demonstrates a strong willingness to help.

4. Suggestion (Sug.)
- Definition: Assessing if the bot gave helpful suggestions for the user's problems.
- 1 = Irrelevant, harmful, or nonsensical advice.
- 5 = Very practical, actionable, and wise advice.

5. Overall (All.)
- Definition: Analyzing which model provides more effective overall emotional support.
- 1 = A terrible response.
- 5 = An ideal response.

Table 10: Annotation guidelines for human evaluation.

User Input: "I'm so terrified about my presentation tomorrow. I feel like I'm going to freeze up and fail."
Initial Emotion: Fear/Anxiety

Role	Strategy	Response Text	Analysis
Behavior Policy	Reflection of feelings	"It sounds like you are really nervous about the presentation. That is a scary feeling."	**Bias:** The human follows the statistical correlation in the dataset, choosing a "safe" reflection. While empathetic, it merely validates the fear without offering relief, leading to a sub-optimal outcome (user remains anxious).
Causal-ESC	Self-disclosure	"I completely understand. I used to get stage fright before every speech too, but I realized that once I started talking, the fear went away. You've got this!"	**Win:** Our model analyzes the Initial Context and identifies that mere reflection provides low utility here. It correctly estimates that Self-disclosure (sharing a vulnerability) creates a bond and offers reassurance, leading to a higher predicted emotional improvement.

Table 11: Case study comparing the Behavior Policy and our proposed Causal-ESC.