Structure-Aware Dual-Reward Policy Optimization Framework for Explainable Emotion-Cause Generation in Conversations

Anonymous ACL submission

Abstract

Although large language models (LLMs) have advanced conversational emotion-cause generation, standard supervised fine-tuning (SFT) often induces disguised comprehension. Instead of internalizing deep semantic dependencies, models exploit superficial shortcuts, leading to severe factual hallucinations. To overcome these shortcomings, we propose a structure-aware, preference-data-free dual-reward policy optimization framework operating in a two-stage pipeline. In the first stage, the framework establishes a structured reasoning scaffold via causality-guided alignment. In the second stage, joint reinforcement learning optimizes the policy via two distinct signals: an asymmetric emotion alignment to dismantle lexical biases, and an ATOMIC-grounded logic module to penalize causal fabrication. We evaluate the proposed framework on RECCON-CG and ECF-CG, two benchmark extensions augmented with structured causal rationales. Extensive experiments demonstrate that the proposed framework consistently outperforms strong extractive and generative baselines. Specifically, it improves extraction F1-score by 1.61 and 0.73 over the strongest compared methods, reaching 52.71 and 55.85 on RECCON-CG and ECF-CG, respectively. Furthermore, generative ROUGE-L scores reach 0.5490 and 0.3115, exceeding the strongest baselines by 0.0793 and 0.0066. Qualitative analyses further suggest that the proposed approach effectively mitigates reasoning errors such as causal inversion.

1 Introduction

Emotion-Cause Pair Extraction in Conversations (ECPEC) extends static document-level extraction to dynamic conversational scenarios (Li et al., 2023), shifting from span extraction to generative causal reasoning (Ju et al., 2025; Zhao et al., 2026). Adapting ECPE to conversations remains challenging: utterance-level models fail to capture complex many-to-many dependencies and cross-turn causal chains (Chen et al., 2024; Wang et al., 2024a). Moreover, end-to-end large language models (LLMs) suffer from error propagation; conversational redundancy disrupts logical chains, causing hallucinated attribution and opaque reasoning (Ju et al., 2025; Zhao et al., 2026).

Direct supervised fine-tuning (SFT) of LLMs is insufficient. Without rigorous formatting constraints, LLMs exhibit shortcut learning, exploiting spurious lexical correlations rather than deep semantic dependencies (Yuan et al., 2024). As a result, generative paradigms suffer from causal inversion, forced emotion, contextual hallucination, and lexical over-extrapolation. Standard preference optimization methods, including RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), are computationally expensive and vulnerable to annotator divergence. To address these limitations, we propose a structure-aware dual-reward policy optimization framework. This zero-preference-data paradigm follows a progressive two-stage pipeline to regularize generation across structural formatting, semantic consistency, and commonsense reasoning.

In stage 1, for structural alignment, a causality-guided alignment mechanism prevents stylistic shortcuts through a causal trigger dictionary. Applying a sequence-level maximum likelihood anchoring loss to explicit causal anchors, such as "because" or "due to", strictly constrains the hypothesis space of the framework and enforces standardized, highly interpretable reasoning formats.

In stage 2, for semantic and commonsense alignment, we introduce two orthogonal reward signals to overcome superficial lexical matching. As the first signal, an asymmetric emotion alignment decouples genuine comprehension from word-matching bias. Using an emotion dictionary, we deterministically mask explicit emotion tokens in gold references while evaluating com-

plete predictions. Removing predictive shortcut keywords forces robust long-range contextual representations that implicitly convey correct emotional states.

As the second signal in stage 2, a multi-dimensional logic policy module prevents causal hallucinations induced by structural forcing. Instead of relying on static knowledge graphs, an advanced reasoning LLM performs zero-shot knowledge completion across four ATOMIC-inspired cognitive dimensions (Sap et al., 2019). Generated rationales are projected into a dense continuous space through Sentence-BERT (Reimers and Gurevych, 2019) to compute a cosine-similarity logic reward. Joint optimization of these rewards with the baseline autoregressive loss penalizes logical oversimplification and mitigates catastrophic forgetting, grounding causality in real-world psychology through policy gradient optimization.

We evaluate the framework on generative benchmark extensions, RECCON-CG (Gao et al., 2021) and ECF-CG (Wang et al., 2024b), constructed through expert annotation and AI refinement.Data and code will be released upon acceptance.The resulting corpus is substantially enlarged (Table 1): RECCON-CG contains 1,106 dialogues (11,104 utterances and 9,915 rationales), while ECF-CG contains 1,374 dialogues (13,619 utterances and 7,081 rationales), with standard training, validation, and test splits. Experiments show that the proposed framework consistently surpasses LLM baselines. On ECF-CG, it effectively mitigates dataset bias and generates causal explanations more consistent with human reasoning than standard SFT.

In summary, the main contributions of this work are three-fold:

- We propose the **Structure-Aware Dual-Reward Policy Optimization** framework for **explainable emotion-cause generation in conversations**. This zero-preference-data paradigm incorporates a causality-guided alignment mechanism to enforce structural alignment and systematically mitigate failure modes arising from shortcut learning.

- As the first optimization signal, we devise an **Asymmetric Emotion Alignment** that employs a deterministic masking strategy to eliminate lexical-overlap bias and promote deep semantic alignment.

- As the second optimization signal, we introduce a **Multi-dimensional Logic Policy Module** powered by an advanced reasoning LLM. By projecting ATOMIC-inspired rationales into a continuous space, it provides robust commonsense alignment and explicitly penalizes factual hallucinations through joint policy gradient optimization.

2 Related Work

Emotion-Cause Reasoning in Conversations. Emotion-Cause Reasoning in Conversations has progressed from text-level Emotion-Cause Pair Extraction (ECPE; Xia and Ding, 2019) to multi-turn settings (Poria et al., 2021) and multimodal frameworks incorporating visual and acoustic signals (Zhao et al., 2022; Wang et al., 2024c). Early approaches based on Graph Neural Networks and multi-task learning suffer from error propagation and annotation bottlenecks, prompting a shift toward LLM-based generative reasoning. However, zero-shot LLMs alone fail to model fine-grained conversational causal chains. Without strict output constraints, they often produce unstructured responses and severe logical hallucinations in complex emotion attribution (Wang et al., 2024c).

Shortcut Learning and Disguised Comprehension. These hallucinations largely arise from shortcut learning, where LLMs exploit spurious lexical correlations instead of learning deep causal semantics (Yuan et al., 2024). In emotion reasoning, models frequently overfit to explicit lexical overlap, leading to disguised comprehension (Yuan et al., 2024) that fails when explicit cues are absent. Preventing reliance on superficial lexical features and encouraging genuine long-range causal reasoning remains a key challenge. This observation motivates the asymmetric masking mechanism proposed in this work.

Preference-Free Policy Optimization. Although alignment methods such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) are effective for general model alignment, applying them to fine-grained emotion-cause reasoning introduces severe annotation bottlenecks. Building high-quality preference pairs for conversational causal chains requires extensive expert effort and is vulnerable to annotator disagreement. Moreover, fitting implicit preference distributions

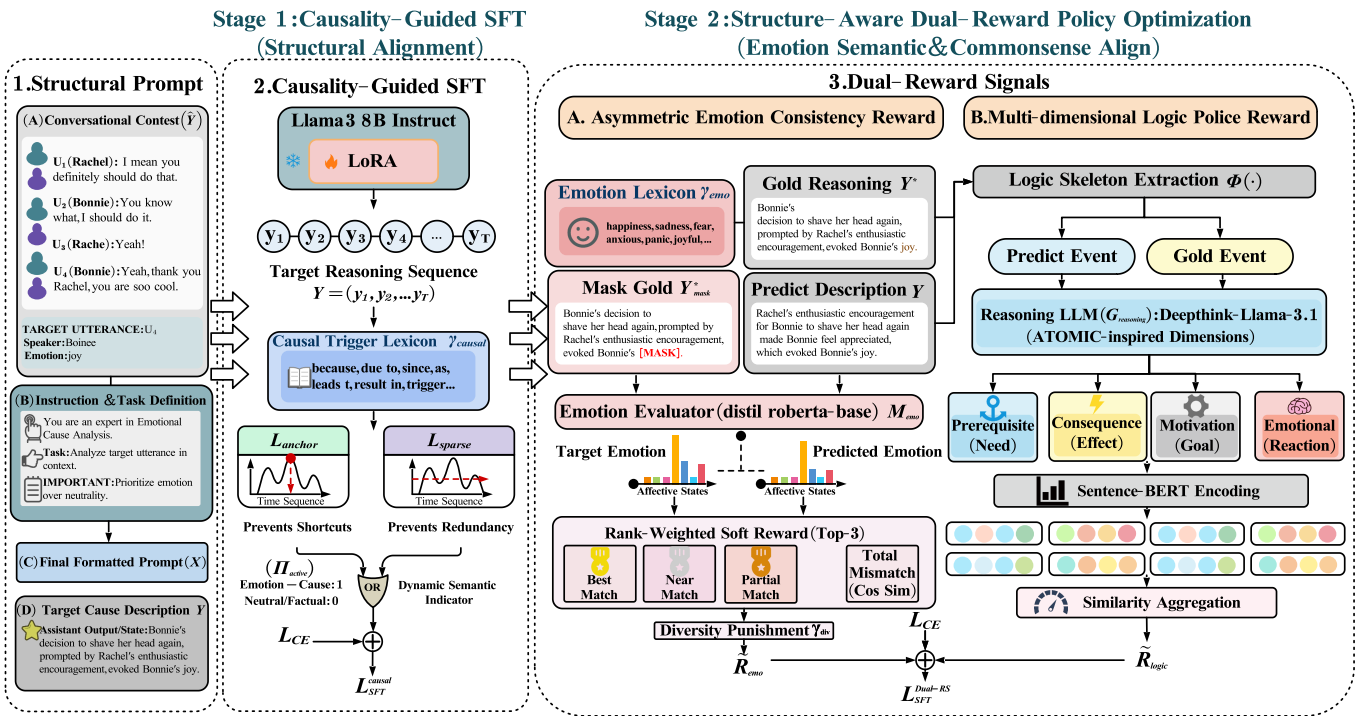

Figure 1: Overall Architecture of the Proposed Method Comprising Stage 1 Causality-Guided Alignment and Stage 2 Structure-Aware Dual-Reward Policy Optimization

hinders the incorporation of explicit commonsense constraints. To avoid dependence on costly paired annotations, we develop a structure-aware policy optimization framework. Leveraging an advanced reasoning LLM guided by cognitive dimensions in the ATOMIC knowledge graph (Sap et al., 2019) and an asymmetric emotion discriminator as explicit rewards, the framework enforces causal consistency and emotion alignment through sequence-level policy gradients without preference data.

3 Structure-Aware Dual-Reward Policy Optimization Framework

This section details the proposed structure-aware dual-reward policy optimization framework, as illustrated in Figure 1. To ensure both stable convergence and highly interpretable outputs, the training methodology operates through a progressive, two-stage pipeline. Following task formulation, stage 1 applies a causality-guided supervised fine-tuning (SFT) mechanism to regularize the generation towards a standardized causal reasoning format, thereby establishing a robust structural foundation. Stage 2 subsequently introduces a dual-reward policy optimization phase to overcome the inherent limitations of superficial lexical matching, effectively aligning the generative process with deep semantic comprehension and human commonsense logic. Furthermore, to maintain a focus on the core architectural design, all specific hyperparameter configurations and thresholds governing these components are deferred to the experimental setup section.

3.1 Task Formulation and Prompt Construction

Before applying structural constraints, we cast the emotion-cause extraction task into a generative sequence-to-sequence format. For a given dialogue, we construct a unified prompt X that integrates the conversational context, the target utterance, and a specific system instruction. This instruction designates the framework as an emotion cause analysis expert and explicitly directs it to prioritize emotional attribution over neutrality. The expected output is a target reasoning sequence $Y = (y_1, y_2, \ldots, y_T)$ that articulates the causal chain explaining why the target emotion was triggered in the given context.

3.2 Causality-Guided Alignment for Structural Alignment

As stage 1 of our training pipeline, standard SFT optimizes large language models via next-token

prediction using the cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid X, y_{<t}). \quad (1)$$

However, in fine-grained emotion-cause generation, pure SFT frequently leads to shortcut learning, wherein the framework mimics stylistic distributions without internalizing underlying causal dependencies. To structurally mitigate reliance on such superficial mappings, we introduce a causality-guided alignment mechanism. Our objective is to compel the framework to explicitly generate causal anchors (such as "because" or "due to") as structural triggers. We construct a causal trigger dictionary $\mathcal{V}_{causal}$ comprising token IDs of predefined causal conjunctions, curated by domain experts for deterministic retrieval.

Instead of uniformly shifting target distributions, we propose a causality-guided distribution constraint. At each step t, we compute the aggregated probability mass allocated to the causal dictionary:

$$S_t = \sum_{v \in \mathcal{V}_{causal}} P(y_t = v \mid X, y_{<t}). \quad (2)$$

Recognizing that valid causal explanations necessitate anchors only at logical turning points, we design a dual-component objective. First, by applying a max-pooling operation across all T time steps, we formulate the existence maximization loss to minimize the negative logarithm of this peak probability:

$$\mathcal{L}_{anchor} = -\log\big(\max_{t\in[1,T]}\{S_t\}\big). \quad (3)$$

Second, to prevent logical redundancy where frameworks continuously generate causal tokens to artificially lower loss, we introduce a dynamic sparsity constraint. We apply a truncation penalty if the average causal probability exceeds a predefined sparsity threshold τ:

$$\mathcal{L}_{sparse} = \max\big(0, \tfrac{1}{T}\textstyle\sum_{t=1}^{T} S_t - \tau\big). \quad (4)$$

Furthermore, conversational dynamics inherently contain factual utterances; enforcing causal triggers on neutral statements induces hallucinated causality. Let $\mathbb{I}_{active} \in \{0, 1\}$ be a dynamic semantic indicator, where 1 denotes valid emotion-cause semantics and 0 indicates neutral statements. The structural alignment objective is formulated as:

$$\mathcal{L}_{SFT}^{causal} = \mathcal{L}_{CE} + \beta \mathbb{I}_{active}(\mathcal{L}_{anchor} + \mathcal{L}_{sparse}), \quad (5)$$

where β is a scaling hyperparameter. This optimization successfully lays a structurally formatted foundation, concluding stage 1.

3.3 Asymmetric Emotion Consistency for Semantic Alignment

Building upon the formatted outputs of stage 1, stage 2 initiates a dual-reward policy optimization to align the generation with actual semantics. Generative frameworks frequently exploit explicit lexical overlaps rather than comprehending underlying emotional semantics, causing disguised comprehension. To fundamentally dismantle this shortcut, we propose an asymmetric emotion alignment as the first signal.

We construct an emotion dictionary $\mathcal{V}_{emo}$ compiling high-frequency emotion words. Deploying a pre-trained emotion evaluator $\mathcal{M}_{emo}$ and this dictionary, we deterministically mask explicit emotional tokens in the gold reasoning sequence Y^*, yielding an emotionally implicit sequence Y^*_{mask}. Conversely, the policy-generated predicted sequence $\hat{Y}$ remains unmasked. This asymmetric evaluation deprives the framework of predictive shortcut keywords, compelling the construction of robust long-range contexts.

Obtaining probability distributions $P^* = \mathcal{M}_{emo}(Y^*_{mask})$ and $\hat{P} = \mathcal{M}_{emo}(\hat{Y})$, we replace rigid binary matching with a rank-weighted soft reward based on a semantic tier system. Let $\mathcal{C}_{top\text{-}k} = \{c_1, c_2, \ldots, c_k\}$ be the top-k emotion classes from P^*, and $\hat{c}$ with confidence $\hat{p}$ be the predicted dominant class from $\hat{P}$. The base emotional reward R_{emo} is defined as:

$$R_{emo} = \begin{cases} \mu_i + \nu_i \hat{p}, & \text{if } \hat{c} = c_i \in \mathcal{C}_{top\text{-}k} \\ \min(\epsilon, \epsilon\cos(P^*, \hat{P})), & \text{else} \end{cases} \quad (6)$$

where μ_i and ν_i denote the rank-specific base rewards and scaling coefficients for the i-th ranked class, ϵ is a minimal fallback threshold, and the tier size k is a hyperparameter. To regularize degenerative behaviors like token repetition, we apply a diversity decay factor γ_{div} based on the unique unigram ratio of $\hat{Y}$, yielding the final semantic reward $\tilde{R}_{emo} = \gamma_{div} \cdot R_{emo}$.

3.4 Multi-Dimensional Logic Policy for Commonsense Alignment

While emotion consistency ensures semantic alignment, generative frameworks often hallucinate plausible but factually flawed causal links. As the second reward signal of stage 2, we devise a multi-dimensional logic policy module. Abandoning static knowledge graph embeddings, we power this module with an advanced reasoning LLM.

To eliminate noise, we parse Y^* and $\hat{Y}$ using a syntactic dependency parser, extracting a logic skeleton $\Phi(\cdot)$ by retaining structural nouns and verbs while explicitly masking emotional variables. Inspired by ATOMIC knowledge graphs, we query the reasoning engine across four cognitive dimensions: $\mathcal{D} = \{\text{Prerequisite}, \text{Consequence}, \text{Motivation}, \text{Emotional Reaction}\}$. Constructing dimension-specific prompts instructs the engine to complete causal subgraphs, generating rationales r_d^* and $\hat{r}_d$ in a zero-shot manner. We project these rationales into dense continuous spaces via Sentence-BERT $\mathcal{E}$. The logic reward R_{logic} represents the weighted sum of cosine similarities:

$$R_{logic} = \frac{\sum_{d \in \mathcal{D}} w_d \cos(\mathcal{E}(r_d^*), \mathcal{E}(\hat{r}_d))}{\sum_{d \in \mathcal{D}} w_d}, \quad (7)$$

where w_d denotes dimension-specific weights. This dimension-wise validation guarantees generated causality remains bounded within rational commonsense domains.

3.5 Unified Dual-Reward Policy Optimization

Finally, as the culmination of stage 2, we integrate the causal structural constraints with the semantic and logical rewards into a unified policy gradient framework. We employ the REINFORCE algorithm to optimize the policy framework π_θ, initialized with weights pre-aligned by $\mathcal{L}_{SFT}^{causal}$.

Given the sequence log-probability $\log \pi_\theta(\hat{Y} \mid X)$, we compute advantages for both rewards by subtracting respective moving-average baselines (b_{emo} and b_{logic}). Policy gradient objectives are formulated as:

$$\mathcal{L}_{PG}^{emo} = -\mathbb{E}_{\hat{Y}}\big[(\tilde{R}_{emo} - b_{emo}) \log \pi_\theta(\hat{Y}|X)\big], \quad (8)$$

$$\mathcal{L}_{PG}^{logic} = -\mathbb{E}_{\hat{Y}}\big[(R_{logic} - b_{logic}) \log \pi_\theta(\hat{Y}|X)\big]. \quad (9)$$

We jointly optimize the policy gradient objectives with the baseline autoregressive loss $\mathcal{L}_{CE}$ to prevent reward hacking and mitigate catastrophic forgetting. The overarching objective function establishes progressive fusion:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{PG}^{emo} + \gamma \mathcal{L}_{PG}^{logic}, \quad (10)$$

where α and γ govern the trade-offs between structural imitation, emotional accuracy, and logical rigor.

Dataset	Split	Dialogues	Utterances	Rationales
RECCON-CG	Train	834	8,206	7,630
	Val	47	493	352
	Test	225	2,405	1,933
	Total	**1,106**	**11,104**	**9,915**
ECF-CG	Train	1,001	9,966	5,123
	Val	112	1,087	614
	Test	261	2,566	1,344
	Total	**1,374**	**13,619**	**7,081**

Table 1: Detailed statistics of conversational emotion-cause benchmarks. The 'Rationales' column represents the volume of gold-standard causal reasoning chains leveraged for fine-grained alignment.

4 Experiments

To evaluate the proposed framework, we conduct experiments on two conversational benchmarks: RECCON-CG and ECF-CG. These datasets extend the original RECCON (Poria et al., 2021) and Emotion-Cause in Friends (ECF) benchmark released in the SemEval-2024 Multimodal Emotion Cause Analysis in Conversations shared task (Wang et al., 2024b). Based on the original causal annotations, expert annotators first constructed the reasoning chains, which were subsequently refined by an AI model for linguistic fluency while preserving the underlying causal logic.

As summarized in Table 1, our corpus scales significantly. RECCON-CG comprises 1,106 dialogues, consisting of 11,104 utterances and 9,915 gold-standard causal reasoning chains. ECF-CG encompasses 1,374 dialogues, containing 13,619 utterances and 7,081 rationales. Both datasets are partitioned into standard training, validation, and test splits to ensure robust evaluation.

The evaluation spans two distinct paradigms. First, to facilitate comparisons with traditional extractive models, the generated reasoning chains are parsed into discrete pairs and evaluated on the emotion-cause pair extraction in conversations

Methods	RECCON-CG			ECF-CG		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Traditional Extractive Models						
ECPE-2steps (Indep)	50.23	44.30	46.73	54.39	46.80	50.13
ECPE-2steps (Inter-EC)	49.69	44.64	46.88	54.40	45.32	49.26
ECPE-2steps (Inter-CE)	50.04	45.81	47.67	54.43	46.15	49.79
RANKCP	55.46	19.56	28.92	**61.89**	22.07	32.53
ECPE-MLL	51.82	50.47	51.10	55.06	48.53	51.59
PairGCN	**59.80**	42.64	49.77	61.67	41.11	49.32
TransECPE	44.47	25.92	32.75	33.85	51.09	40.72
UTOS	31.41	29.60	30.48	23.83	46.13	31.43
MC-ECPE-2steps	–	–	–	56.67	49.91	51.74
UECA-Prompt	41.65	27.63	33.18	48.63	39.67	43.69
UECA-Prompt-M2M (M=5)	51.95	39.02	44.52	53.70	45.18	49.02
PRG-MoE	54.86	41.88	47.50	57.30	46.85	51.55
SHARK	–	–	–	48.19	53.83	50.87
EmoPrompt-ECPE	45.07	45.41	45.06	48.21	47.79	47.84
CENTER	47.39	46.88	47.13	47.90	43.65	44.75
MultiCauseNet	–	–	–	53.27	59.10	55.12
Comparable Generative LLMs						
GMEC	–	–	–	60.41	50.03	54.73
Mistral-7B-v0.3	45.45	51.43	48.25	52.49	54.06	53.26
Qwen2-7B-Instruct	34.18	49.40	40.41	41.21	**62.47**	49.66
Llama-3-8B	47.73	**54.09**	50.71	48.82	60.08	53.87
Ours	52.26	53.71	**52.71**	51.51	60.98	**55.85**

Table 2: Performance comparison on the Emotion-Cause Pair Extraction in Conversations (ECPEC) task across the RECCON-CG and ECF-CG datasets.

(ecpec) and emotion-cause tuple extraction in conversations (ectec) tasks utilizing precision, recall, and the f1-score. Second, to assess the quality of the generated natural language rationales, we employ a comprehensive suite of generative metrics: rouge and bleu for structural integrity, meteor and bertscore for deep semantic overlap, and dist for lexical diversity.

4.1 Implementation Details

The training process adopts a rigorous, two-stage paradigm. In stage 1, the base model, initialized with the `Meta-Llama-3-8B-Instruct` architecture, undergoes Causality-Guided Supervised Fine-Tuning (SFT). To optimize memory efficiency without sacrificing representational capacity, we employ Low-Rank Adaptation (LoRA) with 4-bit NormalFloat and double quantization, maintaining the compute data type at `bfloat16`. LoRA adapters (rank 16, alpha 32, dropout 0.1) are injected into all linear projection layers, encompassing query, key, value, output, up, down, and gate projections. We utilize the `AdamW` optimizer with a peak learning rate of 2×10^{-5}, scheduled via a cosine strategy with a 0.03 warmup ratio. The structural constraint scaling factor β is set to 0.5, and the dynamic sparsity threshold τ is set to 0.5, determined via grid search on the validation set.

In stage 2, to prevent early policy divergence, the converged SFT weights initialize the policy model for unified dual-reward reinforcement learning. This stage incorporates two orthogonal reward signals: the asymmetric emotion alignment and the multi-dimensional logic policy module. For the semantic alignment signal, we utilize a pre-trained `emotion-english-distilroberta-base` model as the emotion evaluator $\mathcal{M}_{emo}$, setting the tier size $k = 3$ for the semantic tier system. The rank-specific base rewards μ_i are $\{0.65, 0.40, 0.15\}$ for $i \in \{1, 2, 3\}$ respectively, with corresponding scaling coefficients ν_i as $\{0.35, 0.20, 0.10\}$ and a minimal fallback threshold $\epsilon = 0.1$. The multi-dimensional logic policy module leverages `Deepthink-Llama-3-8B-Preview` for zero-shot rationale generation across four ATOMIC-inspired dimensions. The dual-reward balancing weights for the policy gradient objectives are empirically set to $\alpha = 0.5$ and $\gamma = 0.3$.

All implementations utilize PyTorch 2.5.1 (CUDA 12.1), trained on a single NVIDIA GeForce RTX 4090 GPU (24GB VRAM) running Ubuntu 20.04 with an Intel Xeon Gold 6133 CPU.

Dataset	Methods	Anger	Disgust	Fear	Happiness	Sadness	Surprise	6-Avg	4-Avg
RECCON-CG	ECPE-2steps	17.19	1.33	0.00	45.81	17.54	25.01	34.25	36.08
	ECPE-2steps + CSK	19.62	0.61	0.95	45.34	20.34	25.80	34.71	36.57
	ECPE-2D	22.58	0.00	0.00	47.72	12.82	36.10	36.58	38.59
	ECPE-2D + CSK	21.51	0.35	0.00	47.82	13.78	32.95	36.32	38.30
	UECA-Prompt	24.11	11.38	31.20	47.53	21.40	35.09	38.63	39.63
	UECA-Prompt + CSK	22.28	6.98	27.43	45.14	23.42	35.74	36.92	38.06
	PRG-MoE	23.20	17.72	44.90	48.40	22.88	40.66	40.11	40.65
	Mistral-7B-v0.3	37.73	19.71	29.09	49.34	34.06	33.33	43.97	45.13
	Qwen2-7B-Instruct	26.70	9.88	19.23	42.92	34.00	22.57	36.54	37.75
	Llama-3-8B	34.48	20.28	41.37	52.15	32.94	43.56	46.57	47.53
	Ours	37.06	20.58	45.61	53.88	35.22	46.39	**48.42**	**49.41**
ECF-CG	ECPE-2steps	24.39	0.00	0.71	38.84	21.60	40.24	29.32	31.92
	ECPE-2steps + CSK	23.86	0.00	3.71	38.41	23.13	40.71	29.50	32.03
	ECPE-2D	25.13	0.00	0.00	41.25	21.62	43.24	30.80	33.55
	ECPE-2D + CSK	25.36	0.00	0.00	38.89	22.62	43.26	30.37	33.09
	UECA-Prompt	27.37	12.85	7.91	37.96	22.51	39.53	30.75	32.49
	UECA-Prompt + CSK	23.57	7.70	9.36	34.95	22.44	38.69	28.49	30.30
	MC-ECPE-2steps	25.16	1.07	0.60	37.37	21.97	40.71	29.32	31.86
	GMEC	28.50	5.93	0.00	41.73	25.90	47.77	33.81	36.48
	Mistral-7B-v0.3	41.73	15.94	13.22	42.57	30.18	45.55	38.73	41.26
	Qwen2-7B-Instruct	24.92	9.87	26.31	42.89	36.17	20.16	36.53	37.71
	Llama-3-8B	35.43	18.18	13.40	43.62	31.76	48.80	37.72	40.29
	Ours	40.69	17.77	13.61	46.15	31.02	47.13	**39.34**	**42.34**

Table 3: F1-score performance comparison on the Emotion-Cause Tuple Extraction in Conversations (ECTEC) task across discrete emotion categories.

Dataset	Methods	R-1 ↑	R-L ↑	D-1 ↑	D-2 ↑	B-1 ↑	B-2 ↑	MET. ↑	BERT. ↑
RECCON-CG	Mistral-7B-v0.3	0.4974	0.4697	0.1060	0.3046	0.4273	0.3642	0.4350	0.9246
	Qwen2-7B-Instruct	0.3041	0.2704	0.1009	0.2947	0.2118	0.1195	0.2378	0.8997
	Llama-3-8B	0.4060	0.3809	0.0920	0.2595	0.3032	0.1889	0.3569	0.9154
	Ours	**0.5739**	**0.5490**	**0.1094**	**0.3123**	**0.5116**	**0.4585**	**0.5192**	**0.9356**
ECF-CG	Mistral-7B-v0.3	0.3310	0.2993	**0.1007**	0.3732	0.2600	0.1782	0.2509	0.8915
	Qwen2-7B-Instruct	0.3131	0.2815	0.1006	**0.3932**	0.2326	0.1494	0.2262	0.8892
	Llama-3-8B	0.3378	0.3049	0.0960	0.3778	0.2503	0.1625	0.2447	0.8959
	Ours	**0.3425**	**0.3115**	0.0971	0.3857	**0.2704**	**0.1868**	**0.2685**	**0.8967**

Table 4: Performance comparison on natural language generation metrics. R-1/L, D-1/2, B-1/2, MET., and BERT. represent ROUGE-1/L, Distinct-1/2, BLEU-1/2, METEOR, and BERTScore, respectively.

Consequently, the stage 1 and stage 2 optimization phases required approximately 3 and 7 hours, respectively.

4.2 Compared Baselines

We compare against a comprehensive set of baselines. Traditional extractive architectures include pipeline models (Xia and Ding, 2019), ranking method (Wei et al., 2020), end-to-end graph and multi-label models (Ding et al., 2020; Chen et al., 2020), transition-based frameworks (Fan et al., 2020; Cheng et al., 2021), prompt-tuning approaches (Zheng et al., 2022; Gu et al., 2024), and conversation-specific designs (Jeong and Bak, 2023; Wang et al., 2024a; Ma et al., 2025). Generative baselines include task-specific models such as SHARK (Wang et al., 2023) and GMEC (Ju et al., 2025), alongside representative open-weight LLMs: Llama-3-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Qwen2-7B-Instruct (Yang et al., 2024). To ensure a strictly fair comparison against our method, these open-weight LLMs were fully fine-tuned on the training splits of RECCON-CG and ECF-CG using standard Supervised Fine-Tuning (SFT). They serve as robust baselines to demonstrate how standard SFT alone is insufficient and inherently prone to shortcut learning.

Dataset	Variant	R-1 ↑	R-L ↑	D-1 ↑	D-2 ↑	B-1 ↑	B-2 ↑	MET. ↑	BERT. ↑	Pair-F1	6-Avg
RECCON-CG	**Ours**	**0.5739**	**0.5490**	0.1094	**0.3123**	**0.5116**	**0.4585**	**0.5162**	**0.9356**	0.5271	0.4842
	w/o AEA	0.5686	0.5433	**0.1096**	0.3122	0.5047	0.4500	0.5102	0.9347	0.5355	0.4875
	w/o MLP	0.5166	0.4904	0.1007	0.2830	0.4475	0.3810	0.4631	0.9261	**0.5509**	**0.5092**
	w/o DRL	0.4309	0.4058	0.0914	0.2565	0.3331	0.2227	0.3873	0.9192	0.4864	0.4438
	w/o CGA	0.4060	0.3809	0.0920	0.2595	0.3032	0.1889	0.3569	0.9154	0.5071	0.4657
ECF-CG	**Ours**	**0.3425**	**0.3115**	**0.0971**	**0.3857**	**0.2704**	**0.1868**	**0.2685**	**0.8967**	**0.5585**	**0.3934**
	w/o AEA	0.3415	0.3090	0.0963	0.3836	0.2685	0.1829	0.2683	0.8957	0.5344	0.3817
	w/o MLP	0.3194	0.2792	0.0801	0.3140	0.2332	0.1489	0.2674	0.8897	0.5401	0.3803
	w/o DRL	0.3017	0.2605	0.0850	0.3337	0.2148	0.1337	0.2356	0.8841	0.4772	0.3316
	w/o CGA	0.3378	0.3049	0.0960	0.3778	0.2503	0.1625	0.2447	0.8959	0.5387	0.3771

Table 5: Ablation study of the proposed framework. CGA denotes Causality-Guided Alignment, DRL denotes Dual-Reward Learning, AEA denotes Asymmetric Emotion Alignment, and MLP denotes the Multi-dimensional Logic Policy Module.

4.3 Main Results

Table 2 reports emotion-cause pair extraction performance. Our framework achieves state-of-the-art F1 scores of 52.71 (RECCON-CG) and 55.85 (ECF-CG), outperforming traditional extractive architectures and competitive LLM baselines. Specifically, it surpasses the strongest generative baseline (`Llama-3-8B`) by 2.00 and 1.98 F1 points, respectively. For emotion-cause tuple extraction (Table 3), the framework achieves the highest six-category average F1 (48.42 on RECCON-CG and 39.34 on ECF-CG), demonstrating robustness in recognizing infrequent emotions such as "Fear" and "Disgust", which standard LLMs often struggle with because of data sparsity and shortcut bias.

For rationale generation (Table 4), our approach achieves ROUGE-L scores of 0.5490 and 0.3115, exceeding the strongest generative baseline by 0.0793 on RECCON-CG and 0.0066 on ECF-CG. A paired bootstrap resampling test confirms that the ROUGE-L gain on RECCON-CG is statistically significant ($p < 0.01$). Moreover, a BERTScore of 0.9356 on RECCON-CG indicates strong semantic faithfulness, narrowing the gap between lexical overlap and causal understanding.

4.4 Ablation Study

Table 5 presents ablation results. Removing the causality-guided alignment stage (w/o CGA) causes a substantial performance drop (e.g., RECCON-CG ROUGE-L decreases from 0.5490 to 0.3809), confirming its role in establishing a structural scaffold. Removing the dual-reward learning stage (w/o DRL) further degrades performance, reducing the model to a vanilla SFT baseline prone to hallucinated attribution.

A trade-off emerges for the multi-dimensional logic policy module (w/o MLP) and asymmetric emotion alignment reward (w/o AEA). On RECCON-CG, removing either module slightly improves pair extraction F1 (e.g., w/o MLP reaches 0.5509), while consistently degrading rationale generation quality and semantic fidelity, reflected by a decline in ROUGE-1 from 0.5739 to 0.5166. This observation suggests that both reward components primarily improve generation fidelity and semantic consistency rather than extraction accuracy. Consistently, the complete framework achieves the best overall performance on the more challenging ECF-CG dataset across both extraction and generation metrics. Overall, causality-guided alignment provides structural guidance, whereas dual-reward learning enhances logical fidelity in complex multi-turn dialogues.

5 Conclusion

We identify *disguised comprehension* as a shortcut-learning phenomenon in conversational emotion-cause reasoning under standard SFT. To mitigate this issue, we propose a structure-aware dual-reward policy optimization framework without paired preference data. The framework combines causality-guided alignment, asymmetric emotion consistency, and a multi-dimensional logic policy module grounded in ATOMIC cognitive dimensions. Experiments on RECCON-CG and ECF-CG demonstrate strong performance in both extraction and generation, while human evaluation suggests improved robustness against causal hallucinations and shortcut reasoning. These findings indicate that preference-free reward alignment is a promising direction for enhancing causal reasoning in conversational LLMs.

Limitations

While effective, our framework presents several limitations that necessitate future investigation. First, regarding computational overhead, while our paradigm avoids the prohibitive costs of curating large-scale paired preference data for rlhf or dpo, the multi-dimensional logic policy module relies on an external reasoning llm to generate atomic-style rationales. Although this operation is decoupled from the online policy gradient updates, it still imposes an offline computational budget compared to standard sft paradigms. Second, concerning modality constraints, our framework is rigorously evaluated within a pure-text modality. However, real-world human interactions are intrinsically multimodal, heavily relying on acoustic prosody and visual micro-expressions. While we utilize the textual annotations of multimodal datasets, extending our asymmetric masking strategy to continuous audio-visual signals remains an open challenge. Finally, regarding linguistic and cultural bias, our benchmarks (reccon-cg and ecf-cg) are predominantly english-centric and reflect specific western cultural contexts. Given that emotional expression and causal attribution are highly culturally dependent, the generalizability of our trained policy and the generalization capability of the logic policy module to low-resource languages or divergent cultural norms requires further empirical validation.

Ethics Statement

This work strictly adheres to the ACL Code of Ethics. The datasets utilized in our experiments (RECCON-CG and ECF-CG) are constructed based on publicly available dialogue corpora. They have been thoroughly anonymized and contain no Personally Identifiable Information (PII). We commit to making these processed datasets and our source code publicly available upon the publication of this work.

Furthermore, our proposed model is intended purely for academic research in interpretable emotion reasoning and must not be employed for clinical diagnosis or psychological assessment. Finally, as the model outputs may reflect biases inherent in the pre-training data, human oversight is strictly required prior to any real-world deployment.

References

Xinhao Chen, Chong Yang, Changzhi Sun, Man Lan, and Aimin Zhou. 2024. From coarse to fine: a distillation method for fine-grained emotion-causal span pair extraction in conversation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, N. Li, and Qing Gu. 2021. A unified target-oriented sequence-to-sequence model for emotion-cause pair extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2779–2791.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583, Online. Association for Computational Linguistics.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717, Online. Association for Computational Linguistics.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xue Gu, Zhihan Zhou, Ziyao Meng, Jian Li, Tiago Gomes, Adriano Tavares, and Hao Xu. 2024. EmoPrompt-ECPE: Emotion knowledge-aware prompt-tuning for emotion-cause pair extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-*

COLING 2024), pages 5678–5688, Torino, Italia. ELRA and ICCL.

DongJin Jeong and JinYeong Bak. 2023. Conversational emotion-cause pair extraction with guided mixture of experts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3288–3298, Dubrovnik, Croatia. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xincheng Ju, Dong Zhang, Junhui Li, Shoushan Li, and Guodong Zhou. 2025. Enhanced generative framework with llms for multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Multimedia*, 27:4924–4935.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2023. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1754–1765.

Jun-chi Ma, Hassan Nazeer Chaudhry, Farzana Kulsoom, Guihua Yang, Sajid Ullah Khan, Sujit Biswas, Zahid Ullah Khan, and Faheem Khan. 2025. MULTICAUSENET temporal attention for multimodal emotion cause pair extraction. *Scientific Reports*, 15(1):19372.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing Emotion Cause in Conversations. *Cognitive Computation*, 13(5):1317–1332.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Cogles, Anton Sowndararajan, Choi Yejin, and 1 others. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Botao Wang, Keke Tang, and Peican Zhu. 2024a. Enhancing emotion-cause pair extraction in conversations via center event detection and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10773–10783, Miami, Florida, USA. Association for Computational Linguistics.

Fanfan Wang, Heqing Ma, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2024b. Observe before generate: Emotion-cause aware video caption for multimodal emotion cause generation in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 58205828, New York, NY, USA. Association for Computing Machinery.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024c. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico. Association for Computational Linguistics.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963, Singapore. Association for Computational Linguistics.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

Yulin Zhao, Dayan Wu, Zheng Lin, and Weiping Wang. 2026. Dualecot: Toward explainable emotion causality analysis in conversations. *IEEE Transactions on Audio, Speech and Language Processing*, 34:650–663.

Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. Uecaprompt: Universal prompt for emotion cause analysis. In *International Conference on Computational Linguistics*.

A Human Evaluation Protocol

All human evaluations were conducted under a strict double-blind protocol. Three senior graduate students with prior research experience in natural language processing and dialogue understanding served as annotators. The annotators were unaware of model identities, experimental hypotheses, and the objectives of the proposed framework throughout the evaluation process.

Before formal annotation, all annotators completed a calibration stage using a small validation subset and were provided with detailed evaluation guidelines for each metric. During evaluation, model outputs were anonymized and randomly shuffled to eliminate potential presentation bias.

For the comprehensive human evaluation in Section D, we randomly sampled 200 test instances for each model variant from the corresponding benchmark. Each generated rationale was independently scored by all annotators using a five-point Likert scale. Final scores were computed by averaging ratings across annotators.

For the Logic Policy evaluation in Section B, we randomly sampled 600 generated rationales for each cognitive dimension (Prerequisite, Consequence, Motivation, and Emotional Reaction) from the training corpus. These rationales were independently assessed using the same evaluation protocol.

Inter-annotator agreement was measured using Krippendorff's α. Across all evaluation settings, α exceeded 0.75, indicating substantial agreement and strong annotation reliability.

To assess annotation consistency, we computed Krippendorffs alpha, obtaining values of 0.79 and 0.82 on RECCON-CG and ECF-CG respectively, indicating substantial agreement.

Dataset	Cog.Dim	Fluency	Coherence	Relevance
RECCON-CG	Prerequisite	4.02	3.99	3.89
	Consequence	4.53	4.01	3.09
	Motivation	4.70	4.34	3.91
	Emotional Reaction	4.65	4.29	3.75
ECF-CG	Prerequisite	3.86	3.89	3.95
	Consequence	4.52	3.98	3.23
	Motivation	4.58	4.28	3.79
	Emotional Reaction	4.76	4.38	3.81

Table 6: Human evaluation results of the Multi-dimensional Logic Policy across four cognitive reasoning dimensions. For each dimension, 600 randomly sampled generated rationales were evaluated by three independent annotators under a double-blind protocol using a five-point Likert scale.

B Human Evaluation of the Logic Policy Module

To assess the reliability of the reasoning model employed as the Multi-dimensional Logic Policy, we conducted a dedicated human meta-evaluation of the generated cognitive rationales. Specifically, for each cognitive dimension (Prerequisite, Consequence, Motivation, and Emotional Reaction), we randomly sampled 600 generated rationales from the training corpus. The evaluation was performed independently on both RECCON-CG and ECF-CG.

Following common practices in dialogue reasoning evaluation, all rationales were assessed by three senior graduate students with prior research experience in natural language understanding and dialogue reasoning. The evaluation adopted a strict double-blind protocol: annotators were not informed of the study hypotheses, model identities, or experimental objectives, and all outputs were anonymized and randomly shuffled before scoring.

Error Type	Full Comparative Rationale Analysis
Causal Inversion *(Temporal Flaw)*	**Context:** Ross: Second date! → Chandler: I do not know. → Rachel: You do not know?! **Gold:** Chandler's unexpected hesitation about a second date after Danielle's considerate visit provoked Rachel's surprise. **SFT ✗:** Rachel's playful teasing provoked his hesitant response, which triggered her surprise. **Ours ✓:** Rachel's expectation of his joy about the date was met with his **unexpected uncertainty**, which triggered her surprise.
Forced Emotion *(Over-sensitivity)*	**Context:** Ross: ...hard to forget. → Joey: Things happened that we should never talk about. **Gold:** The speaker's statement is purely factual. **SFT ✗:** Joey's frustration with Ross's dwelling on past issues provoked Joey's **anger**. **Ours ✓:** The utterance currently **lacks distinct emotional semantics**.
Logic Fabrication *(Contextual Flaw)*	**Context:** Monica: ...if you ready now I let you play at wedding. → Phoebe: Really?! **Gold:** Monica's unexpected offer to let Phoebe play her song at the wedding if she got ready immediately provoked Phoebe's surprise. **SFT ✗:** Phoebe's song being accepted for the wedding after her initial refusal to get ready caused surprise. **Ours ✓:** Monica's **offer** to let Phoebe play her song **if she got ready now** triggered Phoebe's surprise.
Over-extrapolation *(Commonsense Flaw)*	**Context:** Rachel: ...my side hurts from when you crashed into me. → Monica: Oh God, sorry. **Gold:** Monica crashing into Rachel yesterday caused Rachel's side to still hurt, which resulted in Monica's sadness. **SFT ✗:** Rachel's side still hurts from the **car accident** caused Monica's apology... **Ours ✓:** Rachel's pain from the previous day's **collision** caused Monica's apology, resulting in sadness.

Table 7: Qualitative Taxonomy of Representative Reasoning Errors and Comparative Case Analysis.

Each rationale was rated on a five-point Likert scale across three criteria:

- **Fluency**: grammatical correctness and linguistic naturalness;
- **Coherence**: internal logical consistency and reasoning continuity;
- **Relevance**: contextual alignment with the original dialogue situation.

Final scores were obtained by averaging ratings across annotators. The complete results are presented in Table 6.

High Linguistic Quality and Reasoning Consistency Across both datasets, the generated rationales consistently achieve high Fluency and Coherence scores. Motivation and Emotional Reaction obtain the strongest overall performance, reaching Fluency scores between 4.58 and 4.76 and Coherence scores above 4.28. These results indicate that the Logic Policy can generate natural language explanations that remain logically structured while capturing latent psychological states and causal dependencies. Importantly, no cognitive dimension exhibits severe degradation in coherence, suggesting that the generated reasoning chains maintain stable internal consistency and provide reliable supervisory signals for downstream policy optimization.

Challenges in Future Consequence Prediction A consistent pattern emerges in the Relevance metric. Rationales describing historical conditions (Prerequisite), internal intentions (Motivation), and emotional outcomes (Emotional Reaction) maintain strong contextual alignment, with scores ranging from 3.75 to 3.95. In contrast, Consequence exhibits noticeably lower relevance scores (3.09 on RECCON-CG and 3.23 on ECF-CG).

We attribute this discrepancy to the inherently open-ended nature of future event prediction. Unlike retrospective causal reasoning, which is constrained by observed dialogue evidence, consequence generation requires forecasting plausible future developments under substantial uncertainty. Nevertheless, the relatively high Fluency and Coherence scores indicate that these future-oriented rationales remain logically plausible even when exact contextual specificity is difficult to guaran-

Dataset	Methods	Linguistic Quality			Error Avoidance			
		Flu. ↑	Coh. ↑	Rel. ↑	C.I. ↑	F.E. ↑	C.H. ↑	L.O. ↑
RECCON-CG	**Ours**	**4.990**	4.055	**3.745**	**4.785**	**4.160**	**4.280**	**4.750**
	w/o AEA	4.985	3.955	3.595	4.695	4.080	4.105	4.575
	w/o MLP	4.970	**4.115**	3.635	4.640	4.000	4.175	4.590
ECF-CG	**Ours**	4.915	**3.990**	**3.425**	**4.810**	**3.885**	**4.090**	**4.370**
	w/o AEA	4.925	3.880	3.330	4.735	3.760	3.795	4.265
	w/o MLP	**4.945**	3.865	3.205	4.770	3.670	3.780	4.115

Table 8: Comprehensive human evaluation results on RECCON-CG and ECF-CG. For each model variant, 200 randomly sampled test instances were independently evaluated by three annotators under a double-blind protocol. Higher scores indicate better generation quality and stronger resistance to shortcut reasoning failures.

tee.

Robust Cross-Dataset Generalization The evaluation results remain remarkably consistent across RECCON-CG and ECF-CG despite differences in dialogue sources and annotation protocols. For example, Motivation achieves Coherence scores of 4.34 and 4.28 on RECCON-CG and ECF-CG respectively, while Emotional Reaction reaches Fluency scores of 4.65 and 4.76. Such consistency suggests that the Logic Policy captures transferable commonsense reasoning patterns rather than dataset-specific artifacts. Consequently, the generated cognitive rationales constitute stable and trustworthy reward signals for reinforcement learning across diverse conversational settings.

C Comprehensive Qualitative Case Analysis

To provide an interpretable understanding of the behavioral differences between conventional supervised fine-tuning and our reward-guided alignment framework, we present four representative examples sampled from the held-out test set (Table 7). These examples were selected because they expose distinct shortcut reasoning patterns frequently observed in standard language models, including temporal inversion, emotion over-attribution, contextual fabrication, and lexical over-extrapolation.

Cases 1 and 2 reveal the limitations of purely likelihood-driven learning in maintaining temporal causality and emotional calibration. In Case 1, the baseline incorrectly reverses the causal direction of the dialogue and attributes the speaker's hesitation to the listener's subsequent reaction, violating the temporal ordering of events. In Case 2, the baseline assigns an explicit anger state to an otherwise factual utterance, demonstrating an over-sensitivity bias toward emotionally charged interpretations. By contrast, our framework correctly preserves causal chronology and avoids introducing unsupported emotional states. These improvements indicate that the Multi-dimensional Logic Policy and the asymmetric emotion alignment jointly constrain reasoning trajectories toward evidence-supported explanations.

Cases 3 and 4 further demonstrate the framework's robustness against hallucinated reasoning. In Case 3, the baseline invents an unobserved premise ("initial refusal") to artificially complete a causal chain. In Case 4, lexical cues associated with the word "crashed" trigger an exaggerated interpretation involving a car accident despite the absence of such evidence in the dialogue context. Our framework successfully suppresses both forms of hallucination by grounding generation in dialogue-consistent commonsense reasoning. The resulting explanations remain faithful to the available evidence while preserving causal completeness.

D Comprehensive Human Evaluation on Generative Quality and Error Avoidance

Automatic metrics alone are insufficient for assessing whether generated explanations genuinely avoid shortcut reasoning and hallucinated causal chains. Therefore, we conducted a comprehensive human evaluation comparing the full framework with two ablated variants.

For each model and each dataset, we randomly sampled 200 generated explanations from the test set. All samples were anonymized and randomly shuffled. Evaluation was performed under a double-blind protocol by three senior graduate stu-

dents experienced in natural language processing research. Annotators were unaware of model identities, ablation settings, and experimental hypotheses throughout the evaluation process.

Each explanation was independently scored on a continuous five-point scale across seven dimensions. The first three dimensions evaluate linguistic quality:

- **Fluency (Flu.)**
- **Coherence (Coh.)**
- **Relevance (Rel.)**

The remaining four dimensions explicitly measure resistance to shortcut reasoning behaviors:

- **Causal Inversion (C.I.)**
- **Forced Emotion (F.E.)**
- **Contextual Hallucination (C.H.)**
- **Lexical Over-extrapolation (L.O.)**

For all dimensions, higher scores indicate superior generation quality and stronger robustness against the corresponding reasoning error. Final results were obtained by averaging annotator scores and are reported in Table 8.

Effectiveness of the Complete Framework The full framework consistently achieves the strongest overall performance across both datasets. Improvements are particularly evident in the error-avoidance dimensions, where our model obtains the highest scores for Causal Inversion, Contextual Hallucination, and Lexical Over-extrapolation. These findings suggest that the proposed DRL objective effectively discourages shortcut reasoning and promotes dialogue-grounded causal explanations.

Contribution of the Multi-dimensional Logic Policy (MLP) Removing MLP (w/o MLP) leads to systematic degradation in logical faithfulness. On RECCON-CG, the Causal Inversion score decreases from 4.785 to 4.640, while Contextual Hallucination drops from 4.280 to 4.175. Similar trends are observed on ECF-CG. These results indicate that MLP plays a crucial role in enforcing causal consistency and preventing unsupported inference during generation.

Contribution of the Asymmetric Emotion Alignment (AEA) Ablating AEA (w/o AEA) primarily affects emotion-related robustness. The Forced Emotion score decreases on both datasets, confirming that the asymmetric masking mechanism effectively suppresses the tendency to over-attribute emotional states when explicit emotional evidence is absent. Furthermore, the decline in relevance scores suggests that emotionally calibrated supervision also contributes to maintaining contextually appropriate explanations.

Overall, the complementary strengths of AEA and MLP validate the design principle underlying our framework: reliable emotion-cause generation requires simultaneous control of logical faithfulness and emotional consistency.

E Annotation Protocol and Reproducibility

To facilitate reproducibility, all evaluation guidelines, annotation rubrics, prompt templates used by the Multi-dimensional Logic Policy, and implementation details are provided in the supplementary materials. The complete training and inference codebase, together with processed data splits and evaluation scripts, will be publicly released upon acceptance.

All human evaluations followed a double-blind annotation protocol. Annotators received only anonymized model outputs and corresponding dialogue contexts, without access to model identities, experimental conditions, or expected outcomes. Disagreements were resolved through independent scoring aggregation rather than post-hoc discussion to minimize confirmation bias and preserve annotation objectivity.