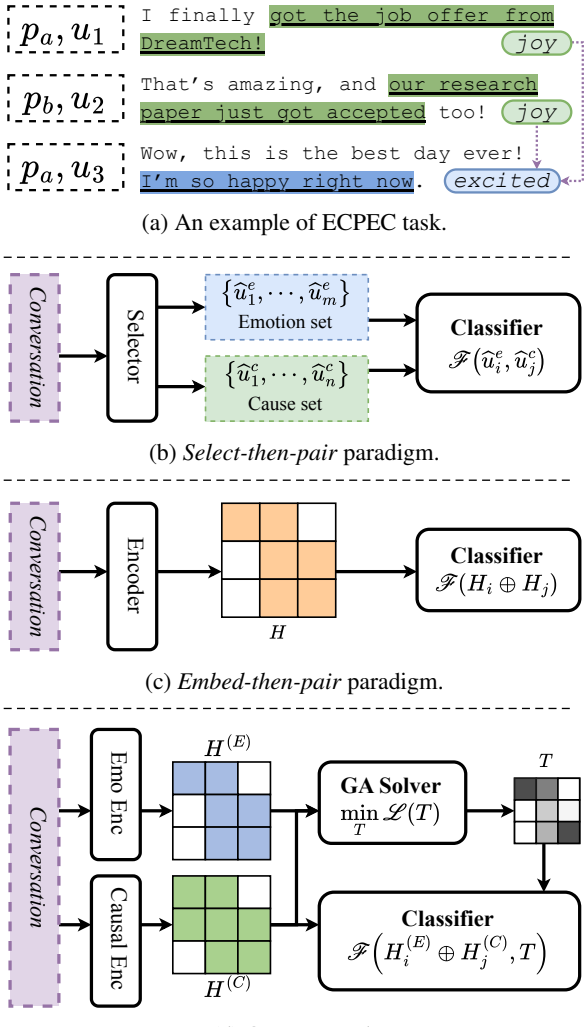
Emotion–Cause Pair Extraction in Conversations via Semantic Decoupling and Graph Alignment

Tianxiang Ma*, **Weijie Feng***†, **Xinyu Wang, Zhiyong Cheng**
Hefei University of Technology
matianxiang@mail.hfut.edu.cn, wjfeng@hfut.edu.cn

Abstract

Emotion-Cause Pair Extraction in Conversations (ECPEC) aims to identify the set of causal relations between emotion utterances and their triggering causes within a dialogue. Most existing approaches formulate ECPEC as an independent pairwise classification task, overlooking the distinct semantics of emotion diffusion and cause explanation, and failing to capture globally consistent many-to-many conversational causality. To address these limitations, we revisit ECPEC from a semantic perspective and seek to disentangle emotion-oriented semantics from cause-oriented semantics, mapping them into two complementary representation spaces to better capture their distinct conversational roles. Building on this semantic decoupling, we naturally formulate ECPEC as a global alignment problem between the emotion-side and cause-side representations, and employ optimal transport to enable many-to-many and globally consistent emotion-cause matching. Based on this perspective, we propose a unified framework SCALE that instantiates the above semantic decoupling and alignment principle within a shared conversational structure. Extensive experiments on several benchmark datasets demonstrate that SCALE consistently achieves state-of-the-art performance. Our codes are released at https://github.com/CoCoSphere/SCALE.



(a) An example of ECPEC task.

(b) *Select-then-pair* paradigm.

(c) *Embed-then-pair* paradigm.

(d) Our proposal.

Figure 1: Comparison between the existing ECPEC paradigm (b-c) and SCALE (d).

1 Introduction

Emotion-Cause Pair Extraction in Conversations (ECPEC) aims to identify the set of causal relations between emotion utterances and their triggering causes within a dialogue, which often exhibit complex and many-to-many dependencies. Figure 1a illustrates a representative example, where the emotion utterance u_3 is jointly triggered by multiple preceding utterances u_1 and u_2. Unlike emotion recognition in conversation (ERC) (Wang et al., 2024b; Fu et al., 2023; Majumder et al., 2019), which focuses on assigning discrete emotion labels to individual utterances (Fu et al., 2021; Hu et al., 2021), ECPEC provides a causal perspective for dialogue understanding, enabling more fine-grained analysis of emotional dynamics. As highlighted by Poria et al. (2021), ECPEC has broad applica-

*Equal Contribution.
†Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12697–12711
July 2-7, 2026 ©2026 Association for Computational Linguistics

bility across multiple domains, including dialogue systems (Rashkin et al., 2019; Zhong et al., 2020), conversational recommendation (Liang et al., 2024; Li et al., 2018), mental health analysis (Cambria et al., 2018; Pontiki et al., 2016), and social media opinion mining (Alexander Pak and Patrick Paroubek, 2010; Liu, 2022).

Early studies on ECPEC predominantly followed the *select-then-pair* paradigm (Ding et al., 2020; Wang et al., 2023a), which independently identifies candidate emotion utterances and cause utterances before pairing them through heuristic or classifier-based matching, as shown in Figure 1c. While intuitive and easy to integrate with existing ERC models (Gao et al., 2023; Nguyen et al., 2024; Ghosal et al., 2019), this pipeline is prone to error propagation during candidate selection and fails to fully exploit contextualized utterance representations. To alleviate these issues, subsequent studies shifted towards the *embed-then-pair* paradigm (An et al., 2023; Li et al., 2023; Jeong and Bak, 2023; Wang et al., 2024a), where utterance embeddings are directly concatenated and classified as emotion-cause pairs in an end-to-end manner. Although this paradigm better leverages utterance-level semantics and avoids explicit candidate construction, it still treats emotion-cause inference as a collection of independent pairwise decisions.

Despite their procedural differences, existing ECPEC approaches share two fundamental ***limitations***. ***L1)*** Most methods encode emotion-related and cause-related information within a unified representation space or interaction structure, implicitly assuming that emotion diffusion and cause explanation follow homogeneous relational patterns. However, in real conversations, emotional states tend to propagate through contextual and speaker-dependent dynamics, whereas causes are grounded in explanatory and often asymmetric dependencies. Conflating these distinct semantics obscures their respective roles in conversational causality. ***L2)*** Existing methods typically formulate ECPEC as independent one-to-one pair classification with binary judgments. Such pairwise formulations are inherently inadequate for modeling globally consistent many-to-many causal structures, where multiple interdependent causes may jointly trigger an emotion and a single cause may influence multiple emotional outcomes.

To address these limitations, we revisit ECPEC from a semantic perspective and argue that emotion diffusion and cause explanation, while grounded in the same conversational structure, should be characterized by different semantic focuses. Rather than duplicating dialogue structures or enforcing task-level separation, we seek to disentangle emotion-oriented and cause-oriented semantics by mapping them into two complementary representation spaces induced from a shared conversation graph. Building on this semantic decoupling, we naturally formulate ECPEC as a global alignment problem between emotion-side and cause-side representations, which enables holistic reasoning over many-to-many emotion–cause relations. Based on this perspective, we propose **SCALE** (**S**emantic **C**ausal **AL**ignment for **E**CPEC), a unified framework that instantiates semantic decoupling and global alignment within conversational contexts. Extensive experiments on multiple benchmark datasets demonstrate that SCALE consistently outperforms existing state-of-the-art approaches. Overall, the main contributions of this work are summarized as follows:

- We revisit ECPEC from a semantic perspective and highlight the necessity of disentangling emotion diffusion and cause explanation while preserving shared conversational structure.
- We propose SCALE, a unified framework that induces emotion-side and cause-side representations from a shared conversation graph and formulates ECPEC as a global alignment problem to support many-to-many and globally consistent inference.
- Extensive experiments on several public ECPEC benchmarks demonstrate that SCALE consistently achieves state-of-the-art performance.

2 Methodology

Formally, given a conversation $\mathcal{C} = \{(u_1, p_{\pi(u_1)}), \cdots, (u_N, p_{\pi(u_N)})\}$ consisting of N utterances, each utterance u_j is associated with a speaker $p_{\pi(u_j)}$, where π denotes a mapping from an utterance u_i to the index of its corresponding speaker. In ECPEC, the goal is to identify a set of emotion-cause pairs $\mathcal{P} = \{(u_e, u_c) \mid u_e \text{ is caused by } u_c\}$ that characterizes the underlying conversational causality. To address the limitations of existing approaches, we propose a framework termed SCALE that provides a unified solution for the ECPEC task.

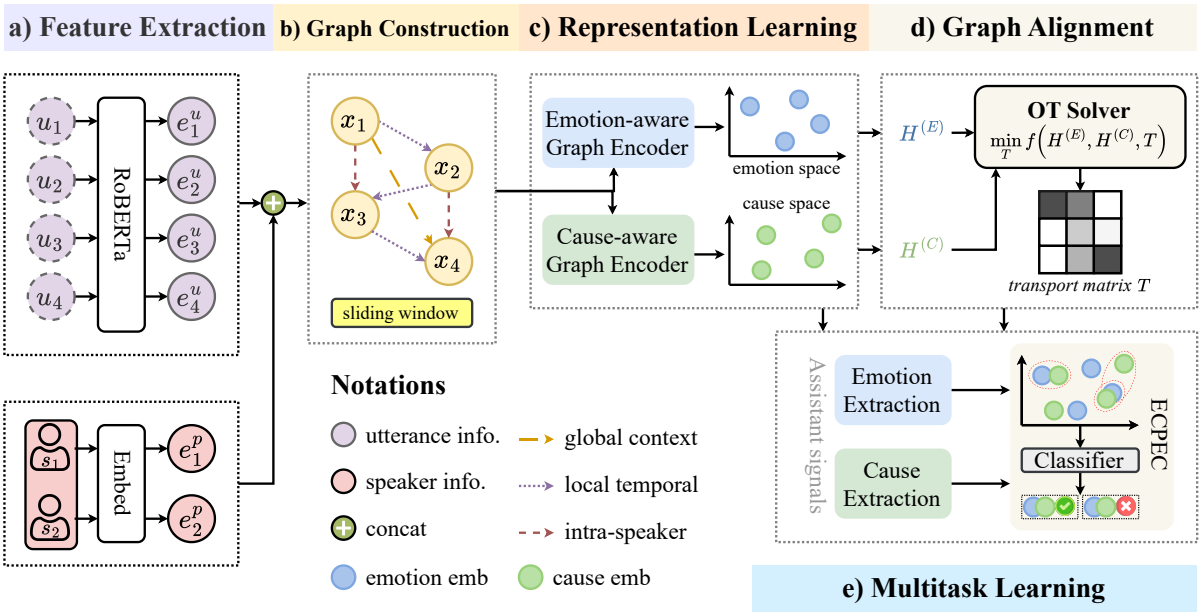


Figure 2: Overall architecture of our SCALE.

An overview of the proposed SCALE is illustrated in Figure 2. Specifically, each conversation is first encoded into utterance-level representations and organized as a conversation graph (§2.1). Then, SCALE induces two complementary semantic views of the dialogue, namely emotion-oriented and cause-oriented representations, by applying semantic-specific graph encoding mechanisms (§2.2). To explicitly model the correspondence between emotional utterances and their underlying causes, SCALE further formulates emotion-cause inference as a global alignment problem between the emotion-side and cause-side representations, which is solved via an optimal transport framework to enable many-to-many and globally consistent matching (§2.3). All components are optimized under a unified learning objective, where emotion extraction and cause extraction are introduced as auxiliary supervision signals to facilitate representation learning and ultimately improve ECPEC performance (§2.4).

2.1 Encode Conversation as a Graph

To establish relationships between utterances while capturing both inter- and intra-speaker dependencies (Ghosal et al., 2019; Li et al., 2023; Gao et al., 2023), we represent each conversation as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$. Each node $v_i \in \mathcal{V}$ corresponds to an utterance u_i, edges $e_{ij} \in \mathcal{E}$ are constructed according to three types of dependencies, and $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix.

Nodes. Each utterance u_i is represented as a node v_i, initialized with a contextual embedding $\boldsymbol{x}_i^u \in \mathbb{R}^{d_u}$ obtained from a pretrained RoBERTa model (Liu et al., 2019). To incorporate speaker information, the corresponding speaker $p_{\pi(u_i)}$ is first represented as a one-hot vector and then projected into a speaker embedding $\boldsymbol{x}^s_{\pi(u_i)} \in \mathbb{R}^{d_s}$ through a learnable embedding layer. The final speaker-aware utterance-level feature is defined as:

$$\boldsymbol{x}_i = \boldsymbol{x}_i^u \oplus \boldsymbol{x}^s_{\pi(u_i)}, \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{d_h}$ is the representation of node v_i, $\oplus$ denotes the concatenation operation.

Edges. We establish three types of relationships between utterance nodes to capture both global and local contextual dependencies in the conversation graph $\mathcal{G}$. The global contextual edge (v_i, v_j) captures long-range semantic dependencies between utterance nodes v_i and v_j. Such an edge exists if $\cos(\boldsymbol{x}_i, \boldsymbol{x}_j) + 1 > \tau_s$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the node features of v_i and v_j, respectively, and τ_s is a similarity threshold. The local contextual edge models short-range emotional dynamics. An edge (v_i, v_j) exists if $|i - j| \leq W$, where W denotes the size of the sliding temporal window. The intra-speaker edge captures speaker-specific emotional consistency. We connect nodes v_i and v_j if $\pi(u_i) = \pi(u_j)$ and $i \neq j$.

Edge weights. We initialize edge weights according to the corresponding dependency type. For

global contextual edges, we initialize weights based on utterance-level semantic similarity:

$$e_{ij} = \frac{\cos(\boldsymbol{x}_i, \boldsymbol{x}_j) + 1}{2}. \tag{2}$$

For local temporal edges, weights decay exponentially with conversational distance:

$$e_{ij} = \exp(-|i-j|/\tau_{\mathrm{e}}), \tag{3}$$

where τ_{e} controls temporal sensitivity. For intra-speaker edges, weights reflect speaker-specific emotional consistency across turns:

$$e_{ij} = \frac{\exp(-|i-j|/\tau_{\mathrm{e}}) + 1}{2}. \tag{4}$$

All initialized edge weights are integrated into a single weighted adjacency matrix $\boldsymbol{A}$, which is treated as learnable and jointly optimized with other model parameters.

2.2 Graph Representation Learning

Given the conversation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$ constructed for emotion-side and cause-side modeling, respectively, along with the same feature matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d_h}$, we employ two independent graph encoders, GNN_E and GNN_C, to learn two sets of node representations in different semantic spaces:

$$\boldsymbol{H}^{(S)} = \mathrm{GNN}_S\left(\mathcal{G}, \boldsymbol{X}\right), \tag{5}$$

where $S \in \{E, C\}$ denotes the emotion or cause semantic space, $\boldsymbol{H}^{(E)} \in \mathbb{R}^{N \times d_h}$ and $\boldsymbol{H}^{(C)} \in \mathbb{R}^{N \times d_h}$ denote the resulting emotion-aware and cause-aware node representations. Following Veličković et al. (2018), at each encoder layer, node representations $\boldsymbol{h}^{(S)} \in \boldsymbol{H}^{(S)}$ are updated via attention-weighted message passing:

$$\begin{aligned} \psi_{ij}^{(S)} &= \phi^{(S)}\left(\boldsymbol{h}_i^{(S)}, \boldsymbol{h}_j^{(S)}\right) A_{ij}, \\ \alpha_{ij}^{(S)} &= \mathrm{softmax}_{j \in \mathcal{N}(i)}\left(\psi_{ij}^{(S)}\right), \\ \boldsymbol{h}_i^{(S)} &= \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(S)} \boldsymbol{W}^{(S)} \boldsymbol{h}_j^{(S)}, \end{aligned} \tag{6}$$

where $\mathcal{N}(i)$ denotes the neighborhood of node v_i in $\mathcal{G}$, $\alpha_{ij}^{(S)}$ is the normalized attention coefficient, $\boldsymbol{W}^{(S)} \in \mathbb{R}^{d_h \times d_h}$ is a learnable matrix, and $\phi^{(S)}(\cdot)$ is a learnable scoring function. Importantly, the attention coefficient $\alpha_{ij}^{(S)}$ learned by each encoder can be viewed as a semantic-specific edge weight between v_i and v_j over the shared graph. Specifically, we define a task-specific weighted adjacency matrix $\boldsymbol{A}^{(S)} \in \mathbb{R}^{N \times N}$ as

$$A_{ij}^{(S)} = \alpha_{ij}^{(S)}, \tag{7}$$

which captures the relative importance of edge (v_i, v_j) under semantic space S. As a result, the emotion-aware and cause-aware encoders implicitly induce two refined adjacency structures, denoted as $\boldsymbol{A}^{(E)}$ and $\boldsymbol{A}^{(C)}$, respectively.

2.3 Graph Alignment via Optimal Transport

To capture the many-to-many relations inherent in emotion-cause pair extraction, we formulate ECPEC as a global alignment problem between emotion-aware and cause-aware representations. Formally, given the emotion representations $\boldsymbol{H}^{(E)} = \{\boldsymbol{h}_i^{(E)}\}_{i=1}^N$ and the cause representations $\boldsymbol{H}^{(C)} = \{\boldsymbol{h}_i^{(C)}\}_{i=1}^N$, our goal is to learn a transport plan $\boldsymbol{T} \in \mathbb{R}^{N \times N}$, where each entry $T_{ij} \geq 0$ indicates the soft correspondence strength between the i-th emotion representation $\boldsymbol{h}_i^{(E)}$ and the j-th cause representation $\boldsymbol{h}_j^{(C)}$. Unlike independent pairwise scoring, the alignment is learned globally over the entire matrix $\boldsymbol{T}$, such that multiple causes can be jointly associated with the same emotion and the correspondences across different emotion-cause pairs are mutually constrained. The alignment matrix $\boldsymbol{T}$ can be directly interpreted as a soft emotion–cause pairing matrix, where $\boldsymbol{T_{ij}}$ measures the association strength between emotion utterance i and cause utterance j.

Alignment objective. We seek to learn the alignment matrix $\boldsymbol{T}$ by minimizing the following objective with respect to $\boldsymbol{T}$:

$$\begin{aligned} \min_{\boldsymbol{T} \geq 0} \mathcal{L}(\boldsymbol{T}) = \alpha \left\langle \boldsymbol{C}_{\text{attr}}, \boldsymbol{T} \right\rangle \\ + (1-\alpha)\mathcal{L}_{\text{struct}}(\boldsymbol{T}) \end{aligned} \tag{8}$$

where $\boldsymbol{C}_{\text{attr}} \in \mathbb{R}^{N \times N}$ denotes the attribute-level cost matrix, $\mathcal{L}_{\text{struct}}(\boldsymbol{T})$ denotes a structure consistency term. $\boldsymbol{C}_{\text{attr}}$ measures semantic compatibility between emotion representation $\boldsymbol{h}_i^{(E)} \in \boldsymbol{H}^{(E)}$ and cause representation $\boldsymbol{h}_j^{(C)} \in \boldsymbol{H}^{(C)}$, with each entry defined as:

$$\boldsymbol{C}_{\text{attr}}(i,j) = 1 - \cos\left(\boldsymbol{h}_i^{(E)}, \boldsymbol{h}_j^{(C)}\right), \tag{9}$$

where smaller values indicate higher semantic affinity. The structure-level term $\mathcal{L}_{\text{struct}}(\boldsymbol{T})$ is defined

as a structure consistency loss that measures the discrepancy between relational patterns encoded in the emotion-side and cause-side dialogue graphs. Specifically, it is formulated as:

$$\mathcal{L}_{\text{struct}}(\boldsymbol{T}) = \sum_{i,k,j,l} \left| A^{(E)}_{ik} - A^{(C)}_{jl} \right|^2 T_{ij} T_{kl}. \quad (10)$$

Therefore, minimizing $\mathcal{L}(\boldsymbol{T})$ corresponds to finding emotion–cause pairs that are both semantically reasonable and structurally coherent within the dialogue.

Optimization. The resulting objective $\mathcal{L}(\boldsymbol{T})$ is non-linear due to the quadratic structure consistency term $\mathcal{L}_{\text{struct}}(\boldsymbol{T})$. To efficiently minimize it in a differentiable manner, we adopt an entropy-regularized Sinkhorn scheme (Zeng et al., 2024) based on the standard fused Gromov-Wasserstein optimization strategy (Tang et al., 2023). Starting from a uniform initialization $\boldsymbol{T}^{(0)}$, we iteratively linearize the structure consistency term around the current solution and solve a sequence of entropic optimal transport subproblems. At iteration t, the linearized approximation of $\mathcal{L}_{\text{struct}}(\boldsymbol{T})$ induces an effective structure-aware cost matrix:

$$\boldsymbol{C}^{(t)}_{\text{struct}}(i,j) = \sum_{k,l} \left| \boldsymbol{A}^{(E)}_{ik} - \boldsymbol{A}^{(C)}_{jl} \right|^2 T^{(t)}_{kl}, \quad (11)$$

which leads to the following cost matrix used to construct a linear surrogate of $\mathcal{L}(\boldsymbol{T})$:

$$\boldsymbol{C}^{(t)} = \alpha \boldsymbol{C}_{\text{attr}} + (1-\alpha)\boldsymbol{C}^{(t)}_{\text{struct}}. \quad (12)$$

The updated alignment matrix $\boldsymbol{T}$ is iteratively updated via Sinkhorn normalization:

$$\boldsymbol{T}^{(t+1)} = \mathcal{S}\left(\exp(-\boldsymbol{C}^{(t)}/\varepsilon)\right), \quad (13)$$

where $\mathcal{S}$ denotes the Sinkhorn operator with standard marginal constraints, and ε is the entropy regularization coefficient controlling the smoothness of the transport plan. After convergence, we apply a row-wise softmax with temperature τ_{r} to emphasize dominant alignments:

$$\widetilde{\boldsymbol{T}} = \text{softmax}(\boldsymbol{T}/\tau_{\text{r}}). \quad (14)$$

The resulting $\widetilde{\boldsymbol{T}}$ can be interpreted as a normalized alignment distribution, which reflects potential many-to-many associations between emotions and causes within each dialogue.

2.4 Multitask Joint Learning

SCALE adopts a multitask joint learning framework, where all task-specific objectives are optimized simultaneously with a shared encoder and task-specific prediction heads.

Emotion-Cause Pair Prediction. Given the row-normalized alignment matrix $\widetilde{\mathbf{T}}$, each entry $\widetilde{T}_{ij}$ represents the global correspondence strength between an emotion utterance u_i and a candidate cause utterance u_j. To incorporate local discriminative evidence, we compute a pairwise matching score by applying a lightweight classifier to the concatenated emotion-aware and cause-aware representations of each utterance pair:

$$s_{ij} = \mathcal{F}_{\text{ECPEC}}\left(\left[\boldsymbol{H}^{(E)}_i, \boldsymbol{H}^{(C)}_j\right]\right), \quad (15)$$

where $\mathcal{F}_{\text{ECPEC}}$ is implemented as a lightweight MLP with a sigmoid output layer. The final prediction score for emotion-cause pairs is obtained by combining global alignment and local evidence:

$$\widehat{y}^{(\text{ECPEC})}_{ij} = \beta \widetilde{T}_{ij} + (1-\beta)\, s_{ij}, \quad (16)$$

where β controls the relative contribution of global correspondence and local evidence.

Emotion and Cause Extraction. We perform utterance-level emotion extraction (EE) and cause extraction (CE) using two parallel lightweight classifiers:

$$\begin{aligned} \widehat{\boldsymbol{Y}}^{(E)} &= \mathcal{F}_{\text{EE}}(\boldsymbol{H}^{(E)}), \\ \widehat{\boldsymbol{Y}}^{(C)} &= \mathcal{F}_{\text{CE}}(\boldsymbol{H}^{(C)}), \end{aligned} \quad (17)$$

where $\mathcal{F}_{\text{EE}}$ and $\mathcal{F}_{\text{CE}}$ are implemented as lightweight MLPs with softmax output layers.

Joint Optimization. The learning of SCALE is performed by minimizing $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{ECPEC}} + \lambda_{EE}\mathcal{L}_{\text{EE}} + \lambda_{CE}\mathcal{L}_{\text{CE}}, \\ \mathcal{L}_{\text{EE}} &= \text{CE}(\widehat{\boldsymbol{Y}}^{(E)}, \boldsymbol{Y}^{(E)}), \\ \mathcal{L}_{\text{CE}} &= \text{CE}(\widehat{\boldsymbol{Y}}^{(C)}, \boldsymbol{Y}^{(C)}), \end{aligned} \quad (18)$$

where $\mathcal{L}_{\text{EE}}$ and $\mathcal{L}_{\text{CE}}$ are cross entropy losses, λ_{EE} and λ_{CE} controlling their contributions. The ECPEC loss $\mathcal{L}_{\text{ECPEC}}$ is defined as:

$$\mathcal{L}_{\text{ECPEC}} = \mathcal{L}_{\text{pair}} + \lambda_{\text{OT}}\, \mathcal{L}_{\text{OT}}. \quad (19)$$

The pair-level supervision term is computed with binary cross-entropy over the predicted pair scores:

$$\mathcal{L}_{\text{pair}} = \text{BCE}\left(\widehat{y}^{(\text{ECPEC})}, y\right), \quad (20)$$

Dataset	#Dlg.	#Utt.	#Pairs	Partition
RECCON-DD	1,106	11,104	5,861	75/5/20
RECCON-IE	16	665	1154	test only
ECF	1,374	13,619	9,794	70/10/20

Table 1: Dataset statistics.

where $y \in \{0, 1\}$ is the ground-truth label for the emotion-cause pair. The OT consistency regularizer encourages the local pair-wise prediction to match the OT-derived alignment score:

$$\mathcal{L}_{\text{OT}} = D_{\text{KL}}\Big(\text{Bern}(s_{ij}) \| \text{Bern}(\widetilde{T}_{ij})\Big). \quad (21)$$

Here D_{KL} denotes the Kullback–Leibler divergence, Bern is a Bernoulli distribution, s_{ij} is the local pair-wise prediction score in Equation (15), and $\widetilde{T}_{ij} \in \boldsymbol{T}$ denotes the corresponding OT-derived alignment score.

3 Experiments

To comprehensively evaluate the proposed SCALE, we formulate the following *Research Questions* to guide our experiments:
RQ1: How does SCALE compare with existing state-of-the-art approaches on ECPEC datasets?
RQ2: How robust is SCALE in handling multi-cause scenarios compared to prior methods?
RQ3: How do key components of SCALE contribute to ECPEC performance?
RQ4: Can SCALE provide interpretable emotion–cause alignments and meaningful insights through qualitative analysis?

3.1 Experimental Setups

Datasets. We evaluate our method on three representative datasets, described below. RECCON (Poria et al., 2021) is a widely used benchmark that consists of two subsets: **RECCON-DD**, annotated from DailyDialog (Li et al., 2017), serves as the main corpus for model training and evaluation, while **RECCON-IE**, annotated from IEMOCAP (Busso et al., 2008), is a smaller subset used exclusively to test the generalization ability. **ECF** (Wang et al., 2023a) is a multimodal benchmark derived from the sitcom *Friends*, providing annotated emotion–cause pairs across text, audio, and visual modalities, with many emotions triggered by multiple utterances. The dataset statistics are illustrated in Table 1.

Baselines. We compare our method against seven representative baselines, which can be broadly grouped into three categories:
1) Method based on general pre-trained model: Following Poria et al. (2021), we adopt a pretrained RoBERTa (Liu et al., 2019) model with a classification layer to identify emotion–cause pairs, which we refer to as **RECCON**, serving as a benchmark baseline.
2) Methods with sequential modeling: **MECPE-2steps** (Wang et al., 2023a) adopts a two-step pipeline that first extracts candidate emotions and causes sets by a shared BiLSTM and then filters valid pairs with another BiLSTM. **PRG-MoE** (Jeong and Bak, 2023) constructs relational graph with a mixture-of-experts, where a gating network aggregates diverse relational patterns. **Joint-Xatt** (Li et al., 2023) utilize cross-attention to model emotion–cause dependencies.
3) Methods based on graph modeling: **Joint-GCN** (Li et al., 2023) extends Joint-Xatt by replacing cross-attention with graph convolutional network to model inter-utterance relations. **MRC** (Liu et al., 2023) reformulates ECPEC as machine reading comprehension and employs GNNs to encode dialogue structure. **CENTER** (Wang et al., 2024a) builds a center event–aware graph with contrastive objectives for pair-level discrimination. **Multi-CauseNet** (Ma et al., 2025) leverages a multimodal graph and employs Graph Attention Networks with temporal attention to prioritize relevant features across text, audio, and video modalities.
4) Methods based on generative frameworks and LLMs: **GMEC** (Ju et al., 2025) transforms the extraction task into a generative question-answering paradigm, utilizing Large Language Models as implicit knowledge engines to capture both linguistic and visual cues.

Metrics. Following previous work (Xia and Ding, 2019), we adopt F1-score (F1), Precision (P), and Recall (R) as evaluation metrics.

Implementation Details. We derive textual features for each utterance using a pretrained RoBERTa model. Unless otherwise specified, hyperparameters are set as follows: the temporal window size $W = 5$, utterance and speaker embedding dimensions $d_u = 768$ and $d_s = 50$, similarity and decay parameters $\tau_{\text{s}} = 0.5$, $\tau_{\text{e}} = 2.0$, and $\tau_{\text{r}} = 1.0$, and weighting coefficients $\alpha = 0.8$, $\beta = 0.4$, $\varepsilon = 0.5$, $\lambda_{\text{EE}} = 0.2$, $\lambda_{\text{CE}} = 0.4$, and $\lambda_{\text{OT}} = 1.0$. Model training is conducted using the

Method	RECCON-DD			RECCON-IE			ECF		
	P	R	F1	P	R	F1	P	R	F1
RECCON	49.31	33.19	39.68	51.04	11.00	18.10	30.26	37.58	33.52
MECPE-2steps	49.34	47.37	48.34	27.31	6.30	10.24	57.64	48.72	52.71
PRG-MoE	**58.95**	55.67	57.26	51.95	20.02	28.90	47.11	55.27	50.86
Joint-Xatt	28.64	40.43	33.53	28.37	12.30	17.16	42.65	39.19	40.85
Joint-GCN	30.79	36.88	33.56	27.49	16.67	20.75	40.29	42.33	41.28
MRC	52.19	52.86	52.47	**59.59**	16.08	20.96	44.46	57.65	50.20
CENTER	47.39	46.88	47.13	34.92	24.38	28.71	47.90	43.65	44.75
MultiCauseNet	-	-	-	-	-	-	53.27	59.10	55.12
GMEC	55.97	50.46	53.07	46.79	20.24	28.25	**60.41**	50.03	54.73
Ours	56.31	**61.60**	**58.83**	42.54	**29.29**	**34.69**	55.01	**60.67**	**57.70**

Table 2: Comparison results of ECPEC task.

AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $1e-4$. We employ a ReduceLROnPlateau scheduler and early stopping based on validation performance. All experiments are implemented in PyTorch (Paszke et al., 2019) and executed on a single NVIDIA RTX 4090 GPU. For a consistent comparison, all baselines are reimplemented based on their publicly released code or original descriptions, and trained under the same experimental settings.

Method	RECCON-DD	RECCON-IE	ECF
RECCON	28.27	8.76	23.75
MECPE-2steps	33.61	11.48	28.71
PRG-MoE	37.84	21.62	33.71
Joint-Xatt	23.18	4.62	25.15
Joint-GCN	23.73	5.08	26.96
MRC	33.39	3.22	25.64
CENTER	34.09	9.73	28.20
SCALE	**38.33**	**25.33**	**35.55**

Table 3: Comparison of F1-scores on the multi-cause scenario.

3.2 Results and Discussion

3.2.1 Overall Performance (RQ1)

We evaluate all methods on the ECPEC task across three benchmarks, namely RECCON-DD, RECCON-IE, and ECF. As reported in Table 2, SCALE consistently achieves the highest recall and F1-score on all three datasets. On RECCON-DD, SCALE attains an F1 score of 58.83, outperforming the strongest baseline in terms of F1, PRG-MoE (57.26), by a relative improvement of $+2.7\%$. On the smaller and cross-domain RECCON-IE dataset, SCALE achieves the best F1 score of 34.69, surpassing the strongest baseline (28.90) by a substantial relative gain of $+20.0\%$, which highlights its strong generalization capability. Similarly, on ECF, SCALE delivers a notable relative improvement of $+9.5\%$ over MECPE-2steps (52.71) in terms of F1-score. We also observe that SCALE does not achieve the highest precision on most datasets. This behavior is consistent with the design of the soft optimal transport alignment, which encourages broader semantic matching between emotion and cause representations and therefore favors higher recall at the potential expense of precision.

3.2.2 Multi-Cause Study (RQ2)

As mentioned above, multi-cause scenarios introduce additional challenges for ECPEC. To evaluate model robustness under such settings, we therefore construct three multi-cause test subsets from RECCON-DD, RECCON-IE, and ECF, by selecting all dialogues in the original test splits where a single target emotion is annotated with two or more distinct causes. We then re-evaluate all models on these subsets to assess their robustness in capturing multiple causal triggers. As shown in Table 3, all models suffer from a notable performance drop, confirming the inherent difficulty of multi-cause prediction. Nevertheless, SCALE consistently achieves the best F1 scores across all three subsets. We attribute these improvements to the global alignment formulation in SCALE, which models emotion–cause relations as soft many-to-many correspondences between emotion-oriented and cause-oriented representations, enabling joint reasoning over dispersed causal evidence for each emotion.

	RECCON-DD	RECCON-IE	ECF
Full model	**58.83**	**34.69**	**57.70**
w/o SRL	56.72	31.22	55.77
w/o GA	55.26	29.08	53.82
w/o SRL & GA	53.66	28.18	52.54
w/o EE	58.44	34.43	57.43
w/o CE	57.99	34.07	57.11
w/o CE & EE	57.15	33.57	56.54

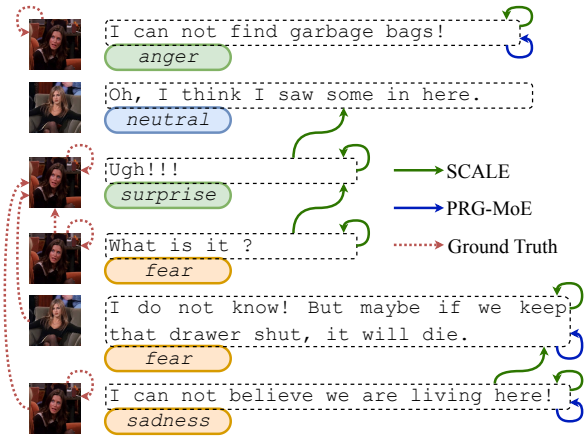
Table 4: Ablation study.

3.2.3 Ablation Study (RQ3)

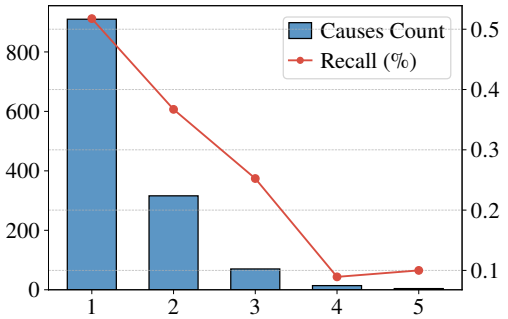
To verify the effectiveness of the key design principles in SCALE, we conduct ablation experiments by removing separated representation learning (SRL; see §2.2), global alignment (GA; see §2.3), and auxiliary supervision (i.e., EE and CE). Specifically, w/o SRL collapses emotion-oriented and cause-oriented encoders into a single graph encoder that learns a shared representation for all utterances, while w/o GA removes the alignment module and performs emotion–cause prediction based solely on independent pairwise scores. As shown in Table 4, removing either SRL or GA consistently degrades ECPEC F1 performance across all datasets, with a more pronounced drop observed when GA is disabled, underscoring the importance of soft many-to-many alignment for modeling complex emotion-cause relations. We further observe that auxiliary supervision is beneficial to ECPEC. Removing EE or CE individually leads to mild performance drops, while jointly removing both results in a more noticeable degradation, indicating that EE and CE provide complementary support for representation learning.

3.2.4 Qualitative Analysis (RQ4)

Case Study Figure 3a presents a qualitative comparison between SCALE and the strongest baseline PRG-MoE on a dialogue sampled from the ECF dataset. Among the seven ground-truth emotion–cause pairs, SCALE correctly predicts five, whereas PRG-MoE identifies only two, indicating a clear performance gap. For long-distance dependencies, such as (u_6, u_3) and (u_5, u_3), both models fail to recover the correct relations, suggesting that capturing causal cues separated by large conversational gaps remains challenging. In contrast, for short-distance dependencies, including (u_1, u_1), (u_3, u_3), (u_4, u_4), (u_4, u_3), and (u_5, u_5), SCALE achieves perfect predictions. Regarding multi-cause scenarios, SCALE successfully identifies the dual-cause emotion $[(u_4, u_3), (u_4, u_4)]$, but fails to capture both causes in $[(u_6, u_6), (u_6, u_3)]$. These observations suggest that while the global alignment mechanism enables flexible one-to-many reasoning, modeling long-range causal dependencies remains an open challenge.



(a) Qualitative comparison.



(b) Cause-number distribution and corresponding recall on ECF dataset.

Figure 3: Qualitative analysis.

Error Analysis. To further analyze the behavior of SCALE in multi-cause scenarios, we examine its performance on the ECF dataset. As shown in Figure 3b, most instances involve a single cause, while samples with multiple causes are increasingly scarce. Despite achieving the best overall performance on the multi-cause subset of ECF, SCALE exhibits decreasing recall as the number of causes increases, since recall requires all causes associated with an emotion to be correctly identified, indicating that exhaustive cause retrieval in complex multi-cause settings remains challenging.

3.2.5 Auxiliary Analysis

Performance on EE and CE. To provide additional context on the intermediate subtasks, we evaluate the performance of SCALE on emotion extraction (EE) and cause extraction (CE). As shown

Method	RECCON-DD		RECCON-IE		ECF	
	EE	CE	EE	CE	EE	CE
MECPE-2steps	71.30	65.81	42.52	44.49	**79.10**	**70.13**
PRG-MoE	73.86	-	57.29	-	71.82	-
Joint-Xatt	58.71	51.93	47.34	40.26	67.75	62.73
Joint-GCN	61.87	51.35	46.58	35.75	68.31	64.05
MRC	**75.49**	-	39.54	-	74.08	-
CENTER	68.32	-	49.61	-	67.07	-
SCALE	73.23	**67.87**	55.57	**54.90**	76.10	61.81

Table 5: Comparative results of EE and CE subtasks (F1-score) across three datasets.

Method	RECCON-DD	ECF
DeepSeek-V3.2	47.11	42.81
GPT-5.1 Instant	55.26	54.76
Gemini-3-pro-preview	56.08	55.42
SCALE	**58.83**	**57.70**

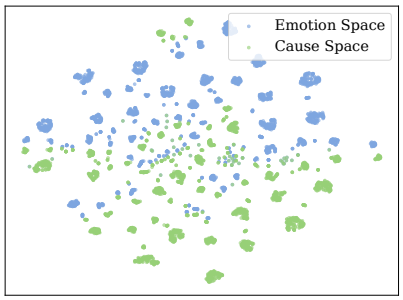
Table 6: Comparison with recent LLMs.

in Table 5, SCALE yields reasonable performance on both EE and CE across datasets, without relying on task-specific architectural designs. It is worth noting that SCALE is primarily optimized for the ECPEC objective, while EE and CE are incorporated as auxiliary supervision during training rather than standalone targets.

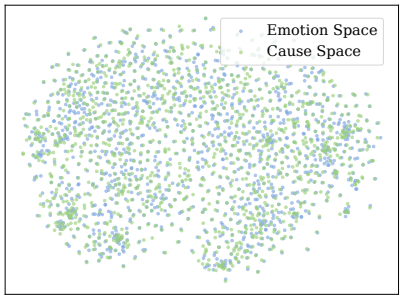
Comparison with recent LLMs. We compare SCALE with three recent LLMs,using a few-shot prompting strategy with four in-context examples and strict JSON output constraints to ensure fair evaluation. The complete prompt template and example dialogues are provided in Appendix B. As shown in Table 6, SCALE achieves higher F1-scores than the best-performing LLM on both datasets. These results suggest that explicit modeling of conversational structure and emotion–cause relations remains advantageous for ECPEC, even in the presence of strong prompt-based LLM baselines.

3.2.6 Visualization Analysis

We further visualize the latent features of the last hidden layer before the classifier using t-SNE, illustrating the distributions of emotion space and cause space for MECPE-2steps and SCALE. As illustrated in Figure 4b, the node features are uniformly intertwined, forming a dense cluster without clear decision boundaries. This phenomenon indicates a severe semantic confusion, where the model fails to explicitly distinguish whether an utterance serves as an emotion carrier or a cause



(a) t-SNE visualization of SCALE.



(b) t-SNE visualization of MECPE-2steps.

Figure 4: t-SNE visualization of emotion and cause representations.

trigger during the encoding stage. In contrast, Figure 4a demonstrates that SCALE maps emotion and cause features into distinct, highly discriminative subspaces, validating that dual-graph module achieves effective semantic decoupling. Furthermore, instead of forming a single massive cluster, our features spontaneously aggregate into dozens of compact, high-cohesion local clusters, suggesting that SCALE captures more fine-grained structural patterns.

4 Conclusion

In this paper, we proposed **SCALE**, a semantic alignment framework for emotion–cause pair extraction in conversations. By decoupling emotion-oriented and cause-oriented semantics and modeling their interactions through global alignment, SCALE reformulates ECPEC as a many-to-many reasoning problem over conversational structure. This design enables more holistic modeling of complex causal dependencies beyond independent pairwise prediction. Extensive experiments on three benchmark datasets demonstrate that SCALE consistently outperforms existing approaches, particularly in challenging multi-cause scenarios, validating its effectiveness and robustness.

Limitations

The proposed **SCALE** focuses on textual conversations and does not incorporate other modalities such as acoustic or visual signals. In real-world scenarios, emotion expression and causal cues may span multiple modalities, which could provide complementary information beyond text alone. While the alignment-based formulation of SCALE is in principle compatible with multimodal representations, extending the framework to fully multimodal ECPEC settings is beyond the scope of this work and remains an interesting direction for future research.

Additionally, emotion-cause relations in conversation are often ambiguous and context-dependent, and models trained on annotated datasets may inherit annotation bias or incomplete causal assumptions. Therefore, the proposed framework should be used as an assistive tool rather than a definitive explanation of emotion causality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. PA2023GDGP0109) and the Research Startup Fund of Hefei University of Technology (Grant No. JZ2023HGQA0470).

References

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, Valletta, Malta. European Language Resources Association (ELRA).

Jiaming An, Zixiang Ding, Ke Li, and Rui Xia. 2023. Global-view and speaker-aware emotion cause extraction in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3814–3823.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Shunjie Chen, Xiaochuan Shi, Jingye Li, Shengqiong Wu, Hao Fei, Fei Li, and Donghong Ji. 2022. Joint alignment of multi-task feature and label spaces for emotion cause pair extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6955–6965, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

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, Lin Gui, Yue Zhang, and Ruifeng Xu. 2021. Multi-task sequence tagging for emotion-cause pair extraction via tag distribution refinement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2339–2350.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. Consk-gcn: Conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22):4714.

Yingxue Gao, Huan Zhao, Yufeng Xiao, and Zixing Zhang. 2023. Gcformer: A graph convolutional transformer for speech emotion recognition. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, pages 307–313, Paris France. ACM.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, Hong Kong, China. Association for Computational Linguistics.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online. Association for Computational Linguistics.

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.

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.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Trans. Multimedia Comput. Commun. Appl.*

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP 2017*.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. Llm-redial: A large-scale dataset for conversational recommender systems created from user behaviors with llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8926–8939, Bangkok, Thailand. Association for Computational Linguistics.

Bing Liu. 2022. *Sentiment Analysis and Opinion Mining*. Springer Nature.

Chen Liu, Changyong Niu, Jinge Xie, Yuxiang Jia, and Hongying Zan. 2023. Emotion-cause pair extraction in conversations based on multi-turn mrc with position-aware gcn. In *2023 International Conference on Asian Language Processing (IALP)*, pages 25–30.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Junchi Ma, Hassan Nazeer Chaudhry, Farzana Kulsoom, Guihua Yang, Sajid Ullah Khan, Sujit Biswas, Zahid Ulalh Khan, and Faheem Khan. 2025. Multicausenet temporal attention for multimodal emotion cause pair extraction. *Scientific Reports*, 15.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Hermina Petric Maretic and Mireille EL Gheche. 2019. Got: An optimal transport framework for graph comparison. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. 2024. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9330–9339, Melbourne VIC Australia. ACM.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

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.

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, Florence, Italy. Association for Computational Linguistics.

Shruti Saxena and Joydeep Chandra. 2024. A survey on network alignment: Approaches, applications and future directions. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 8216–8224, Jeju, South Korea. International Joint Conferences on Artificial Intelligence Organization.

Konstantinos Skitsas, Karol Orłowski, Judith Hermanns, Davide Mottin, and Panagiotis Karras. 2023. Comprehensive evaluation of unrestricted graph alignment algorithms.

Jianheng Tang, Kangfei Zhao, and Jia Li. 2023. A fused Gromov-Wasserstein framework for unsupervised knowledge graph entity alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3320–3334, Toronto, Canada. Association for Computational Linguistics.

Huynh Thanh Trung, Nguyen Thanh Toan, Tong Van Vinh, Hoang Thanh Dat, Duong Chi Thang, Nguyen Quoc Viet Hung, and Abdul Sattar. 2020. A comparative study on network alignment techniques. *Expert Systems with Applications*, 140:112883.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

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, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024b. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Yejiang Wang, Yuhai Zhao, Zhengkui Wang, and Ling Li. 2023b. Galopa: Graph transport learning with optimal plan alignment. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

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.

Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable gromov-wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. 2024. Hierarchical multi-marginal optimal transport for network alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16660–16668.

Zhichen Zeng, Si Zhang, Yinglong Xia, and Hanghang Tong. 2023. Parrot: Position-aware regularized optimal transport for network alignment. In *Proceedings of the ACM Web Conference 2023*, pages 372–382, Austin TX USA. ACM.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Peican Zhu, Botao Wang, Keke Tang, Haifeng Zhang, Xiaodong Cui, and Zhen Wang. 2024. A knowledge-guided graph attention network for emotion-cause pair extraction. *Knowledge-Based Systems*, 286:111342.

A Related Work

A.1 Emotion-Cause Pair Extraction.

Research on emotion-cause analysis originated from the Emotion Cause Extraction (ECE) task (Lee et al., 2010; Gui et al., 2016; Li et al., 2018), which aims to identify the cause span corresponding to a given emotion. To overcome the limitation of requiring emotion annotation before cause extraction, Xia and Ding (2019) proposed the Emotion-Cause Pair Extraction (ECPE) task, which jointly extracts emotions and their causes, and inspired a line of subsequent studies (Ding et al., 2020; Chen et al., 2020; Fan et al., 2021).

While remarkable progress has been made in ECPE research, the majority of existing approaches (Fan et al., 2021; Chen et al., 2022; Zhu et al., 2024) are confined to document-level corpora, where emotions and their corresponding causes are expressed within a single, coherent narrative flow. Such clause-level formulations inherently neglect the distinctive characteristics of dialogues, such as speaker role alternation, intertwined emotional events, and long-range conversational dependencies, thereby highlighting the necessity of extending ECPE into conversational contexts.

A.2 Emotion-Cause Pair Extraction in Conversation.

Early efforts on conversational cause analysis began with RECCON (Poria et al., 2021), which focused on cause recognition rather than emotion-cause pair extraction. Li et al. (2023) formally introduced the ECPEC task and released the ConvECPE dataset, along with a two-step framework that explicitly models conversational properties such as context dependence and speaker interactivity. Subsequent works explored more sophisticated modeling, such as pair-relations-guided mixture-of-experts system PRG-MOE (Jeong and Bak, 2023), machine reading comprehension-based method MRC (Liu et al., 2023), global-view speaker-aware frameworks GSESE (An et al., 2023), and event-guided ECPEC frameworks CENTER (Wang et al., 2024a). Beyond text, multimodal ECPEC has been studied with the ECF dataset (Wang et al., 2023a) and improved by cross-modality interaction mechanisms such as HiLo (Li et al., 2024). Despite these advances, existing methods still follow the pairwise classification paradigm and thus fail to capture global many-to-many alignments between emotions and causes in dialogues, leaving robust causal modeling an open challenge.

A.3 Graph Alignment and Optimal Transport

Graph alignment aims to identify correspondences between nodes across related graphs (Saxena and Chandra, 2024; Skitsas et al., 2023; Trung et al., 2020), a problem widely studied in network analysis and data integration. As a global matching problem, it can be naturally addressed by Optimal Transport (OT) (Zeng et al., 2024; Wang et al., 2023b; Xu et al., 2019), which computes a global coupling between two sets that minimizes transportation cost and captures many-to-many correspondences under a global optimization objective (Zeng et al., 2023; Maretic and Gheche, 2019). This property makes OT particularly suitable for aligning emotions and causes in dialogues, where multiple candidates may coexist and local decisions can be insufficient. However, OT has not yet been explored in ECPEC, which motivates reformulating our task as a global graph alignment problem through the integration of dual graph learning and OT.

B LLMs Baselines

To improve transparency and reproducibility, the complete prompt template is provided in Figure 5.

C Key Hyperparameters Experiments

To further investigate the effects of key hyperparameters, we conduct additional experiments, as shown in Table 7. The results reveal that neither attribute-only nor structure-only alignment yields optimal performance. Notably, enforcing pure alignment ($\beta = 1$) results in a significant degradation in performance. Furthermore, the model maintains stable performance over a reasonably wide range of window sizes.

	P	R	F1
0.0	54.77	54.05	54.41
0.2	55.01	60.67	57.70
0.4	54.05	59.97	56.85
0.6	55.93	57.99	56.94
0.8	55.11	60.49	57.67
1.0	55.04	58.09	56.52

(a) Effect of α on performance.

	P	R	F1
0.0	56.01	57.04	56.52
0.2	55.53	58.95	57.18
0.4	55.01	60.67	57.70
0.6	55.39	57.14	56.25
0.8	56.30	43.41	49.02
1.0	41.25	10.58	16.84

(b) Effect of β on performance.

	P	R	F1
3	54.93	59.32	57.04
4	55.90	58.44	57.14
5	55.01	60.67	57.70
6	54.35	59.76	56.93
8	53.97	60.75	57.15
10	53.63	60.22	56.73

(c) Effect of $window\_size$ on performance.

Table 7: Hyperparameter experiments on the ECF dataset with respect to α, β, and window_size.

D Computational Complexity Analysis

Theoretically, for a conversation with N utterances and $|\mathcal{E}|$ edges, one GNN layer costs $O(|\mathcal{E}|d)$. With L layers and two encoders, encoding cost is $O(2L|\mathcal{E}|d)$, i.e., only a constant-factor increase. The Sinkhorn alignment operates on an $N \times N$ matrix with per-iteration cost $O(N^2)$. Since dialogue length is small ($N < 30$), the alignment remains lightweight. Empirical results in Table 8 compare parameter size, FLOPs, and peak memory usage, demonstrating that SCALE achieves competitive performance with substantially lower computational cost.

Model	**Params**	**FLOPs**	**Memory**
PRG-MoE	110M	∼220G	22G
GMEC	450M	∼360G	20G
SCALE	**8.2M**	**2.15G**	**7G**

Table 8: Comparison of Model Efficiency.

E Multimodal Extension

Although SCALE is not primarily designed for multimodal modeling, it can be easily extended to handle multimodal inputs. As the ECF dataset provides text, audio, and video modalities, we follow prior work by concatenating unimodal features as a simple fusion strategy, since multimodal fusion is not the main focus of this paper. As shown in Table 9, SCALE achieves consistent improvements over the text-only variant when additional modalities are incorporated, demonstrating its adaptability and robustness across multimodal settings.

Text	Audio	Video	F1
+	-	-	57.70
+	+	-	58.13
+	-	+	58.07
+	+	+	**58.63**

Table 9: Multimodal evaluation of SCALE on the ECF dataset.

You are a professional emotion analysis expert. Your task is to identify emotion-cause pairs from conversations.

Definition of emotion-cause pairs:
- Emotion utterance: A sentence expressing a certain emotion (e.g., anger, happiness, sadness)
- Cause utterance: A sentence containing the reason that triggered the emotion
- One emotion may have multiple causes, and one cause may trigger multiple emotions

Output requirements:
1. Output ONLY a JSON list: [[emotion_id, cause_id], ...]
2. emotion_id and cause_id are utterance numbers (starting from 1)
3. If no emotion-cause pairs exist, output an empty list: []
4. Do NOT output any explanation or extra text, ONLY output the JSON list

===== Example 1 =====
Conversation:
{example_1_dialogue}
Emotion-cause pairs: {example_1_output}
===== Example 2 =====
Conversation:
{example_2_dialogue}
Emotion-cause pairs: {example_2_output}
===== Example 3 =====
Conversation:
{example_3_dialogue}
Emotion-cause pairs: {example_3_output}
===== Example 4 =====
Conversation:
{example_4_dialogue}
Emotion-cause pairs: {example_4_output}

Figure 5: The system prompt template.