# Affective Reasoning at Utterance Level in Conversations: A Causal Discovery Approach

**Anonymous ACL submission**

## Abstract

The affective reasoning task is a set of emerging affect-based tasks in conversation, including Emotion Recognition in Conversation (ERC), Emotion-Cause Pair Extraction (ECPE), and Emotion-Cause Span Recognition (ECSR). Existing methods make various assumptions on the apparent relationship while neglecting the essential causal model due to the nonuniqueness of skeletons and unobservability of implicit causes. This paper settled down the above two problems and further proposed Conversational Affective Causal Discovery (CACD). It is a novel causal discovery method showing how to discover causal relationships in a conversation via designing a common skeleton and generating a substitute for implicit causes. CACD contains two steps: (i) building a common *centering one graph node* causal skeleton for all utterances in variable-length conversations; (ii) Causal Auto-Encoder (CAE) correcting the skeleton to yield causal representation through generated implicit causes and known explicit causes. Comprehensive experiments demonstrate that our novel method significantly outperforms the SOTA baselines in six affect-related datasets on the three tasks.

## 1 Introduction

Recently, increasing attempts (Xia and Ding, 2019; Bi and Liu, 2020; Turcan et al., 2021; Uymaz and Metin, 2022) have been made to explore the relationship between emotion and corresponding cause in texts. Hence, the conversation sentiment field has involved cause-related datasets (Poria et al., 2021; Feng et al., 2022) and studies (Gao et al., 2021; Zhao et al., 2022; Li et al., 2022). Until now, three affective reasoning tasks have been widely explored: (i)Emotion Recognition in Conversation (**ERC**) (Chen et al., 2018). (ii)Emotion-Cause Pair Extraction (**ECPE**) Xia and Ding (2019). (iii)Emotion-Cause Span Recognition (**ECSR**) Poria et al. (2021). These three tasks require proper handling of why emotions appear, so most published methods (Kumar and Jain, 2022; Chang et al., 2022; Bao et al., 2022) have strived at reasoning about the **explicit cause**. It is a concept that for the emotion in conversations, the corresponding cause is an utterance from the preceding dialogue.

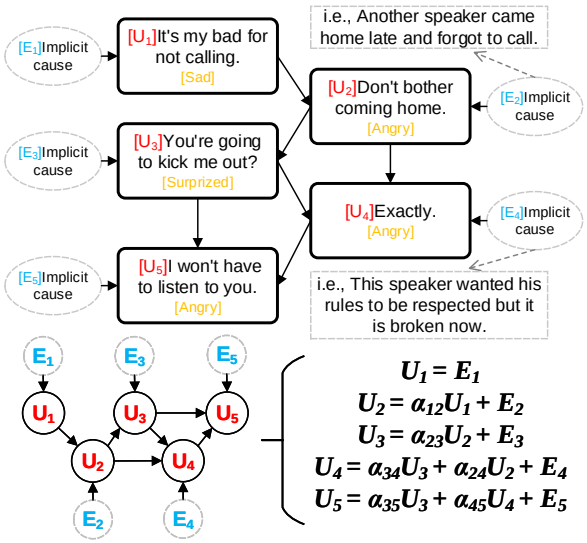



Figure 1: The conversation case with five utterances. In the SCM, we assume that each utterance $U_i$ has a corresponding implicit cause $E_i$, and has several explicit causes. i.e., $U_4$ has an implicit cause $E_4$ and two explicit causes $U_3$ and $U_2$. In the lower part of the figure, SCM adopts SEM $U_t = \sum \alpha_{it} U_i + E_t$ to denote these relationships and formalize the conversation as a DAG, though there is no information about $E$.

Moreover, Causal Discovery (Guo et al., 2020; Shen et al., 2020) focuses on finding a generic relationship between an effect and the cause. Besides explicit cause, this field additionally supported the **implicit cause**, in conversational cases, representing some desires or facts are not straightforwardly mentioned in the conversation. Compared with explicit causes, implicit causes are unobservable (Cheng et al., 2022; Nogueira et al., 2022). Furthermore, A number of cross-sectional studies (Houlihan et al., 2022; Ying et al., 2022; Teo

et al., 2022) have suggested the necessity of implicit cause when we analyze interlocutor's emotions.

Despite the importance of implicit causes and Causal Discovery, to our knowledge, there is no work applying them to affective reasoning in conversation texts. In Figure 1, we use a conversation case's linear Structural Causal Model (SCM)[1] to instantiate the causal discovery processing. To yield a causal representation of each variable, Causal Discovery consists of Causal Skeleton Estimation (CSE) and Causal Direction Identification (CDI) (Glymour et al., 2019). However, due to the reliance on sampling, these two steps have become intractable when applied to conversation datasets.

Specifically, in **CSE**, there is no common skeleton for all conversation texts. While much of the research into Causal Discovery (Yuan and Shou, 2022; Scetbon et al., 2022; Ma et al., 2022) has focused on the mapping from variables to nodes in structured data, an intractable problem is how to process this variable injective with variable-length and unstructured data. Moreover, in **CDI**, implicit causes are dependent variables. For the structured dataset, the existing body of research (Shimizu and Bollen, 2014; Shimizu, 2019) on implicit causes showed the effectiveness of assuming them as given i.i.D. variables. However, due to the same context, it is intractable to hold this hypothesis of independence.

Hence, to build a common skeleton and representing implicit causes, we proposed a novel method called Conversational Affective Causal Discovery (CACD). Instead of focusing on a common skeleton for all variable-length conversation sequences, CACD builds a common skeleton, named *centering one graph node (cogn)* skeleton, for each utterance from some broadly accepted prior hypotheses. That is, a variable-length conversation can be composed of *cogn* skeletons separately corresponding to each utterance. Moreover, CACD designs a novel Causal Auto-Encoder (CAE) to discover causal relationships given generated implicit causes and known explicit causes. Specifically, a graph attention module was proposed to yield an autoregression matrix of the skeleton. Then heterogeneous GNNs encode the information passing from utterances to substitute for implicit causes and decode the information from substitute (implicit causes) and utterances (explicit causes) to causal representations, respectively.

CACD returns the causal graph and the causal representation from a given conversation, reasoning about the relations between emotions and both explicit and implicit causes. The design ethos of CACD can easily extend to all affective computing tasks in conversation. To gauge the usefulness of using our causal representation to analyze conversation sentiment, we conduct downstream experiments in ERC, ECPE, and ECSR respectively.

Our contribution is four-fold:

- To the best of our knowledge, we are the first to apply causal discovery to the conversation emotion field, with CACD learning the causal graph and causal representation of a given conversation.
- We proposed the theoretical *cogn* skeleton common for variable-length conversations and designed six empirical skeleton instances of it from those of all recent works.
- We proposed CAE to be the first to generate the substitute for implicit causes, overcoming adverse conditions of utterances dependent on each other.
- From the extrinsic evaluation of three affect-based tasks, our proposed model, CAE, significantly outperforms SOTA baseline models. From intrinsic evaluation, we empirically demonstrate the effectiveness of implicit causes and SCMs.

## 2 Related Works

### 2.1 Affective Reasoning in Conversation

Chen et al. (2018) has proposed the initial work on ERC due to the growing availability of public conversational data. Then Xia and Ding (2019) proposed ECPE that jointly identifies both emotions and their corresponding causes. Therefore, many works (Wei et al., 2020; Chen et al., 2020) have been involved with causal relationships between emotions and causes. Moreover, Poria et al. (2021) has extended ECPE into conversations and proposed a novel ECSR task recognizing emotion-cause spans. More recently, increasing works have indicated that some prior knowledge plays a critical role in the complex reasoning of these tasks, such as the assumption about interlocutors (Zhang

[1] SCM has become the model of choice in causal discovery since Shimizu et al. (2006) proposed it with the Structural Equation Model (SEM) integrated with Bayesian network.

et al., 2019; Lian et al., 2021; Shen et al., 2021b) and context (Ghosal et al., 2019; Shen et al., 2021a; Chen et al., 2022).

Our work is a natural extension of those works. By leveraging the extensively accepted and adequate prior knowledge of numerous existing works, *cogn* skeleton can be instantiated and evaluated by comprehensive experiments. Additionally, CACD is the first attempt to reason implicit causes of emotions in conversation, aligned with the fact that these causes are often implicit and thus need to be inferred more.

### 2.2 Causal Discovery

To alleviate the Causal Markov and Faithfulness Assumption (Spirtes et al., 2000; Colombo et al., 2012; Ogarrio et al., 2016), several SCMs (Shimizu et al., 2006; Shimizu and Bollen, 2014; Sanchez-Romero et al., 2019) based on the independent non-Gaussian assumptions of implicit causes have been proposed for continuous structured variables. More recently, the causal skeleton in CSE has been defined as a (partially) directed acyclic graph based on some reasonable prior assumptions (Dai et al., 2021; Fonollosa, 2019). And then, we can obtain causal relationships by leveraging some CDI methods (Ding et al., 2020a; Duong and Nguyen, 2022) processing causal skeletons.

Our work is an extension of these methods from structured data to unstructured data characterized by non-sampling and variable length. Based on GNN, the condition for data has been successfully addressed by *cogn* skeleton and CAE.

## 3 Task Definition

For notational consistency, we use the following terminology. The **target utterance** $U_t$ is the $t^{th}$ utterances of a conversation $\mathcal{D} = (U_1, U_2, U_3, \ldots, U_N)$ where $N$ is the maximum number of utterance in this conversation and $0 < t \leqslant N$. The **emotion label** $Emo_t$ denotes the emotion type of $U_t$. The **emotion-cause pair (ECP)** is a pair $(U_t, U_i)$, where $U_i$ is the $i^{th}$ utterance of this conversation. In the ECP, $U_t$ represents the emotion utterance and $U_i$ is the corresponding cause utterance. Moreover, the **cause label** $C_{t,i}$ denotes the cause span type of the ECP $(U_t, U_i)$.

Thus, in a given text, **ERC** is the task of identifying all $Emo_t$. Moreover, **ECPE** aims to extract a set of ECPs and **ECSR** aims to identify all $C_{t,i}$.

## 4 Methodology

In this section, we first detailed *cogn* skeletons with six instantiations from some accepted prior knowledge hypotheses in Section 4.1. Then to learn the implicit causes based on these *cogn* skeletons, we introduced the CAE in Section 4.2. Finally, we introduced an auxiliary loss as well as the optimization of the training process in Section 4.3.

### 4.1 Causal Skeleton Estimation

As their data is variable-length and unstructured, utterances differ from the variables that causal discovery often uses. Specifically, each conversation has a different amount ($N$) of utterances, and different inter-utterances relationships related to the context. Hence, it is intractable to build a general causal skeleton with fixed nodes and edges to describe all conversation samples. (One of the problems is how to define the V-structures of those padding nodes in some cases while originally existing in others.)

Fortunately, several published GNN-based approaches (Shen et al., 2021b; Ishiwatari et al., 2020; Ghosal et al., 2019; Lian et al., 2021; Zhang et al., 2019) in ERC have proposed and verified a common hypothesis to settle down this issue, though they did not involve causal discovery. The hypothesis is:

**Hypothesis 0.** $\forall U_i \in \mathcal{D}$, *it has the same causal skeleton as other utterances.*

By regarding **Hypothesis 0** as the prior knowledge, a common causal skeleton containing a target variable and a fixed number of related variables can reason about the relations between the target utterance and other considered utterances. We denote this skeleton of $U_t$ by $S(U_t)$. There are $\forall U_i, U_j \in \mathcal{D}, S(U_i) = S(U_j)$.

Additionally, there are some other empirical hypotheses from the above approaches. These hypotheses can be divided into two categories: one is about the "order" of utterances (**Hypotheses 1, 2, 3**), and the other is about intermingling dynamics among the interlocutors (**Hypotheses 4, 5, 6**).

**Hypothesis 1.** (Majumder et al., 2019) *Under the sequential order, the target utterance receives information only from the previous utterance.*

**Hypothesis 2.** (Wei et al., 2020) *Under the graph order, the target utterance receives information from all other utterances.*

**Hypothesis 3.** (Ghosal et al., 2019) *Under the local graph order, target utterance receives local information from k surround utterances.*

| Category | Hypothesis | Original work |
|---|---|---|
| **I** | 1 | Majumder et al. (2019) |
| **II** | 2 | Veličković et al. (2017) |
| **III** | 2,4 | Chen et al. (2022) |
| **IV** | 3, 4 | Ghosal et al. (2019) |
| **V** | 3($k = 1$), 4, 5 | Lian et al. (2021) |
| **VI** | 3, 4, 5, 6 | Shen et al. (2021b) |

Table 1: Statistics of 6 *cogn* skeletons. We detailed the hypotheses each *cogn* skeleton adopted and the original works from which we designed them.

**Hypothesis 4.** (Zhang et al., 2019) *The influence between two utterances can be discriminated by whether the two utterances belong to the same speaker identity.*

**Hypothesis 5.** (Lian et al., 2021) *Target utterance only receives information from the predecessor utterances.*

**Hypothesis 6.** (Shen et al., 2021b) *Between two utterances both related to the target utterance, there is also information passing, often dubbed as a partial order.*

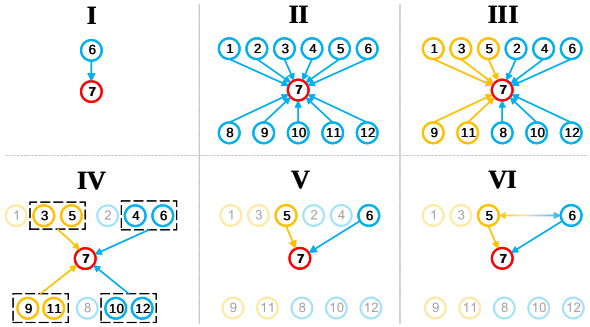



Figure 2: Six *cogn* skeletons from a conversation case with 12 utterances. We adopted the 7-th utterance as the target utterance (Red). Orange nodes denote the utterances of the same speaker as the target utterance, and blue ones denote those belonging to other speakers. Arrow represents the information propagated from one utterance to another, and the bi-way arrow represents the influence-agnostic relationship. The black dash box represents the slide windows.

In Figure 2, we designed six *cogn* skeletons to instantiate those hypotheses from all most recent works which adopted one or more hypotheses introduced above. The statistic is shown in Table 1. The specific algorithms and the whole skeleton of a conversation consisting of *cogn* skeletons are shown in Appendix B.

Note that we only conduct experiments for **II-VI** because our CAE is hard to apply with the recurrent-based skeleton. We detailed this limitation in Section 5.8.

### 4.2 Causal Direction Identification

From a given causal skeleton, a linear Structural Causal Model (SCM) can be equivalently represented as a linear Structural Equation Model (SEM):

$$U_t = \sum_{i \in rel_t} \alpha_{i,t} U_i + E_t \tag{1}$$

where $rel_t$ denotes a set of utterances that point to the $U_t$ (7-th utterance) in Figure 2, $E_t$ represents the exogenous variable towards the $U_t$ in SCM, the noise vector towards the $U_t$ in SEM (often written as $\varepsilon_t$), and the implicit cause towards the $U_t$ in CACD. Furthermore, we denote the word embedding of $U$ by $H = h_1, h_2, \ldots, h_N$, and the relationships between utterances in rows can also be written as :

$$H = A^T H + E \tag{2}$$

where $A_{i,t} \neq 0$ stands for a directed edge from $U_i$ to $U_t$ in SCM. Thus we can define the Causal Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency matrix $A_{i,i} = 0$ for all $i$. However, to obtain $\mathcal{G}$, we can not follow the common practices (Shimizu et al., 2006; Monti et al., 2020; Ng et al., 2022): adopt Independent Component Analysis (ICA) with a hypothesis that $E$ is i.i.d. and has non-Gaussian distributions. Because $E_i$ and $E_j$ are dependent on each other based on the same context.

Hence, we treat $A^T$ as an autoregression matrix of the $\mathcal{G}$, and then $E$ can be yielded by an auto-encoder model. Moreover, with the contribution of Yu et al. (2019) in the generalized nonlinear SEM model, auto-encoding processing can be formalized as:

$$E = f((I - A^T)H) \tag{3}$$

$$\widehat{H} = g((I - A^T)^{-1}E) \tag{4}$$

where $f$ and $g$ represent the encoder and decoder neural networks respectively. Encoder aims to generate a implicit cause matrix $E$ and Decoder devotes to yield a causal representation $\widehat{H}$. From the Equation 1, causal representation $\widehat{H}_t$ reasons about the fusion relations of heterogeneous explicit causes $\sum_{i \in rel_t} H_i$ and implicit cause $E_t$. The whole processing of CACD is shown in Figure 3.

**Encoder.** We use the graph attention mechanism to learn the adjacency matrix $A$ and construct a hierarchical GNN to instantiate the $f$. And

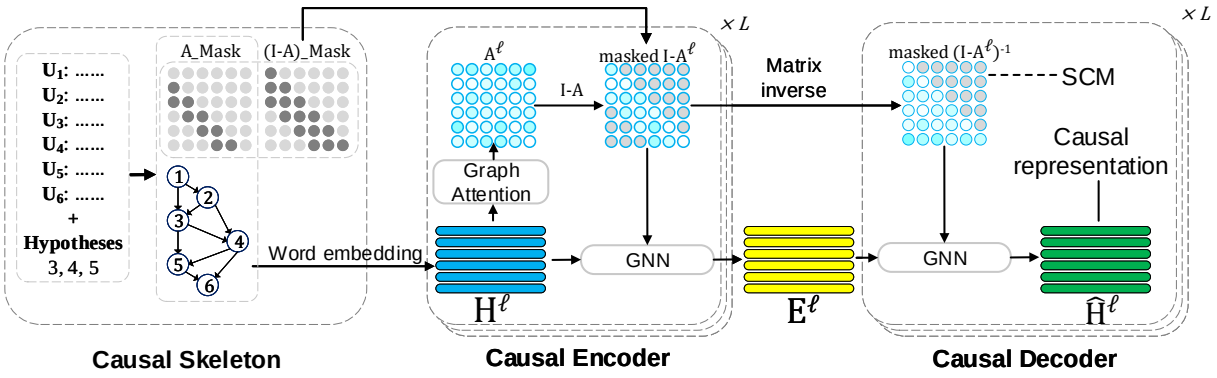



Figure 3: Processing of CACD, with a six-utterances conversation case as the input. Causal skeleton indicates which utterances (nodes) are used for aggregation. For each layer $\ell$, we collect representations $H^\ell$ for all utterances where each row represents one utterance. Causal Encoder yields the implicit causes $E^\ell$, the input for Decoder learning the causal representation.

$\ell = 1, 2, \ldots, L-1$ represents the layer of GNN. Thus, for each utterance at the $\ell$-th layer, the $A^\ell_{i,t}$ computed by attention mechanism is a weighted combination of $h^\ell_t$ for each directly related utterance $U_i (i \in rel_t)$:

$$A^\ell_{i,t} = \frac{LeakyReLU(e^\ell_{i,t})}{\sum_{j \in rel_t} LeakyReLU(e^\ell_{j,t})} \quad (5)$$

$$e^\ell_{i,t} = \overrightarrow{h}_i W^\ell_{i(row)} + (\overrightarrow{h}_t W^\ell_{t(col)})^T \quad (6)$$

where $W^\ell_{row} \in \mathbb{R}^{N \times 1}$ and $W^\ell_{col} \in \mathbb{R}^{N \times 1}$ are the learnable parameters in the graph attention. Moreover, the GNN aggregates the information from the neighbor utterances as following:

$$H^{\ell+1} = eLU((I - (A^\ell)^T) H^\ell W^\ell) \quad (7)$$

where $W^\ell$ stands for parameters in the corresponding layer. From the final layer of the evaluation process, by extracting $A^{L-1}$ computed in Equation 5, the marginal or conditional "distribution" of $H$ is obtained, showing CACD how to discover Causal Graph $\mathcal{G}$ from $\mathcal{D}$. Besides, by extracting $H^L$ in Equation 7, we can have the susbtitude for the implicit causes $E = MLP(H^L)$.

Note that for skeleton **VI** or other skeletons with **Hypothesis 6**, depending on RNNs is necessary for learning processing aligned with the partial order between related utterances:

$$H^{\ell+1} = GRU^\ell(H^\ell, m^\ell) \quad (8)$$

where $H^\ell$ is the input and $m^\ell$ is the state of GRU model, with $m$ computed by self-attention proposed by Thost and Chen (2021).

**Decoder.** We utilize the $A$ and $E$ computed from Encoder to generate the causal representation $\widehat{H}$. With a fixed adjacency matrix, the GNN aggregates the information of implicit causes from neighbor nodes as following:

$$\widehat{E}^{\ell+1} = eLU((I - (A^L)^T)^{-1} E^\ell M^\ell) \quad (9)$$

where $M^\ell$ is parameters in the corresponding layer. As the same architecture as the encoder, $\widehat{H} = MLP(E^L)$. Additionally, the plug-in RNN is also integrated with GNN to address the appetite of **Hypothesis 6**:

$$\widehat{E}^{\ell+1} = GRU^\ell(\widehat{E}^\ell, p^\ell) \quad (10)$$

### 4.3 Optimization

In the affective computing of the conversation, causal representation is an affect-based utterance representation. Under the **linear** SEM model, $\widehat{H}$ and $H$ represent one and the same. Thus, from the definition that implicit causes $E_t$ denote some mental causes (some desires or goals) leading to emotion $Emo_t$, $\widehat{H}$ should be aligned with $H$ in emotion dimensions under the **non-linear** SEM model. In short, we adopt an auxiliary loss measuring the Kullback-Leibler (KL) divergence of $\widehat{H}$ and $H$ when mapped into the exact emotion dimensions. This loss aims to impose the constraint that $E$ is the implicit cause we need.

$$Loss_{KL} = \sum_t \sum_{e \in Emo_t} p_e(\widehat{U_t}) \log \frac{p_e(\widehat{U_t})}{p_e(U_t)} \quad (11)$$

| Dataset | conversations | | | tasks | | |
|---|---|---|---|---|---|---|
| | Train | Val | Test | ERC | ECPE | ECSR |
| DailyDialog | 11118 | 1000 | 1000 | √ | – | – |
| MELD | 1038 | 114 | 280 | √ | – | – |
| EmoryNLP | 713 | 99 | 85 | √ | – | – |
| IEMOCAP | 100 | 20 | 31 | √ | – | – |
| RECCON-DD | 833 | 47 | 225 | – | √ | √ |
| RECCON-IE | – | – | 16 | – | √ | √ |
| Synthetic data | 833 | 47 | 225 | – | √ | √ |

Table 2: The statistics of seven datasets

where $e$ is one of all emotion types in $Emo_t$, $p_e(\cdot)$ denotes the probability that the target representation is labeled with emotion $e$. When $Loss_{KL}$ achieves its lower bound, $E$ is pure implicit cause. In the whole process of three tasks, we followed (Shen et al., 2021b; Wei et al., 2020; Poria et al., 2021) to add several losses of ERC, ECPE, and ECSR respectively.

Furthermore, if we treated implicit causes $E$ as latent variables $Z$, CAE seems extremely similar to Variational Auto-Encoder (VAE) (Kingma and Welling, 2014). However, Equation 7 outlines some essential differences between implicit causes and latent variables. The outputs of the encoder in VAE are $\mu^i$ and $\sigma^i$, describing the distribution of $q_\phi(Z)$. With this estimation of $\widehat{q}_\phi(Z)$, we can measure the variation $\xi(q_\phi(Z))$ (also called $\nabla_\phi ELBO(\widehat{q}_\phi(Z))$) to obtain the approximation estimation of $ELBO(q)$. In contrast, our outputs are $A$ and $E$, two critical components of a fixed graph rather than a distribution.

In other words, the variation depends on the prior over that the latent variables are multivariate distribution, whereas CAE has a dependency on the prior about conversation SCM which is non-sampling and non-distributive.

# 5 Experiments

In this section, we conduct extensive experiments to answer the research questions:

**RQ1:** How effective is our method in three downstream tasks?

**RQ2:** How can we justify the implicit causes in our method?

**RQ3:** How do the implicit causes help to improve performance?

## 5.1 Datasets

We use six real datasets for three affective reasoning tasks and one synthetic dataset for justifying $E$ in our model. The statistics of them are shown in Table 2. Appendix A depicts the detailed introductions of each dataset.

| Skt | Model | DailyDialog | MELD | EmoryNLP | IEMOCAP |
|---|---|---|---|---|---|
| **II** | DialogXL | 54.93 | 62.41 | 34.73 | 65.94 |
| | CAE | **59.51** | **63.62** | **39.16** | **66.47** |
| **III** | EGAT† | 59.23 | 63.51 | 38.77 | 66.76 |
| | CAE | **59.68** | **63.71** | **39.62** | **68.18** |
| **IV** | RGAT | 54.31 | 60.91 | 34.42 | 65.22 |
| | CAE | **59.65** | **63.69** | **39.22** | **67.65** |
| **V** | DECN† | 59.08 | 63.78 | 39.44 | 67.41 |
| | CAE | **59.28** | **63.91** | **40.11** | **67.61** |
| **VI** | DAG-ERC | 59.33 | 63.65 | 39.02 | 68.03 |
| | CAE | **59.53** | **63.81** | **39.54** | **69.17** |

Table 3: Overall performance in ERC task. † denotes the results implemented in this paper. The better scores in the same skeleton are in bold, and the best of all skeletons is in red.

## 5.2 Implementation

We adopt the consistent benchmarks of the SOTA methods in three tasks, including the pre-training language model, hyper-parameters, $t$-test and metrics. The implementation details are shown in Appendix B.

## 5.3 Baseline Models

According to the hypotheses of these baselines, for each *cogn* skeleton, we choose the one recent SOTA work:

**II**: **DialogXL** (Shen et al., 2021a).
**III**: **EGAT** (Chen et al., 2022).
**IV**: **RGAT** (Ishiwatari et al., 2020).
**V**: **DECN** (Lian et al., 2021).
**VI**: **DAG-ERC** (Shen et al., 2021b).
**Ours**: **CAE**.

## 5.4 Overall Performance (RQ1)

Table 3 reports the overall results in ERC task, and Table 4 reports the results in ECPE and ECSR.

In Table 3, CAE performs better than the corresponding baseline under all skeletons in four datasets. Hence, using a causal auto-encoder to find the implicit causes benefits this task. Besides, CAE improves significantly under skeletons **II**, **III**, and **IV**. From Figure 2, these three skeletons have more relevant nodes than others, so there are more redundant edges to be corrected by CAE, which is demonstrated again in Appendix E. In contrast, **V** and **VI** achieve the best results in MELD, EmoryNLP, and IEMOCAP datasets, which indicates that **Hypothesis 5** is more probably a strong inductive bias that conversation enjoys.

In Table 4, with $p$ <0.01 in the $t$-test, best improvement and best performance both concentrate on **VI**. With the visualization of Appendix E, we infer that the upper triangular adjacency matrix of DAG-ERC, not restricted by the backpropagation,

| Skt | model | ECPE in RECCON | | ECSR in RECCON | |
|---|---|---|---|---|---|
| | | DD($\pm\sigma_{10}$) | IE | DD($\pm\sigma_{10}$) | IE |
| – | RANK-CP† | 63.51±2.1 | 41.56 | 26.57±0.8 | 18.99 |
| | ECPE-2D† | 64.35±1.7 | **47.42** | 34.41±0.1 | 22.03 |
| | ECPE-MLL† | **65.72**±1.7 | 44.61 | **37.79**±0.4 | **26.19** |
| **II** | DialogXL† | 61.92±1.7 | 50.31 | **35.79±0.5** | 21.78 |
| | CAE | **64.74±1.6** | **51.23** | 34.63±0.2 | **27.92** |
| **III** | EGAT | 68.05±1.5 | 53.43 | 29.68±0.7 | 16.42 |
| | CAE | **69.16±1.2** | **53.81** | **30.5±0.2** | **18.55** |
| **IV** | RGAT† | 69.02±1.9 | 52.48 | **30.39±0.4** | 17.49 |
| | CAE | **70.12±2.1** | **53.93** | 30.24±0.5 | **19.31** |
| **V** | DECN† | 68.32±1.5 | 51.73 | 30.7±0.9 | 18.47 |
| | CAE | **68.84±1.7** | **53.89** | **31.88±0.2** | **20.13** |
| **VI** | DAG-ERC† | 70.36±1.5 | 55.7 | 40.12±0.7 | 24.89 |
| | CAE | **73.17±1.1** | **56.67** | **42.14±0.1** | **30.41** |

Table 4: Overall performance in ECPE and ECSR tasks. We additionally compare three baselines not belonging to any skeleton: RANK-CP (Wei et al., 2020), ECPE-2D (Ding et al., 2020b), and ECPE-MLL (Ding et al., 2020c).

| Model | CAE Categories | | | | |
|---|---|---|---|---|---|
| | **II** | **III** | **IV** | **V** | **VI** |
| Ours | 64.74 | 69.16 | 70.12 | 68.84 | 73.17 |
| $BCE$ | ↓0.62 | ↓0.04 | ↓0.15 | ↓0.16 | ↓0.29 |
| w/o $Loss_{KL}$ | ↓2.18 | ↓1.95 | **↓2.42** | ↓1.33 | ↓1.58 |
| w/o Decoder | **↓3.59** | **↓2.79** | ↓2.11 | **↓2.83** | **↓4.14** |
| w/o Hypo 6 | – | – | – | – | ↓1.59 |
| w/o Hypo 5 | – | – | – | ↓2.34 | ↓1.88 |
| w/o Hypo 4 | – | **↓3.67** | **↓2.72** | **↓3.15** | **↓4.19** |

Table 5: Ablation results

benefits from **Hypothesis 6**. Moreover, **II** lags farthest behind in the ECPE while achieving the second best in the ECSR, showing that the reliance on a hypothesis is not equal in different tasks. Furthermore, without **Hypotheses 1** and **6**, **III**, **IV**, and **V** are far from the best performance since **Hypothesis 1** has the maximum identifying space, and **Hypothesis 6** supplies the highest number of information passing. Finally, it is worth noting that three skeleton-agnostic baselines perform poorly in the RECCON-IE dataset, indicating that our models have stronger representation learning capabilities.

We also conducted the sensitivity experiments to analyse how our model performs in different $L$ and $k$ in Appendix C.

### 5.5 Ablation Study (RQ1)

To further evaluate the contribution of neural components and auxiliary Hypotheses (**Hypo 4, 5, 6**), we conducted six sets of ablation experiments to study the effects of different parts. In Table 5, we summarized results under the following cases: replacing $Loss_{KL}$ with $BCE$ loss function ($BCE$); removing the $Loss_{KL}$ (w/o $Loss_{KL}$); replacing Decoder module with a Linear layer (w/o Decoder); removing the RNN module (w/o **Hypo 6**); adding the edges from successors to predecessors (w/o **Hypo 5**); reducing the speaker types to one (w/o **Hypo 4**).

| Skt | DailyDialog | MELD | EmoryNLP | IEMOCAP |
|---|---|---|---|---|
| **II** | 51.48 (↓8.03) | **58.41 (↓5.21)** | 34.97 (↓4.19) | 59.71 (↓6.76) |
| **III** | 54.37 (↓5.31) | 58.19 (↓5.52) | 36.55 (↓3.07) | 63.42 (↓4.76) |
| **IV** | **55.62 (↓4.03)** | 57.22 (↓6.47) | **36.91 (↓2.31)** | 62.34 (↓5.31) |
| **V** | 54.62 (↓4.66) | 58.19 (↓5.72) | 35.49 (↓4.62) | 63.13 (↓4.48) |
| **VI** | 53.27 (↓6.26) | 58.39 (↓5.42) | 34.98 (↓4.56) | **65.18 (↓3.99)** |

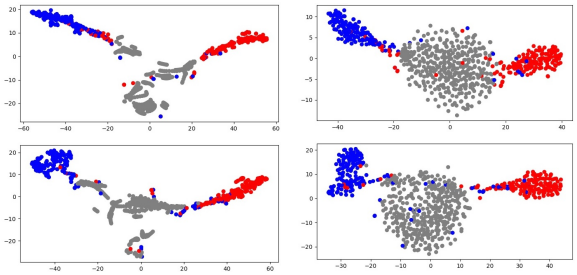
Table 6: Overall performance of implicit causes $E$ in ERC task.

As shown in Table 5, $BCEloss$ performs similarly to $Loss_{KL}$; thus, we empirically demonstrate that our auxiliary loss is essentially different from $Loss_{KL}$ in VAE. Since the F1 score decreases heavily without auxiliary loss or decoder, these two are necessary ingredients for building complete processing to learn the causal representation via $E$. Besides, in the study of **Hypotheses 4, 5**, and **6**, they are all critical but removing **Hypothesis 4** leads to the highlight degradation in 4 skeletons. This result corroborates the theory of Lian et al. (2021) and Shen et al. (2021b), who state that speaker identity is the most strong inductive bias in conversation. Finally, with **Hypothesis 3** being the most effective order skeleton from the results in Tables 3, 4, it is interesting to see that skeleton with **Hypotheses 3, 4, 5, and 6** should be the closest to perfection when the DAG-ERC indeed achieves the SOTA.

### 5.6 Implicit Causes(RQ2)

All experiments above are extrinsic evaluations; thus, to further gain insights into $E$, we perform two experiments to justify $E$. First, we test the F1 scores in ERC task by replacing $\widehat{H}$ with $E$ from a consensus that implicit causes should be aligned with utterances in the emotion types (Proof see Appendix F). Second, we trained our model in a synthetic dataset to observe how the $E$ is similar to the ground truth implicit causes distributions. Note that although human evaluation labels are better for proving the reasonable performance, it is intractable to assess implicit causes due to their unobservability. So we create a synthetic dataset given a set of fixed i.i.d. implicit causes.

In Table 6, we reported the overall results of $E$ in ERC task. Among five skeletons and four datasets, almost all results achieve 90% scores of corresponding performances of $\widehat{H}$, which indicates that $E$ is practically aligned with $\widehat{H}$ in the affective dimension. Furthermore, Figure 4 (a-b) shows the



(a) without noise (b) with noise

Figure 4: Visualization of $E$ (upper subfigures) and implicit causes (lower subfigures) with colors in the simulated datasets. The gray cluster means padding utterances in each dialogue, the blue cluster corresponds to the non-emotion utterances, and the red cluster corresponds to emotion utterances.

projection of $E$ and implicit causes, respectively, using t-SNE (Knyazev et al., 2019) on the synthetic dataset without and with noise $\xi$. We observe that $E$ and implicit causes are similarly clustered into three parts. $E$ is consistent with the implicit causes in the samples with or without noise through the distribution properties and independence. The results indicate that $E$ successfully learns the implicit causes.

### 5.7 Case analysis (RQ3)

We also concern how $E$ improves the model performance, especially in dialogue samples with different causal models. So we show the adjacent matrices of our model and current SOTA methods in Appendix E and conduct the case analysis to visualize and investigate the effect of $E$ in Appendix G. Appendix E shows that our model can more freely explore the relationship between different utterances via adjacent matrices shifting rather than being limited to a fixed structure (i.e., attention module). Besides, Appendix G demonstrates that our model has significant improvement in three causal models while failing in the causal model under the presence of confounders (we discuss this limitation in section 5.8).

### 5.8 Results and Discussions

**Extrinsic evaluation**: In Tables 3 and 4, CAE outperforms the SOTA work by 3.84% in ERC task, 2.81% in ECPE task, and 2.02% in ECSR task respectively. Moreover, Table 5 demonstrates that the decoder and the auxiliary loss function contributes more than other components (3.09% and 1.89% average decrease in ablation).

**Intrinsic evaluation**: To justify $E$ is the implicit causes we need, We theoretically estimate the consistency of implicit causes and utterances in emotional embedding, and Table 6 practically provides that $E$ satisfies this estimation. Besides, Figure 4 shows the $E$ and implicit causes are highly similar in low-dimension projection. Moreover, Appendix E and G show the visualization and case anlaysis about how implicit causes improve model training and performance.

Nevertheless, we detailed limitations as follows:

**GNN-based model**: To ameliorate the non-sampling of conversation datasets, we proposed *cogn* skeleton and designed a GNN-based CAE to learn it. However, some intractable issues exist: (i) CAE can not apply to all Recurrence-based Methods because GNN can not process the sequential information. (ii) CAE needs some plug-in technologies (i.e., RNN module) to satisfy the complex prior knowledge (i.e., **Hypothesis 6**).

**Confounder**: In Tables 3, 4, and 6, skeletons **II**, **III**, and **IV** generally lag far behind **V** and **VI**. This unsatisfactory performance of these skeletons indicates that excessive adding-edge leads to serious confounders. Although **V** and **VI** alleviate introducing confounders, they lose some practical information due to reducing edges. Therefore, processing confounders in the neural network plays an indispensable role in CACD.

## 6 Conclusion

We have addressed the problem of affective causal discovery in conversation and proposed CACD, a novel causal discovery method containing two steps, CSE and CDI. First, we proposed a *cogn* skeleton in the CSE to avoid non-sampling and variable-length unstructured conversation cases. In the CDI, we proposed CAE, a simple but broadly applicable GNN-based model, reasoning the substitute for implicit causes and returning favorable causal representation for downstream affective tasks. Further, we detailed our comprehensive, empirical experiments for extrinsic and intrinsic evaluation on seven datasets. From the expected results, we thus argue that CACD is our *de facto* need for affective reasoning in conversation, which deserves more future research. In future work, we will focus on how to represent and handle the confounder by designing a more effective model.

## References


Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. 2021. The decamfounder: Non-linear causal discovery in the presence of hidden variables. *arXiv preprint arXiv:2102.07921*.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Multi-granularity semantic aware graph model for reducing position bias in emotion-cause pair extraction. *arXiv preprint arXiv:2205.02132*.

Hongliang Bi and Pengyuan Liu. 2020. Ecsp: A new task for emotion-cause span-pair extraction and classification. *arXiv preprint arXiv:2003.03507*.

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.

Ting Wei Chang, Yao-Chung Fan, and Arbee LP Chen. 2022. Emotion-cause pair extraction based on machine reading comprehension model. *Multimedia Tools and Applications*, pages 1–21.

Hang Chen, Xinyu Yang, and Chenguang Li. 2022. Learning a general clause-to-clause relationships for enhancing emotion-cause pair extraction.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Xinhong Chen, Qing Li, and Jianping Wang. 2020. Conditional causal relationships between emotions and causes in texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3111–3121, Online. Association for Computational Linguistics.

Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, Kasim Selcuk Candan, and Huan Liu. 2022. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321.

Haoyue Dai, Rui Ding, Yuanyuan Jiang, Shi Han, and Dongmei Zhang. 2021. Ml4c: Seeing causality through latent vicinity. *arXiv preprint arXiv:2110.00637*.

Rui Ding, Yanzhi Liu, Jingjing Tian, Zhouyu Fu, Shi Han, and Dongmei Zhang. 2020a. Reliable and efficient anytime skeleton learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10101–10109.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170, Online. Association for Computational Linguistics.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020c. 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.

Bao Duong and Thin Nguyen. 2022. Bivariate causal discovery via conditional divergence. In *Conference on Causal Learning and Reasoning*, pages 236–252. PMLR.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašic. 2022. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems.

José AR Fonollosa. 2019. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*, pages 339–347. Springer.

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.

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.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

Sean Dae Houlihan, Desmond Ong, Maddie Cusimano, and Rebecca Saxe. 2022. Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory of mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370, Online. Association for Computational Linguistics.

D. P. Kingma and M. Welling. 2014. Auto-encoding variational bayes. *arXiv.org*.

Boris Knyazev, Graham W Taylor, and Mohamed Amer. 2019. Understanding attention and generalization in graph neural networks. *Advances in neural information processing systems*, 32.

Anshul Kumar and Ankit Kumar Jain. 2022. Emotion detection in psychological texts by fine-tuning bert using emotion–cause pair extraction. *International Journal of Speech Technology*, 25(3):727–743.

Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. *arXiv preprint arXiv:2205.00759*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Decn: Dialogical emotion correction network for conversational emotion recognition. *Neurocomputing*, 454:483–495.

Pingchuan Ma, Rui Ding, Haoyue Dai, Yuanyuan Jiang, Shuai Wang, Shi Han, and Dongmei Zhang. 2022. Ml4s: Learning causal skeleton from vicinal graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1213–1223.

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.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. 2020. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR.

Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. 2022. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1449.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. 2016. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pages 368–379. PMLR.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2015. Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. 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, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. 2019. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306.

Meyer Scetbon, Laurent Meunier, and Yaniv Romano. 2022. An asymptotic test for conditional independence using analytic kernel embeddings. In *International Conference on Machine Learning*, pages 19328–19346. PMLR.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixiang Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *AAAI*.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. 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 1551–1560, Online. Association for Computational Linguistics.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. 2020. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 10(1):1–12.

Shohei Shimizu. 2019. Non-gaussian methods for causal structure learning. *Prevention Science*, 20(3):431–441.

Shohei Shimizu and Kenneth Bollen. 2014. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *J. Mach. Learn. Res.*, 15(1):2629–2652.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. 2000. Constructing bayesian network models of gene expression networks from microarray data.

Chandler Squires, Annie Yun, Eshaan Nichani, Raj Agrawal, and Caroline Uhler. 2022. Causal structure discovery between clusters of nodes induced by latent factors. In *Conference on Causal Learning and Reasoning*, pages 669–687. PMLR.

Dennis WH Teo, Zheng Yong Ang, and Desmond Ong. 2022. Modeling causal inference from emotional displays. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Veronika Thost and Jie Chen. 2021. Directed acyclic graph neural networks. *arXiv preprint arXiv:2101.07965*.

Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. 2021. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3975–3989, Online. Association for Computational Linguistics.

Hande Aka Uymaz and Senem Kumova Metin. 2022. Vector based sentiment and emotion analysis from text: A survey. *Engineering Applications of Artificial Intelligence*, 113:104922.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

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.

Lance Ying, Audrey Michal, and Jun Zhang. 2022. A bayesian drift-diffusion model of schachter-singer's two-factor theory of emotion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR.

Alex Eric Yuan and Wenying Shou. 2022. Data-driven causal analysis of observational biological time series. *eLife*, 11:e72518.

Sayyed Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, AFFCON'18, pages 44–51, New Orleans, LA.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*.

## A Datasets

**DailyDialog** (Li et al., 2017): A Human-written dialogs dataset with 7 emotion labels (*neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, and *fear*). We follow Shen et al. (2021b) to regard utterance turns as speaker turns.

**MELD** (Poria et al., 2019): A multimodel ERC dataset with 7 emotion labels as the same as DailyDialog.

**EmoryNLP** (Zahiri and Choi, 2018): A TV show scripts dataset with 7 emotion labels (*neutral*, *sad*, *mad*, *scared*, *powerful*, *peaceful*, *joyful*).

**IEMOCAP** (Busso et al., 2008): A multimodel ERC dataset with 9 emotion labels (*neutral*, *happy*, *sad*, *angry*, *frustrated*, *excited*, *surprised*, *disappointed*, and *fear*). However, models in ERC field are often evaluated on samples with first six emotions due to the too few samples of latter three emotions. 20 dialogues for validation set is following Shen et al. (2021b).

**RECCON** (Poria et al., 2021): The first dataset for emotion cause recognition of conversation including RECCON-DD and RECCON-IE (emulating an out-of-distribution generalization test). RECCON-DD includes 5380 labeled ECPs and 5 cause spans (*no-context*, *inter-personal*, *self-contagion*, *hybrid*, and *latent*).

**Synthetic dataset**: We create a synthetic dataset by following the benchmark of the causal discovery field (Agrawal et al., 2021; Squires et al., 2022). To minimize sample bias, we did not randomly draw causal graphs as samples. Inversely, the number of samples in the synthetic dataset and the number of utterances and labels per sample are restricted to be consistent with RECCON. We use Causal Additive Models (CAMs), Specifically SCM structure for our datasets. As shown in Algorithm 1, first, we assume that each $i.i.d.$ implicit causes $E \sim \|^{50}\mathcal{N}(1,1)$ if it is an emotion utterance in the original dataset, and $E \sim \|^{50}\mathcal{N}(-1,1)$ if it not. Then, we update each utterance via speaker turns $S$: if there is a emotion-cause pair $(U_i, U_j) \in L$, then $U_i = \alpha_{j,i}U_j + E_i$ ($\alpha_{j,i} \sim Unifrom([0.7,1])$), and for those pairs without emotion-cause label, $\alpha_{j,i} \sim Unifrom([0,0.3])$. Finally, we randomly select a noise $\xi \sim Unifrom([-0.25,0.25])$ for each utterance $U_i = U_i + \xi_i$.

**Algorithm 1:** Creating Non-noise Synthetic dataset

**Input:** $\mathcal{D}, S, L$
**Output:** $SCM_{\mathcal{D}}$

**forall** $i \in 2,3,\ldots,S$ **do**
  **if** $Emotion(U_i)$ **then**
    $E_i \sim \|^{50}\mathcal{N}(1,1)$
  **else**
    $E_i \sim \|^{50}\mathcal{N}(-1,1)$
  $U_i \leftarrow E_i$
**forall** $i \in 1,2,3,\ldots,S$ **do**
  **forall** $j \in 1,2,\ldots,i$ **do**
    **if** $(U_i, U_j) \in L$ **then**
      $U_i = \alpha_{j,i}U_j + E_i(\alpha_{j,i} \sim Unifrom([0.7,1]))$
    **else**
      $U_i = \alpha_{j,i}U_j + E_i(\alpha_{j,i} \sim Unifrom([0,0.3]))$
$SCM_{\mathcal{D}} \leftarrow U_1, U_2, \ldots, U_S$ return $SCM_{\mathcal{D}}$

## B Implementation Details

In the word embedding, we adopt the affect-based pre-trained features[2] proposed by Shen et al. (2021b) for all baselines and models.

Although there are different pre-trained models in these skeleton baselines, the SOTA work DAG-ERC and EGAT have investigated their performances in a consistent pre-trained model. Therefore, for a fair and direct comparison, we continue this benchmark using the pre-trained embedding published by DAG-ERC for three tasks.

In the hyper-parameters, we follow the setting of Shen et al. (2021b) in the ERC task. Moreover, in the ECPE and ECSR, the learning rate is set to 3e-5, batch size is set to 32, and epoch is set to 60. Further in CAE, we set $L$ to 1, and implicit cause size is set to 192, hidden size of GNN is set to 300, and dropout rate is 0.3.

Meanwhile, because there is only one training dataset for ECPE and ECSR, we evaluated our method ten times with different data splits by following Chen et al. (2022) and then performed paired sample $t$-test on the experimental results.

Finally, we adopted downstream task modules consistent with the SOTA baselines: Wei et al.

[2] https://drive.google.com/file/d/1R5K_2PlZ3p3RFQ1Ycgmo3TgxvYBzptQG/view?usp=sharing

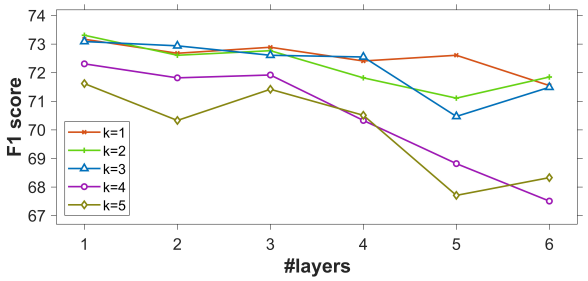



Figure 5: Further layers $L$ and related node number $K$ with **VI** skeleton CAE model in ECPE task.

(2020) in ECPE and ECSR, and Shen et al. (2021b)for the ERC task.

For evaluation metrics, we follow Shen et al. (2021b) towards ERC, Xia and Ding (2019) towards ECPE, and Poria et al. (2021) towards ECSR. Specifically, we adopt the macro F1 score in ECPE and ECSR tasks, micro F1 score for DailyDialog, and macro F1 score for the other three datasets in ERC task.

## C Sensitivity Analysis

In this section, we investigate how the number of layers and the variants of causal skeletons would affect the performance of CAE. So we further conducted several contrasts with $k$ up to 5 and $L$ up to 6, as shown in Figure 5. One observation is that the best performance occurs at either $k = 1$, 2, or 3, which indicates that $k \geqslant 4$ offers no advantage and even leads to confounding. Moreover, $L = 1$ achieves the best performance under all $k$ values. In other words, one layer is sufficient to yield the most effective implicit causes.

## D Algorithms for 6 Skeletons

A *cogn* skeleton is denoted by $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$. The $\mathcal{V} = U_1, U_2, U_3, ..., U_N$ represents a set of utterances in a conversation, and the edge $(i, j, m_{i,j}) \in \mathcal{E}$ denotes the influence from $U_i$ to $U_j$, where $m_{i,j} \in \mathcal{M}$ is the type of the edge depending on whether $U_i$ and $U_j$ belong to one and the same speaker. Thus $\mathcal{M} = 0, 1$, where 1 for that they are the same speaker and 0 for different. Then we denote the speaker type of $U_i$ by a function $p(U_i)$. At last, we show the process of building 6 *cogn* skeletons in Algorithms $1 - 6$.

Finally, in Figure 6, we show the adjacency matrix of each *cogn* skeleton by inputting a binary alternating conversation case with 6 utterances. But note that adjacency can not indicate all the dif-

**Algorithm 2:** Buliding **I** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
**forall** $i \in 2, 3, \ldots, N-1$ **do**
  $\mathcal{E} \leftarrow \mathcal{E} \cup (i, i+1)$
return $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

**Algorithm 3:** Buliding **II** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E})$
$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
**forall** $i \in 2, 3, \ldots, N$ **do**
  **forall** $j \in 2, 3, \ldots, N$ **do**
    **if** $i! = j$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (j, i)$
    **else**
      *Continue*
return $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

**Algorithm 4:** Buliding **III** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$
$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
$\mathcal{M} \leftarrow 0, 1$
**forall** $i \in 2, 3, \ldots, N$ **do**
  **forall** $j \in 2, 3, \ldots, N$ **do**
    **if** $p(U_j) = p(U_i)$ *and* $i! = j$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (j, i, 1)$
    **else if** $p(U_j)! = p(U_i)$ *and* $i! = j$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (j, i, 0)$
    **else**
      *Continue*
return $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

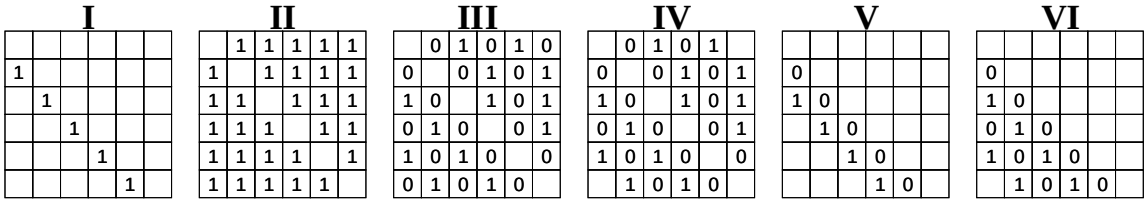



Figure 6: Adjacency matrices towards 6 *cogn* skeletons when $k = 2$. $(i, j) \neq None$ represents that $U_i$ is influenced by $U_j$.

**Algorithm 5:** Buliding **IV** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
$\mathcal{M} \leftarrow 0, 1$
**forall** $i \in 2, 3, \ldots, N$ **do**
  **forall** $j \in 2, 3, \ldots, N$ **do**
    **if** $p(U_j) = p(U_i)$ *and* $0 < |i - j| < k$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (j, i, 1)$
    **else if** $p(U_j)! = p(U_i)$ *and* $0 < |i - j| < k$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (j, i, 0)$
    **else**
      $Continue$
return $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

**Algorithm 6:** Buliding **V** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
$\mathcal{M} \leftarrow 0, 1$
**forall** $i \in 2, 3, \ldots, N$ **do**
  $\gamma \leftarrow i - 1$
  **if** $p(U_\gamma) = p(U_i)$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup (\gamma, i, 1)$
  **else**
    $\mathcal{E} \leftarrow \mathcal{E} \cup (\gamma, i, 0)$
  $\gamma \leftarrow \gamma - 1$
return $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

**Algorithm 7:** Buliding **VI** *cogn* skeleton

**Input:** $\mathcal{D}, p(\cdot), k$
**Output:** $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

$\mathcal{V} \leftarrow U_1, U_2, U_3, ..., U_N$
$\mathcal{E} \leftarrow \emptyset$
$\mathcal{M} \leftarrow 0, 1$
**forall** $i \in 2, 3, \ldots, N$ **do**
  $c \leftarrow 0$
  $\gamma \leftarrow i - 1$
  **while** $\gamma > 0$ *and* $c < k$ **do**
    **if** $p(U_\gamma) = p(U_i)$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (\gamma, i, 1)$
      $c \leftarrow c + 1$
    **else**
      $\mathcal{E} \leftarrow \mathcal{E} \cup (\gamma, i, 0)$
    $\gamma \leftarrow \gamma - 1$
return $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$

ferences among these skeletons, for example, Hypothesis 6 takes effect when the model learns the relationship based on the **VI** skeleton.

## E Visualization of Causal Graph

In the Figure 8 to 12, we showed the Visualization of the adjacency matrix $(I - A^T)^{-1}$. When the auxiliary loss $Loss_{KL}$ achieves the lower bound, $(I - A^T)^{-1}$ represents the relationship matrix between utterances and implicit causes.

In the ECPE task, we extracted 10 samples from test sets in different folds. To facilitate comparison and contrasting, we selected five 7-utterances cases and five 8-utterances cases. The IDs are as following:

**7-utterances cases**: 110, 170, 224, 372, 500.

**8-utterances cases**: 62, 74, 104, 177, 584.

To obtain the non-negative value, we adopted the $T = sigmoid(\cdot) - 0.05$ to process the original tensors $(I - A^T)^{-1}$ outputted from the encoder. We follow a common practice: set the threshold

as $0.05$ to delete some unimportant edges. And to highlight which implicit cause contributes the each utterance best, we adopted the $sofmax(\cdot)$ to process columns afterward and labeled the block with value $> 0$.

It is excepted that: (i) when skeletons construct overage edges, our model is able to degrade the influences of some negligible utterances by deleting the corresponding edges from their implicit causes. (ii) when skeletons construct insufficient edges, our model can add some edges to obtain more information.

## F Case Analysis

We want to investigate how the implicit causes $E$ influence the evaluation, which could help to explain implicit causes. From the causal analysis view, we observed four causal models about two utterances, $U_i$ and $U_j$, and their corresponding implicit causes, $E_i$ and $E_j$, shown in Figure 7.

The first model is that the ground truth (emotion, cause) label is $(j, j)$ (the same utterance), while the baseline model outputs the incorrect prediction result $(j, i)$. In the SCM model, $E_i$ and $E_j$ do not have a direct influence, so there is no spurious correlation between $U_i$ and $U_j$. In evaluation statistics, the $precision$ of the pairs whose emotion and cause belong to the same utterance has a significant difference between CAE and Baseline. CAE has a higher $precision$ ($4.6\%$ improvement) than Baseline, demonstrating that CAE can correct some spurious relationships in these cases.

The second model is that the ground truth (emotion, cause) label is $(j, i)$ (the different utterances), and the baseline model outputs the true prediction result $(j, i)$. In the SCM model, $E_j$ directly causes by $E_i$, so there is a backdoor path (Pearl, 2009) $U_i \leftarrow E_i \rightarrow E_j \rightarrow U_j$. In evaluation statistics, the $precision$ of the pairs whose emotion and cause belong to the different utterances performs similarly between CAE and Baseline. However, CAE has a higher $precision$ ($1.8\%$ improvement) than Baseline, demonstrating that CAE can not break down the right relationship in these cases.

The third model is similar to the Second model, while there is no ground truth label about $U_j$ because $U_j$ is not an emotion utterance. However, due to the backdoor path $U_i \leftarrow E_i \leftarrow E_j \rightarrow U_j$, Baseline learns a wrong emotion-cause relationship $(j, i)$. In evaluation statistics, the $recall$ of the pairs with $U_j$ being not an emotion utterance performs differently between CAE and Baseline. CAE has better performance ($3.3\%$ improvement) than Baseline, demonstrating that CAE enables the recognition of emotion utterance because implicit causes are the replacement of emotion desires.

The fourth model is the confounder we mentioned in 5.8. With only utterances and implicit causes, the SCM can not handle this trouble. Specifically, baseline outputs a wrong pair $(j, i)$ when the $U_j$ is an emotion utterance while $U_i$ is not its cause utterance. This trouble often represents a common implicit cause (in Figure 7, we use $=$ to stand for it) between $U_j$ and $U_i$, which leads to a backdoor path $U_i \leftarrow E_i = E_j \rightarrow U_j$. Therefore, CAE could learn two conclusions: i) $U_i$ and $U_i$ are causally related, and ii)$U_j$ is an emotion utterance. We investigated the $recall$ of these pairs and found that CAE indeed can not have a better performance than baseline.

These four models and evaluation statistics are consistent, demonstrating that implicit causes $E$ computed by SCM and CAE are the expected replacement.

## G Proof of emotional consistency of implicit causes and utterances

We want to explain why implicit causes and utterances are consistent in emotion from both theory and euqation.

We define the implicit causes as the unobservable emotional desire and the utterances as the observable emotional expression. This definition is proposed in Ong et al. (2019, 2015), which also argues that emotional expression is affected by desires and event outcomes. Moreover, the desire and the expression generally belong to the same emotion because the outcomes often have little effect on emotional expression. Our paper can also deduce this conclusion from the SCM (Equation 1). Considering there is a linear map $f(\cdot)$ from representation space to emotion space. Then we can obtain the following:

$$f((I - A)U) = f(E) \tag{12}$$

$$(I - A)f(U) = f(E) \tag{13}$$

$$f(U) = W^T f(E) \tag{14}$$

Note that $W = (I - A)$ and $A_{i,i} = 0$. So in $W$, the value of the elements on the diagonal is

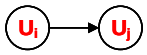
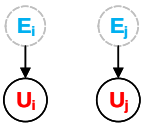
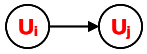
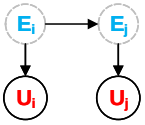
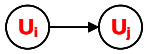
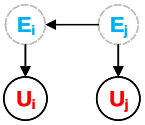
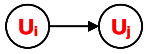
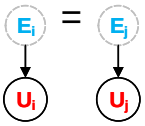
| | | Causal model | | Prediction | | Evaluation statistic | |
|---|---|---|---|---|---|---|---|
| **Case** | **Label** | **Baseline** | **CAE** | **Baseline** | **CAE** | **Baseline** | **CAE** |
| **1** | (j, j) |  |  | (j, i) | (j, j) | 2137/3203 (66.7%) | 2285/3203 (71.3%) |
| **2** | (j, i) |  |  | (j, i) | (j, i) | 2376/3291 (72.2%) | 2435/3291 (74.0%) |
| **3** | |  |  | (j, i) | | 1223/1770 (69.1%) | 1253/1732 (72.4%) |
| **4** | (j, z) |  |  | (j, i) | (j, i) | 958/1694 (56.6%) | 907/1596 (56.8%) |

Figure 7: Four causal models and corresponding evaluation results between baseline and CAE. Based on cross validation, there are 3203 pairs whose emotion and cause are the same utterance and 3291 pairs whose emotion and cause are different utterances. Baseline is DAG-ERC in ECPE task. $F1$ score is $69.5\%$, the number of all predicted pairs is $6511$ (true pairs $4513$ and false pairs $1998$). CAE is **VI** in ECPE task. $F1$ score is $72.7\%$, the number of all predicted pairs is $6565$ (true pairs $4720$ and false pairs $1845$).

constant at $1$ and is a constant maximum of each column. Naturally, $f(E)$ is an approximate estimate of $f(U)$, which is why we think implicit causes are reasonable when the F1 score of Table 6 is high.

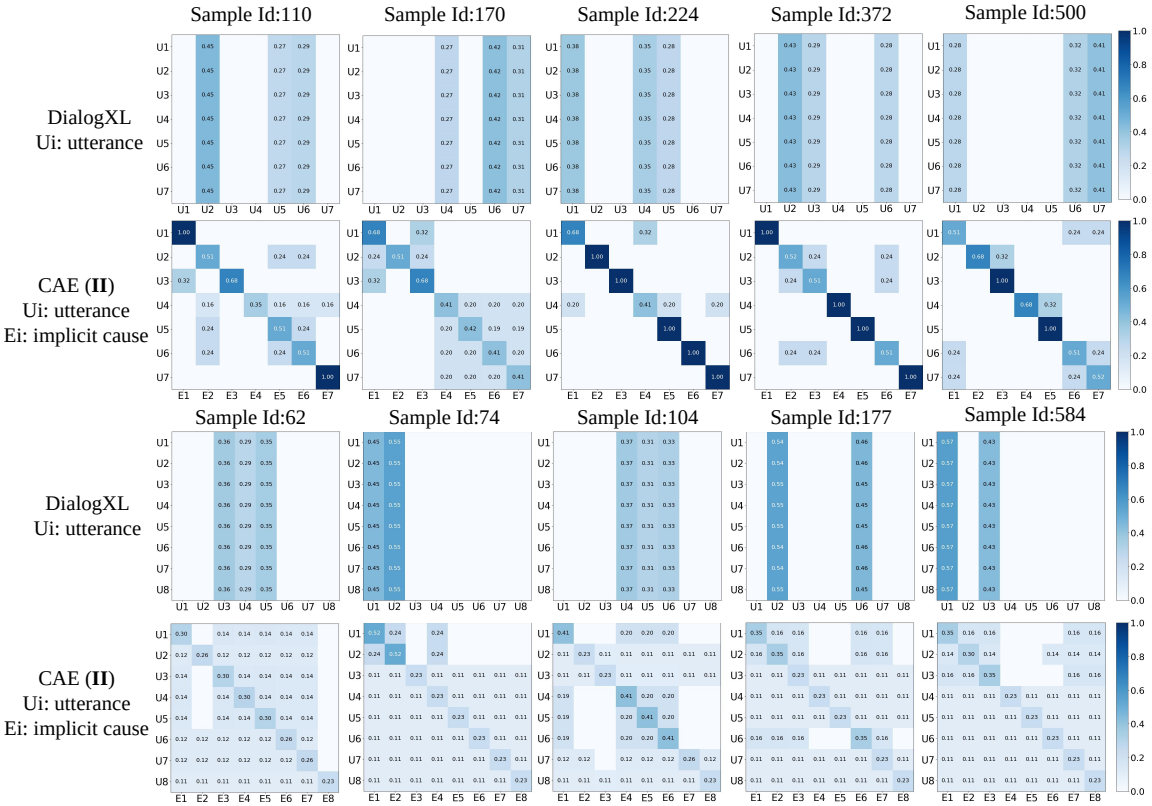



Figure 8: Causal Graph cases of DialogXL and CAE (**II**).

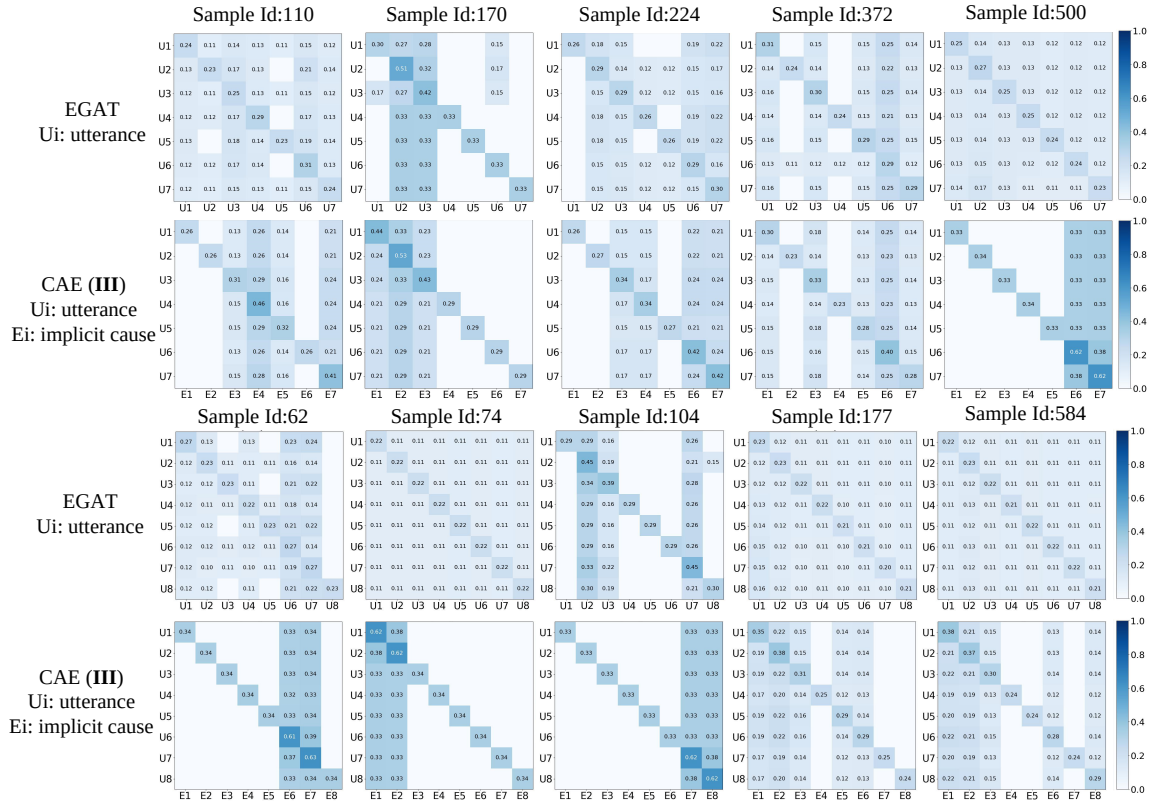



Figure 9: Causal Graph cases of EGAT and CAE (**III**).

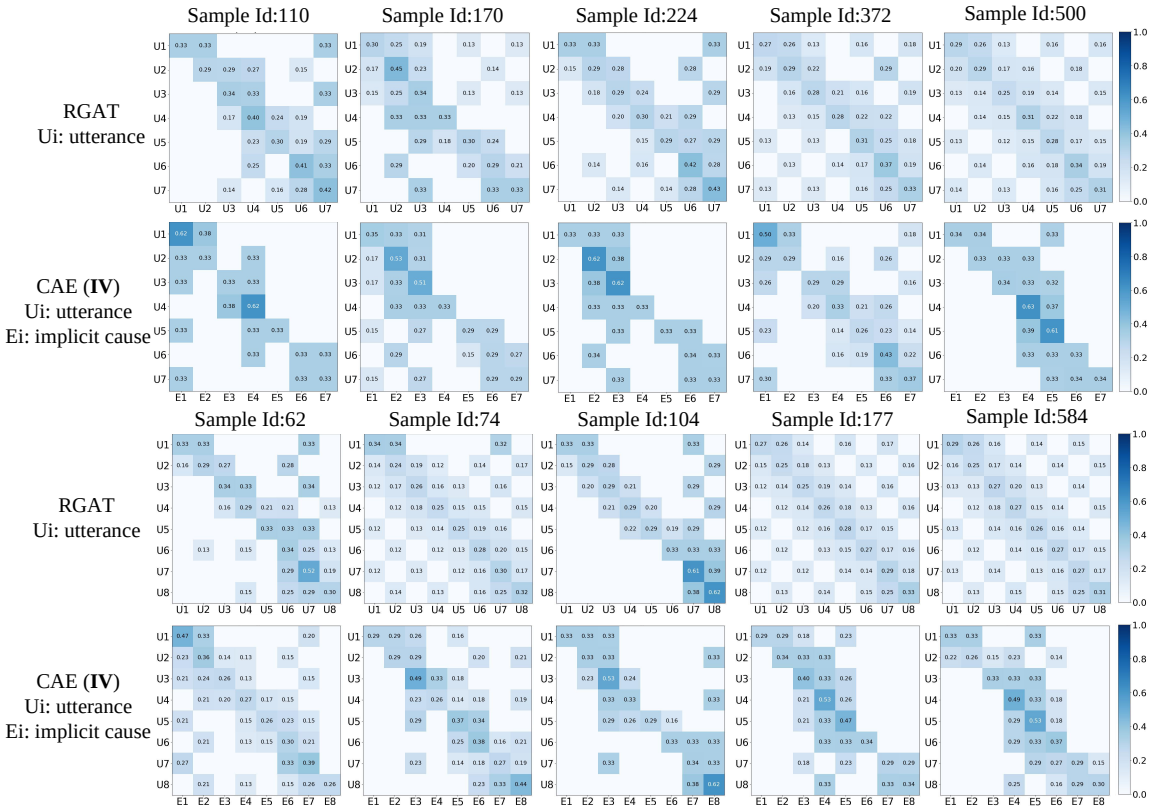



Figure 10: Causal Graph cases of RGAT and CAE (**IV**).

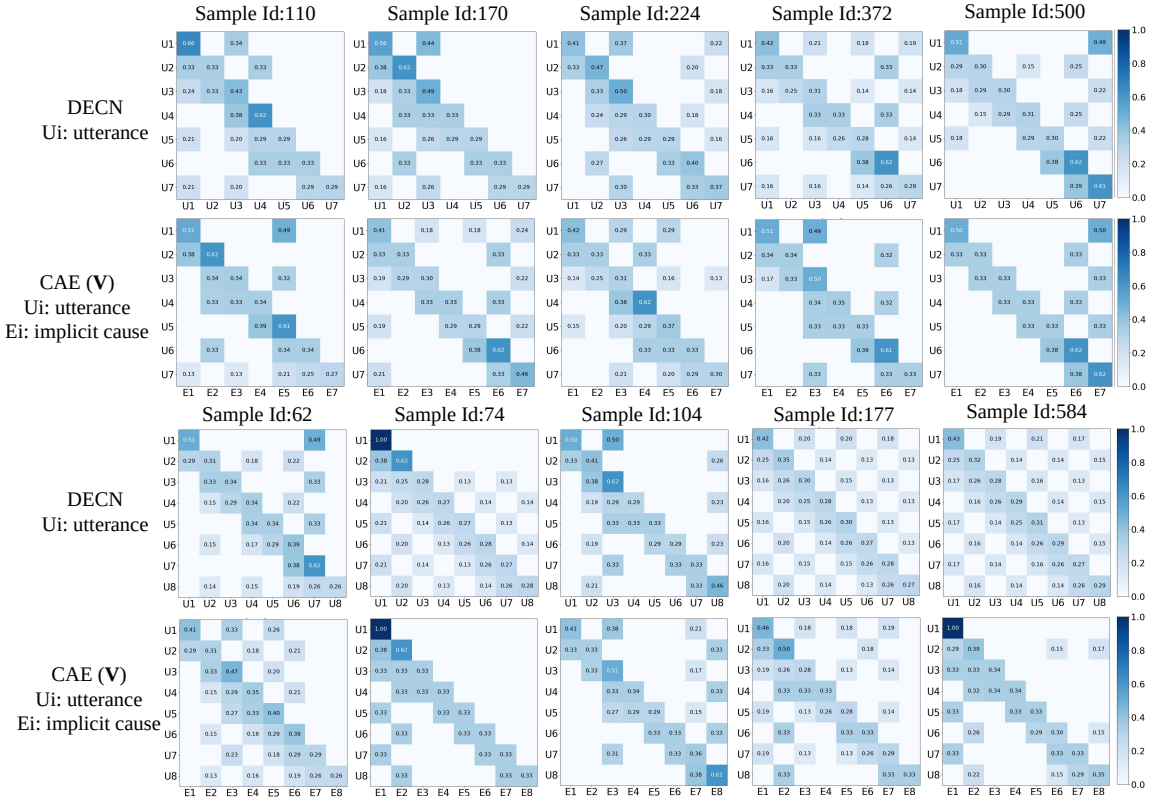



Figure 11: Causal Graph cases of DECN and CAE (**V**).

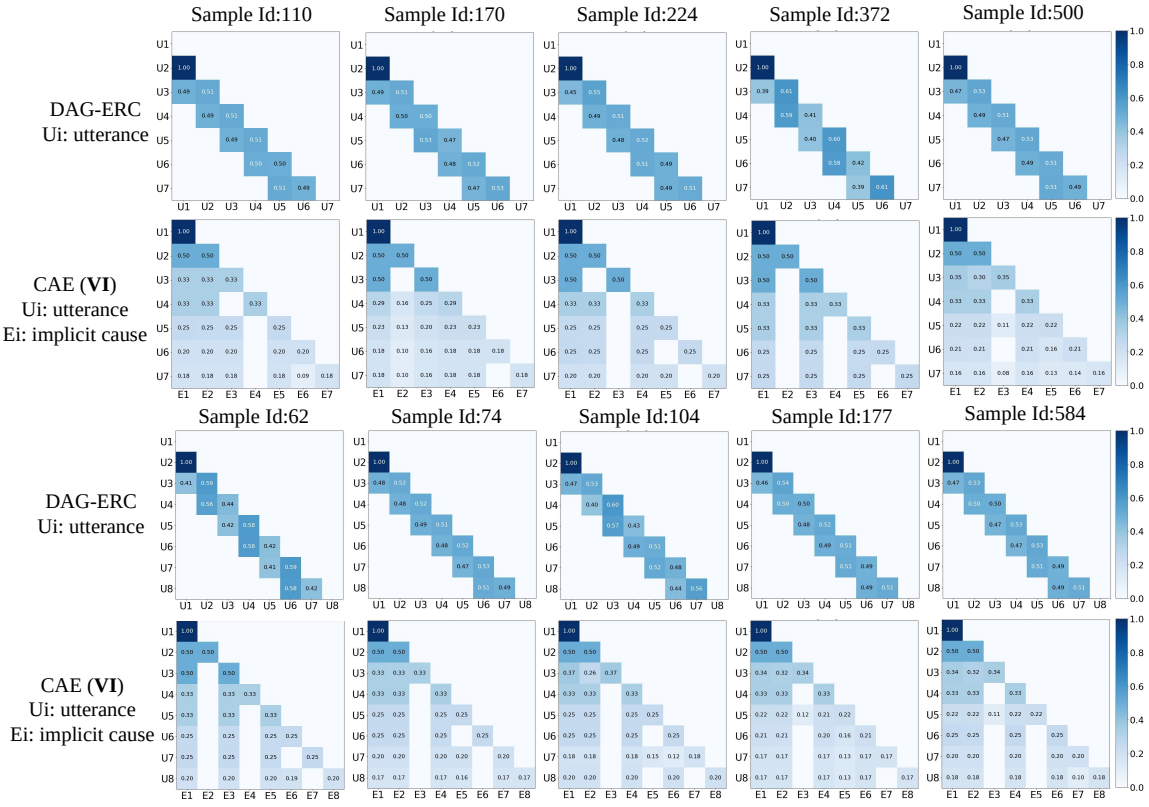



Figure 12: Causal Graph cases of DAG-ERC and CAE (**VI**).