# Incorporating Causal Analysis into Diversified and Logical Response Generation

**Jiayi Liu[1], Wei Wei[2,3]*, Zhixuan Chu[4], Xing Gao[1], Ji Zhang[1], Tan Yan[4], Yulin Kang[4]**
[1]Alibaba Group, China
[2]CCIIP Laboratory, Huazhong University of Science and Technology, China
[3]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL), China
[4]Ant Group, China
{ljy269999,chuzhixuan.czx,gaoxing.gx,zj122146}@alibaba-inc.com
weiw@hust.edu.cn, tanyan.ty@gmail.com, yulin.kyl@antgroup.com

## Abstract

Although the Conditional Variational Auto-Encoder (CVAE) model can generate more diversified responses than the traditional Seq2Seq model, the responses often have low relevance with the input words or are illogical with the question. A causal analysis is carried out to study the reasons behind, and a methodology of searching for the mediators and mitigating the confounding bias in dialogues is provided. Specifically, we propose to predict the mediators to preserve relevant information and auto-regressively incorporate the mediators into generating process. Besides, a dynamic topic graph guided conditional variational auto-encoder (TGG-CVAE) model is utilized to complement the semantic space and reduce the confounding bias in responses. Extensive experiments demonstrate that the proposed model is able to generate both relevant and informative responses, and outperforms the state-of-the-art in terms of automatic metrics and human evaluations.

## 1 Introduction

With recent advances in deep learning and readily available large-scale dialogue data, generation-based methods have become one of the most prevailing methods for building dialogue systems. Based on the Seq2seq framework (Sutskever et al., 2014; Cho et al., 2014), generation-based models learn to map the input post to its corresponding response through an encoding-decoding strategy and are trained in end-to-end manners (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015). However, Seq2seq model tends to produce generic and safe responses (Li et al., 2015) such as "So am I" or "I don't know". Researchers conjecture that the cause of this phenomenon is that one certain post can be replied by multiple responses (*i.e.,* one-to-many mapping), and the maximum likelihood estimation (MLE) training would average out these responses and produce a more bland and generic candidate.

| Post: | Have you had dinner? |
|---|---|
| Response1: | Yeah, sure! |
| Response2: | Yes, I had it at McDonald's. |
| Response3: | Nope, I'm busy with my work. |
| Response4: | Yes, I've had it. I tried a nearby restaurant that features Thai food. |

Table 1: An illustration of a general question and its multiple valid responses. The *direct responding semantics* (marked in red) are semantically homogeneous because they have to reply the issue directly. The *supplementary semantics* are more diversified because they add more information to explain or supplement the corresponding *direct responding semantics*.

To tackle this problem and model the one-to-many mapping relationships in dialogues, (Zhao et al., 2017) firstly leverages Conditional Variational Auto-Encoder (CVAE) model to map the input post into a semantic distribution, instead of a fixed vector as used in the vanilla Seq2seq model. The decoder then decodes the sampled points from the semantic distribution to generate corresponding responses. This model significantly increases the diversity of responses, but it is hard to train as the valid responses are too few to shape a clear semantic distribution for each post. As a result, the CVAE model is inclined to learn some spurious statistical cues for predicting diversified words, which may have very low relevance with the input post. Other studies focus on re-using the model's components to fit the multiplicity of dialogues, for instance, the multiple mechanisms used in (Zhou et al., 2017) and (Zhou et al., 2018), the multi-head attention used in (Tao et al., 2018; Liu et al., 2022), and reinforced methods(Qiu et al., 2021). The most-related work in this line is the Multi-Mapping and Posterior Mapping Selection (MMPMS) (Chen et al., 2019) model, which directly builds multiple mapping modules to learn diversified semantics and

*Corresponding author.



*Proceedings of the 29th International Conference on Computational Linguistics*, pages 378–388
October 12–17, 2022.

generate responses. However, these studies haven't considered the intrinsic nature of this one-to-many phenomenon in dialogues.

We always face the trade-off between the accuracy of response and diversity of semantics, and cannot directly generate relevant and diversified responses from the original input post. To solve this dilemma and examine the nature of dialogues, we introduce the causal inference analysis (Pearl, 1995, 2000) into the dialogue generation task. Here, we assume between the input post and outcome response, there exists one mediator. The mediator can easily capture the relevant but simple response from the input post (input post $\rightarrow$ mediator) and also can pass the learned information to the outcome so as to preserve the relevance. In addition, when generating the diversified responses, the sampling steps in prior and posterior distributions of CVAE will act as the confounders between the input and outcome response. Therefore, we establish one causal graph including the mediator, confounder, input post, and segmented responses, i.e, *direct responding semantics* and *supplementary semantics*, to facilitate the information transmission and enrichment, and preserve the relevance and logicality.

Based on the above causal analysis, this work presents a unified end-to-end sentence-level auto-regressive model (SLARM) to predict the mediator and mitigate the confounding bias in generating diverse responses. We concrete the mediator by predicting the direct responding semantics, and leverage this mediator in an auto-regressive manner for response generation. A dialogue topic graph enhanced CVAE model with a larger semantic space is proposed to reduce the confounding bias in CVAE model, and thus make sure the transition is smooth and natural. In conclusion, the contributions of this work are three-fold:

1. It provides an in-depth analysis of the underlying causality involved in the dialogue generation task, and proposed a methodology of searching for the mediators and mitigating the confounding bias in dialogues.

2. It proposes an innovative dialogue generation model based on the established causal graph with mediator and confounder. The model predicts the *direct responding semantics* as mediators and generate the *supplementary semantics* in a unified auto-regressive manner using the proposed TGG-CVAE part to mitigate the confounding bias.

3. It conducts broad experiments on a real-world dialogue dataset, which demonstrates that our proposed approach outperforms the state-of-the-art methods and has the capability of enhancing the diversity of responses without the sacrifice of relevance.

## 2 Related Works

**Diversified Generation models.** Some researchers suggest that the maximum-likelihood training objective used in the seq2seq model will average out the targets and result in safe and commonplace responses. Several attempts have been made to tackle this problem by proposing diversity-promoting objective functions, such as Maximum Mutual Information (MMI) (Li et al., 2015), Inverse Token Frequency Loss (ITF) (Nakamura et al., 2018). Although these studies help mitigate the safe response problem, their performance is far from satisfactory. Recently, researchers have discovered that incorporating additional information can lead to more diverse responses. Such methods include predicting keywords to guide the generation process (Mou et al., 2016; Yao et al., 2017), and using latent variables such as (Zhao et al., 2017; Gao et al., 2019a,b; Wei et al., 2019, 2021). Some recent studies focus on the one-to-many relationship between a certain post and its multiple valid responses, which is a common phenomenon in real dialogues. For instance, (Zhou et al., 2017) and (Zhou et al., 2018) model the one-to-many mapping relationships through multiple latent mechanisms and leverage diverse mechanisms to enhance the diversity of generated responses. (Tao et al., 2018) leverages the multi-head attention to focus on different parts of the input post and generate diverse responses. The state-of-the-art model in this line is the Multi-Mapping and Posterior Mapping Selection (MMPMS) (Chen et al., 2019) model, which directly builds multiple mapping modules to learn diversified semantics and generate responses.

**Causal Inference.** Causal inference (Pearl, 2000; Rubin, 2005) has been an attractive research topic for a long time since it provides an effective way to uncover causal relationships in real-world problems. Nowadays, the combination of the incisive ideas in the causal inference and various deep learning model can help improve existing methodologies in a wide range of fields, such as treatment effect es-

timation with observational data (Li and Fu, 2017; Chu et al., 2020b, 2022b), causality analysis of graph networked data (Chu et al., 2021), continual learning (Hu et al., 2021; Chu et al., 2020a), natural language processing task (Yang et al., 2021; Niu et al., 2021; Abbasnejad et al., 2020), few-shot learning (Yue et al., 2020, 2021), domain adaptation (Bengio et al., 2019), clinical trials (Chu et al., 2022c), finance (Atanasov and Black, 2016), accounting (Gow et al., 2016), marketing campaigns (Chu et al., 2022a) and so on. It is very challenging to choose or define proper confounders and mediators so as to construct one reasonable causal graph for different new tasks. A confounder is related to both cause and effect in a study, and a mediator explains the process by which cause and effect are related. In this work, we aim to incorporate causal inference into the dialogue generation task to help the model balance the relevance and diversity of response semantics.

## 3 Causal Analysis

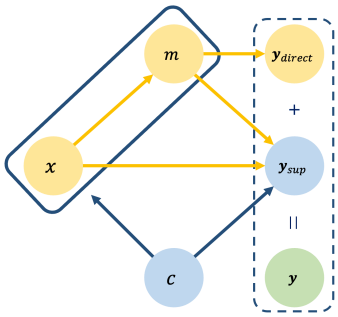


Figure 1: The causal graph of dialogue generation task. The dialogue generation task can be naturally abstracted to one causal graph involving input post $x$, confounder $c$, mediators $m$, direct responding semantics $\boldsymbol{y}_{direct}$, and supplementary semantics $\boldsymbol{y}_{sup}$. The direct responding semantics $\boldsymbol{y}_{direct}$ is the proxy variable of mediator $m$. The complete response $y$ consists of direct responding semantics $\boldsymbol{y}_{direct}$, and supplementary semantics $\boldsymbol{y}_{sup}$.

In this section, we introduce the causal inference analysis (Pearl, 1995, 2000; Yao et al., 2021) into this task and define the mediator and confounder in the dialogue generation causal graph. A mediator is determined by input post and has causal effects on outcome response, and a confounder has causal effects on both input post and outcome response. Our objective is to leverage the causal relationship involved in the established causal graph to increase the diversity of response semantics, but at the same time, not to reduce the relevance of response to input post.

We assume there exists one mediator between the input post and outcome response. The mediator can easily capture the relevant but simple response from the input post (input post $\rightarrow$ mediator) and also can pass the learned information to outcome so as to preserve the relevance (mediator $\rightarrow$ outcome response). Except for the path via mediator, the input post is also directly predictive of the outcome response (input post $\rightarrow$ outcome response). In addition, we propose to use the CVAE to increase the diversity of responses. However, the sampling steps in prior and posterior distributions of CVAE will act as the confounder between the input combination (input post and mediator) and outcome response (input combination $\leftarrow$ confounder $\rightarrow$ outcome response). Because the input combination and response pairs maybe do not conform to the assumed prior or posterior distributions of CVAE, this confounding bias may make the model learn the spurious statistical cues for the prediction of diversified response, resulting in some linguistically similar but inconsistent or irrelevant expressions in the generated sentences. Therefore, reducing the confounding bias is essential for the dialogue generation task.

Corresponding to the above causal relationship, we split the complete response into two parts, i.e., *direct responding semantics* and *supplementary semantics*, as shown in Figure 1. The *direct responding semantics* represents the semantic part that can be directly leveraged to answer the input question. The direct responding semantics is the proxy variable of the mediator. The *Supplementary semantics* represents the peripheral semantic part that is either an explanation, a supplement, or an extension of the *direct responding semantics*. The *direct responding semantics* is semantically homogeneous because it has to solve the issue directly, and the *supplementary semantics* is more diversified because it adds more information to explain or supplement the *direct responding semantics*, or even change the topics to make conversation continue.

Although the high-quality observations of the mediators can reduce the confounding bias hidden in the causal structure by reducing the possibility of counting on the confounders, it is not enough to attain one relevant and diversified response in the complex dialogue generation task. In addition, unlike the mediator that can be represented by direct responding semantics, it is very challenging to define and construct the exact confounders clearly. Therefore, due to the complex

causal graph and hidden confounders, the front-door and back-door adjustments (Glymour et al., 2016; Pearl and Mackenzie, 2018) for reducing the confounding bias cannot be easily applied. Therefore, instead of the conventional causal intervention based on Pearl's do-calculus (Pearl and Mackenzie, 2018), we propose to exploit the dialogue topic graph to complement the semantic space and assign more relevant information into CVAE, which can enhance the diversity and keep the relevance of input post simultaneously.

## 4 Proposed Model

Our response generation task is defined as follows. Given an input post $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$, the problem is to generate the corresponding response sequence $\boldsymbol{y} = \{y_1, y_2, \cdots, y_{T''}\}$, where $T$ is the length of the post and $T''$ is the length of response.

To address this problem, we propose to generate the response sequence in a sentence-level auto-regressive manner. Firstly, we predict the proxy variable of the mediator by maximizing the log-likelihood of the following formula:

$$\boldsymbol{y}^*_{direct} = \arg\max \Pr(\boldsymbol{y}|\boldsymbol{x}). \quad (1)$$

As mentioned above, this process produce general responses, but they are closely related to the input post and may help determine where the conversation should go. Hence, we preserve the causal path (input post $\rightarrow$ mediator) and then we can transmit the learned information to outcome so as to preserve the relevance (mediator $\rightarrow$ outcome response).

Then, an Sentence Level Auto-Regressive generating Model (SLARM) is proposed to produce diverse and informative responses based on the mediator and the dialogue topic graph. We first propose to utilize the predicted mediator in an auto-regressive manner:

$$\boldsymbol{y}^*_{sup} = \arg\max \Pr(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{y}^*_{direct}), \quad (2)$$

and then build a topic graph enhanced CVAE model to mitigate the confounding bias in traditional CVAE models. The auto-regressive training manner serves like a prompt to naturally inject the mediator into generation process, and the topic graph provides dynamic guidance to prevent the CVAE model from off-the-topic deviation and complement the semantic space.

### 4.1 Mediator Predictor

As aforementioned, we need to capture the relevant information with the input post and thus we need to predict the mediators in dialogue. Here, we propose to leverage Seq2seq-model with attention mechanism as the mediator predictor to generate *direct responding semantics*. This deterministic model can easily capture this simple semantic responding pattern and produce relevant response for our further processing.

### 4.2 Auto-Regressive Response Generator

So far, we have utilized the direct responding semantics generator to attain the mediator. Except for the path via mediator, the input post is also directly predictive of the diversified response. Now, based on the combination of input post and *direct responding semantics*, we aim to learn the *supplementary semantics*. The *supplementary semantics* is of great importance to provide useful information for interlocutors, and it can be rendered as an explanation, supplement, or extension of the previous *direct responding semantics*. This semantic part has great diversity and contains many relevant entities. Although the high-quality observations of the mediators can reduce the confounding bias hidden in CVAE, it is not enough to attain one relevant and diversified *supplementary semantics* in the complex dialogue generation task. Following the previous causal analysis, we propose to exploit the dialogue topic graph to complement the semantic space and assign more relevant information into CVAE. Therefore, we design a novel model, *i.e.,* topic graph guided CVAE model (TGG-CVAE), to extend the semantic space in the conversation and sample more diversified and relevant sentences, and leverage the dynamic guidance from the dialogue topic graph to provide smooth and natural transition from the *direct responding semantics* to the *supplementary semantics*. The model structure is depicted in Figure 2.

To generate the supplementary semantics, the proposed TGG-CVAE model takes in the input post and previously generated direct semantics response. We denote the input $\hat{\boldsymbol{x}}$ as:

$$\hat{\boldsymbol{x}} = \{x_1, x_2, ..., x_T, [SEP], y_1, y_2, ..., y_{T'}\}, \quad (3)$$

where the [SEP] token is a special token to separate the two sentences (Devlin et al., 2018). The goal of this model is to generate the *supplementary*

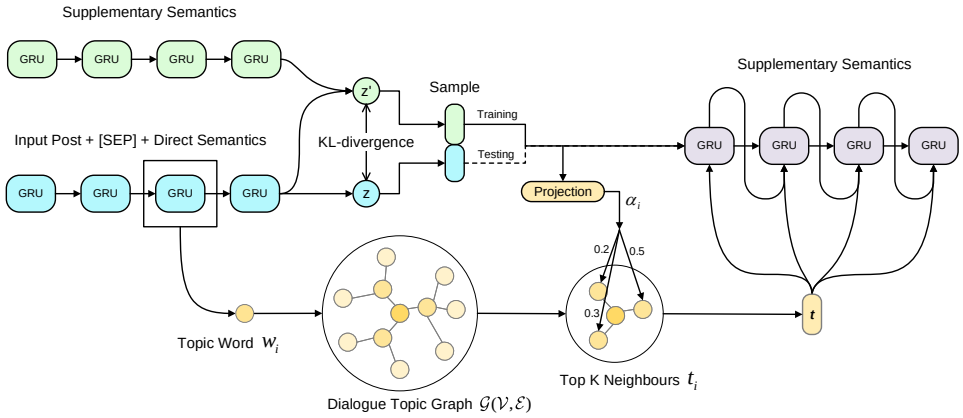


Figure 2: The architecture of our proposed TGG-CVAE model.

*semantics*:

$$\boldsymbol{y}_{sup} = \{y_{T'+1}, y_{T'+2}, \cdots, y_{T''}\}. \quad (4)$$

This model mainly consists of four components: a prior network, a posterior network, a topic guide network, and a decoding network. The prior network is trained to approximate $p_\theta(z|\hat{\boldsymbol{x}})$ while the posterior network is trained to approximate $q_\psi(z|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup})$, where $\theta$ and $\psi$ are the network parameters and $z$ is the latent variable. Here, $z$ is assumed to follow multivariate Gaussian distribution (Zhao et al., 2017) and then we have:

$$p_\theta(z|\hat{\boldsymbol{x}}) \sim \mathcal{N}(\mu, \sigma^2\mathbf{I}) \quad (5)$$
$$q_\psi(z|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup}) \sim \mathcal{N}(\mu', \sigma'^2\mathbf{I}) \quad (6)$$

Typically, the prior network and the posterior network are RNN-based encoders that transform the input $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}_{sup}$ into hidden states:

$$\boldsymbol{h}_x^i = f(\boldsymbol{h}_x^{i-1}, \hat{\boldsymbol{x}}_i) \quad (7)$$

$$\boldsymbol{h}_y^j = f(\boldsymbol{h}_y^{j-1}, \boldsymbol{y}_j), \quad (8)$$

where $i \in [1, T+T'+1]$ and $j \in [T'+1, T'']$. The last hidden states from the prior/posterior network are denoted as $\boldsymbol{h}_x$ and $\boldsymbol{h}_y$ respectively. The latent variable is estimated by parameterizing its mean and log variance:

$$\begin{bmatrix} \mu \\ \log(\sigma^2) \end{bmatrix} = \boldsymbol{W}_p(\boldsymbol{h}_x) + b_p \quad (9)$$

$$\begin{bmatrix} \mu' \\ \log(\sigma'^2) \end{bmatrix} = \boldsymbol{W}_r \begin{bmatrix} \boldsymbol{h}_x \\ \boldsymbol{h}_y \end{bmatrix} + b_r, \quad (10)$$

where the $\boldsymbol{W}_p$, $\boldsymbol{W}_r$ and $b_p$, $b_r$ are the weights and biases for the prior network and posterior network respectively. Reparametrization trick (Kingma and Welling, 2014) is used to keep the gradient propagate successfully in networks via a differentiable transformation of an auxiliary noise variable $\epsilon$:

$$z = \mu + \sigma\epsilon \quad (11)$$
$$z' = \mu' + \sigma'\epsilon \quad (12)$$

Then we can sample the latent variables $z$ or $z'$ from the prior network or the posterior network. However, in the testing stage, as the ground-truth response is not available, the posterior latent variable $\boldsymbol{z}'$ cannot be properly estimated. Therefore, we need to make sure that the prior network can fully acquire useful information from the posterior network by homogenizing $\boldsymbol{z}$ and $\boldsymbol{z}'$. Here, KL-divergence loss is leveraged in our model to minimize the discrepancy between the two latent distributions:

$$\begin{aligned} \mathcal{L}_{KL} =& KL(q_\psi(\boldsymbol{z}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup})||p_\theta(\boldsymbol{z}|\hat{\boldsymbol{x}})) \\ =& \int q_\psi(\boldsymbol{z}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup}) log \frac{q_\psi(\boldsymbol{z}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup})}{p_\theta(\boldsymbol{z}|\hat{\boldsymbol{x}})} dz, \end{aligned} \quad (13)$$

from which we can derive the final formula for calculating KL-divergence loss:

$$\mathcal{L}_{KL} = log\frac{\sigma}{\sigma'} + \frac{\sigma'^2 + (\mu - \mu')^2}{2\sigma^2} - \frac{1}{2} \quad (14)$$

For the vanilla CVAE model, $\boldsymbol{z}$ or $\boldsymbol{z}'$ is directly fed as the input of the decoder for decoding from the semantic space. However, as aforementioned, the latent semantic space is too large to train well and the sampling steps in prior and posterior distributions of vanilla CVAE will act as the confounder

between the input combination (input post and *direct responding semantics*) and *supplementary semantics*. Because the input post and response pairs in the real data maybe do not conform to the assumed prior or posterior distributions of CVAE, this confounding bias may make the model learn the spurious statistical cues for prediction of diversified response, resulting in some linguistically similar but inconsistent or irrelevant expressions in the generated sentences. Therefore, reducing the confounding bias is essential for the *supplementary semantics* generation. We exploit the dialogue topic graph to complement the semantic space and assign more accurate and relevant relationship into CVAE so as to mitigate the confounding bias. Details of this strategy are as follows:

Firstly, the topic words $w_1, w_2, ..., w_m$ are extracted from $\hat{\boldsymbol{x}}$ using the TF-IDF method, and then they are placed into the dialogue topic graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ to find their nearest $n$ neighbours $w_{11}, w_{12}, ..., w_{mn}$ according to the weight, where $w_{ij}$ is the $j$-th neighbours of the topic word $w_i$. We then choose the neighbour with the highest probabilities:

$$t_1, ..., t_K = \underset{i\in[1,m], j\in[1,n]}{\text{argmax}_K} (\Pr(w_{ij}|w_i)), \quad (15)$$

to select top K topic words, namely, $t_1, ..., t_K$.

Secondly, since these topic words contribute differently to the generation of a response, we leverage the sampled latent variables to formulate a dynamic prior/posterior selection of the topic words. The sampled latent variables $\boldsymbol{z}$ (testing) or $\boldsymbol{z}'$ (training) are passed through a projection layer to produce a distribution over the K topic words, namely $\boldsymbol{\alpha} = \alpha_1, \alpha_2, ..., \alpha_K$. The final representation of the topic information is formulated as a weighted summation of the topic embeddings:

$$\boldsymbol{t} = \alpha_i \cdot \boldsymbol{t_i}, \quad i = 1, 2, 3, ..., k \quad (16)$$

where $\boldsymbol{t_i}$ is the embedding of the word $t_i$.

Thirdly, the topic information $\boldsymbol{t}$ and the sampled latent variable $\boldsymbol{z}$ or $\boldsymbol{z}'$ are fed into the decoder for generating the supplementary semantics:

$$\Pr(y_t|y_{1:t-1}, \boldsymbol{x}, \boldsymbol{y}_{direct}) = g(\boldsymbol{y}_{t-1}, \boldsymbol{s_t}, \boldsymbol{z}, \boldsymbol{t}), \quad (17)$$

from which we can find that each supplementary semantic word is conditioned on both the topic information and the sampled latent variable, and thus the sentences can be related to the previous words and have more diversity. Following (Sohn et al., 2015), we train the proposed model by maximizing the variational lower bound of the conditional log likelihood:

$$\begin{aligned}\mathcal{L}_{ELBO} = &- KL(q_\psi(\boldsymbol{z}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup})||p_\theta(\boldsymbol{z}|\hat{\boldsymbol{x}})) \\ &+ \mathrm{E}_{q_\psi(\boldsymbol{z}|\hat{\boldsymbol{x}}, \boldsymbol{y}_{sup})}[\log p_\theta(\boldsymbol{y}_{sup}|\boldsymbol{z}, \hat{\boldsymbol{x}}, \boldsymbol{t})],\end{aligned} \quad (18)$$

where the $KL(.,.)$ denotes the KL-divergence of two distributions. Since the latent semantic distribution is easy to collapse (*a.k.a.,* KL-collapse problem), we add a bag-of-words loss $\mathcal{L}_{BOW}$ and use KL-annealing strategy to deal with this problem (Zhao et al., 2017). The final loss function of this proposed model is formulated as:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{BOW} + \mathcal{L}_{direct} \quad (19)$$

Note that, we also consider the possible circumstance where the responses do not contain any *supplementary semantics* by leveraging the [EOS] token as the placeholder. If the TGG-CVAE model predicts [EOS] token at the first step, this indicates that the *direct responding semantics* is already complete and it does not need any *supplementary semantics*.

## 5 Experimental Results

### 5.1 Dataset

We conduct experiments on a large-scale real-world dialogue dataset, *i.e.,* Short-Text Conversation (STC) dataset (Shang et al., 2015). This dataset is publicly available and is cleaned by the data publishers. It consists of 4,433,853 post-comment pairs collected from Chinese Weibo, a social media platform where people can chat online.

### 5.2 Evaluation Metrics

**Automatic Evaluation**. We adopted two widely-used metrics, *BLEU-n* (Papineni et al., 2002) and *Distinct-n* (Li et al., 2015), to automatically evaluate the dialogue generation models. *BLEU-n* score is a referenced evaluation metric to measure word overlap between the generated response and the reference. Note that in our experiment we apply smoothing function 7 (Chen and Cherry, 2014) to avoid the problem when no $n$-gram overlaps are found. *Distinct-n* score (Li et al., 2015) is used to determine word-level diversity of the generated response. It is measured by calculating the percentage of distinct $n$-grams in the generated responses.

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Distinct-1 | Distinct-2 |
|---|---|---|---|---|---|---|
| **Seq2seq** (Sutskever et al., 2014) | 0.2392 | 0.1937 | 0.1646 | 0.1304 | 0.0549 | 0.1859 |
| **CVAE** (Zhao et al., 2017) | 0.2223 | 0.1808 | 0.1541 | 0.1222 | 0.0936 | 0.4208 |
| **MMPMS** (Chen et al., 2019) | 0.2246 | 0.1868 | 0.1612 | 0.1289 | **0.0972** | **0.4214** |
| **DCVAE** (Gao et al., 2019b) | 0.2124 | 0.1700 | 0.1436 | 0.1134 | 0.0405 | 0.1681 |
| **SLARM** (ours) | **0.2657** | **0.2169** | **0.1850** | **0.1469** | 0.0879 | 0.3685 |
| **SLARM** w/o TGG (ours) | 0.2569 | 0.2099 | 0.1792 | 0.1423 | 0.0967 | 0.4088 |
| **SLARM** w/o CVAE (ours) | 0.2544 | 0.2069 | 0.1763 | 0.1398 | 0.0881 | 0.2195 |

Table 2: Automatic evaluation results on STC dataset. The best results are in boldface and the second best results are underlined.

| Models | Relevance | Informativeness | Fluency | Average |
|---|---|---|---|---|
| **Seq2seq** (Sutskever et al., 2014) | 1.52 | 1.63 | **2.68** | 1.94 |
| **CVAE** (Zhao et al., 2017) | 1.45 | 1.73 | 2.49 | 1.89 |
| **MMPMS** (Chen et al., 2019) | 1.54 | **2.02** | 2.00 | 1.85 |
| **DCVAE** (Gao et al., 2019b) | **1.96** | 1.53 | 2.48 | 1.99 |
| **SLARM** (ours) | 1.57 | 1.82 | 2.67 | **2.02** |

Table 3: Human evaluation results on STC dataset. The best results are in boldface and the second best results are underlined.

**Human Evaluation**. We randomly sampled 100 posts from the test set and let the models generate corresponding responses. Three annotators were invited to rate the post-response pairs from three aspects: relevance (whether the response is relevant to the input post), informativeness (whether the response is informative) and fluency (whether the response has no grammar mistakes). A three-point scale (0,1,2) is used in the evaluation for the above aspects. When contradiction occurs between the first two annotators, the third annotator will resolve the disagreement. Fleiss' kappa (Fleiss and Cohen, 1973) is calculated to measure the inter-rater agreement between the first two annotators.

### 5.3 Baseline Models

**Seq2seq** (Bahdanau et al., 2014): it is a canonical seq2seq model with the attention mechanism. **CVAE** (Zhao et al., 2017): it is a conditional variational auto-encoder model. During testing, we randomly sample latent variables from the prior network and generate corresponding responses. **MMPMS** (Chen et al., 2019): it is a multi-mapping and posterior mapping selection model. We use their original implementation and hyper-parameter settings. **DCVAE** (Gao et al., 2019b): it is a discrete CVAE model. We use their original implementation and adopt the two-stage sampling strategy during testing.

### 5.4 Implementation Details

For our approach, we use 2-layers GRU units for encoders in the prior network/posterior network and the hidden size is set to 256. The embedding size and vocabulary size are set to 200 and 40,000 respectively. Word embeddings are randomly initialized and OOV (out-of-vocabulary) words are replaced with a special token UNK. Adam optimizer (Kingma and Ba, 2014) is used for optimization and the training batch size is 128. The initial learning rate is set to 0.5 and a learning rate decay operation is employed when the validation loss stops decreasing for three consecutive epochs. The decay rate is 0.99. The top 5 neighbors of the topic words in the dialogue graph are chosen and fed into the decoder.

### 5.5 Results

**Automatic evaluation** results are shown in Table 2. Notably, our SLARM model outperforms all of the baselines in terms of BLUE metric (with p-value $< 0.05$) and its performance is 11.2% ahead of the second best model. This verifies our assumption that splitting the to-be-generated responses into different semantic parts and separately generating them with suitable methods will enhance the overall performance. As for the Distinct metric, the performance of our model is moderate compared to the CVAE model and MMPMS model. This is because our main objective is not only boosting the diversity of responses but also promoting relevance

between posts and generated responses.

To further analyze the results, we conduct ablation studies by removing the Topic Graph Guided module (*i.e.,* SLARM w/o TGG) or replacing the CVAE module with traditional GRUs (*i.e.,* SLARM w/o CVAE). After removing TGG, the Distinct performance increases and the BLEU performance decreases. This indicates that our dynamic topic graph guiding strategy is effective in providing relevant information from posts and thus can increase BLEU scores. However, this strategy gets slightly lower Distinct scores because the restrained topics would reduce possibilities in choosing more diversified words. When the CVAE module is removed, the Distinct-2 score drops by a large margin, indicating the CVAE module is effective in extending the semantic space and sampling diversified phrases. The BLUE scores also decrease because the posterior network is essential in providing additional information to dynamically weigh the contribution of topic words. Hence, each component of our model complements each other, and thus the model has the self-adaptive capability to reach a balance between diversity and relevance.

**Human evaluation** results are shown in Table 3. The DCVAE model surpasses other models in relevance metric. This is owing to the fact that DCVAE model tends to re-use the words in the posts to generate a response, which makes the annotators give high relevance scores. The discrete latent variables from the prior and posterior network are pretrained to predict keywords in the post, and thus sampling from these variables tends to produce the same words in the post. However, the latent variables constrain the generation process, which leads to low informative scores. The MMPMS model performs the best from the informativeness aspect. This is because the auxiliary loss (*i.e.,* matching loss) is effective in encouraging the selection module to choose different and diverse mapping modules. However, some mapping modules are not well-trained and they generate ungrammatical sentences. Therefore their fluency score is rather low. The Seq2seq model gets the highest fluency score, as it often generates common and simple sentences.

Our proposed SLARM model outperforms all the baseline models in terms of the average score. For every single aspect, the SLARM model consistently obtains the second best scores. The second best relevance score indicates that first generating the *direct responding semantics* will assure the relevance with the post because it directly answers the question. The second best informative score shows that the proposed SLARM model can enhance the diversity and generate informative sentences. Our fluency score is also the second best and is close to the Seq2seq model's, which verifies that our methods can alleviate the grammatical problems when concatenating two semantic parts.

Note that the Fleiss' kappa for relevance, informativeness, and fluency are 0.4153, 0.4188, and 0.4378, respectively, indicating "moderate agreement" among the annotators.

### 5.6 Case Study

We present sampled 4 cased in our Appendix. As is shown in the figure, the Seq2seq model tends to generate safe and generic responses, such as case 1, 2, and 3. The response pattern generated by Seq2seq models often starts with "I also like..." or "Haha...", which makes the responses dull and boring. However, in cases 1 and 3, although these responses are generic, they are semantically appropriate and relevant according to the post. Therefore, this model is suitable for searching for the mediators in dialogue generation, which is actually observed in our model by *direct responding semantics*. The CVAE model is better at choosing diversified words such as "punchline" in case 3 and "autumn" in case 4. However, the confounding bias makes this model learn some spurious statistical cues for predicting diversified words, and thus these words are not logical with regard to the input question. The MMPMS model can produce informative sentences, such as in cases 1, 3, and 4. In case 3, the MMPMS model produces a response that is not only informative but also relevant to the input post, but responses in case 1 are irrelevant. Besides, another major problem is that some of the mapping modules are not well-trained and thus in case 2 we can see the generated sentence is ungrammatical and irrelevant. The DCVAE model tends to copy the input post, such as in case 4. This is the reason why the relevant score for DCVAE model is higher than other models. In some circumstances, DCVAE produces the same results as Seq2seq, such as in case 1 and 2.

We can conclude that the SLARM model performs the best and reaches a balance between relevance and diversity. The *direct responding semantics* (marked in red) in case 1, 2, 3, and 4 are very relevant to the input post, and the *supplementary se-*

| Case | Posts | Seq2seq | CVAE | MMPMS | DCVAE | SLARM |
|---|---|---|---|---|---|---|
| 1 | 我只喜欢雨天游泳。痛快。I only like to swim on rainy days. Nice! | 我也喜欢这种感觉。I also like this feeling. | 不错哦，可以注意哦。Very good, you should notice that. | 看小图我以为我是沙发 I thought that was a couch according to the small picture. | 我也喜欢这样的感觉 I also like this feeling. | 游泳训练不是很好，注意安全 It is not very good to do such swimming exercise. Be careful. |
| 2 | 很多人是复杂的动物，摸不透，搞不清 Many people are very complicated, and you can't figure them out. | 哈哈，我也喜欢，我也喜欢 Haha, I like that too, I like that too. | 这就是我们的生活了 This is our life. | 他花生的花生是花生的花生和花生的花生 His peanuts peanuts peanuts peanuts. | 我也是这么想的 I think so. | 我承认，我是这样的。I should admit, that I'm of such kind. |
| 3 | 《学习之神》,男主角不错，93 年的 The leading man in *The God of Learning* is very nice. He was born in 1993. | 我也是这么认为的。I think so. | 这是你的笑点吗? Is that your punchline? | 哈哈，钟汉良的发色也很帅啊。Haha, Hanliang Zhong's hair style also looks handsome. | 看一次笑一次…… Every time I see that I can't help laughing. | 这个是我的最爱！我就被他打到了 He is my favorite! I'm only struck by him. |
| 4 | 漓江的春天拍得真美！This photo of the Li River in spring is so beautiful! | 春天在哪里？春天在哪里？ Where is spring? Where is spring? | 我喜欢！我的秋天！I like that! This is my autumns! | 美得让人不敢恭维！It is so beautiful that I dare not to compliment! | 美！美！美！美！ Beautiful! Beautiful! Beautiful! Beautiful! | 好美啊！求推荐！It is so beautiful! Please recommend me more! |

Figure 3: Case study of the sampled 4 cases. For the SLARM model, words in red are generated by the *direct responding semantics* generator, while the rest are generated by *supplementary semantics* generator.

*mantics* provide more and diversified information to complete the response. In case 1, the *supplementary semantics* is generated to provide further instruction of being careful when swimming on rainy days. In case 3, the *supplementary semantics* re-emphasizes that the interlocutor is fond of the actor. Additionally, we can observe from the cases that with the dialogue topic graph guiding strategy, the transition from *direct responding semantics* to *supplementary semantics* is smooth and natural. Therefore, these cases fully demonstrate the model's capacity for generating the relevant and diversified responses via searching for the direct responding semantic parts as mediators in dialogues and then utilizing our proposed SLARM model to mitigate the confounding bias and thus enhance the diversity without the loss of relevance.

## 6 Conclusion

In this paper, we incorporate the causal analysis into the dialogue generation task by searching for the mediators and mitigating the confounding bias in dialogues. We thus propose a sentence level auto-regressive response generation model to first generate mediators to preserve relevance with the input post, and then generate the diversified semantics based on our proposed (SLARM) model. Extensive experimental results demonstrate the effectiveness of our approach. For future work, we are exploring more complicated and self-adaptive methods for locating mediators, and we are trying to leverage de-confounding methods to deal with the CVAE problem.

## 7 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No.61602197, Grant No.L1924068, Grant No.61772076, Grant No.62276110, in part by CCF-AFSG Research Fund under Grant No.RF20210005, and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL).

## References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054.

Vladimir A Atanasov and Bernard S Black. 2016. Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review*, 5:207–304.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014*, pages 362–367.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 4918–4924.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734.

Zhixuan Chu, Hui Ding, Guang Zeng, Yuchen Huang, Tan Yan, Yulin Kang, and Sheng Li. 2022a. Hierarchical capsule prediction network for marketing campaigns effect. *arXiv preprint arXiv:2208.10113*.

Zhixuan Chu, Stephen Rathbun, and Sheng Li. 2020a. Continual lifelong causal effect inference with real world evidence.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2020b. Matching in selective and balanced representation space for treatment effects estimation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 205–214.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2021. Graph infomax adversarial learning for treatment effect estimation with networked observational data. *arXiv preprint arXiv:2106.02881*.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2022b. Learning infomax and domain-independent representations for causal effect inference with real-world data. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 433–441. SIAM.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2022c. Multi-task adversarial learning for treatment effect estimation in basket trials. In *Conference on Health, Inference, and Learning*, pages 79–91. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019a. Generating multiple diverse responses for short-text conversation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6383–6390.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019b. A discrete cvae for response generation on short-text conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019*, pages 1898–1908.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Ian D Gow, David F Larcker, and Peter C Reiss. 2016. Causal inference in accounting research. *Journal of Accounting Research*, 54(2):477–523.

Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. 2021. Distilling causal effect of data in class-incremental learning.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT 2016*, pages 110–119.

Sheng Li and Yun Fu. 2017. Matching on balanced nonlinear representations for treatment effects estimation. In *NIPS*.

Yuhang Liu, Wei Wei, Daowan Peng, Xianling Mao, Zhiyong He, and Pan Zhou. 2022. Depth-aware and semantic guided relational attention network for visual question answering. *IEEE Transactions on Multimedia*.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *26th International Conference on Computational Linguistics, COLING 2016*, pages 3349–3358.

Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2018. Another diversity-promoting objective function for neural dialogue generation. *arXiv*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2000. *Causality: models, reasoning and inference*, volume 29. Springer.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Minghui Qiu, Xinjing Huang, Cen-Chieh Chen, Feng Ji, Chen Qu, Wei Wei, Jun Huang, and Yin Zhang. 2021. Reinforced history backtracking for conversational question answering. In *AAAI*.

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015*, pages 1577–1586.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28, NIPS 2015*, pages 3483–3491.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT 2015*, pages 196–205.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27, NIPS 2014*, pages 3104–3112.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4418–4424.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Workshop*.

Wei Wei, Jiayi Liu, Xian-Ling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Wei Wei, Jiayi Liu, Xian-Ling Mao, Guibing Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. 2021. Target-guided emotion-aware chat machine. *ACM Transactions on Information Systems (TOIS)*.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2190–2199.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 654–664.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 3400–3407.

Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2018. Elastic responding machine for dialog generation with dynamically mechanism selecting. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5730–5737.