CE-VDG: Counterfactual Entropy-based Bias Reduction for Video-grounded Dialogue Generation

Hongcheng Liu [1], Pingjie Wang[1,2], Zhiyuan Zhu [1], Yu Wang[1,2]*, Yanfeng Wang[1,2]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory
{hongcheng_liu,pingjiewang,zzysjtu_iwct,yuwangsjtu,wangyanfeng622}@sjtu.edu.cn

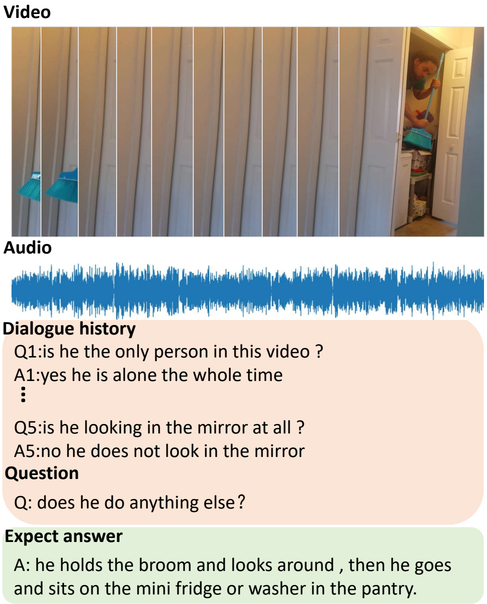
Abstract

The Video-Grounded Dialogue generation (VDG) is a challenging task requiring a comprehensive understanding of the multi-modal information to produce a pertinent response. However, VDG models may rely on dataset bias as a shortcut and fail to learn the multi-modal knowledge from both video and audio. Counterfactual reasoning is an effective method that can estimate and eliminate bias on some special aspects of classification tasks. However, conventional counterfactual reasoning cannot be applied to VDG tasks directly due to the BPE algorithm. In this paper, we reformulate the counterfactual reasoning from the information entropy perspective and extend it from the classification task to the generative task, which can effectively reduce the question-related bias in the auto-regressive generation task. We design CE-VDG to demonstrate the effectiveness in bias elimination of the reformulated counterfactual reasoning by using the proposed counterfactual entropy as an external loss. Extensive experiment results on two popular VDG datasets show the superiority of CE-VDG over the existing baseline method, demonstrating the effective debiasing capability in our model considering counterfactual entropy.

Keywords: counterfactual reasoning, video-grounded dialogue generation, information entropy

1. Introduction

Video-grounded dialogue generation (VDG) task aims to develop a system that can establish the capability to see (i.e., understand video scenes), listen (i.e., perceive audio state), read (i.e., comprehend dialogue), and write (i.e., generate responses) simultaneously (AlAmri et al., 2019), which is depicted in Figure 1. However, despite the recent advancements, the VDG dataset often contains inherent bias, which can cause VDG models to learn spurious correlations between questions and answers (Liu et al., 2022). To illustrate this problem, Figure 2 demonstrates the bias distribution example present in the AVSD-DTSC dataset, which reveals an extremely unbalanced distribution of counting questions. It is important to note that this bias may cause the system to only focus on questions when answering counting questions, disregarding other useful information. In extreme cases, even when the multi-modal information is completely different, the VDG system consistently generates the same response because the questions are similar to other counting questions.



Q1:is he the only person in this video ?
A1:yes he is alone the whole time
⋮
Q5:is he looking in the mirror at all ?
A5:no he does not look in the mirror

Q: does he do anything else？

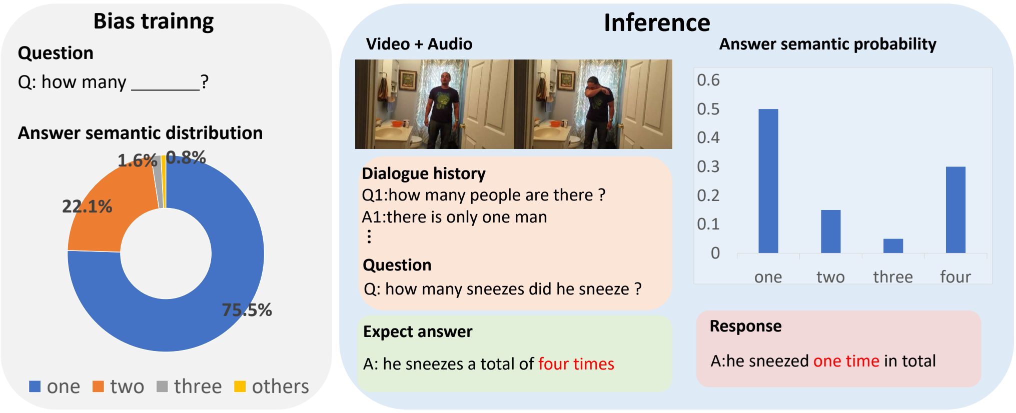
A: he holds the broom and looks around , then he goes and sits on the mini fridge or washer in the pantry.

Figure 1: The demonstration of video-grounded dialogue generation task.

To address this problem, several works have made an effort to mitigate the spurious correlations by forcing the model to concentrate more on information other than the question (Li et al., 2021; Chen et al., 2023a). However, these methods solely focus on the video and dialogue history, disregarding any bias in the dataset, which can result in biased generation. One effective approach to reduce this bias is through counterfactual reasoning. Niu et al. (2021) introduced the VQA, QA, and VA models, which utilize different modalities for answer selection and reduce bias through counterfactual subtraction. Moreover, counterfactual reasoning is widely applied in classification tasks, such as scene graph generation (Tang et al., 2020), multi-modal fake news detection (Chen et al., 2023b), and multi-modal sentiment analysis (Sun et al., 2022). However, the conventional counterfactual

[*] Corresponding authors

LREC-COLING 2024, pages 2958–2968
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Q1:how many people are there ?
A1:there is only one man
⋮

Q: how many sneezes did he sneeze ?

A: he sneezes a total of four times

A:he sneezed one time in total

Figure 2: The bias training and inference in AVSD-DSTC when the question type is counting number. As shown in the answer semantic histogram, even though the system becomes aware of the correct answer “four" through model training using semantic information from video, audio, and dialogue history, there is still a significant bias that leads the system to answer “one" during inference.

reasoning method, typically used in classification tasks as $P_{final} = P_A - P_B$, cannot directly apply to the VDG task due to the byte pair encoding (BPE) algorithm (Gage, 1994). Although the BPE can reduce computation costs by dividing the words into different common byte pairs, it also brings inherent context correlation, such as the high probability of the ‘llo’ after ‘he’ in ‘hello’. However, the inherent correlation will be drastically reduced by conventional counterfactual reasoning for the probabilities of the two models are both high in the same context and the entire generated sentence will deviate from the ground truth.

In this paper, we propose a new approach to counterfactual reasoning using information entropy. We introduce modifications to the total indirect effect (TIE). By applying this reformulated counterfactual reasoning algorithm, we extend the conventional classification task to auto-regressive generation tasks. By incorporating this formulation into generative tasks, counterfactual reasoning can effectively reduce bias related to questions in the dataset. This is achieved by eliminating spurious causal effects and leveraging multimodal information. To demonstrate the effectiveness of our approach, we introduce a new application of counterfactual reasoning in generative tasks, called counterfactual entropy. Based on this, we propose a model called CE-VDG (counterfactual entropy-based video-grounded dialogue generation), which consists of a bias estimation model and a generation model. Specifically, CE-VDG integrates counterfactual entropy as an additional training loss to mitigate the inherent bias related to questions in the final output.

To evaluate the performance of CE-VDG, we compare it with a range of state-of-the-art methods on two popular VDG benchmarks: AVSD-DSTC and NExT-OE datasets. The results show that the performance of CE-VDG significantly surpasses the existing baseline methods in all benchmarks, demonstrating the superiority and effectiveness of the counterfactual entropy.

Our main contributions are three folds:

1) We reformulate counterfactual reasoning from an information entropy perspective, extending its application from traditional classification tasks to generative tasks.
2) We propose CE-VDG, a method designed to address the bias associated with questions in datasets and effectively leverage multimodal information. This is achieved by incorporating the reformulated counterfactual entropy as an additional loss.
3) We conduct extensive experiments on two popular VDG benchmarks. The results show that our proposed CE-VDG outperforms existing methods, demonstrating the superior performance and effectiveness of our approach.

2. Related Works

The video-grounded dialogue task is to answer questions based on the content of the video and the dialogue history and requires perceiving the complex information among the visual, auditory, and text modalities. Compared to other visual dialogue tasks, VDG has a greater diversity in responses (AlAmri et al., 2019), which requires strong information integration ability between video and dialogue. To better extract semantic information, some methods choose to understand dialogue history semantics more comprehensively, such as the DialogMCF (Chen et al., 2023a) and multimodal pointer network (Le and Chen, 2020). Instead of focusing on the text modality, many methods tend to fully improve the capability of video reasoning. Many methods

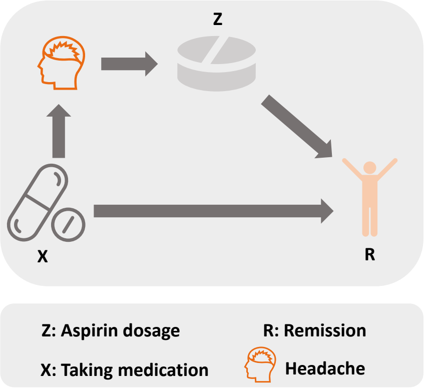


Figure 3: The illustration of the causal effect among the medicine, aspirin, and remission.

adopt approaches that provide additional video information by various extractors, such as video action recognition (Huang et al., 2022) and structured entity graph generation (Geng et al., 2021). Others adopt various mechanisms to improve the utilization of the video, such as memory network (Xie and Iacobacci, 2020) and video-audio sequence modeling (Li et al., 2021). However, these methods make the system more comprehensive understanding of specific aspects without considering the bias in the dataset and there are always huge spurious correlations in the trained models (Liu et al., 2022). In this paper, we introduce the counterfactual reasoning from the information perspective to the VDG task and propose CE-VDG to reduce the questions-related bias in the datasets.

3. Preliminaries

This section provides an introduction to the fundamental concepts of counterfactual reasoning and information entropy that will be used in this paper.

3.1. Counterfactual Reasoning

Causal Graph The causal relationship between different variables can be represented as a causal graph $\mathcal{G} = \{\nu, \varepsilon\}$, where the ν denotes the set of the variables and ε represents the causal relationships between them. Figure 3 illustrates the causality between medicine X and remission R. Besides the direct effect $X \rightarrow R$, headache is a side effect after medication and prompts the consumption of aspirin Z, which provides further pain remission R. Therefore, exploring the causal effect of X on R requires considering the factors of Z.

To enable computational reasoning and inference, We formulate the causal graph by setting $X = x$ and $Z = z$ as

$$R_{x,z} = R(do(X = x), Z = z), \tag{1}$$

where $z = Z_x = Z(X = x)$, and the do operator denotes the intervention operations in causal reasoning. In addition, the do operator can be omitted in the absence of any confounder in X, and the $R_{x,z}$ can be represented as:

$$R_{x,z} = R_{x,Z_x} = R(X = x, Z = z). \tag{2}$$

Regarding counterfactual reasoning, X is set to different values to determine the impact on R. Specifically, we use x and x^* to denote the condition with and without input X, which leads to R_{x,Z_x} and $R_{x^*,Z_{x^*}}$ respectively.

Causal Effects Causal effects quantify the impact on R under different treatments (e.g. $X = x$ and $X = x^*$), and we use the total effect (TE) as the measurement. Concretely, TE can be formulated as

$$TE = R_{x,Z_x} - R_{x^*,Z_{x^*}}. \tag{3}$$

To assess the causal effects in a more fine-grained manner, TE can be further divided into the total indirect effect (TIE) and natural direct effect (NDE). Compared to TE, TIE reflects the effect of Z on R, while NDE reflects the effect of X on R given Z_{x^*}, and they are defined respectively as

$$TIE = R_{x,Z_x} - R_{x,Z_{x^*}}, \tag{4}$$

and

$$NDE = R_{x,Z_{x^*}} - R_{x^*,Z_{x^*}}. \tag{5}$$

Considering a dataset with bias, the TIE can reduce the bias related to X as the reduced portion $R_{x,Z_{x^*}}$ can reflect the causal effect of x on R without Z. Therefore, it is widely used in classification tasks to reduce bias as these tasks only need to complete a single inference without considering the BPE in sequential inference tasks.

3.2. Basics of Information Entropy

Given a discrete random variable K with n possible values, the information entropy $H(K)$ is defined as the average level of the uncertainty inherent to the possible outcomes of K, which is formulated as

$$H(K) = -\sum_{i=1}^{n} p(k_i) * \text{log}p(k_i), \tag{6}$$

where $p_(k_i)$ denotes the probability of k_i. For the sake of convenience, we simplify the formulation as:

$$H(K) = -\sum_{K} P_K * \text{log}P_K, \tag{7}$$

Given a set of variables K and T, the conditional entropy $H(T|K)$ is defined to quantify the uncertainty of T in the condition of K, and formulated as

$$H(T|K) = -\sum_{KT} P_{KT} * \text{log}P_{T|K}. \tag{8}$$

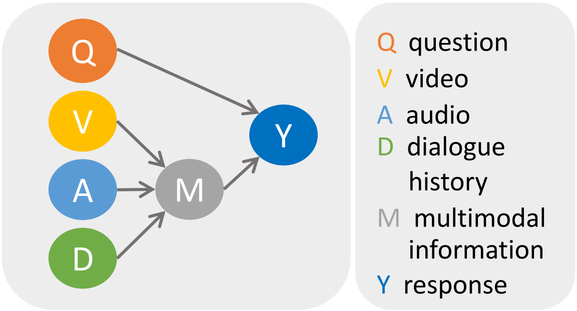


Figure 4: The causal graph of the video-grounded dialogue generation task.

To measure the amount of information in $T|C$ by observing K, we introduce the conditional mutual information $I(K;T|C)$ as

$$I(K;T|C) = \sum_{KTC} P_{KTC} * \text{log}(\frac{P_{K,T|C}}{P_{K|C}P_{T|C}}). \tag{9}$$

4. Method

In this section, we first introduce the video dialogue generation task, then reformulate the counterfactual reasoning problem from the information entropy perspective, and finally describe the implementation details of utilizing counterfactual entropy as an external training loss.

4.1. Task Formulation

The video dialogue generation task aims at generating an appropriate and fluent response $Y = \{y_1, y_2, ..., y_t\}$ containing t words, given the question Q, video V, audio A, and the dialogue history D as system inputs, among which the dialogue history comprises multiple rounds of questions and answers. The task can be formulated as

$$P(Y|V, A, D, Q) = \prod_{i=1}^{t} P(y_i|V, A, D, Q, Y^{<i}). \tag{10}$$

4.2. Counterfactual Entropy

Causal Graph Definition To apply counterfactual reasoning, we first translate the VDG task into a causal graph. As shown in Figure 4, the generation of final response Y is determined by the question Q and multimodal information M, which consists of the video V, audio A, and dialogue history D. M is also denoted as M_{vad} when setting V, A, D as v, a, d respectively.

To assess the impact of the multimodal information (e.g. video input V) on the final result, we set $V = v^*$ to represent the condition without the video as input, and the other modalities follow a similar manner. As described in Section 3.1, we use the TIE to reduce the question-related bias by eliminating question-related causal effect and enhancing multi-modal information, which is formulated as

$$TIE = Y_{q,M_{vad}} - Y_{q,M_{v^*a^*d^*}}, \tag{11}$$

For conciseness, we denote $Y_{q,M_{vad}}$ and $Y_{q,M_{v^*a^*d^*}}$ as $Y_{q,M}$ and Y_{q,M^*} respectively in the following parts of this paper.

Formulation Based on the definition of causal effects, the TIE we have formulated in Equation (11) reflects the importance of multimodal information provided by video, audio, and dialogue history inputs. However, the conventional strategy to utilize the TIE cannot be applied to generative tasks directly as the result of the BPE in sequential inference. Therefore, we provide a detailed derivation of counterfactual reasoning from the perspective of information entropy and utilize Equation (11) as an example to demonstrate the application of our proposed approach to VDG tasks.

As counterfactual reasoning and information entropy share similar semantics, we utilize the information entropy as the measurement of the causal effect. Specifically, the subparts of Equation (11) can be formulated from the information entropy perspective with the same semantics, and we use $\rightarrow$ to denote the transformation from the statistical perspective to the information entropy field. The transformations are formulated as:

$$Y_{q,M^*} \rightarrow H(Y|Q), \tag{12}$$

and

$$Y_{q,M} \rightarrow H(Y|Q, M). \tag{13}$$

As shown in Equations (12) and (13), Y is generated given only Q, or both Q and M respectively. To assess the effect of the above two conditions on the result Y, we calculate the conditional information entropy $H(Y|Q)$ and $H(Y|Q, M)$ respectively, which have similar semantics with the original form of the counterfactual concept. Therefore, Equation (11) can be written as:

$$\begin{aligned} TIE &= Y_{q,M} - Y_{q,M^*} \\ &\rightarrow H(Y|Q, M) - H(Y|Q). \end{aligned} \tag{14}$$

Moreover, to analyze the feasibility of utilizing this transformation approach, we further explore the exact meaning of $H(Y|Q, M) - H(Y|Q)$. For Equation (12), it can be rewritten as follows:

$$\begin{aligned} H(Y|Q) &= \sum_{YQ} P_{YQ} * \text{log} P_{Y|Q} \\ &= -\sum_{YQ}(-\sum_{M} P_{YQM}) * \text{log} P_{Y|Q} \\ &= -\sum_{YQM} P_{YQM} * \text{log} P_{Y|Q} \\ &= -\sum_{YQM} P_{YQM} * \text{log} \frac{P_{YQ}}{P_Q}. \end{aligned} \tag{15}$$

For Equation (13), it can be calculated as follows:

$$\begin{aligned} H(Y|Q,M) &= -\sum_{YQM} P_{YQM} * \log P_{Y|QM} \\ &= -\sum_{YQM} P_{YQM} * \log(\frac{P_{YQM}}{P_{QM}}). \end{aligned} \tag{16}$$

Thus, Equation (11) can be calculated as:

$$\begin{aligned} & H(Y|Q,M) - H(Y|Q) \\ &= -\sum_{YQM} P_{YQM} * \log(\frac{P_{YQM}}{P_{QM}}) + \sum_{YQM} P_{YQM} * \log \frac{P_{YQ}}{P_Q} \\ &= -\sum_{YQM} P_{YQM} * \log(\frac{P_{YQM}P_Q}{P_{QM}P_{YQ}}) \\ &= -\sum_{YQM} P_{YQM} * \log(\frac{P_{YQM}}{P_Q} * \frac{P_Q}{P_{QM}} * \frac{P_Q}{P_{YQ}}) \\ &= -\sum_{YQM} P_{YQM} * \log(\frac{P_{M,Y|Q}}{P_{M|Q}P_{Y|Q}}) \\ &= -I(M;Y|Q). \end{aligned} \tag{17}$$

Finally, we transfer the Equation (14) as:

$$TIE \to -I(M;Y|Q), \tag{18}$$

which reflects that $I(M;Y|Q)$ and TIE both represent the influence of M on $Y|Q$. As a result, both equation derivation and semantic similarity can demonstrate that $I(M;Y|Q)$ can replace the TIE to measure the causal effect.

After the reformulation, we establish the connection between the causal effect TIE and the conditional mutual information $I(M;Y|Q)$, which can prove the effectiveness of the TIE in the continuous domain. Furthermore, we can introduce TIE as the external training loss combined with the information theory. By doing so, we mitigate the biases related to the question by eliminating the question-related causal effect and enhancing the significance of the multimodal information. Compared to conventional applications on classification tasks, such a method utilizes the TIE via information entropy and therefore avoids considering the BPE.

Definition Therefore, we provide two methods for counterfactual reasoning via information theory. For counterfactual reasoning can be utilized in generative tasks, we use information theory to measure the changes in counterfactual reasoning. As the common setting in generative tasks, we use the cross-entropy as the measurement, which is formulated as:

$$\underbrace{-\sum_{t=1}^{T} \log P\left(y_t \mid M, Q, Y_{<t}\right) + \sum_{t=1}^{T} \log P\left(y_t \mid Q, Y_{<t}\right)}_{\mathrm{TIE}_{\mathrm{CR}}}, \tag{19}$$

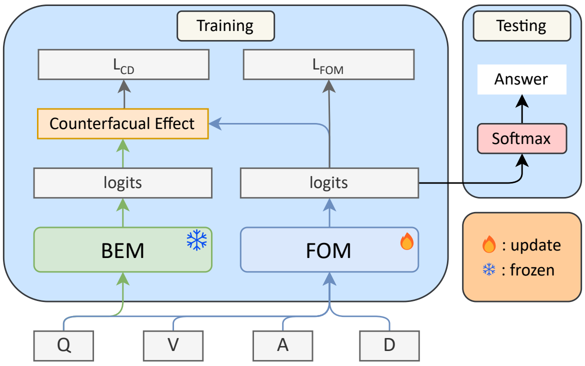


Figure 5: The overview of the CE-VDG, which consists of the BEM and FOM.

Nonetheless, this formulation can reflect counterfactual changes within the ground truth, it cannot reflect the effect on the whole vocabulary. Therefore, we further adopt the conditional entropy without ground truth as the measurement, which is formulated as:

$$\underbrace{-\sum_{QMY} P_{QMY} * \log P_{Y|QM} + \sum_{QY} P_{QY} * \log P_{Y|Q}}_{\mathrm{TIE}_{\mathrm{CE}}}, \tag{20}$$

In this way, we can utilize the counterfactual reasoning in the generative tasks through these two formulations, as both $\mathrm{TIE}_{\mathrm{CR}}$ and $\mathrm{TIE}_{\mathrm{CE}}$ can reflect the counterfactual effect which we can use as the measurement.

4.3. Implementation

4.3.1. Overview

As described in Section 4.2, we utilize the counterfactual entropy $I(M;Y|Q)$ to establish the counterfactual effect $Y_{q,M} - Y_{q,M^*}$ on generative tasks. Specifically, we mitigate the bias of Q and force the system more attentive to M by minimizing the counterfactual entropy, which is named counterfactual entropy-based video-grounded dialogue generation (CE-VDG). The overview of CE-VDG is shown in Figure 5. We prepare two models to eliminate the bias: factual output model (FOM) and bias estimation model (BEM) responsible for $Y_{q,M}$ and Y_{q,M^*} respectively. To enhance the generation ability of the FOM, we adopt the counterfactual entropy to eliminate the question-related bias provided by the BEM.

For FOM and BEM, we select BART as our backbone model, named VideoBART. Considering the encoder and decoder are more powerful in semantic extraction and complex reasoning respectively, we utilize the BART encoder for feature extraction of the video and dialogue history, and employ the BART decoder for response generation according to the question. In detail, visual and auditory features are extracted from the video respectively, and concatenated with dialogue history embedding into a joint representation. Finally, the response is generated by the BART decoder using the encoder state and the input question.

4.3.2. Feature Extraction

The video information is composed of visual information and auditory information. To further utilize the semantics from the video, we extract the visual features and auditory features from the video respectively by pre-trained feature extractors.

Visual Feature The video is segmented into n consecutive video frames $F = \{f_1, f_2, \ldots, f_n\}$ in 1 frame per second, and then the visual feature F_V is obtained by ActionCLIP (Wang et al., 2021) image encoder as

$$F_V = \mathrm{Concat}(\mathrm{ActionCLIP}(f_i))\mid_{i=0}^{n}. \tag{21}$$

Auditory Feature The audio is divided into n continuous waves $W = \{w_1, w_2, \ldots, w_n\}$ where the duration of the w is 1 second, and then the auditory feature F_A is obtained by using Wav2CLIP (Wu et al., 2022) as

$$F_A = \mathrm{Concat}(\mathrm{Wav2CLIP}(w_i))\mid_{i=0}^{n}. \tag{22}$$

The visual feature F_V and auditory feature F_A are concatenated into the video feature as the input to the model.

4.3.3. Causal Effect Establishment

To establish different causal effects and estimate the question-related bias, we employ a final output model (FOM) to generate $Y_{q,M}$ and bias estimation model (BEM) to generate Y_{q,M^*}.

Factual Output Model FOM is responsible for accurate response generation. We train the FOM with information from all the modalities, including the video V, dialogue history D, and question Q. Additionally, FOM uses the cross-entropy loss as the training loss:

$$\mathcal{L}_{\mathrm{FOM}} = -\sum_{t=1}^{T} \log P\left(y_t \mid V, A, D, Q, Y_{<t}\right). \tag{23}$$

Bias Estimation Model BEM aims to estimate the question-related bias in the dataset. To achieve this, we restrict the input to only the question Q. Specifically, the input of the BART encoder is replaced with Q, and the BART decoder is trained with the same settings as FOM. The training loss of BEM is the cross-entropy loss, which is formulated as

$$\mathcal{L}_{\mathrm{BEM}} = -\sum_{t=1}^{T} \log P\left(y_t \mid Q, Y_{<t}\right), \tag{24}$$

4.3.4. Counterfactual Debias Process

We utilize the FOM and BEM and compare the output of them to represent TIE. Specifically, we maximize TIE by minimizing the counterfactual entropy for better question-related bias reduction, in which the counterfactual entropy loss is:

$$\mathcal{L}_{\mathrm{CE}} \in \{TIE_{CR}, TIE_{CE}\}. \tag{25}$$

To further enhance the accurate generation ability of the FOM, we combine the cross entropy loss and the counterfactual loss and obtain the final objective function as:

$$\mathcal{L}_{\mathrm{OBJ}} = \mathcal{L}_{\mathrm{FOM}} + \alpha \mathcal{L}_{\mathrm{CE}}, \tag{26}$$

where α is a hyper-parameter considering the trade-off between the $\mathcal{L}_{\mathrm{FOM}}$ and $\mathcal{L}_{\mathrm{CE}}$.

5. Experiments

5.1. Datasets and Metrics

Experiments were conducted on two popular VDG benchmarks, which include AVSD-DSTC datasets and NExT-OE datasets.

AVSD-DSTC AVSD-DSTC datasets (AlAmri et al., 2019) were expanded from the Charades dataset (Sigurdsson et al., 2016) with question-answer pairs, which consist of the AVSD-DSTC7, AVSD-DSTC8, and AVSD-DSTC10 dataset. The question-answer pairs are closely correlated with the relationship between the person and objects, and the length of the answers is between 5 to 9 words. Specifically, each new version of the AVSD-DSTC dataset is expanded by additional follow-up questions, which require more detailed video understanding abilities. Therefore, conducting experiments on different versions of AVSD-DSTC datasets enables more comprehensive evaluation of the capability, including comprehensive understanding and detailed recognition. In this paper, we use BLEU, METEOR, ROUGE-L, and CIDEr as the evaluation metrics.

NExT-OE NExT-OE dataset (Xiao et al., 2021) is constructed based on YFCC-100M (Thomee et al., 2015), in which the types of questions are divided into casualty, temporary reasoning, and descriptive ability. Different from the AVSD-DSTC dataset, the answers are mostly shorter than 4 words and the Wu-Palmer Similarity (WUPS) score is used to evaluate the semantic similarity between the generated answer and the ground truth.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGH-L	CIDEr
AVSD-DSTC7 official test set							
Naive Fusion (2019a)	0.644	0.500	0.395	0.318	0.220	0.500	0.779
MTN* (2019)	0.692	0.556	0.459	0.368	0.259	0.537	0.964
JST (2019b)	0.686	0.557	0.458	0.382	0.254	0.537	1.005
BiST (2020a)	0.715	0.560	0.477	0.390	0.259	0.552	1.030
CE-VDG (Ours)	**0.741**	**0.615**	**0.513**	**0.431**	**0.280**	**0.583**	**1.216**
AVSD-DSTC8 official test set							
DMN (2020)	-	-	-	0.270	0.208	0.482	0.714
VideoGPT* (2021)	0.677	0.556	0.462	0.387	0.249	0.544	1.022
SCGA (2021)	0.675	0.559	0.459	0.377	0.269	0.555	1.024
MED-CAT (2022)	-	-	-	0.376	0.247	0.547	0.982
CE-VDG (Ours)	**0.711**	**0.585**	**0.484**	**0.404**	**0.271**	**0.567**	**1.110**
AVSD-DSTC10 official test set							
AV-trans (2022)	-	-	-	0.247	0.191	0.437	0.566
NLM (2022)	0.641	0.489	0.379	0.298	0.225	0.502	0.804
MAG (2022)	0.646	0.489	0.380	0.299	0.225	0.499	0.787
TSF(ensemble)* (2022)	0.680	0.558	0.461	0.385	0.247	0.539	0.957
DialogMCF (2023a)	0.693	0.556	0.450	0.369	0.249	0.536	0.912
CE-VDG (Ours)	**0.721**	**0.588**	**0.481**	**0.397**	**0.267**	**0.559**	**1.008**

Table 1: Evaluation results of our model compared with baseline approaches on AVSD-DSTC7, AVSD-DSTC8 and AVSD-DSTC10 official test sets. The * denotes the best performance in the corresponding AVSD-DSTC challenge.

Methods	WUPS
HCRN (2020b)	23.92
HME (2019)	24.06
UATT (2017)	24.25
HGA (2020)	25.18
ClipBERT (2021)	24.17
KcGA (2023)	28.20
CE-VDG(Ours)	**28.71**

Table 2: WUPS scores of our model compared with baseline approaches on NExT-OE dataset.

5.2. Experimental Settings

In our experiments, we initialize our model by BART-base [1] and the `$\alpha$` is 0.01. We choose the TIE_{CE} as the final counterfactual loss. We adopt an AdamW optimizer with a learning rate of 6.25e-5 for causal effect establishment and the counterfactual debias process and the batch size is 32. During the inference phase, we use the factual output model for response generation and the beam search as the generation algorithm. The beam size is 6 and the penalty factor is 0.6.

5.3. Baseline Methods

We use the following methods as our baseline system. (i) **VideoGPT** which chooses GPT-2 as the backbone and involves integrating visual-audio features and dialogue text as a combined input.

Methods	BLEU4	CIDEr
VideoBART	0.384	0.987
VideoBART+$\mathrm{TIE}_{\mathrm{CR}}$	0.389	0.998
VideoBART+$\mathrm{TIE}_{\mathrm{CE}}$	0.397	1.008

Table 3: Evaluation results of definitions ablation experiments on AVSD-DSTC10 official test set.

(ii) **TSF** which extracts the visual feature through TimeSformer and improves performance through the ensemble. (iii) **AV-trans** which proposes bi-modal attention to fuse audio-visual features via the encoder-decoder structure without pre-training. (iv) **NLM** which recognizes the visual and auditory actions as video features, and feeds them to the GPT-2 to obtain the answer with the text input.

5.4. Main Results

The main results are presented in Table 1 and Table 2. Compared to other baselines, our proposed CE-VDG demonstrates consistent superiority across all the metrics on both the AVSD-DSTC and NExT-OE datasets. This affirms the effectiveness of counterfactual debiasing in improving the video dialogue generation performance on both long sentences and short texts. As shown in Table 1, CE-VDG substantially outperforms other methods across all the metrics, especially on BLUE4 and CIDEr, which reflects the capabilities for accurate and fluent response. Furthermore, the results in Table 2 illustrate that our method can produce closer results semantically to the ground truth, and

[1] https://huggingface.co/facebook/bart-base

Methods	BLEU4	CIDEr
VideoGPT*	0.380	0.937
VideoGPT+CE	**0.392**	**0.988**
TSF(ensemble)	0.385	0.957
TSF(unensemble)+CE	**0.389**	**0.965**
AV-trans	0.247	0.566
AV-trans+CE	**0.287**	**0.660**
NLM	0.298	0.804
NLM+CE	**0.383**	**0.972**
VideoBART	0.384	0.987
VideoBART+CE	**0.397**	**1.008**

Table 4: Evaluation results of baseline ablation experiments on AVSD-DSTC10 official test set. The * denotes the results we reproduced and the CE is an abbreviation of the counterfactual entropy.

achieve debiased inference from biased training by reducing the question-related causal effect.

5.5. Ablation Study

5.5.1. Ablation Study on Definitions

To validate the effectiveness of $\mathrm{TIE}_{\mathrm{CE}}$ than $\mathrm{TIE}_{\mathrm{CR}}$, we apply the $\mathrm{TIE}_{\mathrm{CR}}$ as the $\mathcal{L}_{CE}$ in ASVD-DSTC dataset, which is shown in Table 3. As shown in this table, both the $\mathrm{TIE}_{\mathrm{CE}}$ and $\mathrm{TIE}_{\mathrm{CR}}$ can improve the effectiveness of the system. Specifically, the $\mathrm{TIE}_{\mathrm{CE}}$ can reflect the counterfactual effect among the whole vocabulary, which can enhance the impact of counterfactual effects for better bias mitigation.

5.5.2. Ablation Study on Baselines

To validate the robustness of our proposed counterfactual entropy, we apply it to various baseline models of the AVSD-DSTC challenge and evaluate the performance on the AVSD-DSTC10 official test set, which is shown in Table 4. Specifically, we conduct experiments on five baseline models that differ in feature extractors, architectures, and modality fusion strategies. Considering the results are absent on the AVSD-DSTC10 official test set, we reproduced the result on the VideoGPT by official code. As shown in Table 4, the inclusion of additional counterfactual entropy loss enhances the performance of all models, particularly the baselines that rely solely on the decoder architecture. Furthermore, our model demonstrates comparable performance compared to the ensemble model, even without utilizing ensemble learning, thanks to the proposed counterfactual debiasing process.

5.5.3. Ablation Study on Modality

To validate the effectiveness of our method on bias reduction for different modalities, we conduct a series of experiments on a variety of modalities on BEM. The results are shown in Table 5. Specifically, It is important to note that all the different settings on BEM lead to performance improvement. Among the settings, VideoBART+Q shows significant performance gains, indicating that the counterfactual debias process effectively improves response accuracy by reducing question-related bias and utilizing multi-modal information more. Additionally, BEM with QV outperformed BEM with QD, suggesting that the QVA-related bias is more severe and leveraging dialogue history can effectively mitigate this bias.

BEM	BLEU4	CIDEr
VideoBART	0.384	0.987
VideoBART+Q	0.397	1.008
VideoBART+QD	0.395	1.008
VideoBART+QVA	0.400	1.015
VideoBART+VAD	0.397	1.004

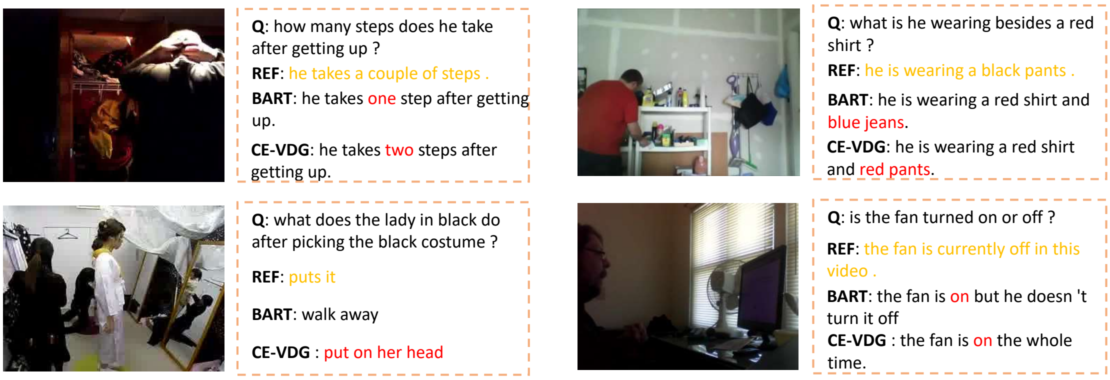
Table 5: Evaluation results of BEM ablation experiments on AVSD-DSTC10 dataset. The VideoBART + X denotes the BEM is established with X and is proposed to reduce X-related bias.

5.5.4. Ablation Study on Questions Types

To explore the effect of different types of questions, we evaluate the performance of our method with various question types. The results are presented in Table 6. These results demonstrate that the counterfactual debias process effectively reduces biases introduced by the training set for most question types. Specifically, the performance of descriptive questions shows a significant improvement, indicating enhanced utilization of video information through the counterfactual debias process. However, the performance gains for temporal and causal questions are not as satisfactory, likely because the BEM does not specifically consider temporal and causal information.

Question Types	VideoBART	CE-VDG
Casualty Why	19.25	**19.77**
Casualty How	25.06	**25.32**
Temporal Next	**16.41**	15.74
Temporal Current	26.51	**27.93**
Descriptive Choice	69.91	**71.94**
Descriptive Counting	42.58	**44.43**
Descriptive Location	46.77	**48.41**
Descriptive Open-Form	52.80	**53.85**

Table 6: The WUPS score on different types of questions ablation study in NExT-OE dataset.



Q: how many steps does he take after getting up ?
REF: he takes a couple of steps .
BART: he takes one step after getting up.
CE-VDG: he takes two steps after getting up.

Q: what does the lady in black do after picking the black costume ?
REF: puts it
BART: walk away
CE-VDG : put on her head

Q: what is he wearing besides a red shirt ?
REF: he is wearing a black pants .
BART: he is wearing a red shirt and blue jeans.
CE-VDG: he is wearing a red shirt and red pants.

Q: is the fan turned on or off ?
REF: the fan is currently off in this video .
BART: the fan is on but he doesn 't turn it off
CE-VDG : the fan is on the whole time.

(a) Successful Case

(b) Failed case

Figure 6: Four examples are taken from the AVSD-DSTC10 and NExT-OE datasets. The reference response is highlighted in yellow, and noteworthy parts are highlighted in red. It is important to note that the questions about counting and coarse-grained actions can be improved a lot but the questions about color and object state are not satisfactory for us.

5.6. Case Study

We also evaluate the ability of the proposed method to generate dialogues using two datasets. As shown in Four examples in Figure 6, the CE-VDG model, which incorporates counterfactual entropy, is compared to the original model without it. The left results demonstrate that CE-VDG can mitigate the bias present in the training dataset related to the question and the right results illustrate that there is still a huge room to be improved. Specifically, the first example in the left part shows that CE-VDG can address the inherent bias in counting numbers and generate responses that are more relevant and appropriate based on the video content. However, the second example in the right part shows that CD-VDG cannot detect the state of the object accurately, compared to the coarse-grained actions in the right part. So, the effect is highly correlated to the question types, the more serious in bias of this type of question the better the performance of the counterfactual entropy.

6. Conclusion

To address the bias in the video-grounded dialogue generation task, we propose a reformulation of the counterfactual reasoning process using information entropy. This allows us to extend the application of counterfactual reasoning to generative tasks. Specifically, we introduce the concept of counterfactual entropy to mitigate question-related bias by eliminating causal effects and enhancing multi-modal information. Additionally, we present CE-VDG, a method that applies counterfactual reasoning to the VDG task by utilizing counterfactual entropy as an external loss. Through extensive experiments on two VDG benchmarks, we demonstrate the effectiveness of our proposed approach compared to various state-of-the-art methods.

Limitations

We propose CE-VDG to reduce the bias introduced by the unbalanced data and achieve superiority in two video-grounded dialogue generation datasets. However, there are still many aspects that require improvement on our end. We provide two definitions of counterfactual entropy and the main difference is the subjects of the distribution, which are the ground truth and the whole vocabulary. However, we can explore the rate between the two objects to obtain the two distributions for bias reduction. Furthermore, the improvement is highly related to the degree of the bias and we fail to enhance the performance for little biased questions, such as color questions, because the original capability lacks this. In the future, we will improve the original capability of the model by focusing more on model architecture.

Acknowledges

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), STCSM (No. 21511101100, No. 22DZ2229005), and State Key Laboratory of UHD Video and Audio Production and Presentation.

Bibliographical References

Huda AlAmri, Vincent Cartillier, Abhishek Das, and et al. 2019. Audio visual scene-aware dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Zhe Chen, Hongcheng Liu, and Yu Wang. 2023a. Dialogmcf: Multimodal context flow for audio visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–13.

Ziwei Chen, Linmei Hu, Weixin Li, and Set al. 2023b. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, and et al. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Shijie Geng, Peng Gao, Moitreya Chatterjee, and et al. 2021. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1415–1423.

Chiori Hori, Huda AlAmri, Jue Wang, and et al. 2019a. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2352–2356.

Chiori Hori, Anoop Cherian, Tim K. Marks, and Takaaki Hori. 2019b. Joint student-teacher learning for audio-visual scene-aware dialog. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 1886–1890.

Xin Huang, Hui Li Tan, Mei Chee Leong, and et al. 2022. Investigation on transformer-based multimodal fusion for audio-visual scene-aware dialog. In *Proceedings of DSTC10 Workshop at AAAI-2022*.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Yao Jin, Guocheng Niu, Xinyan Xiao, and et al. 2023. Knowledge-constrained answer generation for open-ended video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8141–8149.

Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang D Yoo. 2021. Structured co-reference graph attention for video-grounded dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1789–1797.

Hung Le and Nancy F Chen. 2020. Multimodal transformer with pointer network for the DSTC8 AVSD challenge. *ArXiv preprint*, abs/2002.10695.

Hung Le, Nancy F Chen, and Steven CH Hoi. 2021. C3: Compositional counterfactual constrastive learning for video-grounded dialogues. *arXiv preprint arXiv:2106.08914*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2020a. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1846–1859, Online. Association for Computational Linguistics.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020b. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981.

Jie Lei, Linjie Li, Luowei Zhou, and et al. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Zekang Li, Zongjia Li, Jinchao Zhang, and et al. 2021. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483.

Aishan Liu, Huiyuan Xie, Xianglong Liu, and et al. 2022. Revisiting audio visual scene-aware dialog. *Neurocomput.*, 496(C):227–237.

Yulei Niu, Kaihua Tang, Hanwang Zhang, and et al. 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.

Ankit Shah, Shijie Geng, Peng Gao, and et al. 2022. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7732–7736.

Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*.

Teng Sun, Wenjie Wang, Liqaing Jing, and et al. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 15–23, New York, NY, USA. Association for Computing Machinery.

Kaihua Tang, Yulei Niu, Jianqiang Huang, and et al. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and Li-Jia Li. 2015. Yfcc100m. *Communications of the ACM*, 59:64 – 73.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. ActionCLIP: A new paradigm for video action recognition. *ArXiv preprint*, abs/2109.08472.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning robust audio representations from CLIP. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4563–4567. IEEE.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Huiyuan Xie and Ignacio Iacobacci. 2020. Audio visual scene-aware dialog system using dynamic memory networks. *DSTC8 at AAAI2020 workshop*.

Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing*, 26(12):5656–5666.

Yoshihiro Yamazaki, Shota Orihashi, Ryo Masumura, and et al. 2022. Audio visual scene-aware dialog generation with transformer-based video representations. *ArXiv preprint*, abs/2202.09979.

Heo Yoonseok, Kim Gyunyeop, Yoo Eunseok, Lee Seungsoo, Jeong Eunseo, and Kang Sangwoo. 2022. Interpretable multimodal dialogue system with natural language-based multimodal integration. In *Proceedings of DSTC10 Workshop at AAAI-2022*.