

Received 7 October 2023, accepted 1 December 2023, date of publication 7 December 2023,
date of current version 15 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340705

RESEARCH ARTICLE

Dynamic Debiasing Network for Visual Commonsense Generation

JUNGEUN KIM¹, (Graduate Student Member, IEEE),
JINWOO PARK¹, (Student Member, IEEE), JAEKWANG SEOK,
AND JUNYEONG KIM¹, (Member, IEEE)

Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT), Artificial Intelligence Graduate School Program, Chung-Ang University under Grant 2021-0-01341; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT), Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics under Grant 2022-0-00184.

ABSTRACT The task of Visual Commonsense Generation (VCG) delves into the deeper narrative behind a static image, aiming to comprehend not just its immediate content but also the surrounding context. The VCG model generates three types of captions for each image: 1) the events preceding the image, 2) the characters' current intents, and 3) the anticipated subsequent events. However, a significant challenge in VCG research is the prevalent yet under-addressed issue of dataset bias, which can result in spurious correlations during model training. This occurs when a model, influenced by biased data, infers associations that frequently appear in the dataset but may not provide accurate or contextually appropriate interpretations. The issue becomes even more complex in multimodal tasks, where different types of data, such as text and image, bring their unique biases. When these modalities are combined as inputs to a model, one modality might exhibit a stronger bias than others. To address this, we introduce the **Dynamic Debiasing Network (DDNet)** for Visual Commonsense Generation. DDNet is designed to identify the biased modality and dynamically counteract modality-specific biases using causal relationship. By considering biases from multiple modalities, DDNet avoids over-focusing on any single modality and effectively combines information from all modalities. The experimental results on the VisualCOMET dataset demonstrate that our proposed network fosters more accurate commonsense inferences. This emphasizes the critical need for debiasing in multimodal tasks and enhances the reliability of machine-generated commonsense narratives.

INDEX TERMS Multimodal reasoning, visual commonsense generation, VisualCOMET, dataset bias, debiasing, causal inference.

I. INTRODUCTION


Humans, in our cognitive processes, don't rely solely on presented data when reasoning; we naturally employ our commonsense to fill gaps, draw inferences, and make context-aware decisions. Emulating this quintessential aspect of human cognition is crucial for artificial intelligence models, enabling them to enhance their understanding and decision-making capabilities by incorporating commonsense

reasoning. By harnessing commonsense, models can transcend data-driven limitations and adapt more effectively to novel situations, addressing complex challenges across diverse domains.

To address these challenges, [1] introduced the task of Visual Commonsense Generation (VCG). This novel framework aims to understand event beyond provided visual and textual inputs, generating captions about events that might have occurred previously, potential future events, and the current intentions of the appeared people in the image. Unlike earlier works in visual commonsense inference, such

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang¹.

TABLE 1. An example from the VCG dataset: When provided with an image and two types of corresponding textual information, the model generates commonsense inferences about three aspects: before, after, and intent.

Inputs	Type	Outputs
 <p>Event : 2 holds an egg roll and looks disappointed as she reads a paper in front of her Place : in a bedroom</p>	before	order chinese takeout,
		go over work with 1,
		ask 1 to help her do her taxes
	intent	eat some dinner,
		show her feelings towards what she read,
		show she was upset she didn't find what she wanted
after	argue with 1 about her being disorganized,	
	drop the egg roll, point out what she is disappointed by to 1	

as Visual Commonsense Reasoning (VCR) [2], which mainly focused on comprehension from given data, VCG pushes models to go beyond mere presented inputs and requires inference grounded in commonsense. For example, as shown in Table.1, the model is presented with an image depicting a woman and a man working on the floor in front of a bed. Along with the image, the model receives two types of textual information about the woman: a description of an event and a place. The model is asked to generate three types of inferences: (1) past events (e.g., order Chinese takeout), (2) future events (e.g., argue with a man about her being disorganized), and (3) the woman’s intent (e.g., eat some dinner).

Nevertheless, current research on VCG tasks [1], [3], [4] has yet to tackle the crucial concern of dataset bias, which can lead to spurious correlations during training. Commonly, their approaches rely on Empirical Risk Minimization (ERM), which over-depends on the frequently occurring co-occurrences in the dataset. This over-reliance tends to create shortcuts from input to output, resulting in models that are biased towards frequently occurring patterns and potentially neglect the comprehensive information necessary for accurate results. Unaddressed, this bias can lead to inaccurate or unfair predictions. In particular, given that multi-modal models handle various data types like text and images, each of which may potentially carry different biases, it is crucial to take extra precautions to address bias in VCG, three modality task. When integrating multiple modalities as inputs, certain modality might exhibit a stronger bias than others. Identifying and understanding the modality-specific biases associated with each type of input data is an essential step in the process of bias mitigation in multimodal tasks.

Fig.1 illustrates examples of spurious correlations observed in the two modalities within the VCG dataset. In the top example, because of the frequent pairing of the object ‘room’ with certain output sentences in the training set, the model generate potentially inaccurate sentences, relying on the object ‘room’ without considering the whole description or the image, as seen. This shows the case where the input data is biased more toward the text modality than the image modality. On the other hand, the bottom example reveals a different kind of bias. When ‘multiple people’ appear in an

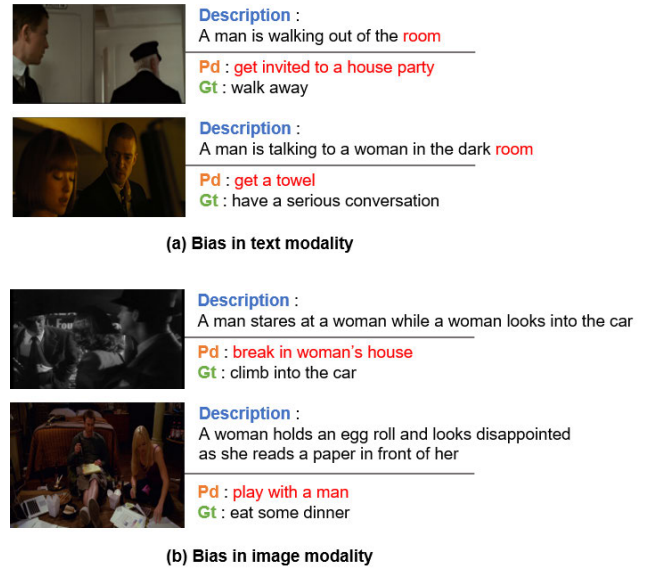


FIGURE 1. Examples about the spurious correlation in a VCG dataset. Two examples for each modality : (a) text and (b) images. Given an image and its corresponding description, the model generates a biased sentence (represented as ‘Pd’ in the figure) compared to the ground truth (represented as ‘Gt’ in the figure).

image, the model tends to generate biased prediction like ‘break in woman’s house’ or ‘play with a man’, even if the accompanying text modality suggests otherwise. In this particular case, a bias towards the image modality becomes evident. These examples demonstrate that due to the biases present in the dataset, the model tends to overly focus on one particular biased modality, consequently overlooking other crucial information that it should consider. Furthermore, it’s evident that each modality exhibits a distinct type of bias. In other words, the components that co-occur frequently vary between modalities. Some data show a bias towards the text modality, while others demonstrate a stronger inclination towards the image modality.

Fig. 2 provides quantitative evidence demonstrating the distinct modality-specific bias present in each data sample. As depicted in (a), when the object ‘truck’ appears in the input text, the predicted sentence by the baseline model (represented as ‘pred’ in Fig. 2) closely aligns with co-occurrence patterns observed in the training dataset (represented as ‘train’ in Fig. 2) rather than with the ground truth of the test set. For example, when the word ‘truck’ is provided in text modality, the sentence ‘get out’ appears frequently in the training dataset (29%). Due to this bias associating ‘truck’ with ‘get out’ in the training data, the phrase emerges with a similar frequency in the predicted sentences of the baseline model (35%), even though it is less common in the test ground truth (5%). This can be attributed to the model’s strong reliance on co-occurrence patterns observed during training. Conversely, in the example shown in (b), when the object ‘truck’ is represented visually and not in text, the influence of dataset bias is diminished. As a result, the baseline model produces predictions that are

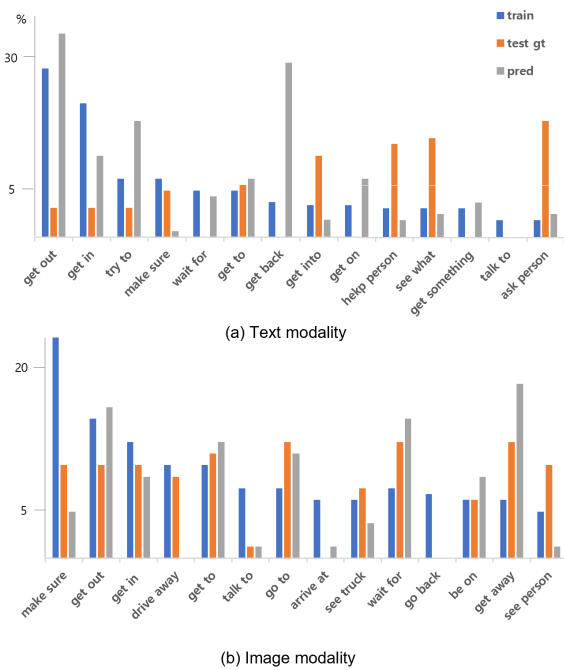


FIGURE 2. Comparison of the co-occurrence distributions between the word ‘truck’ and its corresponding output sentence (first bigram) across three different sources: (1) the training dataset, (2) the ground truth in test dataset, and (3) the predicted sentences from the baseline model of the VisualCOMET dataset. (a) When ‘truck’ is represented in text modality, the distribution of the baseline model’s predictions tends to align more with the training dataset’s distribution than with the ground truth of the test dataset. (b) Conversely, when the same object is represented in image modality, the distribution of predictions aligns more closely with the ground truth of the test dataset. To maintain clarity, only the top 14 pairs are visualized.

more aligned with the test ground truth. This suggests that the objects causing bias, known as confounder, differ for each modality.

Drawing from our observation, we propose **Dynamic Debiasing Network (DDNet)** for Visual Commonsense Generation, addressing the above crucial issue of modality-specific biases that may affect the accuracy and fairness of the model. The DDNet is designed to discern the biased modality of input data and subsequently remove bias for the corresponding modality, considering modality-specific bias. The operations of DDNet is divided into two stages: (1) Modality-specific Bias Detection, (2) Dynamic Bias Mitigation. Specifically, to effectively identify and address bias in input data, we initially focus on determining which modality in the input data exhibits bias. To do so, we measure bias as a direct causal effect on the output of each modality, motivated by [5]. To assess a direct causal effect in each modality, we first train a biased model for each individual modality. Rather than feeding all modalities simultaneously, we input a single modality at a time. This ensures that the model emphasizes and magnifies the inherent biases of that specific modality. The output of this specialized model, which reveals its intrinsic bias, is considered a bias score. Once we have identified the biased modality through its bias score, the next step is to mitigate this bias. This is achieved by disrupting

the ‘shortcut path’ often created by frequent co-occurrence patterns in the dataset. This disruption is specifically tailored to the identified biased modality, ensuring a more targeted and effective debiasing process. We perform this process using causal inference [6], [7]. The detailed process for this will be explained in the following sections.

DDNet’s dynamic approach adeptly identifies and counteracts biases, by carefully considering the characteristics of each input data and modality, resulting in commonsense inferences that are both precise and minimally affected by modality-specific biases. Our contributions are summarized as follows:

- (1) To the best of our knowledge, we are the first to delve into the exploration of biases in the dataset for VCG. We offer both qualitative and quantitative insights, establishing the existence of biases within the VCG dataset.
- (2) Considering the fluctuating nature of biased modalities across distinct input data, we introduce a novel debiasing framework, DDNet, tailored for the VCG.
- (3) DDNet, our proposed model-agnostic solution, discerns and dynamically addresses biases across different modalities. This approach marks a notable stride towards generating more precise and impartial commonsense inferences.

II. RELATED WORK
A. CAUSAL INFERENCE

Causal inference [6], [7] has emerged as a vital aspect of deep learning, aiming to discern and model the intricate cause-and-effect relationships inherent within neural networks. Instead of relying solely on traditional statistical approaches, where relationships are often inferred from correlations stemming from unbalanced co-occurrences in datasets, causal inference delves deeper, seeking to uncover genuine causal effect.

As causal inference techniques in deep learning have evolved, recent research in multimodal tasks has leveraged them to effectively debias models, eliminating spurious correlations stemming from inherent biases. In Visual Question Answering (VQA), a task that merges visual and textual information, [5] proposed a counterfactual inference framework, focusing on the fact that the VQA model easily relies on language bias. This framework, influenced by causal effects, identifies language bias as a direct result of the questions influencing the answers and subtract this direct language effect to reduce bias. Recently, [8] delves into a causal representation of VQA data, introducing a framework that capture causally related real association for both visual and textual data to boost model generalization. Recent researches in Video Moment Retrieval (VMR) [9], [10] has also utilized backdoor adjustment [6], [7], one of techniques in causal inference, to remove spurious correlation.

While a few of substantial and valuable research for Visual Commonsense Generation (VCG) has been conducted, the crucial challenge of dataset bias remains largely unaddressed, not truly looking into the multimodal inputs. In this regard, we leverage causal inference to identify sample-specific

modality bias and subsequently remove misleading correlation for each modality.

B. VISUAL COMMONSENSE GENERATION

The Visual Commonsense Generation (VCG) [1] task aims to reason about the intricate stories behind a still image, understanding not just the immediate content but also the context spanning before, after, and beyond the given input. This requires a shift from basic image recognition to a deeper, cognitive-level understanding, leveraging visual commonsense reasoning informed by extensive knowledge of the visual and social worlds. Reference [1] first provide strong baseline, which extend the GPT-2 model [11] to incorporate both visual and textual information. Reference [3] introduced the Knowledge Enhanced Multimodal BART (KM-BART), which adapt BART model [12] for pretraining on vast external datasets to draw knowledge from them. KM-BART was subsequently fine-tuned on the VisualCOMET benchmark. However, these previous studies had two limitations in that they neglect intricate intra and inter-modality relationships and treated each caption (i.e. captions for “before”, “intent” and “after”) as independent. In response, [4] suggested Cause-and-Effect BART (CE-BART) comprised of Structured Graph Reasoner to interpret relationships within and between modalities and Cause-and-Effect Generator to consider the causal relationships among three types of generated captions.

Nevertheless, while these studies provide valuable insights, the models still predominantly rely on statistical patterns in dataset. Instead of understanding the data in a deeper, more semantic or cognitive manner, these frameworks are likely looking for patterns that frequently appear in the dataset and basing their decisions on these patterns. To address the above issue, we propose the novel method which discerns and dynamically addresses biases across different modalities.

III. PRELIMINARIES

In this section, we introduce the foundational concepts of causal inference, setting the groundwork for our proposed method.

A. CAUSAL GRAPH

The causal graph [6], visualized in Fig. 3, serves as a foundational tool in causal inference to represent the interactions between variables. Formally defined, a causal graph is expressed by a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where \mathcal{N} signifies the set of variables in consideration, and \mathcal{E} , represented as arrows, indicates the causal relationships between these variables. In essence, if there’s an arrow from variable X to variable Y , denoted as $X \rightarrow Y$, this implies that X acts as a causal driver for Y . In other words, the outcome or manifestation of Y is influenced or precipitated by X . Such a graphical representation provides an overarching view of the causal interdependencies among the variables, making it a pivotal tool for understanding and inferring causality, especially in complex systems.

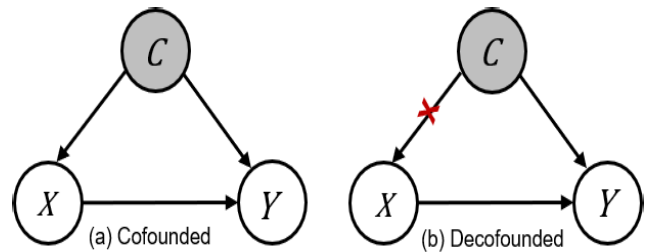


FIGURE 3. Example of the causal graph. The causal intervention $P(Y|do(X))$ cut off the short-cut caused by confounder C .

B. CAUSAL INTERVENTION

In causal inference, a confounder is a lurking variable that can introduce bias, potentially skewing or misrepresenting the actual causal relationship between the primary variables of interest. Taking Fig.3(a) as an illustration, suppose a model is designed to identify a specific feature, let’s call it ‘towel texture’, represented by X . In the training dataset, numerous instances of this feature might frequently appear against a particular backdrop, say, a ‘bedroom’, denoted as confounder $c \in C$. Due to this prevalent co-occurrence, the model could inadvertently associate the presence of c (bedroom) with Y (detection of the towel texture), even if the direct and logical indicator is X . This unintended correlation makes it problematic to rely solely on $P(Y|X)$ to determine the causal effect, as it doesn’t account for the potential influence of confounder $c \in C$.

To address the confounder’s influence and extract the true causal relation between variables, we utilize the backdoor adjustment, as conceptualized in the *do-calculus* from [6]. The “do” operation, a unique form of intervention, severs all incoming connections to the chosen variable, ensuring its independence from antecedent variables. In other words, as visualized in Fig.3(b), executing $do(X)$ entails disregarding external influences on X . The accurate causal relationship between X and Y is best represented as $P(Y|do(X))$, as this operation neutralizes confounding pathways such as $C \rightarrow X$ and $C \rightarrow Y$. The backdoor adjustment formula can be expressed as:

$$\begin{aligned}
 P(Y|do(X)) &= \sum_c P(Y|X, c)P(c|X) \\
 &= \sum_c P(Y|X, c)P(c) \tag{1}
 \end{aligned}$$

Here, the equation averages over all possible states of the confounder $c \in C$, weighing by their likelihood. $P(c|X)$ transitions to $P(c)$, ensuring it remains uninfluenced by X . The term $P(Y|X, c)$ considers the joint effects of X and c on Y , and by summing over c , we are effectively integrating out its effects to focus solely on the influence of X on Y .

In this work, we employ the backdoor adjustment to address dataset biases stemming from such spurious correlations in the task of Visual Commonsense Generation [1].

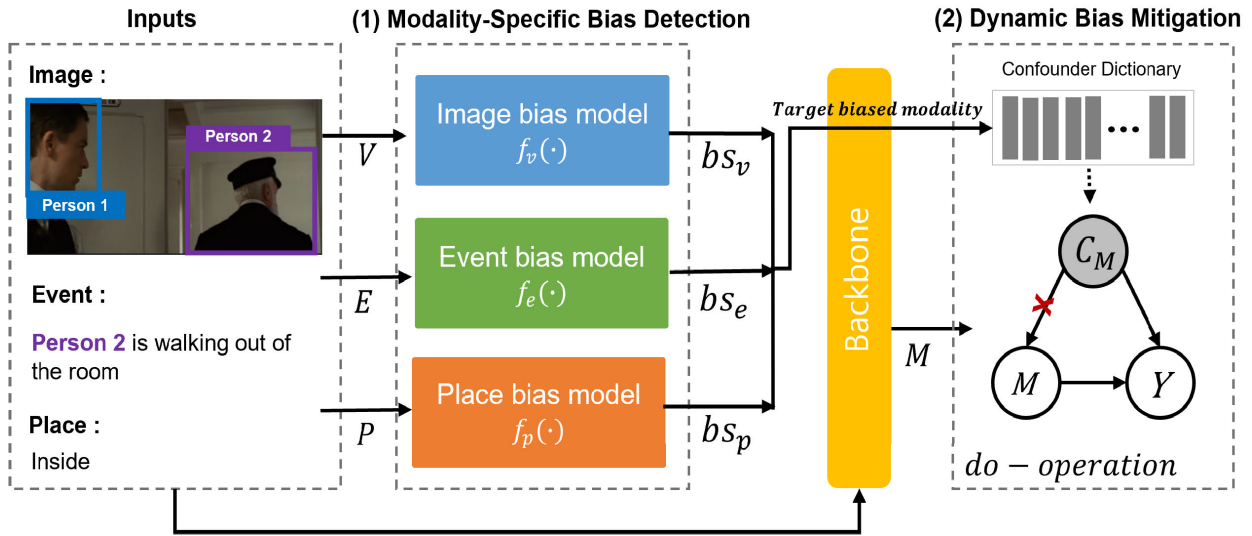


FIGURE 4. A brief overview of Dynamic Debiasing Network (DDNet) for VCG, comprised of (1) Modality-specific Bias Detection, with modality-specific bias model f_v, f_e, f_p and (2) Dynamic Bias Mitigation, containing a causal intervention process. Here, V, E , and P represent the embeddings for image, event, and place modalities, respectively. The outputs of each modality-specific bias model, bs_v, bs_e , and bs_p , denote the bias scores for their respective modalities. Additionally, M refers to the output from the final decoder layer of the backbone, which is utilized during the causal intervention.

IV. METHOD

In this section, we introduce Dynamic Debiasing Network (DDNet), which identifies and counteracts modality-specific biases dynamically. As illustrated in Fig.4, Dynamic Debiasing Network (DDNet) is divided in two main steps: (1) Modality-specific Bias Detection: capture which modality has the bias in input data, (2) Dynamic Bias Mitigation: dynamically mitigate bias to that particular modality, utilizing causal inference, which can remove spurious correlation and uncover true causal effect between variables. We elaborate each steps in following sections.

A. INPUT REPRESENTATION

The Visual Commonsense Generation (VCG) [1] takes three types of inputs comprised of an image, the event description, and the place description to generate three captions about *before, after, character’s intent*. In terms of input representation, we adopt approaches from prior research on VCG for both visual and textual inputs. For each image, a sequence of visual embeddings, V , is constructed. This sequence includes a representation of the entire image as well as separate representations for each person identified within the image. The Region of Interest (RoI) Align features [13] from Faster-RCNN [14] are employed as the visual embeddings. The final sequence, V , is expressed as $V = \{v_0, v_1, \dots, v_k\}$ where $v_0 \in \mathbb{R}^{d^v}$ represents the embedding of the whole image and $\{v_1, \dots, v_k\} \in \mathbb{R}^{d^v}$ are embeddings for each detected person. Here, d^v is the embedding dimension, and k denotes the total number of people detected in the image.

For two textual inputs, event and place descriptions, we derive the embeddings of event $E = \{e_0, e_1, \dots, e_{L^e}\}$, and place $P = \{p_0, p_1, \dots, p_{L^p}\}$ from word-embedding layer of pre-trained GPT-2 [11]. Here, L^e, L^p are the length of event,

place description, and $e_i \in \mathbb{R}^{d^e}, p_i \in \mathbb{R}^{d^p}$ are the embeddings for i -th token in event and place description, where d^e, d^p is the dimension of embedding.

B. MODALITY-SPECIFIC BIAS DETECTION

Our DDNet focus on the fact that when integrating multiple modalities as inputs, certain modality might exhibit a stronger bias than other. Therefore, we introduce a method to measure the bias of each modality in input data and identify the modality with the pronounced bias. We measure bias as a direct causal effect on the output of each modality, motivated by [5], which perceives language bias as the direct causal effect of questions on answers, disregarding other inputs. To assess the direct causal effect, our approach involves training a biased model for each modality, biased specifically towards that modality, and then comparing their outputs to discern which modality exhibits the most significant bias. Instead of feeding all modalities into the model simultaneously, we input only one modality individually at a time. This strategy ensures the model learns and produces outputs primarily influenced by that particular modality, highlighting its inherent biases. For instance, when training for image modality bias model f_v , we use only the visual embedding V as input data. During training bias model, the output of model’s last layer is then compared to the ground truth sentence to determine its degree of bias, which we term as the bias score bs_m , where m represent the type of modality. The same process is applied for other modalities, namely embeddings of event description E and place description P as follow :

$$\begin{aligned}
 bs_v &= f_v(V), \\
 bs_e &= f_e(E), \\
 bs_p &= f_p(P),
 \end{aligned} \tag{2}$$

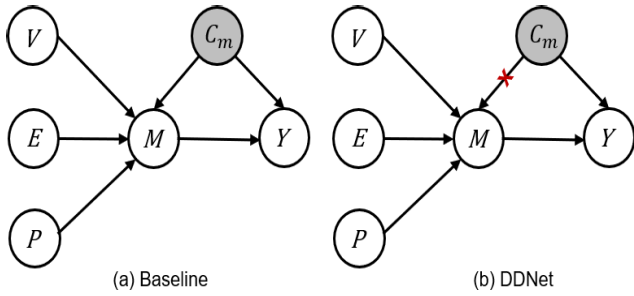


FIGURE 5. (a) A causal graph of baseline for VCG, (b) A causal graph of our proposed DDNet, which applies causal intervention to remove spurious correlation caused by the confounder C_m .

where $f_v, f_e,$ and f_p denote modality-specific bias models learned with only image V , event E , and place P inputs, respectively, and bs_v, bs_e, bs_p is bias score for that modality. Lastly, to ascertain the modality with the highest bias, we compare three bias score and detect a target modality.

$$\text{target modality} = \text{argmax}([bs_v, bs_e, bs_p]) \quad (3)$$

C. DYNAMIC BIAS MITIGATION

After pinpointing the primary modality exhibiting bias in input data, illustrated in section IV-B, we employ causal intervention to that data, as detailed in section III, targeting the confounder associated with this identified modality. In this section, our initial step is to formulate the task of VCG within the framework of a causal graph. Subsequently, we address biases present in each data instance by employing causal intervention on the confounders specific to the target biased modality, determined in the stage of IV-B.

Further insights into the construction of the causal graph and the process of the intervention process will be provided in subsequent subsections.

1) CAUSAL GRAPH FOR VCG

As depicted in Fig. 5, our causal graph representation for VCG incorporates six variables. These include the image V , event description E , and place description P . Combined, these three elements constitute the unified multimodal inputs M . Additionally, we introduce a confounder C_m , where depends on the type of modality : confounder C_v for image modality, C_e for event modality, C_p for place modality. Its determination is based on the modality that exhibits the most pronounced bias for each dataset instance, as detailed in IV-B. Lastly, the variable Y stands for the generated word in the output sentence. Specifically, the relationship $M \rightarrow Y$ captures the direct causal effect of the unified embeddings from the three inputs (V, E, P) on the generated word in the output sentence. On the other hand, the causal path $M \leftarrow C_m \rightarrow Y$ illustrates that the confounder C_m , stemming from co-occurrence patterns in the training data for the modality m , induces a spurious correlation between M and Y .

2) CAUSAL INTERVENTION FOR VCG

In order to infer the true causal effect from M to Y , we implement causal intervention as discussed in section III

and visualized in Fig.5(b). To tailor (1) to the VCG domain, we reformulate it as:

$$P(Y|do(M)) = \sum_c P(Y|M, c)P(c), \quad (4)$$

where $c \in C_m$ stands for a confounder that pertains to a specific modality m .

Given the challenge of observing every potential confounder, we only focus on the aspect that frequently co-occurrence pattern cause spurious correlation. Therefore, we approximate this by creating three confounder matrices C_m , each specifically designed for a modality m : image, event, and place. Each matrix adheres to the dimensions $N \times d_m$, where N represents the number of manually selected confounders, namely the top- N most frequently appearing objects within each modality’s training set, and d_m represents the feature embedding dimension for those objects. For the image modality’s confounder, inspired by [15], we derive each element using a pre-trained Faster R-CNN model [14], utilizing ground truth bounding boxes to calculate the average of RoI features from each object in the training set. In contrast, for the event and place modalities, we establish textual confounder dictionaries, denoted as C_e, C_p , respectively. These dictionaries are obtained using embeddings from a pre-trained BERT model [16].

3) IMPLEMENTATIONS

From (4), since the last layer of our network for word prediction is the softmax layer : $P(Y|M, c) = \text{softmax}(f(M, c))$, where M is the embedding of unified multimodal inputs, c is the entry of confounder C_m , and f represents the linear layer before softmax layer. Using this, (4) can be rewritten as :

$$\begin{aligned} P(Y|do(M)) &= \mathbb{E}_c[\text{softmax}(f(M, c))] \\ &\approx \text{softmax}(\mathbb{E}_c[f(M, c)]) \end{aligned} \quad (5)$$

Given the computational cost associated with sampling for c in (5), we employ the NWGM approximation [17], [18], allowing us to efficiently integrate the expectation within the softmax function. In our network, we model $f(M, c) = M + c$, where M is the embedding of unified multimodal input, the hidden state of the last decoder layer of backbone network. With this, we can calculate $\mathbb{E}_c[f(M, c)]$ to $M + \mathbb{E}_c[C_m]$. To enhance the overall model’s representational capability, we set c to be conditioned on the multimodal input M [19] : $\mathbb{E}_c[C_m] \approx \mathbb{E}_{[c|M]}[C_m]$. Finally, we use dot-product attention and obtain : $\mathbb{E}_{[c|M]}[C_m] = \text{softmax}(L^T K) \odot C_m$, where $L = w_1 M, K = w_2 C_m$ and w_1, w_2 is learnable parameters, \odot is element-wise product, and M is the embedding of unified multimodal input.

V. EXPERIMENTS

A. DATASET

The VisualCOMET dataset, a benchmark for Visual Commonsense Generation [1] provides a unique and large-scale resource with over 1.4 million textual captions detailing three

types of visual commonsense inferences : before/after and character’s intent in a image. These inferences are annotated across 59,000 images paired with 139,000 event descriptions. The dataset comprises 1,174K training examples, 146K validation examples, and 145K test examples. Following prior works [1], [3], [4], we present our model’s performance on both the validation and test sets.

B. METRICS

To evaluate performance on the VisualCOMET benchmarks, we employed three standard evaluation metrics for generative tasks: BLEU-2 [20], METEOR [21], and CIDEr [22]. The BLEU score measures precision between generated and reference captions using n-gram overlap. The METEOR score computes a weighted F-score considering unigram mappings and penalizes out-of-order correct words. Latly, the CIDEr score evaluates image captioning by calculating cosine similarity between TF-IDF features of generated and reference captions based on n-grams.

C. IMPLEMENTATION DETAILS

Our approach is model-agnostic, emphasizing flexibility and compatibility with various model architectures. We choose the pretrained GPT-2 framework as our backbone, as network proposed by [1]. Not only was it among the first to establish a robust baseline, but its clear design principles also guarantee reproducibility. For three modality-specific bias model, we utilize Transformer model [23] by pretraining for 10 epochs. When these three models are integrated into DDNet with a backbone, their parameters are frozen and not subjected to further updates during training. Our computational framework comprised 4 NVIDIA Quadro RTX 8000 GPUs, each with 48 GB of memory. During the training phase, we set the learning rate as $5e-5$ and utilized the Adam optimizer [24]. The training regimen spanned 30 epochs with a batch size of 256. To decode the output sequences, we employed a beam search strategy, which is well-regarded for producing accurate predictions by analyzing a wide array of sequence hypotheses. For all three pre-defined confounder dictionaries, we set a matrix for N as 100 and for d_m as 768.

D. QUANTITATIVE RESULTS

Table.2 presents the outcomes of applying various confounders to the validation and test sets of the VisualCOMET dataset. Initially, we applied a singular type of confounder to all data without using a modality-specific bias model. Subsequently, by incorporating DDNet, we dynamically mitigated bias across data. Across the board, our results indicate that the introduction of confounders significantly improves the performance of the base models. Moreover, the use of DDNet proves to be more effective in bias reduction compared to the application of a single type of confounder (e.g., only event, place, or image confounder). This underscores the benefit of customizing the choice of

TABLE 2. Results of applying our method to the baseline model on validation and test set of VCG dataset. Evaluation metrics “B@2”, “M”, “C” denote BLEU2, METEOR, CIDEr, respectively. “Event conf”, “Place”, and “Image conf” refer to confounders in each respective modality. The results shown are from applying only one type of confounder to all samples in the dataset. For instance, “+Event conf” indicates the application of the event modality confounder to all data, without a modality-specific bias model.

Models	Validation Set			Test Set		
	B@2	M	C	B@2	M	C
Baseline [1]	13.50	11.55	18.27	12.71	11.13	17.36
Add on baseline						
+ Event confounder	26.59	17.28	42.0	25.01	16.34	40.01
+ Place confounder	24.27	16.13	39.46	23.11	14.99	37.89
+ Image confounder	25.48	17.19	39.99	24.24	15.75	38.79
+ DDNet	29.02	18.88	44.01	28.14	17.89	43.02

TABLE 3. Comparison with state-of-the-arts methods on validation and test set of VCG dataset.

Models	Validation Set			Test Set		
	B@2	M	C	B@2	M	C
Baseline [1]	13.50	11.55	18.27	12.71	11.13	17.36
KM-BART [3]	23.47	15.02	39.76	-	-	-
CE-BART [4]	28.60	19.32	43.58	28.14	18.91	42.64
DDNet	29.02	18.88	44.01	28.14	17.89	43.02

confounder based on the unique requirements of each data sample, as opposed to a uniform application of a single confounder across all data.

In Table 3, we compare DDNet with other state-of-the-art methods [3], [4] on validation and test set of VCG dataset. Notably, we can observe that DDNet achieves the highest BLEU2, and CIDEr scores on both validation and test set. Even when DDNet simply fine-tunes the pre-existing GPT-2 architecture applying causal inference, it surpasses KM-BART in performance across both validation and test sets. It’s noteworthy that KM-BART requires specialized pre-training for VCG using various external data, whereas DDNet does not. On the other hand, while CE-BART necessitates the construction of additional graphs for each modality, DDNet may offers a compelling alternative, boasting both ease of implementation and superior interpretability. Considering that the DDNet proposed in this study is model-agnostic, we foresee potential enhancements in performance when combined with methods like KM-BART and CE-BART.

E. QUALITATIVE ANALYSIS

In Fig.6, we present a qualitative comparison of DDNet with the baseline model. Our objective with DDNet is to address and reduce the modality-specific bias observed in the VCG dataset. For the image modality, the confounder is ‘multiple people’, example mentioned in Fig.1. Unlike the baseline model [1] which generates sentences biased towards the confounder, such as ‘play games in his room’ and ‘play with a man’, DDNet creates captions that are less influenced by this bias, effectively leveraging information from diverse modalities. Notably, when generating the before event, DDNet produce more accurate inference, i.e. be told by a man to read the paper, not relying solely on the image.


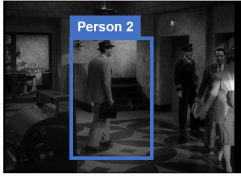
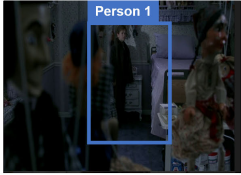
Bias	Inputs	Type	Baseline	DDNet
Image	<p>Image</p>  <p>Event 2 holds an egg roll and looks disappointed as she reads a paper in front of her</p> <p>Place In a bed room</p>	Before	Bring the egg to a man	Be told by a man to read the paper
		Intent	Play games in his room	Read the paper
		After	Play with a man	Tell a man to leave her alone
Event	<p>Image</p>  <p>Event 2 is carrying a notebook and is walking toward a door</p> <p>Place inside</p>	Before	Be interested in the book	Pick up the book
		Intent	Go to school	Go to the door
		After	Open the book	Open the door
Place	<p>Image</p>  <p>Event 1 looks nervous as he stands near the entryway in a room full of creepy puppets</p> <p>Place In a bedroom</p>	Before	Feel afraid	Enter the room
		Intent	Try to break in	See what the puppets are doing
		After	Go to sleep	Tell someone about the room

FIGURE 6. Qualitative comparison between the baseline model and our DDNet. Examples of mitigating the modality-specific bias in three modalities : image, event, place. In inputs of these examples, a red bold box in image, and red bold text in event and place description, highlight one of modality-specific confounders, leading to spurious correlation during training.

It utilizes textual information, like ‘she looks disappointed’, which is crucial for generating accurate predictions.

For event and place modality, it can be seen that DDNet effectively combines information from the three modalities to produce sentences that are more accurate than baseline, removing the influence of specific confounders, namely ‘book’ for the event modality and ‘bedroom’ for the place modality.

VI. LIMITATIONS

We recognize that our DDNet has certain limitations, and we anticipate that future experiments can provide solutions to address these issues. As DDNet focuses on identifying and mitigating the modality with the highest bias for each data sample, it may overlook biases present in other modalities. Consequently, while the dominant bias in a particular modality may be reduced, residual biases in other modalities might remain unaddressed. In future work we extend the debiasing framework to simultaneously address biases in multiple modalities, even if they are not the most dominant. This can be done by employing a weighted debiasing approach, where each modality is debiased based on its degree of bias.

In the context of the visual commonsense generation task, words are generated sequentially. Applying causal intervention at each step of word prediction would significantly increase the computational overhead, leading to prolonged training times. To address this, our future direction

will focus on developing an framework that measure the bias automatically associated with each word during the generation process. This approach aims to provide a more dynamic and efficient means of bias mitigation.

VII. CONCLUSION

In this study, we introduced DDNet, an novel debiasing approach for Visual Commonsense Generation. While previous research has largely neglected the impact of causal perspectives in this domain, DDNet brings this critical aspect to the forefront.

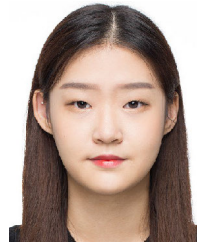
DDNet is architected to identify the modality that exhibits the most prominent bias in data sample. Once detected, it initiates a tailored debiasing process for the identified modality, ensuring that the generated outputs are free from each modality-specific confounder’s influence. A notable advantage of our approach is its model-agnostic nature, which ensures compatibility with a wide array of existing methodologies. This versatility translates to enhanced performance when integrated with other models. Additionally, a distinctive feature of DDNet, setting it apart from its contemporaries, is its ability to dynamically counteract spurious correlations accross each modality. By invoking causal intervention mechanisms, it not only refines the output but also significantly bolsters the model’s resilience against biases.

Our empirical evaluations, conducted on the VCG dataset, provide robust evidence of DDNet in enhancing the fairness

and accuracy of visual commonsense generation tasks. Promising results from extensive experiments on VCG dataset demonstrate the effectiveness of our debiasing method. In essence, this research marks a pioneering stride towards fostering fairness in visual commonsense generation by actively countering dataset-induced biases. As we continue to evolve our methodologies and delve deeper into this domain, we direct readers to section VI for insights into our prospective directions.

REFERENCES

- [1] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi, "VisualCOMET: Reasoning about the dynamic context of a still image," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Manhattan, NY, USA: Springer, 2020, pp. 508–524.
- [2] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6713–6724.
- [3] Y. Xing, Z. Shi, Z. Meng, G. Lakemeyer, Y. Ma, and R. Wattenhofer, "KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation," 2021, *arXiv:2101.00419*.
- [4] J. Kim, J. W. Hong, S. Yoon, and C. D. Yoo, "CE-BART: Cause-and-effect BART for visual commonsense generation," *Sensors*, vol. 22, no. 23, p. 9399, Dec. 2022.
- [5] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12700–12710.
- [6] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.
- [7] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.
- [8] C. Zang, H. Wang, M. Pei, and W. Liang, "Discovering the real association: Multimodal causal reasoning in video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19027–19036.
- [9] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1–10.
- [10] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2764–2774.
- [11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, Feb. 2019.
- [12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
- [15] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 655–666.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhudinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [19] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao, "Show, deconfound and tell: Image captioning with causal inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18041–18050.
- [20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
- [21] A. Lavie and A. Agarwal, "Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 65–72.
- [22] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.



JUNGEUN KIM (Graduate Student Member, IEEE) received the B.S. degree in intelligent mechatronics engineering from Sejong University, Seoul, Republic of Korea, in 2022. She is currently pursuing the M.S. degree with Chung-Ang University.

Her research interests include causal reasoning, visual-language reasoning, and various multi-modal reasoning.



JINWOO PARK (Student Member, IEEE) received the B.A. degree in urban planning and real estate from Chung-Ang University, Seoul, Republic of Korea, in 2018, where he is currently pursuing the M.S. degree.

His research interests include visual-language reasoning and various multi-modal reasoning problem.



JAEKWANG SEOK received the B.S. degree in mathematics from Chung-Ang University, Seoul, Republic of Korea, in 2023. He is currently pursuing the M.S. degree in artificial intelligence.

His research interests include disentangle representation learning, fairness, and multi-modal reasoning.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively.

He was a Postdoctoral Research Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include

visual-language reasoning, visual question answering, and various video-based problem.

• • •