Two Causal Principles for Improving Visual Dialog

Jiaxin Qi*[1] Yulei Niu*[2] Jianqiang Huang[1,3] Hanwang Zhang[1]
[1]Nanyang Technological University, [2]Renmin University of China,
[3]Damo Academy, Alibaba Group,
jiaxin003@e.ntu.edu.sg, niu@ruc.edu.cn, jianqiang.jqh@gmail.com, hanwangzhang@ntu.edu.sg

Abstract

This paper unravels the design tricks adopted by us — the champion team MReaL-BDAI — for Visual Dialog Challenge 2019: ***two causal principles*** *for improving Visual Dialog (VisDial). By "improving", we mean that they can promote almost every existing VisDial model to the state-of-the-art performance on the leader-board. Such a major improvement is only due to our careful inspection on the* ***causality*** *behind the model and data, finding that the community has overlooked two causalities in VisDial. Intuitively,* ***Principle 1*** *suggests: we should remove the direct input of the dialog history to the answer model, otherwise a harmful shortcut bias will be introduced;* ***Principle 2*** *says: there is an unobserved confounder for history, question, and answer, leading to spurious correlations from training data. In particular, to remove the confounder suggested in Principle 2, we propose several* ***causal intervention*** *algorithms, which make the training fundamentally different from the traditional likelihood estimation. Note that the two principles are* model-agnostic*, so they are applicable in any VisDial model. The code is available at* `https://github.com/simpleshinobu/visdial-principles`.

1. Introduction

Given an image I, a dialog history of past Q&A pairs: $H = \{(Q_1, A_1), ..., (Q_{t-1}, A_{t-1})\}$, and the current t-th round question Q, a Visual Dialog (VisDial) agent [9] is expected to provide a good answer A. Our community has always considered VQA [5] and VisDial as sister tasks due to their similar settings: Q&A grounded by I (VQA) and Q&A grounded by (I, H) (VisDial). Indeed, from a technical point view — just like the VQA models — a typical VisDial model first uses *encoder* to represent I, H, and Q as vectors, and then feed them into *decoder* for answer A. Thanks to the recent advances in encoder-decoder frameworks in VQA [22, 38] and natural language processing [39], the performance (NDCG [1]) of VisDial in literature is significantly improved from the baseline 51.63% [2] to the state-of-the-art 64.47% [11].

*Equal contribution

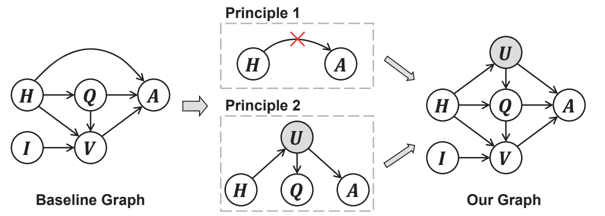


Figure 1. Causal graphs of VisDial models (baseline and ours). H: dialog history. I: image. Q: question. V: visual knowledge. A: answer. U: user preference. Shaded U denotes unobserved confounder. See Section 3.2 for detailed definitions.

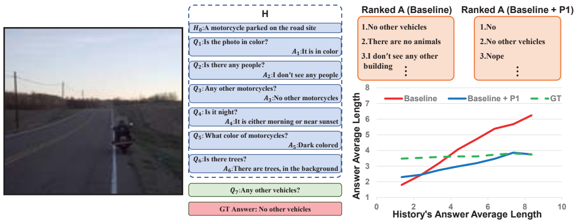
However, in this paper, we want to highlight an important fact: *VisDial is essentially NOT VQA with history!* And this fact is so profound that all the common heuristics in the vision-language community — such as the multimodal fusion [38, 47] and attention variants [22, 25, 26] — cannot appreciate the difference. Instead, we introduce the use of *causal inference* [27, 28]: a graphical framework that stands in the cause-effect *interpretation* of the data, but not merely the statistical *association* of them. Before we delve into the details, we would like to present the main contributions: two causal principles, rooted from the analysis of the difference between VisDial and VQA, which lead to a performance leap — a farewell to the 60%-s and an embrace for the 70%-s — for all the baseline VisDial models[1] in literature [9, 21, 41, 26], promoting them to the state-of-the-art in Visual Dialog Challenge 2019 [2].

Principle 1 ***(P1)****: Delete link* $H \rightarrow A$.

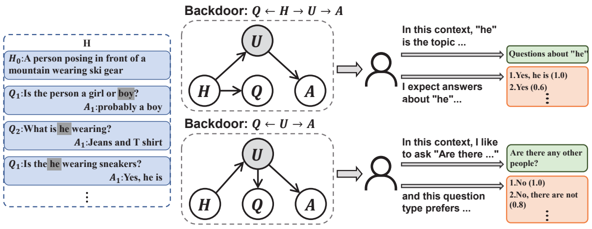
Principle 2 ***(P2)****: Add one new (unobserved) node* U *and three new links:* $U \leftarrow H$, $U \rightarrow Q$, *and* $U \rightarrow A$.

Figure 1 compares the causal graphs of existing VisDial models and the one applied with the proposed two principles. Although a formal introduction of them is given in

[1]Only those with codes&reproducible results due to resource limit.



(a) A Typical $H \rightarrow A$ Bias



(b) User Preference

Figure 2. The illustrative motivations of the two causal principles: (a) P1 and (b) P2.

Section 3.2, now you can simply understand the nodes as data types and the directed links as modal transformations. For example, $V \rightarrow A$ and $Q \rightarrow A$ indicate that answer A is the *effect caused* by visual knowledge V and question Q, through a transformation, *e.g.*, a multi-modal encoder.

P1 suggests that we should remove the *direct* input of dialog history to the answer model. This principle contradicts most of the prevailing VisDial models [9, 15, 41, 26, 43, 16, 11, 32], which are based on the widely accepted intuition: the more features you input, the more effective the model is. It is mostly correct, but only with our discretion of the data generation process. In fact, the VisDial [9] annotators were not allowed to copy from the previous Q&A, *i.e.*, $H \nrightarrow A$, but were encouraged to ask consecutive questions including co-referenced pronouns like "it" and "those", *i.e.*, $H \rightarrow Q$, and thus the answer A is expected to be only based on question Q and reasoned visual knowledge V. Therefore, a good model should reason over the context (I, H) with Q but not to memorize the bias. However, the direct path $H \rightarrow A$ will contaminate the expected causality. Figure 2(a) shows a very ridiculous bias observed in all baselines without P1: the top answers are those with length closer to the average length in the history answers. We will offer more justifications for P1 in Section 4.1.

P2 implies that the model training only based on the association between (I, H, Q) and A is spurious. By "spurious", we mean that the effect on A caused by (I, H, Q) — the goal of VisDial — is *confounded* by an unobserved variable U, because it appears in every undesired causal path (*a.k.a.*, backdoor [28]), which is an indirect causal link from input (I, H, Q) to output A: $Q \leftarrow U \rightarrow A$ and $Q \leftarrow H \rightarrow U \rightarrow A$. We believe that such unobserved U should be *users* as the VisDial dataset essentially brings humans in the loop. Figure 2(b) illustrates how the user's hidden preference confounds them. Therefore, during training, if we focus only on the conventional likelihood $P(A|I, H, Q)$, the model will inevitably be biased towards the spurious causality, *e.g.*, it may score answer "Yes, he is" higher than "Yes", merely because the users prefer to see a "he" appeared in the answer, given the history context of "he". It is worth noting that the confounder U is more impactful in VisDial than in VQA, because the former encourages the user to rank similar answers subjectively while the latter is more objective. A plausible explanation might be: VisDial is interactive in nature and a not quite correct answer is tolerable in one iteration (*i.e.*, dense prediction); while VQA has only one chance, which demands accuracy (*i.e.*, one-hot prediction).

By applying P1 and P2 to the baseline causal graph, we have the proposed one (the right one in Figure 1), which serves as a *model-agnostic* roadmap for the causal inference of VisDial. To remove the spurious effect caused by U, we use the *do-calculus* [28] $P(A|do(I, H, Q))$, which is fundamentally different from the conventional likelihood $P(A|I, H, Q)$: the former is an active *intervention*, which cuts off $U \rightarrow Q$ and $H \rightarrow Q$, and sample (where the name "calculus" is from) every possible $U|H$, seeking the true effect on A only caused by (I, H, Q); while the latter likelihood is a passive *observation* that is affected by the existence of U. The formal introduction and details will be given in Section 4.3. In particular, given the fact that once the dataset is ready, U is no longer observed, we propose a series of effective approximations in Section 5.

We validate the effectiveness of P1 and P2 on the most recent VisDial v1.0 dataset. We show significant boosts (absolute NDCG) by applying them in 4 representative baseline models: LF [9] ($\uparrow$16.42%), HCIAE [21] ($\uparrow$15.01%), CoAtt [41] ($\uparrow$15.41%), and RvA [26] ($\uparrow$16.14%). Impressively, on the official test-std server, we use an ensemble model of the most simple baseline LF [9] to beat our 2019 winning performance by 0.2%, a more complex ensemble to beat it by 0.9%, and lead all the single-model baselines to the state-of-the-art performances.

2. Related Work

Visual Dialog. Visual Dialog [9, 10] is more interactive and challenging than most of the vision-language tasks, *e.g.*, image captioning [46, 44, 4] and VQA [5, 38, 37, 36]. Specifically, Das *et al.* [9] collected a large-scale free-form visual dialog dataset VisDial [7]. They applied a novel protocol: during the live chat, the questioner cannot see the picture and asks open-ended questions, while the answerer gives free-form answers. Another dataset GuessWhat?! proposed by [10] is a goal-driven visual dialog: questioner should locate an unknown object in a rich image scene by asking a sequence of closed-ended "yes/no" questions. We apply the first setting in this paper. Thus, the key difference is that the users played an important role in the data collection process.

All of the existing approaches in the VisDial task are based on the typical encoder-decoder framework [15, 12, 33, 11, 32, 48]. They can be categorized by the usage of history. 1) Holistic: they treat history as a whole to feed into models like HACAN [43], DAN [16] and CorefNMN [18]. 2) Hierarchical: they use a hierarchical structure to deal with history like HRE [9]. 3) Recursive: RvA [26] uses a recursive method to process history. However, they all overlook the fact that the history information should not be directly fed to the answer model (*i.e*., our proposed Principle 1). The baselines we used in this paper are LF [9]: the earliest model, HCIAE [21]: the first model to use history hierarchical attention, CoAtt [41]: the first one to a co-attention mechanism, and RvA [26]: the first one for a tree-structured attention mechanism.

Causal Inference. Recently, some works [24, 6, 23, 34, 40, 45] introduced causal inference into machine learning, trying to endow models the abilities of pursuing the cause-effect. In particular, we use the Pearl's structural causal model (SCM) proposed by [28] to hypothesize the data generation process, which is a model-agnostic framework that reflects the nature of the data.

3. Visual Dialog in Causal Graph

In this section, we formally introduce the visual dialog task and describe how the popular encoder-decoder framework follows the baseline causal graph shown in Figure 1. More details of causal graph can be found in [28, 29].

3.1. Visual Dialog Settings

Settings. According to the definition of VisDial task proposed by Das *et al*. [9], at each time t, given input image I, current question Q_t, dialog history $H = \{C, (Q_1, A_1), \ldots, (Q_{t-1}, A_{t-1})\}$, where C is the image caption, (Q_i, A_i) is the i-th round Q&A pair, and a list of 100 candidate answers $A_t = \{A_t^{(1)}, \ldots, A_t^{(100)}\}$. A VisDial model is evaluated by ranking candidate answers A_t.

Evaluation. Recently, the ranking metric Normalized Discounted Cumulative Gain (NDCG) is adopted by the VisDial community [1]. It is different from the classification metric (*e.g*., top-1 accuracy) used in VQA. It is more compatible with the relevance scores of the answer candidates in VisDial rated by humans. NDCG requires to rank relevant candidates in higher places, rather than just to select the ground-truth answer.

3.2. Encoder-Decoder as Causal Graph

We first give the definition of causal graph, then revisit the encoder-decoder framework in existing methods using the elements from the baseline graph in Figure 1.

Causal Graph. Causal graph [28], as shown in Figure 1, describes how variables interact with each other, expressed by a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ consisting of nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ (*i.e*., arrows). $\mathcal{N}$ denote variables, and $\mathcal{E}$ (arrows) denote the causal relationships between two nodes, *i.e*., $A \rightarrow B$ denotes that A is the cause and B is the effect, meaning the outcome of B is caused by A. Causal graph is a highly general roadmap specifying the causal dependencies among variables.

As we will discuss in the following part, all of the existing methods can be revisited in the view of the baseline graph shown in Figure 1.

Feature Representation and Attention in Encoder. Visual feature is denoted as node I in the baseline graph, which is usually a fixed feature extracted by Faster-RCNN [31] based on ResNet backbone [13] pre-trained on Visual Genome [19]. For language feature, the encoder firstly embeds sentence into word vectors, followed by passing the RNN [14, 8] to generate features of question and history, which are denoted as $\{Q, H\}$.

Most of existing methods apply attention mechanism [42] in encoder-decoder to explore the latent weights for a set of features. A basic attention operation can be represented as $\tilde{\boldsymbol{x}} = Att(\mathcal{X}, \mathcal{K})$ where $\mathcal{X}$ is the set of features need to attend, $\mathcal{K}$ is the key (*i.e*., guidance) and $\tilde{\boldsymbol{x}}$ is the attended feature of $\mathcal{X}$. Details can be found in most visual dialog methods [21, 41, 43]. In the baseline graph, the subgraph $\{I \rightarrow V, Q \rightarrow V, H \rightarrow Q \rightarrow V\}$ denotes a series of attention operations for visual knowledge V. Note that the implementation of the arrows are not necessarily independent, such as co-attention [41], and the process can be written as *Input* : $\{I, Q, H\} \Rightarrow$ *Output* : $\{V\}$, where possible intermediate variables can be added as mediator nodes into the original arrows. However, without loss of generality, these mediators do not affect the causalities in the graph.

Response Generation in Decoder. After obtaining the features from the encoder, existing methods will fuse them and feed the fused ones into a decoder to generate an answer. In the baseline graph, node A denotes the answer model that decodes the fused features from $\{H \rightarrow A, Q \rightarrow A, V \rightarrow A\}$ and then transforms them into an answer sentence. In particular, the decoder can be generative, *i.e*., to generate an answer sentence using RNN; or discriminative, *i.e*., to select an answer sentence by using candidate answer classifiers.

4. Two Causal Principles

4.1. Principle 1

When should we draw an arrow from one node pointing to another? According to the definition in Section 3.2, the criterion is that if the node is the cause and the other one is the effect. Intrigued, let's understand P1 by discussing the rationale behind the "double-blind" review policy. Given three variables: "Well-known Researcher" (R), "High-quality Paper" (P), and "Accept" (A). From our community common sense, we know that $R \to P$ since top researchers usually lead high-quality research, and $P \to A$ is even more obvious. Therefore, for the good of the community, the double-blind prohibits the direct link $R \to A$ by author anonymity, otherwise the bias such as personal emotions and politics from R may affect the outcome of A.

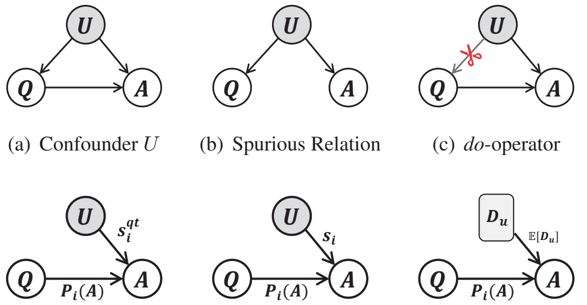
The story is similar in VisDial. Without loss of generality, we only analyze the path $H \to Q \to A$. If we inspect the role of H, we can find that it is to help Q resolve some co-references like "it" and "their". As a result, Q listens to H. Then, we use Q to obtain A. Here, Q becomes a mediator which cuts off the direct association between H and A that makes $P(A|Q,H) = P(A|Q)$, like the "High-quality Paper" that we mentioned in the previous story. However, if we set an arrow from H to A: $H \to A$, the undesirable bias of H will be learned for the prediction of A, that hampers the natural process of VisDial, such as the interesting bias illustrated in Figure 2(a). Another example is discussed in Figure 4 that A prefers to match the words in H, even though they are literally nonsense about Q if we add the direct link $H \to A$. After we apply P1, these phenomena will be relieved, such as the blue line illustrated in Figure 2(a), which is closer to the NDCG ground truth average answer length , denoted as the green dashed line. Please refer to other qualitative studies in Section 6.5.

4.2. Principle 2

Before discussing P2, we first introduce an important concept in causal inference [28]. In causal graph, the fork-like pattern in Figure 3(a) contains a *confounder* U, which is the common cause for Q and A (*i.e.*, $Q \leftarrow U \to A$). The confounder U opens a *backdoor* path started from Q, making Q and A spuriously correlated even if there is no direct causality between them.

In the data generation process of VisDial, we know that not only both of the questioner and answerer can see the dialog history, but also the answer annotators can look at the history when annotating the answer. Their preference after seeing the history can be understood as a part of the human nature or subtleties conditional on a dialog context, and thus it has a causal effect on both Q and A. Moreover, due to the fact that the preference is nuanced and uncontrollable, we consider it as an *unobserved* confounder for Q and A.

It is worth noting that the confounder hinders us to find the true causal effect. Let's take the graph in Figure 3(b) as an example. The causal effect from Q to A is 0; however, we can quickly see that $P(A|Q) - P(A)$ is nonzero because Q and A are both influenced by U and thus are correlated (thanks to Reichenbach's common cause principle [28]). That is, if we are given Q, the any likelihood change for A will be sensible compared to nothing is given.



(a) Confounder U (b) Spurious Relation (c) *do*-operator

(d) Question Type (e) Score Sampling (f) Hidden Dictionary

Figure 3. Example of confounder, *do*-operator and sketch causal graphs of our three attempts to de-confounder

Therefore, if we consider $P(A|Q)$ as our VisDial model, it will still predict nonsense answers even if Q has nothing to do with A. As illustrated in Figure 2(b), model will prefer the candidates about "he" even though Q is not given, that means it captures the confounder U but not the true rationale between Q and A. Next, we will introduce a powerful technique that makes the Q and A in Figure 3(b) "independent", *i.e.*, no causal relation.

4.3. *do*-calculus.

The technique is *do*-calculus introduced in [28, 29]. Specifically, $do(Q = q)$ denotes that we deliberately assign a value q to variable Q (*i.e.*, intervention), rather than passively observe $Q = q$. As illustrated in Figure 3(c), $do(Q = q)$ can be understood as cutting all the original incoming arrows to Q, and then making Q and U independent. Therefore, we can have the well-known backdoor adjustment [28]: $P(A|do(Q = q)) = \sum_u P(A|Q = q,u)P(u)$. Note that this is different from Bayes rule $P(A|Q = q) = \sum_u P(A|Q = q,u)P(u|Q = q)$ thanks to the independence $P(u|Q = q) = P(u)$ introduced by *do*-calculus. Let's revisit Figure 3(b) by using *do*-calculus. We can find that $P(A|do(Q = q)) - P(A) = 0$, that is to say, any intervention of Q will not influence the probability of A, meaning the correct relation between Q and A: no causal relation. Therefore, $P(A|do(Q = q))$ should be the objective answer model in VisDial.

For our proposed graph of VisDial shown in Figure 1, we can use intervention $do(Q,H,I)$ and the backdoor adjustment to obtain our overall model. Here, we slightly abuse the notation $do(Q,H,I)$ as $do(Q = q, H = h, I = i)$:

$$\begin{aligned} & P(A|do(Q,H,I)) \\ & = \sum\nolimits_u P(A|do(Q,H,I),u)P(u|do(Q,H,I)) \\ & = \sum\nolimits_u P(A|do(Q),H,I,u)P(u|H) \\ & = \sum\nolimits_u P(A|Q,H,I,u)P(u|H). \end{aligned} \tag{1}$$

The detailed derivation and proof can be found in supplementary materials.

So far, we have provided all the ingredients of the baseline causal graph, two proposed principles and their theoretical solution: *do*-calculus. Next, we will introduce some implementations for the proposed solution in Eq. (1).

5. Improved Visual Dialog Models

It is trivial to implement P1 and we will provide its training details in Section 6.3. For P2, since U is unobserved, it is impossible to sample u in Eq. (1) directly. Therefore, our technical contribution is to introduce 3 approximations. For notation simplicity, we first re-write Eq. (1) as:

$$P(A|do(Q,H,I)) = \sum\nolimits_u P_u(A)P(u|H), \tag{2}$$

where $P_u(A) := P(A|Q,H,I,u)$.

5.1. Question Type

Since we cannot directly sample u from the unobserved confounder, we use the i-th answer candidate a_i as a delegate for sample u. That is because a_i is a sentence observed from the "mind" of user u during dataset collection. Then, $\sum_u P_u(A)P(u|H)$ can be approximated as $\sum_i P_i(A)P(a_i|H)$. We further use $p(a_i|QT)$ to approximate $P(a_i|H)$ because of two reasons: First, $P(a_i|H)$ essentially describes a prior knowledge about a_i without comprehending the whole $\{Q,H,I\}$ triplet. A similar scenario is that if we know the QT (question type), *e.g.*, "what color", the answer candidates denoting colors have higher probabilities without even comprehending the question details. Second, QT is extracted from question Q, which is a descendent of history H in our graph, indicating that QT partially reveals H [28]. In practice, we manually define some question types, each of which has a certain answer frequency. For each dialog round, a normalized score $s_i^{qt} := p(a_i|QT)$ (*i.e.*, $\sum_i s_i^{qt} = 1$) of each candidate a_i will be calculated according to the frequency of a_i under question type qt. More details are given in Section 6.3. Finally, we have the approximation for Eq. (2):

$$\sum\nolimits_u P_u(A)P(u|H) \approx \sum\nolimits_i P_i(A) \cdot s_i^{qt}, \tag{3}$$

where $P_i(A) = \text{softmax}(f_s(\boldsymbol{e}_i, \boldsymbol{m}))$, f_s is a similarity function, $\boldsymbol{e}_i$ is the embedding of candidate a_i, $\boldsymbol{m}$ is the joint embedding for $\{Q,I,H\}$, and the sketch graph is shown in Figure 3(d). Since question type is observed from Q, the approximation $p(a_i|QT)$ undermines the prior assumption of the backdoor adjustment in Eq. (1) (*i.e.*, the prior $p(u|H)$ cannot be conditional on Q). Fortunately, QT is only a small part of Q (*i.e.*, the first few words) and thus the approximation is reasonable.

5.2. Answer Score Sampling

Since the question type implementation slightly undermines the backdoor adjustment, we will introduce a better approximation which directly samples from u: Answer Score Sampling. This implementation is also widely known as our previously proposed **dense fine-tune** in community [3].

We still use a_i to approximate u, and we use the (normalized) ground-truth NDCG score s_i annotated by the humans to approximate $P(a_i|H)$. Note that s_i directly reveals human preference for a_i in the context H (*i.e.*, the prior $P(a_i|H)$). In practice, we use the subset of training set with dense annotations to sample s_i. Therefore, we have:

$$\sum\nolimits_u P_u(A)P(u|H) \approx \sum\nolimits_i P_i(A) \cdot s_i, \tag{4}$$

and the sketch graph is illustrated in Figure 3(e). In practice, Eq. (4) can be implemented using different loss functions. Here we give three examples:

Weighted Softmax Loss (R_1). We extend the log-softmax loss as a weighted form, where $P_i(A)$ is denoted by $\log(\text{softmax}(p_i))$, p_i denotes the logit of candidate a_i, and s_i is corresponding normalized relevance score.

Binary Sigmoid Loss (R_2). This loss is close to the binary cross-entropy loss, where $P_i(A)$ represents $\log(\text{sigmoid}(p_i))$ or $\log(\text{sigmoid}(1-p_i))$, and s_i represents corresponding normalized relevance score.

Generalized Ranking Loss (R_3). Note that the answer generation process can be viewed as a ranking problem. Therefore, we derive a ranking loss that $P_i(A)$ is $\log \frac{\exp(p_i)}{\exp(p_i)+\sum_{j\in G}\exp(p_j)}$, where G is a group of candidates which have lower relevance scores than candidate a_i and s_i is normalized characteristic score (*i.e.*, equals to 0 for a_i with relevance score 0 and equals to 1 for a_i with positive relevance score).

More details of the three loss functions are given in supplementary materials. It is worth noting that our losses are derived from the underlying causal principle P2 in Eq. (4), but not from the purpose of regressing to the ground-truth NDCG. The comparison will be given in Section 6.4.

5.3. Hidden Dictionary Learning

The aforementioned two implementations are discrete since they sample specific a_i to approximate u. For better approximation, we propose learning to approximate the unobserved confounder U. As shown in Figure 3(f), we design a dictionary to model U. In practice, we design the dictionary as a $N \times d$ matrix D_u, where N is manually set and d is the hidden feature dimension. Note that given a sample u and a answer candidate a_c, Eq.(2) can be implemented as $\sum_u P_u(a_c)P(u|H)$. Since the last layer of our network for answer prediction is a softmax layer:

$P_u(a_c) = \text{softmax}(f_s(\boldsymbol{e_c}, \boldsymbol{u}, \boldsymbol{m}))$, where $\boldsymbol{e_c}$ is the embedding of candidate a_c, $\boldsymbol{u}$ is sampled from $\boldsymbol{D_u}$, $\boldsymbol{m}$ is the joint embedding for $\{Q, I, H\}$, and f_s is a similarity computation function, the Eq.(2) can be re-written as:

$$P(A|do(Q, H, I)) := \mathbb{E}_{[u|H]}\left[\text{softmax}(f_s(\boldsymbol{e_c}, \boldsymbol{u}, \boldsymbol{m}))\right]. \tag{5}$$

Since Eq. (5) needs expensive samplings for $\boldsymbol{u}$, we use NWGM approximation [42, 35] to efficiently move the expectation into the softmax:

$$\mathbb{E}_{[u|H]}[\text{softmax}(f_s(\boldsymbol{e_c}, \boldsymbol{u}, \boldsymbol{m}))] \approx \text{softmax}(\mathbb{E}_{[u|H]}[f_s(\boldsymbol{e_c}, \boldsymbol{u}, \boldsymbol{m})]). \tag{6}$$

The details of the NWGM approximation can be found in supplementary materials. In this paper, we model $f_s(\boldsymbol{e_c}, \boldsymbol{u}, \boldsymbol{m}) = \boldsymbol{e_c}^T(\boldsymbol{u} + \boldsymbol{m})$. Thanks to the linear additive property of expectation calculation, we can use $\boldsymbol{e_c}^T(\mathbb{E}_{[u|H]}[\boldsymbol{D_u}] + \boldsymbol{m})$ to calculate $\mathbb{E}_{[u|H]}[\boldsymbol{e_c}^T(\boldsymbol{u} + \boldsymbol{m})]$. In practice, we use a dot-product attention to compute $\mathbb{E}_{[u|H]}[\boldsymbol{D_u}]$. Specifically, $\mathbb{E}_{[u|H]}[\boldsymbol{D_u}] = \text{softmax}(\boldsymbol{L}^T\boldsymbol{K}) \odot \boldsymbol{D_u}$, where $\boldsymbol{L} = \boldsymbol{W_1}\boldsymbol{h}$, $\boldsymbol{K} = \boldsymbol{W_2}\boldsymbol{D_u}$ and $\odot$ is element-wise product, $\boldsymbol{h}$ is the embedding of history H, and $\boldsymbol{W_1}, \boldsymbol{W_2}$ are mapping matrices. The training details can be found in Section 6.3.

6. Experiments

6.1. Experimental Setup

Dataset. Our proposed principles are evaluated on the recently released real-world dataset VisDial v1.0. Specifically, the training set of VisDial v1.0 contains 123K images from the COCO dataset [20] with a 10-round dialog for each image, resulting in 1.2M dialog rounds. The validation and test sets were collected from Flickr, with 2K and 8K COCO-like images respectively. The test set is further split into test-std and test-challenge splits, both with the number of 4K images that are hosted on the blind online evaluation server. Each dialog in the training and validation sets has 10 rounds, while the number in the test set is uniformly distributed from 1 to 10. For each dialog, a list of 100 answer candidates is given for evaluation. In the following, the results are reported on the validation and test-std set.

Metrics. As mentioned in Section 3.1, NDCG is recommended by the official and accepted by the community. There are some other retrieval-based metrics like MRR (Mean Reciprocal Rank), where the ground-truth answer is generated by the single user. Note that the only answer may be easily influenced by the single user's preference (*i.e.*, length). We argue that this may be the reason why the models with history shortcut achieve higher MRR, (*e.g.*, due to the bias illustrated in Figure 2) and lower NDCG. Therefore, retrieval-based metrics are not consistent with NDCG. According to the mentioned reasons and space limitation, we only show the results on NDCG in the main paper. For completeness, the further discussion between NDCG and other retrieval-based metrics and the performance on all metrics will be given in the supplementary materials.

6.2. Model Zoo

We report the performance of the following base models, including LF [9], HCIAE [21], CoAtt [41] and RvA [26]:

LF [9]. This naive base model has no attention module. We expand the model by adding some basic attention operations, including question-based history attention and question-history-based visual attention refinement.

HCIAE [21]. The model consists of question-based history attention and question-history-based visual attention.

CoAtt [41]. The model consists of question-based visual attention, image-question-based history attention, image-history-based question attention, and the final question-history-based visual attention.

RvA [26]. The model consists of question-based visual attention and history-based visual attention refinement.

6.3. Implementation Details

Pre-processing. For language pre-processing, we followed the process introduced by [9]. First, we lowercased all the letters in sentences and converted digits to words and removed contractions. After that, we used Python NLTK toolkit to tokenize sentences into word lists, followed by padding or truncating captions, questions, and answers to the length of 40, 20 and 20, respectively. Then, we built a vocabulary of the tokens with the size of 11,322, including 11,319 words that occur at least 5 times in train v1.0 and 3 instruction tokens. We loaded the pre-trained word embeddings from GloVe [30] to initialize all word embeddings, which were shared in encoder and decoder, and we applied 2-layers LSTMs to encode word embedding and set their hidden state dimension to 512. For the visual feature, we used bottom-up-attention features [4] given by the official [1].

Implementation of Principles. For Principle 1 (P1), we eliminated the history feature in the final fused vector representation for all models, while kept other parts unchanged. For HCIAE [21] and CoAtt [41], we also blocked the history guidance to the image. For Principle 2 (P2), we trained the models using the preference score, which can be counted from question type or given by the official (*i.e.*, dense annotations in VisDial v1.0 training set). Specifically, for "question type", we first defined 55 types and marked answers occurred over 5 times as preferred answers, then used the preference to train our model by (R_2) loss proposed in Section 5.2. "Answer score sampling" was directly used to fine-tune our pre-trained model by the proposed loss functions. For "hidden dictionary", we set a matrix for N as 100 and d as 512 to realize D_u. The dictionary is ini-

Model	baseline	QT	S				D
			R_0	R_1	R_2	R_3	
LF [9]	57.21	58.97	67.82	71.27	72.04	72.36	72.65
LF +P1	61.88	62.87	69.47	72.16	72.85	73.42	**73.63**

Table 1. Performance (NDCG%) comparison for the experiments of applying our principles on the validation set of VisDial v1.0. LF is the enhanced version as we mentioned. QT, S and D denote question type, answer score sampling, and hidden dictionary learning, respectively. R_0, R_1, R_2, R_3 denote regressive loss, weighted softmax loss, binary sigmoid loss ,and generalized ranking loss, respectively.

tialized with the features of top-100 popular answers, then trained by dense annotations with R_3 loss. More details can be found in supplementary materials. Note that the implementations following P1 and P2 are flexible.

Training. We used softmax cross-entropy loss to train the model with P1, and used Adam [17] with the learning rate of 4×10^{-3} which decayed at epoch 5, 7, 9 with the decay rate of 0.4. The model was trained for 15 epochs totally. In addition, Dropout [35] was applied with ratio of 0.4 for RNN and 0.25 for fully connected layers. Other settings were set by default.

6.4. Quantitative Results

Table 1 shows the results with different implementations in P2, *i.e.*, question type, answer score sampling, and hidden dictionary learning. Overall, all of the implementations can improve the performances of base models. Specifically, the implementations of P2 can further boost performance by at most 11.75% via hidden dictionary learning. Specifically, our designed loss functions based on Eq. (2) outperform the regressive score, which is implemented as Euclidean distance loss and denoted as R_0. The reason is that the regression fine-tune strategy is not a proper approximation for P2. We also find that the proposed ranking loss (*i.e.*, R_3) performs best since it satisfies the ranking property of VisDial.

Note that our principles are model-agnostic. Table 2 shows the results about applying our principles on four different models (*i.e.*, LF [9], HCIAE [21], CoAtt [41] and RvA [26]). In general, both of our principles can improve all the models in any ablative condition (*i.e.*, P1, P2, P1+P2). Note that the effectiveness of P1 and P2 are additive, which means combining P1 and P2 performs the best.

We finally used the blind online test server to justify the effectiveness of our principles on the test-std split of VisDial v1.0. As shown in Table 3, the top part contains the results of the baseline models implemented with our principles, while the bottom part represents the recent Visual Dialog Challenge 2019 leaderboard [2]. We used the ensemble of the enhanced LF [9] to beat the Visual Dialog Challenge 2019 Winner (*i.e.*, MReaL-BDAI), which can also be regarded as the implementations of P1 and P2. Promisingly,

Model	LF [9]	HCIAE [21]	CoAtt [41]	RvA [26]
baseline	57.21	56.98	56.46	56.74
+P1	61.88	60.12	60.27	61.02
+P2	72.65	71.50	71.41	71.44
+P1+P2	**73.63**	71.99	71.87	72.88

Table 2. Performance (NDCG%) of ablative studies on different models on VisDial v1.0 validation set. P2 indicates the most effective one (*i.e.*, hidden dictionary learning) shown in Table 1. Note that only applying P2 is implemented by the implementations in Section 5 with the history shortcut.

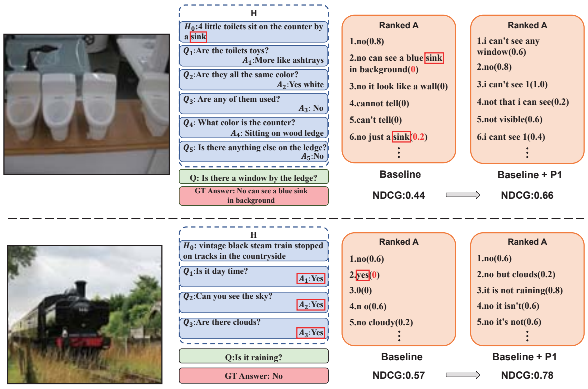


Figure 4. Qualitative results of the baseline and baseline with P1 on the validation set of VisDial v1.0. The numbers in brackets in ranked A denote relevance scores. Red boxes denote that the selected candidates of the baseline model influenced by the shortcut (*e.g.*, word matching) from the dialog history. For the baseline with P1, it does not make such biased shortcut choices. More details can be found in Section 6.5.

by applying our principles, we can promote all the baseline models to the top ranks on the leaderboard.

6.5. Qualitative Analysis

The qualitative results illustrated in Figure 4 and Figure 5 show the following advantages of our principles.

History Bias Elimination. After applying P1, the harmful patterns learned from history are relieved, like the answer-length bias shown in Figure 2(a) as we mentioned. The example on the top of Figure 4 shows the word-match bias in the baseline. From this example, we can observe that the word "sink" from history is literally unrelated to the current question, However, for the baseline model, some undesirable candidates (*i.e.*, with low relevance score) containing the word "sink" can be found in the top-ranked answers due to the wrong direct history shortcut. To further confirm the conjecture, we counted the word-match cases of objective words (*e.g.*, "sink" and "counter") on the validation set for the top-10 candidates of the ranked lists. The statistic indicates that P1 can decrease about 10% word matching from history (from ∼5200 times of baseline to ∼4800 times using P1). The bottom example shows that, when the answer "yes" exists in history, the baseline model will tend to rank

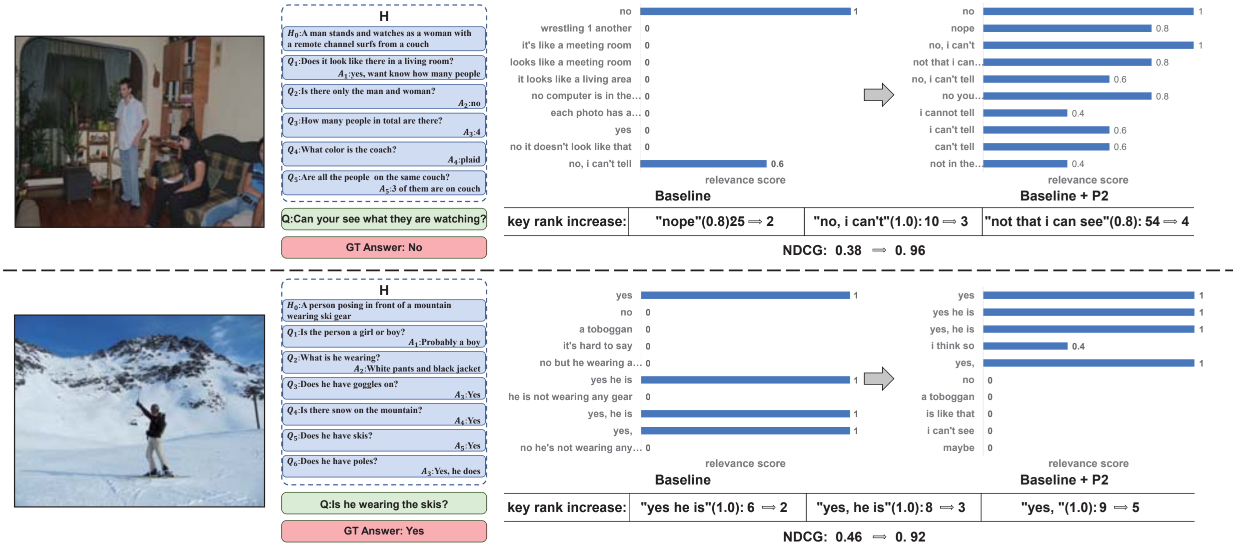


Figure 5. Qualitative examples of the ranked candidates of baseline and baseline with P2. We also give some key rank changes for boosting NDCG performance by implementing P2. These examples are taken from the validation set of VisDial v1.0.

	Model	NDCG(%)
Ours	P1+P2 (More Ensemble)	74.91
	LF+P1+P2 (Ensemble)	74.19
	LF+P1+P2 (single)	71.60
	RvA+P1+P2 (single)	71.28
	CoAtt+P1+P2 (single)	69.81
	HCIAE+P1+P2 (single)	69.66
Leaderboard	VD-BERT(Ensemble)*	75.13
	Tohuku-CV Lab (Ensemble)*	74.88
	MReaL-BDAI*	74.02
	SFCU (Single)*	72.80
	FancyTalk (HeteroFM)*	72.33
	Tohuku-CV Lab (Ensemble w/o ft)*	66.53

Table 3. Our results and comparisons to the recent Visual Dialog Challenge 2019 Leaderboard results on the test-std set of VisDial v1.0. Results are reported by the test server, (*) denotes it is taken from [2]. Note that the top five models in the Leaderboard use the dense fine-tune implementation illustrated in Section 5.2.

"yes" in a high place. However, it is opposite to the real answer "no" in some cases, which will lead to a lower NDCG. After applying P1, this problem can be effectively alleviated. To testify this conclusion, we further calculate the average ranks of "yes" for baseline and baseline with P1 in the above case (*i.e.*, "yes" appears in history and the real answer is "no"). We find that the average ranks are 4.82 for baseline and 6.63 for baseline with P1 respectively. The lower rank means that P1 relieves the "yes" shortcut in history. More examples of these biases can be found in supplementary materials.

More Reasonable Ranking. Figure 5 shows that the baseline model only focuses on ground truth answers like "no" or "yes" and does not care about the rank of other answers with similar meaning like "nope" or "yes, he is". This does not match human's intuition because the candidates with similar semantics are still reasonable. This also leads the baseline model to a lower NDCG. As shown in Figure 5 the model with P2 almost ranks all the suitable answers like "yes, he is", "yes he is" and "I think so" at top places together with the ground truth answer "yes", which significantly improves the NDCG performance.

7. Conclusions

In this paper, we proposed two causal principles for improving the VisDial task. They are model-agnostic, and thus can be applied in almost all the existing methods and bring major improvement. The principles are drawn from our in-depth causal analysis of the VisDial nature, which is however unfortunately overlooked by our community. For technical contributions, we offered some implementations on how to apply the principles into baseline models. We conducted extensive experiments on the official VisDial dataset and the online evaluation servers. Promising results demonstrate the effectiveness of the two principles. As moving forward, we will stick to our causal thinking to discover other potential causalities hidden in embodied Q&A and conversational visual dialog tasks.

Acknowledgement This work was partially supported by the National Natural Science Foundation of China (61573363 and 61832017), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01), and NTU-Alibaba JRI. We would like to thank the anonymous reviewers for their constructive comments.

References

[1] Visual Dialog. https://visualdialog.org/.

[2] Visual Dialog Challenge 2019 Leaderboard. https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483/.

[3] Visual Dialog Challenge 2019 Winner Report. https://drive.google.com/file/d/1fqg0hregsp_3USM6XCHx89S9JLXt8bKp.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[6] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *8rd International Conference on Learning Representations, ICLR*, 2020.

[7] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[8] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[9] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[10] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[11] Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 6463–6474, 2019.

[12] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2018.

[16] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 2024–2033, 2019.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[18] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169, 2018.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324, 2017.

[22] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[23] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. In *Advances in Neural Information Processing Systems Workshop on Causal Machine Learning*, 2019.

[24] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. In *Advances in Neural Information Processing Systems Workshop on Causal Machine Learning*, 2019.

[25] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

[26] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 6679–6688, 2019.

[27] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

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

[29] Judea Pearl and Dana Mackenzie. *THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT*. Basic Books, 2018.

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[32] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048, 2019.

[33] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729, 2017.

[34] Rahul Singh and Liyang Sun. De-biased machine learning for compliers. In *Advances in Neural Information Processing Systems Workshop on Causal Machine Learning*, 2019.

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[37] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019.

[38] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[41] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115, 2018.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[43] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: Gold-critic sequence training for visual dialog. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[44] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[45] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect, 2020.

[46] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.

[47] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.

[48] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6669–6678, 2019.